



Dimension reduction for statistical inference of distribution tails

Alex Podgorny

► To cite this version:

Alex Podgorny. Dimension reduction for statistical inference of distribution tails. Mathematics [math]. Université de Strasbourg, 2025. English. ⟨NNT : 2025STRAD016⟩. ⟨tel-05240773v3⟩

HAL Id: tel-05240773

<https://theses.hal.science/tel-05240773v3>

Submitted on 9 Apr 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

ED 269 : Mathématiques, Sciences de l'information et de l'Ingénieur

Institut de Recherche Mathématique Avancé

UMR 7501

THÈSE présentée par :

Alex PODGORNÝ

soutenue le : 24 septembre 2025

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Mathématiques

**Réduction de dimension pour
l'inférence statistique de queues de
distribution**

THÈSE dirigée par :
Laurent Gardes

Professeur, Université de Strasbourg

RAPPORTEURS :
Anne Sabourin
Clément Dombry

Professeur, Université Paris Cité
Professeur, Université de Franche-Comté

AUTRES MEMBRES DU JURY :

Sébastien Engelke
Laurent Gardes
Stéphane Girard
Emmanuel Franck
Samuel Maistre

Associate Professor, Université de Genève
Professeur, Université de Strasbourg
Directeur de recherche, Inria Grenoble Rhône-Alpes
Chargé de recherche, Inria Nancy Grand-Est
Maître de conférences, Université de Strasbourg

Avertissement au lecteur / Warning to the reader

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition des membres de la communauté universitaire. Il est soumis à la propriété intellectuelle de l'auteur. Cela implique une obligation de citation et de référencement lors de l'utilisation de ce document. D'autre part, toute contrefaçon, plagiat, reproduction ou représentation illicite encourt une poursuite pénale.

This document is the result of a long process approved by the jury and made available to members of the university community. It is subject to the intellectual property rights of its author. This implies an obligation to quote and reference when using this document. Furthermore, any infringement, plagiarism, unlawful reproduction or representation will be prosecuted.

Code de la Propriété Intellectuelle

Article L122-4 : Toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite. Il en est de même pour la traduction, l'adaptation ou la transformation, l'arrangement ou la reproduction par un art ou un procédé quelconque.

Any representation or reproduction in whole or in part without the consent of the author or his successors in title or assigns is unlawful. The same applies to translation, adaptation or transformation, arrangement or reproduction by any art or process whatsoever.

Articles L335-1 à L335-9 : Dispositions pénales / Penal provisions.

Licence attribuée par l'auteur



Réduction de dimension pour l'inférence statistique de queues de distribution

Résumé

Ce manuscrit s'inscrit dans le domaine de la statistique des valeurs extrêmes, et s'intéresse plus particulièrement à l'étude des quantiles extrêmes d'une variable aléatoire. Un quantile est qualifié d'extrême dès lors que son estimation requiert de connaître la forme de la distribution au-delà de l'observation maximale dont on dispose. Le manuscrit repose sur trois contributions principales. La première porte sur la construction d'intervalles de confiance pour les quantiles extrêmes, pour l'ensemble des domaines d'attraction et basés sur des statistiques d'ordre. Cette méthode permet de quantifier rigoureusement l'incertitude autour d'un quantile extrême. La deuxième contribution propose une méthode originale de réduction de dimension pour l'estimation de l'indice de queue conditionnel, en supposant que celui-ci dépend uniquement d'une projection des covariables sur un sous-espace linéaire. Cette approche permet de contourner le fléau de la grande dimension. Enfin, le troisième axe présente une étude comparative de différentes méthodes récentes issues du machine learning et des approches de réduction dimension, pour l'estimation de quantiles extrêmes conditionnels en haute dimension. Cette analyse met en lumière les forces et limites respectives de ces approches.

Mots clés : Théorie des valeurs extrêmes, Quantiles extrêmes, Intervalles de confiance, Estimation non paramétrique, Indice de queue conditionnel, Réduction de dimension, Grande dimension, Apprentissage statistique

Résumé en anglais

This manuscript falls within the field of extreme value statistics and focuses more specifically on the study of extreme quantiles of a random variable. A quantile is considered extreme when its estimation requires knowledge of the distribution's behavior beyond the largest observed value. The manuscript is based on three main contributions. The first concerns the construction of confidence intervals for extreme quantiles, valid across all maximum domains of attraction and based on order statistics. This method enables a rigorous quantification of the uncertainty surrounding an extreme quantile. The second contribution introduces an original dimension reduction method for estimating the conditional tail index, under the assumption that it depends only on a projection of the covariates onto a linear subspace. This approach helps mitigate the curse of dimensionality. Finally, the third part presents a comparative study of several recent methods from machine learning and dimension reduction, applied to the estimation of conditional extreme quantiles in high-dimensional settings. This analysis highlights the respective strengths and limitations of these approaches.

Keywords : Extreme value theory, Extreme quantiles, Confidence intervals, Nonparametric estimation, Conditional tail index, Dimension reduction, High dimensionality, Machine learning



UNIVERSITÉ DE STRASBOURG

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur
Institut de Recherche Mathématique Avancée, UMR 7501

Réduction de dimension pour l'inférence statistique de queues de distribution

Thèse de doctorat en Statistique
Présentée par **Alex PODGORNÝ**

Pour l'obtention du grade de
Docteur de l'Université de Strasbourg

Soutenue le 24 septembre 2025

Directeur de thèse : Laurent Gardes

Membres du jury :

L. Gardes, Directeur de thèse ; A. Sabourin, Rapporteuse ; C. Dombry,
Rapporteur ; S. Engelke, Examineur ; S. Girard, Examineur ; E. Franck,
Examineur ; S. Maistre, Examineur ;

Je tiens à exprimer ma profonde gratitude envers mon directeur de thèse, Laurent Gardes, pour son encadrement attentif, ses conseils éclairés et sa confiance tout au long de ce travail. Toujours disponible et d'une grande aide, il a su m'encourager à donner le meilleur de moi-même. J'ai beaucoup apprécié écrire mes articles avec lui et j'espère que d'autres occasions se présenteront pour collaborer à nouveau. Je remercie également Samuel Maistre, avec qui j'ai rédigé mon tout premier article et qui m'a accompagné dans mes premiers cours donnés aux étudiants. Ma reconnaissance va aussi à tous les permanents de mon équipe et leurs précieux conseils.

Le doctorat est une aventure. Je le sais, la période qui vient de s'achever restera exceptionnelle dans ma vie. Elle m'a donné le goût du savoir, de la rigueur et surtout celui de la liberté. J'ai compris rapidement que pour faire une bonne thèse, il fallait s'ouvrir au monde. Merci à Malaury, Gaëtan, Luc, Pierre, Youna, Séverin, Guillaume, Mathis et Léa pour le club de philosophie, pour toutes les aventures et toutes ces soirées à discuter des heures durant : sans doute les expériences les plus fécondes de ma vie.

Je remercie aussi chaleureusement tous les doctorants. Ce fut un véritable plaisir d'essayer de saisir tant bien que mal vos recherches. Je suis fasciné par ce que vous êtes capables de faire. Merci à ceux qui ont pris le temps de m'expliquer, particulièrement Thomas, avec qui apprendre a toujours été un plaisir. Nous avons aussi partagé bien d'autres moments : les travaux et déménagements avec mes cobureaux Clarence, Tom, Amaury et Killian ; les bonnes soirées avec Adam, Céline, Mathilde, Roxana et Tom ; l'escalade (même si j'ai sûrement trop vite abandonné) ; les débats politiques avec Roméo, Ludovic, Paul et Amaury. Je me suis particulièrement senti bien avec les doctorants de l'équipe statistique, Jean-Pierre, Louise et Jules. Malheureusement je ne peux citer tout le monde, mais ces trois années en votre compagnie resteront inoubliables.

Merci à mes vieux amis, Gabriel, Lucie, Édouard, Louis, Léo, Clément, Louise, Marianne, Alexandra, Raphaël, Louis, Maxime, Stan, Walid et Guillaume qui m'ont toujours soutenu et encouragé.

Je remercie aussi de tout mon cœur ma chère Léa, qui a toujours été d'un grand soutien. Pendant les moments les plus difficiles, elle a toujours été présente, attentive, aidante... Je suis heureux de pouvoir discuter avec toi

tous les soirs d'art, de littérature, d'histoire, de science, de tout ce qui nous passionne. J'ai saisi la chance d'être avec toi et je ne compte pas la lâcher.

Enfin, je remercie ma famille aimante. Il est certain que si je n'avais pas reçu autant d'amour des miens, rien de tout cela n'aurait été possible. Ma famille est mon pilier. Merci à mes frères et sœurs, Caroline, Thomas, Laura et Nina. Vous comptez énormément pour moi. Merci Véronique, Pascal, Maxime et Malaury de m'accueillir toujours avec grand plaisir chez vous. Merci Ludovic, Valérie, pour votre gentillesse. Merci aux plus jeunes Camille, Sidney, Maya et Fanny qui réussiront à être heureux aussi j'en suis sûr. Merci David, et merci Maman.

À mon père, dont le souvenir m'accompagne à chaque instant et à qui je dois mon goût pour la science.

J'ai promis de le rendre fier.

Table des matières

Introduction	7
1 Asymptotic confidence intervals for extreme quantiles in a maximum domain of attraction	28
1.1 Introduction	30
1.2 Construction of the confidence interval	34
1.3 Estimation of the extrapolation factors	38
1.3.1 General estimation procedure	38
1.3.2 Application : Moment Estimator	39
1.4 Simulation study	42
1.5 Real data study	47
1.6 Proofs	55
1.6.1 Preliminary results	55
1.6.2 Proofs of main results	57
1.6.3 Proof of Corollary 1	59
1.6.4 Proof of Proposition 3	60
2 Dimension reduction for the estimation of the conditional tail index	63
2.1 Introduction	65
2.2 Tail Index Dimension Reduction subspace	68
2.2.1 Framework	68
2.2.2 Preliminary result	69
2.2.3 Main definitions	70
2.3 Estimation of the CTI subspace and the conditional tail index	72
2.3.1 Construction and definition of the estimators	73
2.3.2 Main results	74
2.3.3 A local-Hill estimator	75

2.4	Finite sample size behavior	78
2.4.1	Tuning parameters	79
2.4.2	Finding the solution of the optimization problem	80
2.4.3	Models	81
2.4.4	Simulation results	82
2.4.5	Estimation of the dimension of the CTI subspace	84
2.5	Application to a real data set	93
2.6	Proofs	96
2.6.1	Preliminary results	96
2.6.2	Proofs of main results	99
3	Extreme Conditional Quantile Estimation in High Dimensions : A Comparative Study	120
3.1	Introduction	122
3.2	Classical conditional quantile estimation	126
3.3	Tail parameters estimation	128
3.3.1	Extreme random forest	129
3.3.2	Extreme neural network	132
3.3.3	Dimension reduction procedures	137
3.4	Simulation Study	143
3.4.1	General Setup	144
3.4.2	Choice of Tuning Parameters	144
3.4.3	Experimental Results	147

Introduction

Il n'est pas aisé de cerner précisément les contours du champ des statistiques, tant celui-ci est vaste. Les méthodes mobilisées sont d'une grande diversité, allant de techniques descriptives élémentaires aux algorithmes d'intelligence artificielle. Toutefois, une constante subsiste : l'objectif fondamental de la statistique est l'extraction d'informations pertinentes à partir d'un ensemble limité d'observations, appelées échantillon. Cette extraction d'information vise généralement à éclairer une prise de décision dans un contexte d'incertitude. Par exemple, dans le domaine médical, des essais cliniques permettent d'estimer l'efficacité d'un nouveau traitement ; en marketing, des analyses de données aident à anticiper les comportements des consommateurs ; et en politique, les sondages permettent de prévoir les intentions de vote.

Certaines problématiques statistiques ne se focalisent pas sur les comportements typiques ou moyens d'un phénomène, mais sur ses manifestations les plus rares, les plus extrêmes. Ce sont ces événements à faible probabilité de survenance, mais amenant potentiellement à des conséquences majeures, qui exigent une attention particulière. Un exemple emblématique se trouve dans le domaine de l'assurance, où la réglementation européenne Solvabilité II impose aux entreprises du secteur de quantifier précisément les risques extrêmes afin de garantir leur solvabilité même en cas de sinistres exceptionnels. Cette réglementation, ancrée dans une logique de gestion prudente des risques, oblige les assureurs à détenir un capital suffisant pour faire face à des événements rares de sorte, en théorie, de ne faire faillite en moyenne qu'une fois tous les 200 ans...

Au-delà de l'assurance, de nombreux autres domaines rencontrent ce type de problématiques : en climatologie, il s'agit par exemple d'estimer la probabilité d'occurrence de vagues de chaleur ou de froid extrêmes ; en hydrologie, de

prédire les crues centennales ; en finance, de modéliser les krachs boursiers ; ou encore en biologie, d'étudier les limites de longévité humaine. Dans tous ces cas, les outils statistiques classiques, conçus pour décrire le comportement central d'un phénomène, deviennent inadaptés car les données observées sont, par définition, peu représentatives des extrêmes.

C'est précisément à cette difficulté que s'attelle la théorie des valeurs extrêmes (TVE). Ce domaine spécialisé de la statistique propose un cadre théorique rigoureux pour analyser les événements rares en se concentrant sur le comportement de la queue de distribution de la variable aléatoire modélisant au mieux le phénomène. Autrement dit, la TVE permet d'extraire de l'information sur ce qui se passe aux marges d'un phénomène, là où les données sont rares.

Ma thèse s'inscrit dans ce cadre, et porte plus précisément sur l'estimation des quantiles extrêmes d'une loi de probabilité. Un quantile associé à la probabilité $\beta \in (0, 1)$ est la valeur telle que la probabilité que la variable aléatoire soit plus grande que ce seuil est égale à β . C'est une généralisation du concept de médiane et de quartiles. Plus formellement, si Y est une variable aléatoire réelle de fonction de répartition F , la fonction quantile est définie comme l'inverse généralisée de la fonction de survie $\bar{F} := 1 - F$. Autrement dit le quantile d'ordre $\beta \in (0, 1)$ donné par

$$Q(\beta) = \bar{F}^{\leftarrow}(\beta) = \inf\{y \in \mathbb{R} | \bar{F}(y) \leq \beta\}.$$

Pour un échantillon de taille n indépendant et identiquement distribué (i.i.d.) $Y_1, \dots, Y_n$ de même loi Y , un quantile $Q(\beta_n)$ est qualifié d'extrême lorsque la suite $n\beta_n \rightarrow c \geq 0$ pour $n \rightarrow \infty$, de sorte que l'on cherche à estimer des seuils élevés du phénomène étudié, avec peu d'observations disponibles dans cette zone.

Dans un grand nombre d'applications, la variable aléatoire étudiée Y dépend d'une covariable $X \in \mathbb{R}^p$, et l'on cherche à estimer les quantiles extrêmes de la distribution conditionnelle de Y sachant $X = x$, de fonction de répartition $F(\cdot|x)$. Dans ce cadre, pour un point fixé $x \in \mathbb{R}^p$, on définit le quantile extrême conditionnel par :

$$Q(\beta|x) = \bar{F}^{\leftarrow}(\beta|x) = \inf\{x \in \mathbb{R} | \bar{F}(y|x) \leq \beta\}.$$

Toutefois, une haute dimension p de l'espace des covariables X peut rendre cette tâche difficile en pratique. Pour y remédier, plusieurs solutions sont envisageables, comme la réduction dimension ou l'utilisation du machine learning. Mais avant de formaliser correctement le problème, rappelons d'abord quelques généralités sur la théorie des valeurs extrêmes dans le cadre non-conditionnel, puis conditionnel.

Valeurs extrêmes dans le cas non-conditionnel

Introduction à la TVE

Considérons un échantillon i.i.d. $(Y_1, \dots, Y_n)$ de taille n , et de fonction de répartition commune F . En statistique, l'étude des propriétés asymptotiques des estimateurs ou des statistiques est centrale : on cherche à comprendre le comportement de certaines quantités lorsque la taille de l'échantillon n tend vers l'infini, dans l'idée que plus d'observations permettent une meilleure estimation et une convergence vers un comportement limite stable.

Il est par exemple classique d'étudier la somme des observations $\sum_{i=1}^n Y_i$ lorsque $n \rightarrow \infty$. Dans de nombreux cas, cette somme ne converge pas, il convient donc de la normaliser pour espérer obtenir une convergence en loi vers une variable aléatoire non dégénérée (non constante). On considère alors des suites de normalisation (a_n) avec $a_n > 0$ et (b_n) dans $\mathbb{R}$, et on s'intéresse au comportement de

$$\sum_{i=1}^n \frac{Y_i - b_n}{a_n}.$$

Un exemple fondamental est donné par le théorème central limite, qui établit que si la variance $\sigma^2 = \text{Var}(Y)$ existe, alors

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mathbb{E}(Y)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

c'est-à-dire que la somme centrée et réduite converge en loi vers une variable normale. Plus généralement, on peut se demander quelles sont les lois possibles comme limites de telles sommes normalisées. Cela conduit à la théorie des lois α -stables, qui sont les seules lois non dégénérées pouvant apparaître comme limites de sommes normalisées de variables i.i.d.

La théorie des valeurs extrêmes (TVE) s'inscrit dans une démarche exactement analogue, mais au lieu d'étudier le comportement asymptotique de la somme, elle s'intéresse au maximum d'un échantillon i.i.d.,

$$M_n = \max(Y_1, \dots, Y_n),$$

dans le but de comprendre le comportement des valeurs atypiquement grandes, c'est-à-dire des événements situés dans la queue de la distribution. Comme pour les sommes, le maximum doit être normalisé pour espérer obtenir une convergence non triviale. On cherche ainsi des suites de normalisation $(a_n) > 0$ et $(b_n) \in \mathbb{R}$ telles que la variable

$$\frac{M_n - b_n}{a_n}$$

converge en loi, lorsque $n \rightarrow \infty$, vers une variable aléatoire non dégénérée. Si une telle limite existe, on dit que la loi F appartient à un domaine d'attraction. Nous pouvons réécrire cette propriété en terme de fonction de répartition (comme l'échantillon est i.i.d., la fonction de répartition de M_n est F^n).

Définition 1. *Une fonction de répartition F appartient au domaine d'attraction d'une fonction de répartition non dégénérée H si, pour des suites réelles $(a_n)_{n \in \mathbb{N}}$ avec $a_n > 0$ et $(b_n)_{n \in \mathbb{N}} \subset \mathbb{R}$, nous avons pour tout point de continuité y de H :*

$$F^n(a_n y + b_n) \rightarrow H(y).$$

Ce cadre permet de définir rigoureusement les lois limites des maxima normalisés. On peut montrer que si F appartient aux domaines d'attraction de H et G , alors celles-ci sont du même type : il existe des constantes $a > 0$ et $b \in \mathbb{R}$ telles que

$$\forall y \in \mathbb{R}, \quad H(ay + b) = G(y).$$

Le théorème fondamental de la théorie des valeurs extrêmes donne alors la caractérisation complète des distributions limites possibles des maxima. Celles-ci appartiennent à une unique famille paramétrée par un réel $\gamma \in \mathbb{R}$, appelé *indice des valeurs extrêmes*.

Théorème 1 (Théorème des valeurs extrêmes). *Si $F \in DA(G)$, alors il existe un réel $\gamma \in \mathbb{R}$ tel que G est du même type que la fonction*

$$H_\gamma(y) = \begin{cases} \exp(-(1 + \gamma y)^{-1/\gamma}), & \text{si } \gamma \neq 0, 1 + \gamma y > 0, \\ \exp(-e^{-y}), & \text{si } \gamma = 0. \end{cases}$$

Ce résultat est fondamental : il montre que si une loi F appartient à un domaine d'attraction, alors le comportement asymptotique de sa queue de distribution est entièrement caractérisée par un unique paramètre γ . C'est le cas pour la grande majorité des lois usuelles, même si certaines distributions pathologiques peuvent y échapper.

Selon la valeur de γ , appelé indice des valeurs extrêmes, on distingue trois types de comportements asymptotiques, associés à trois classes de distributions :

- **Fréchet** ($\gamma > 0$) : la loi F possède une queue lourde et a un support infini. La décroissance de la queue suit une loi de puissance.
- **Gumbel** ($\gamma = 0$) : la loi F a un point terminal infini, mais la queue est légère, décroissant de manière exponentielle. C'est le cas, par exemple, de la loi normale ou exponentielle.
- **Weibull** ($\gamma < 0$) : la loi F admet un support fini. La queue est dite courte.

Estimation des quantiles extrêmes

Dans le cadre classique, l'estimation d'un quantile d'ordre $\alpha \in (0, 1)$ peut s'effectuer en remplaçant la fonction de survie théorique, définie par

$$\overline{F}(y) = \mathbb{P}(Y > y) = \mathbb{E}[\mathbb{I}\{Y > y\}],$$

par son analogue empirique :

$$\widehat{\overline{F}}_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{Y_i > y\}.$$

En notant les statistiques d'ordre de l'échantillon i.i.d. $(Y_1, \dots, Y_n)$ par

$$Y_{1,n} \leq Y_{2,n} \leq \dots \leq Y_{n,n},$$

on obtient alors l'estimateur empirique du quantile d'ordre α :

$$\widehat{Q}_n(\alpha) = \inf\{y \in \mathbb{R} \mid \widehat{\overline{F}}_n(y) \leq \alpha\} = Y_{n-\lfloor n\alpha \rfloor, n}.$$

Cependant, ce type d'estimateur devient inadapté lorsqu'on cherche à estimer des quantiles extrêmes, c'est-à-dire lorsque $n\beta_n \rightarrow c \geq 0$. Dans ce cas, l'estimateur asymptotiquement équivalent à $Y_{n-\lfloor c \rfloor, n}$ ne sera pas asymptotiquement normal. De plus si $c < 1$, le seul estimateur possible devient alors

le maximum empirique $Y_{n,n}$, qui ne peut pas converger vers le vrai quantile $Q(\beta_n)$.

Ce constat met en évidence la limite des méthodes empiriques classiques dans le cas extrême, et justifie le recours à la théorie des valeurs extrêmes, qui fournit des outils d'extrapolation adaptés à l'estimation des queues de distribution. En effet, d'après le théorème 1.1.6 de [39], nous avons plusieurs caractérisations équivalentes du fait que F appartienne au domaine d'attraction de H_γ . Une particulièrement utile est la suivante :

Proposition 1. *Soit $\gamma \in \mathbb{R}$. Les propositions suivantes sont équivalentes :*

1. $F \in DA(H_\gamma)$;
2. *Il existe une fonction de normalisation positive $a(\cdot)$ telle que, pour tout $t > 0$,*

$$\lim_{\alpha \rightarrow 0} \frac{Q(t\alpha) - Q(\alpha)}{a(\alpha^{-1})} = \frac{t^{-\gamma} - 1}{\gamma}, \quad (1)$$

où $\frac{t^{-\gamma}-1}{\gamma} = \log(t)$ si $\gamma = 0$.

Cette limite offre une stratégie pour l'estimation des quantiles extrêmes : plutôt que d'estimer directement un quantile de niveau très faible β_n , on estime un quantile intermédiaire $Q(\alpha_n)$ pour une suite α_n telle que $\alpha_n \rightarrow 0$ mais $n\alpha_n \rightarrow \infty$ de sorte que la statistique d'ordre $Y_{n-[n\alpha_n],n}$ converge vers le vrai quantile $Q(\alpha_n)$, puis on extrapole $Q(\beta_n)$ à l'aide de la relation asymptotique donnée par (1) :

$$\hat{Q}(\beta_n) := Y_{n-[n\alpha_n],n} + \hat{a}_n(\alpha_n^{-1}) \frac{(\beta_n/\alpha_n)^{-\hat{\gamma}_n} - 1}{\hat{\gamma}_n} \quad (2)$$

où $\hat{\gamma}$ et $\hat{a}(\cdot)$ sont respectivement des estimateurs de l'indice des valeurs extrêmes γ et de la fonction de normalisation $a(\cdot)$. Plusieurs méthodes ont été proposées dans la littérature pour estimer ces deux quantités, telles que l'approche du maximum de vraisemblance (voir [39], section 4.3.1]), la méthode des moments pondérés par les probabilités (Probability-Weighted Moments) introduite par Hosking et al. [46], etc. L'une des approches les plus communes est fondée sur les statistiques de moments, définies par :

$$M_n^{(j)} := \frac{1}{[n\alpha_n]} \sum_{i=1}^{[n\alpha_n]} (\ln Y_{n-i+1,n} - \ln Y_{n-[n\alpha_n],n})^j, \quad j \in \{1, 2\},$$

voir [39, section 4.3.2]. Ces statistiques sont construites à partir des $\lfloor n\alpha_n \rfloor$ plus grandes valeurs de l'échantillon. Il est ici possible d'appliquer le logarithme aux statistiques d'ordre car la variable aléatoire Y est supposée strictement positive.

L'indice des valeurs extrêmes γ est estimé par

$$\hat{\gamma}_n^{(M)} := M_n^{(1)} + 1 - \frac{1}{2} \left(1 - \frac{\{M_n^{(1)}\}^2}{M_n^{(2)}} \right)^{-1},$$

estimateur introduit par Dekkers, Einmahl et de Haan [27]. La fonction de normalisation $a(\alpha_n^{-1})$ est quant à elle estimée par

$$\hat{a}_n^{(M)}(\alpha_n^{-1}) := Y_{n-\lfloor n\alpha_n \rfloor, n} \cdot M_n^{(1)} \cdot (1 - \hat{\gamma}_n^{(M)} + M_n^{(1)}).$$

En remplaçant ces estimations dans l'expression (2), on obtient alors l'estimateur du quantile extrême fondé sur les moments, défini par :

$$\hat{Q}^{(M)}(\beta_n) := Y_{n-\lfloor n\alpha_n \rfloor, n} + \hat{a}_n^{(M)} \left(\frac{n}{\lfloor n\alpha_n \rfloor} \right) \cdot \frac{(\lfloor n\alpha_n \rfloor / (n\beta_n))^{\hat{\gamma}_n^{(M)}} - 1}{\hat{\gamma}_n^{(M)}}. \quad (3)$$

Sous certaines hypothèses portant notamment sur la vitesse de convergence de la limite (1) (détaillées dans le Chapitre 1), on peut montrer que cet estimateur est asymptotiquement normal. Cette vitesse de convergence est quantifiée par un paramètre du second ordre noté $\rho \leq 0$. Plus celui-ci est petit, plus la convergence de (1) est rapide.

Une approche simplifiée pour les lois à queue lourde

Dans le cas particulier où l'indice des valeurs extrêmes γ est strictement positif, on dit que la distribution de la variable aléatoire Y est à queue lourde. Ce cas correspond à une large classe de lois usuelles, telles que les lois de Pareto, de Student ou de Cauchy, et revêt une importance particulière dans de nombreuses applications pratiques.

Dans ce cadre, la fonction de normalisation $a(\alpha^{-1})$ admet le comportement asymptotique simple $a(\alpha^{-1}) \sim \gamma Q(\alpha)$, lorsque $\alpha \rightarrow 0$ (voir [39, Lemme 1.2.9]). La limite (1) s'écrit alors sous la forme simplifiée :

$$\lim_{\alpha \rightarrow 0_+} \frac{Q(t\alpha)}{Q(\alpha)} = t^{-\gamma}, \quad (4)$$

pour tout $t > 0$. Cela signifie que la fonction quantile est à variations régulières en 0, d'indice $-\gamma$. De manière équivalente, la fonction de survie $\overline{F}(y)$ est à variations régulières en $+\infty$ d'indice $-1/\gamma$, c'est-à-dire :

$$\lim_{y \rightarrow \infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = t^{-1/\gamma}, \quad \text{pour tout } t > 0.$$

Cette propriété conduit naturellement à une stratégie d'estimation simplifiée pour les quantiles extrêmes, puisqu'il suffit seulement d'estimer l'indice des valeurs extrêmes γ . Une fois cette estimation effectuée, un quantile d'ordre extrême β_n peut être approché à partir d'un quantile intermédiaire $Q(\alpha_n)$ par la formule dite de Weissman [74] :

$$Q(\beta_n) \approx Q(\alpha_n) \left(\frac{\alpha_n}{\beta_n} \right)^\gamma. \quad (5)$$

L'un des estimateurs les plus classiques et les plus simples de l'indice des valeurs extrêmes est l'estimateur de Hill, introduit par [45]. Il est également construit lui aussi à partir des $[n\alpha_n]$ plus grandes observations :

$$\hat{\gamma}_n^{(H)} := \frac{1}{[n\alpha_n]} \sum_{i=1}^{[n\alpha_n]} (\ln Y_{n-i+1,n} - \ln Y_{n-[n\alpha_n],n}) = M_n^{(1)}. \quad (6)$$

Le choix de la suite (α_n) est crucial, car si $[n\alpha_n]$ est trop petit, l'estimation souffre d'une forte variance ; s'il est trop grand, elle est biaisée par des observations moins extrêmes. Sous des hypothèses de régularité, l'estimateur de Hill est consistant et asymptotiquement normal.

Nous accordons une attention particulière au cas $\gamma > 0$, car il est à la fois le plus fréquent dans les applications réelles. De telles lois apparaissent notamment en finance (modélisation des pertes extrêmes), en assurance (coût de sinistres rares), dans les sciences de l'environnement (niveaux de crues ou de pollution extrêmes), et plus généralement dans toute situation où la probabilité d'occurrence d'événements extrêmes ne peut être négligée car ayant potentiellement de grave conséquence.

Valeurs extrêmes dans le cas conditionnel

Nous considérons à présent un cadre plus général, dans lequel la variable aléatoire d'intérêt Y dépend d'une covariable explicative $X \in \mathbb{R}^p$. On sup-

pose que l'on dispose d'un échantillon i.i.d. $(Y_1, X_1), \dots, (Y_n, X_n)$, où chaque couple suit la même loi jointe (Y, X) , mais où la loi marginale de Y varie conditionnellement à la valeur de X .

Dans ce contexte, notre objectif est d'estimer un quantile extrême de la loi conditionnelle $Y|X = x$, pour un point fixé $x \in \mathbb{R}^p$. On suppose pour cela que, pour tout $x \in \mathbb{R}^p$, la loi de $Y|X = x$ appartient au domaine d'attraction d'une loi des valeurs extrêmes. Cela signifie que pour tout x , il existe $\gamma(x)$ (désignant l'indice des valeurs extrêmes associé au point x) et $a(\cdot|x)$ une fonction de normalisation tels que la fonction quantile conditionnelle $Q(\cdot|x)$ vérifie, pour toute valeur $t > 0$, la convergence :

$$\lim_{\alpha \rightarrow 0^+} \frac{Q(t\alpha|x) - Q(\alpha|x)}{a(\alpha^{-1}|x)} = \frac{t^{-\gamma(x)} - 1}{\gamma(x)}. \quad (7)$$

Comme pour le cas non-conditionnel, il est possible d'estimer le quantile extrême $Q(\beta_n|x)$ avec $\beta_n \ll 1$ par extrapolation :

$$Q(\beta_n|x) \approx Q(\alpha_n|x) + \frac{a(\alpha_n^{-1}|x)}{\gamma(x)} \left[\left(\frac{\alpha_n}{\beta_n} \right)^{\gamma(x)} - 1 \right]. \quad (8)$$

et dans le cas où $\gamma(x) > 0$ pour tout x ,

$$Q(\beta_n|x) \approx Q(\alpha_n|x) \left(\frac{\alpha_n}{\beta_n} \right)^{\gamma(x)}. \quad (9)$$

Cela demande l'estimation de trois quantités (resp. deux quantités pour $\gamma(x) > 0$) : le quantile intermédiaire $Q(\alpha_n|x)$, l'indice des valeurs extrêmes conditionnel $\gamma(x)$ et la fonction de normalisation $a(\cdot|x)$. Un premier défi majeur apparaît ici : contrairement au cadre non conditionnel, les variables Y_i ne suivent pas toutes la même loi. Il n'est donc pas possible d'utiliser indistinctement les observations (Y_i) pour estimer une quantité propre à la distribution de $Y|X = x$. Cela rend inopérantes les approches classiques fondées sur les statistiques d'ordre globales. Pour surmonter cette difficulté, il est nécessaire de supposer une certaine régularité du lien entre Y et X .

Principe de pondération

L'idée la plus simple est de considérer qu'à des valeurs proches de x correspondent des lois conditionnelles similaires : autrement dit, si X_i est proche de x , alors la loi de $Y_i|X = X_i$ est proche de celle de $Y|X = x$. Cette hypothèse de continuité locale ouvre la voie à des méthodes de pondération, qui consistent à attribuer à chaque observation un poids en fonction de la distance entre X_i et x . Ainsi, les observations les plus proches du point cible x contribuent davantage à l'estimation, tandis que les observations éloignées sont peu, voire pas, prises en compte. Ce principe de pondération, que l'on retrouve dans les méthodes à noyau ou des k plus proches voisins, constitue une des approches les plus communes pour l'estimation conditionnelle non paramétriques. Cette approche a été exploré dans plusieurs travaux pour l'estimation des quantiles extrêmes, comme celui de Daouia, Gardes et Girard [25].

Il est important de souligner que le principe de pondération ne repose pas nécessairement sur une distance euclidienne dans l'espace $\mathbb{R}^p$. D'autres formes de régularité peuvent être envisagées pour guider la pondération. Par exemple, on peut restreindre la mesure de proximité à un sous-espace de dimension réduite, ce qui revient à supposer que seule une combinaison linéaire des composantes de x influence la loi de $Y|X = x$, c'est le principe des méthodes de réduction de dimension (linéaire). Plus généralement encore, on peut considérer une notion de proximité avec une distance sur une sous-variété de $\mathbb{R}^p$ (réduction de dimension non-linéaire), ou sur des structures plus complexes encore, lorsque l'information pertinente ne se trouve pas dans la totalité de l'espace initial. Dans les situations où l'on ne souhaite pas formuler a priori d'hypothèse forte sur la structure de régularité, on peut faire appel à des approches issues de l'apprentissage statistique (machine learning), telles que les forêts aléatoires ou les réseaux de neurones. Ces méthodes sont capables de détecter automatiquement des patterns complexes dans les données et peuvent, parfois de manière implicite, construire des pondérations adaptées à la structure locale des lois conditionnelles. Elles ne sont cependant pas dénuées d'hypothèses : leur efficacité repose sur des mécanismes de régularisation ou de pénalisation (par exemple profondeur des arbres, early stopping, pénalisation des poids, etc.), qui permettent de contrôler la complexité effective du modèle.

Exemple : estimation non paramétrique par histogramme Pour illustrer concrètement le principe de pondération dans le cadre conditionnel, considérons une approche simple par histogramme pour l'estimation d'un quantile conditionnel extrême. L'estimation d'un quantile intermédiaire peut être réalisée de manière indirecte, en commençant par estimer la fonction de survie conditionnelle de Y sachant $X = x$:

$$\overline{F}(y|x) = \mathbb{E}[\mathbb{I}\{Y > y\}|X = x].$$

Cette fonction qui est une espérance conditionnelle est approchée par une moyenne empirique pondérée en fonction de x :

$$\widehat{F}_n(y|x) := \sum_{i=1}^n W(x, X_i) \mathbb{I}\{Y_i > y\},$$

où les poids $W(x, X_i)$ sont définis par :

$$W(x, X_i) := \frac{\mathbb{I}\{\|X_i - x\| < h\}}{\sum_{j=1}^n \mathbb{I}\{\|X_j - x\| < h\}}.$$

Ici, $h > 0$ est un paramètre appelé fenêtre, et les poids sont non nuls uniquement pour les observations (Y_i, X_i) dont la covariable X_i appartient à la boule de centre x et de rayon h . Les poids ainsi définis sont positifs et leur somme vaut 1. Pour l'estimation de la fonction de survie conditionnelle, nous n'utilisons donc que les $n_x := \sum_{i=1}^n \mathbb{I}\{\|X_i - x\| < h\}$ observations d'une distance au moins h de x .

L'estimateur du quantile conditionnel intermédiaire d'ordre α_n est alors obtenu comme suit :

$$\widehat{Q}(\alpha_n|x) := \inf \left\{ t \in \mathbb{R} \mid \widehat{F}_n(t|x) \leq \alpha_n \right\} = Y_{n_x - \lfloor n_x \alpha_n \rfloor, n_x}^{(x)},$$

où $Y_{1, n_x}^{(x)} < \dots < Y_{n_x, n_x}^{(x)}$ sont les statistiques d'ordre des Y_i tels que X_i est dans la boule de rayon h autour de x .

Les paramètres $\gamma(x)$ et $a(\alpha_n^{-1}|x)$ peuvent ensuite être estimés à l'aide de statistiques de moments locales, définies à partir des n_x observations sélectionnées

$$\widehat{\gamma}_n^{(M)}(x) := M_n^{(1)}(x) + 1 - \frac{1}{2} \left(1 - \frac{\{M_n^{(1)}(x)\}^2}{M_n^{(2)}(x)} \right)^{-1},$$

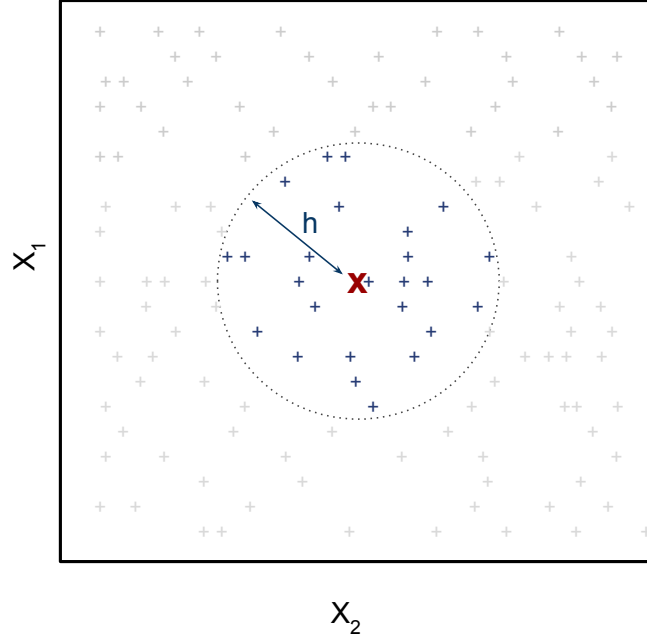


FIGURE 1 – Schéma de la sélection des points en dimension $p = 2$.

et

$$\hat{a}_n^{(M)}(\alpha_n^{-1}) := Y_{n_x - \lfloor n_x \alpha_n \rfloor, n_x}^{(x)} \cdot M_n^{(1)}(x) \cdot (1 - \hat{\gamma}_n^{(M)}(x) + M_n^{(1)}(x)),$$

avec

$$M_n^{(j)}(x) := \frac{1}{\lfloor n_x \alpha_n \rfloor} \sum_{i=1}^{\lfloor n_x \alpha_n \rfloor} \left(\ln Y_{n_x - i + 1, n_x}^{(x)} - \ln Y_{n_x - \lfloor n_x \alpha_n \rfloor, n_x}^{(x)} \right)^j, \quad j \in \{1, 2\}.$$

Fléau de la grande dimension

Le choix de la fenêtre h dans les méthodes à noyau a un impact crucial sur la qualité de l'estimation. Plus la fenêtre h est grande, plus le nombre d'observations sélectionnées pour l'estimation est élevé; en contrepartie, ces observations sont potentiellement plus éloignées du point cible x . À l'inverse, une fenêtre plus petite sélectionne uniquement les observations très proches de x , au prix d'un nombre réduit de données utilisables.

On se retrouve ici dans une situation classique de compromis biais-variance : une grande fenêtre h introduit un biais important, car les observations lointaines suivent des lois trop différentes de celle de $Y|X = x$; à l'inverse, une fenêtre trop petite engendre une variance élevée, car elle repose sur peu d'observations. Pour assurer une convergence asymptotique de l'estimateur (c'est-à-dire un biais qui tend vers zéro), il est nécessaire que $h = h_n \rightarrow 0$ lorsque $n \rightarrow \infty$. En pratique, cela signifie que le choix de la fenêtre doit dépendre de la taille de l'échantillon : plus n est grand, plus on peut se permettre de prendre une fenêtre petite tout en conservant suffisamment d'observations proches de x .

La dimension p de la covariable X joue un rôle majeur dans la performance de l'estimation. En effet, la vitesse de convergence typique d'un estimateur à noyau d'un quantile extrême conditionnel est de l'ordre de $\sqrt{n\alpha_n h^p}$. Cela signifie que, toutes choses égales par ailleurs, une augmentation de la dimension p ralentit exponentiellement la convergence. Ce phénomène est connu sous le nom de fléau de la grande dimension (curse of dimensionality), un obstacle bien documenté en statistique. Intuitivement, dans un espace de grande dimension, les observations X_i sont plus dispersées : la densité de points autour de n'importe quel x devient très faible. Ainsi, la probabilité qu'une réalisation X_i tombe dans la boule de rayon h centrée en x diminue rapidement lorsque p augmente. En conséquence, même avec un échantillon de taille importante, le nombre d'observations réellement utiles pour estimer un quantile de $Y|X = x$ reste faible. Pour maintenir un niveau de précision constant lorsque p augmente, il faut donc accroître de manière exponentielle la taille de l'échantillon. La Figure 2 illustre ce phénomène : elle représente le nombre d'observations n nécessaire pour que $nh^p = 100$ en fonction de la dimension p , pour plusieurs choix de h . On observe une croissance très rapide du nombre d'observations requises. Cet effet rend l'estimation non fiable, voire impossible, dans de nombreux cas pratiques. La prise en compte du fléau de la dimension est donc indispensable.

Pour faire face aux difficultés posées par l'estimation de quantiles en grande dimension, plusieurs stratégies ont été proposées dans la littérature. Parmi elles, on peut citer les méthodes issues du machine learning, telles que les réseaux de neurones, les forêts aléatoires, le gradient boosting ou encore les modèles de régression pénalisée (comme la régression Lasso ou la régression ridge). Elles s'appuient généralement sur une phase d'apprentissage super-

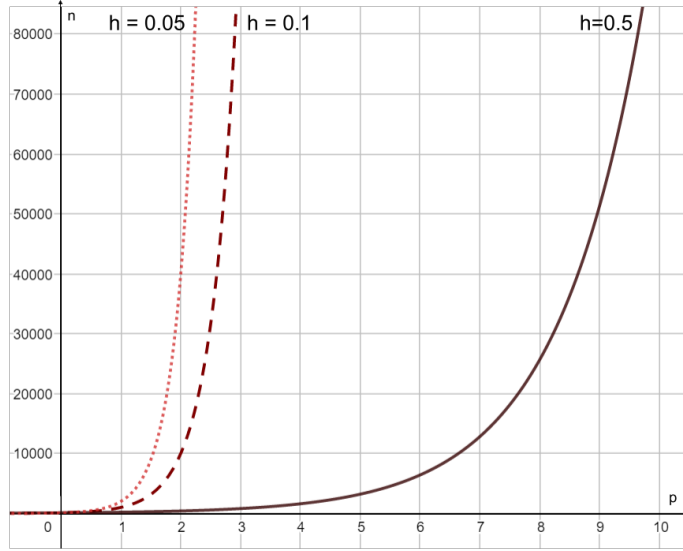


FIGURE 2 – Nombre n d’observations nécessaires pour avoir $nh^p = 100$ en fonction de p , pour $h = 0,5$, $h = 0,1$ et $h = 0,05$.

visé, au cours de laquelle l’algorithme apprend à partir d’un ensemble de données d’entraînement à prédire une grandeur cible en minimisant un critère de perte. Ces méthodes ne reposent pas sur des hypothèses fortes a priori sur la structure de la relation entre les variables, car elles sont capables d’inférer cette structure (si elle existe) automatiquement. Elles nécessitent cependant toujours un procédé de régularisation ou de pénalisation pour garantir de bonnes propriétés de généralisation.

Une autre approche est d’au contraire supposer l’hypothèse structurelle forte qu’il existe un sous-espace linéaire de faible dimension $\mathcal{S} \subset \mathbb{R}^p$ contenant toute l’information pertinente de X pour l’estimation de la quantité cible. Autrement dit, on suppose que la dépendance de X à la quantité cible passe uniquement par une projection $B^\top X$, où $B \in \mathbb{R}^{p \times q}$ est une matrice de rang $q \ll p$. On peut citer par exemple dans le cadre de la régression classique les méthodes Sliced Inverse Regression (SIR) [51] ou la Sliced Average Variance Estimation, (SAVE) [22]. Cette hypothèse structurelle permet d’utiliser des techniques non paramétriques classiques — telles que les méthodes, par histogramme, à noyau ou les k plus proches voisins — dans l’espace réduit une fois celui-ci identifié (voir Figure 3). On peut ainsi contourner en partie les

limitations imposées par la grande dimension. En outre, ces méthodes ont l'avantage de fournir une meilleure interprétabilité des relations entre les co-variables et la variable cible.

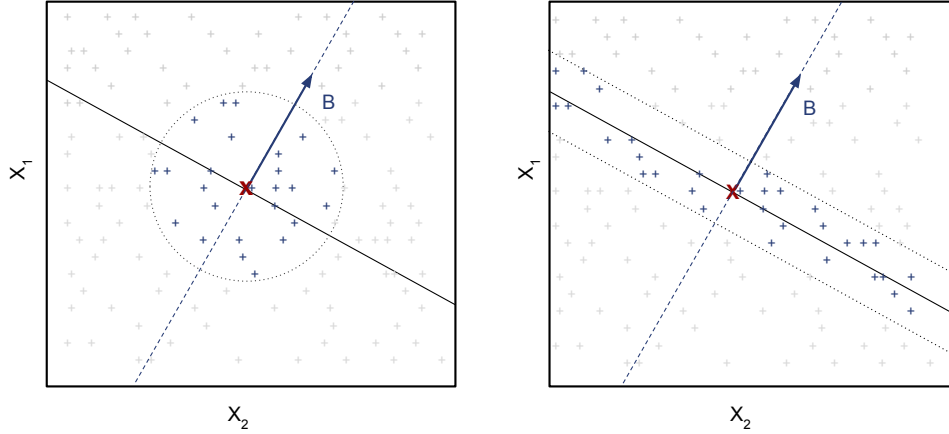


FIGURE 3 – Comparaison de sélection des points dans la méthode par histogramme sans (à gauche) et avec (à droite) réduction de dimension.

Plan du manuscrit et contributions

Ce manuscrit s'articule autour de deux problématiques principales, toutes deux motivées par les limites actuelles des méthodes d'estimation des quantiles extrêmes dans des contextes réalistes.

La première concerne la construction d'intervalles de confiance dans le cadre non conditionnel. Si l'estimation ponctuelle de ces quantiles est bien étudiée, la quantification de l'incertitude reste délicate.

La seconde problématique traite du cadre conditionnel en grande dimension, dans lequel l'estimation de quantiles extrêmes devient particulièrement difficile lorsque la variable explicative X appartient à un espace de grande dimension. Ce contexte souffre du fléau de la dimension, d'autant plus que seul un petit pourcentage des plus grandes observations sont utilisées pour l'estimation des quantiles extrêmes.

Le manuscrit est structuré selon les trois articles de recherche qui en constituent la base.

Chapitre 1 : Intervalles de confiance asymptotiques pour les quantiles extrêmes

L'article présenté dans ce chapitre a fait l'objet d'une publication dans la revue *Electronic Journal of Statistics* :

Gardes, L., Maistre, S., and Podgorny, A. (2025). Asymptotic confidence intervals for extreme quantiles in a maximum domain of attraction. *Electronic Journal of Statistics*, 19(1), 1103-1132.

Ce premier article s'intéresse à la construction d'intervalles de confiance pour les quantiles extrêmes dans le cas non conditionnel. La méthode proposée constitue une généralisation de celle introduite par Gardes et Maistre [35], qui fournit des intervalles pour les quantiles extrêmes uniquement dans le cas $\gamma > 0$. Ici, nous étendons cette approche au cas général $\gamma \in \mathbb{R}$, permettant ainsi de couvrir les trois domaines classiques des lois extrêmes (Fréchet, Gumbel et Weibull). Contrairement à d'autres méthodes basées sur l'approximation normale de l'estimateur du quantile, cette approche repose sur l'extrapolation de statistiques d'ordre. Nous décrivons ici brièvement la procédure utilisée pour construire la borne supérieure de l'intervalle de confiance de niveau $\nu \in (0, 1)$, notée $R_{n,\nu}$. La même approche peut être utilisée de manière symétrique pour la borne inférieure $L_{n,\nu}$.

L'idée est de partir d'une statistique d'ordre extrême $X_{n-j_R,n}$, où j_R est un entier fixé, indépendant de n . Pour un niveau β_n proche de zéro, cette statistique d'ordre n'est en général pas suffisante pour dépasser le quantile extrême $Q(\beta_n)$. Il est donc nécessaire de l'extrapoler à l'aide d'un facteur multiplicatif $t_{n,R} > 0$ de sorte que

$$\mathbb{P}(t_{n,R}X_{n-j_R,n} \geq Q(\beta_n)) \approx \frac{1 + \nu}{2}.$$

Cette construction permet d'obtenir, avec un niveau de confiance asymptotique contrôlé, une borne supérieure pour le quantile extrême.

Il existe des variables aléatoires indépendantes $U_1, \dots, U_n$ de loi uniforme sur $[0, 1]$, telles que le vecteur des statistiques d'ordre $(X_{n-j,n})_{j=0}^{n-1}$ admet la même loi que le vecteur $(Q(U_{j+1,n}))_{j=0}^{n-1}$. Ainsi, pour tout entier j_R , on a l'égalité

$$\mathbb{P}(X_{n-j_R,n} \geq Q(\beta_n)) = \mathbb{P}(U_{j_R+1,n} \leq \beta_n) = F_{\text{Beta}}^{(j_R+1, n-j_R)}(\beta_n),$$

où $F_{\text{Beta}}^{(r,s)}$ désigne la fonction de répartition de la loi Beta de paramètres $r > 0$ et $s > 0$. Cela suggère que le facteur optimal d'extrapolation est donné par

$$t_{n,R} = \frac{Q(\beta_n)}{Q(\mathbf{a}_{n,R}(\nu))},$$

où

$$\mathbf{a}_{n,R}(\gamma) := q_{\text{Beta}}^{(j_R+1, n-j_R)} \left(\frac{1+\nu}{2} \right),$$

et $q_{\text{Beta}}^{(r,s)}$ désigne la fonction quantile (la réciproque de $F_{\text{Beta}}^{(r,s)}$).

Le facteur inconnu $t_{n,R}$ est alors estimé en remplaçant Q_X par un estimateur consistant, comme par exemple l'estimateur des moments [\[3\]](#) :

$$\hat{t}_{n,R}^{(M)} = \frac{\hat{Q}_n^{(M)}(\beta_n)}{\hat{Q}_n^{(M)}(\mathbf{a}_{n,R}(\nu))}.$$

La borne supérieure de l'intervalle de confiance est donc obtenue par extrapolation de la statistique d'ordre $X_{n-j_R,n}$, multipliée par ce facteur estimé, $R_{n,\nu} = X_{n-j_R,n} \hat{t}_{n,R}^{(M)}$.

En résumé, les principales contributions de l'article sont :

- La proposition d'une nouvelle procédure de construction d'intervalles de confiance pour les quantiles extrêmes ;
- l'étude théorique de la convergence asymptotique de cet intervalle ;
- la réalisation de simulations montrant une bonne couverture pour un échantillon fini ;
- l'application à des données réelles de longévité humaine.

Chapitre 2 : Réduction de dimension pour l'estimation de l'indice de queue conditionnel

L'article présenté dans ce chapitre a été accepté pour publication dans la revue *Scandinavian Journal of Statistics* :

Gardes, L., & Podgorny, A. (2025). Dimension reduction for the estimation of the conditional tail-index, *Scandinavian Journal of Statistics*, to appear.

Dans ce second article, nous nous intéressons à la problématique de l'estimation de l'indice des valeurs extrêmes conditionnel $\gamma(x)$ sur un compact $\mathcal{X} \subset \mathbb{R}^p$ dans un cadre de grande dimension. Lorsque la distribution conditionnelle de la variable Y sachant $X = x$ est à queue lourde, l'estimation précise de $\gamma(x)$ est une étape essentielle pour l'estimation des quantiles extrêmes via l'extrapolation [9].

Ce travail propose une méthodologie originale permettant de contourner cette difficulté via une approche de réduction de dimension spécifiquement adaptée à l'estimation de $\gamma(x)$. D'autres méthodes de réduction de dimension existent dans le domaine des extrêmes, comme celle de l'espace TDR de Gardes [33] et la méthode TIREX proposée par Aghbalou, Portier, Sabourin, et Zhou [1]. Toutefois, la première méthode souffre d'un problème de calculabilité qui la rend inutilisable en grande dimension ($p > 6$), tandis que TIREX n'est pas spécifiquement conçu pour l'estimation des quantiles extrêmes.

Notre approche se base sur l'idée que la fonction de répartition conditionnelle $F_B(\cdot|B^\top x)$ de Y sachant $B^\top X = B^\top x$ peut être interprétée comme un mélange des fonctions de répartition $F(\cdot|X)$, dans la mesure où

$$F_B(y|B^\top X) = \mathbb{E}[F(y|X)|B^\top X].$$

Nous montrons alors sous certaines hypothèses supplémentaires que, pour tout $x \in \mathcal{X}$, la distribution conditionnelle de Y sachant $B^\top X = B^\top x$ reste à queue lourde, avec un indice de queue $\xi_B(B^\top x)$ qui correspond à la plus grande valeur de l'indice de queue apparaissant dans ce mélange, c'est-à-dire :

$$\xi_B(B^\top x) := \max_{z: B^\top z = B^\top x} \gamma(z). \quad (10)$$

Dans ce cadre, nous définissons le sous-espace de réduction appelé Tail-Index Dimension Reduction (TIDR) comme suit :

Définition 2. *Un sous-espace vectoriel $\mathcal{S} \subset \mathbb{R}^p$, engendré par une base $B \in \mathbb{R}^{p \times q}$, est dit TIDR si les lois conditionnelles $Y|X$ et $Y|B^\top X$ partagent le même indice de queue conditionnel, c'est-à-dire si*

$$\xi_B(B^\top X) = \gamma(X) \quad p.s.$$

On dit que $\mathcal{S}$ est un sous-espace central de type CTI (Central Tail Index) s'il est contenu dans tous les sous-espaces TIDR.

On suppose que le sous-espace CTI existe. Celui-ci est alors unique, en tant qu'intersection de tous les sous-espaces TIDR. On note $B_0 \in \mathcal{B}_q$ sa base canonique (voir Définition 2 du Chapitre 2), où $\mathcal{B}_q$ désigne l'ensemble des bases canoniques de tous les sous-espaces linéaires de dimension q . On a, pour tout $x \in \mathcal{X}$ et tout $B \in \mathcal{B}_q$:

$$\gamma(x) = \xi_{B_0}(B_0^\top x) \leq \xi_B(B^\top x) = \max_{z: B^\top z = B^\top x} \gamma(z),$$

car $\gamma(x)$ appartient au support de la distribution conditionnelle de $\gamma(X)$ sachant $B^\top X = B^\top x$.

Ainsi, pour un sous-ensemble compact $\mathcal{X}_0$ inclus dans l'intérieur de $\mathcal{X}$, satisfaisant la condition suivante :

(C) Pour tout $B \in \mathcal{B}_q$, si $\xi_B(B^\top x) = \gamma(x)$ pour tout $x \in \mathcal{X}_0$, alors $B = B_0$, on obtient que :

$$\arg \min_{B \in \mathcal{B}_q} \mathbb{E} [\xi_B(B^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}] := \arg \min_{B \in \mathcal{B}_q} \Psi(B, \mathcal{X}_0) = \{B_0\}.$$

Sous la condition (C), nous proposons d'estimer B_0 par la base canonique $\hat{B}_n^{(\text{CTI})}$ minimisant la version empirique de $\Psi(B, \mathcal{X}_0)$ donnée par :

$$\hat{\Psi}_n(B, \mathcal{X}_0) := \frac{1}{n} \sum_{i=1}^n \hat{\xi}_{n,B}(B^\top X_i) \mathbb{I}\{X_i \in \mathcal{X}_0\}.$$

La fonction $x \mapsto \hat{\xi}_{n,B}(B^\top x)$ est un estimateur de $\xi_B(B^\top x)$, l'indice de queue conditionnel de Y sachant $B^\top X$. Lorsque la dimension du sous-espace CTI est $q \ll p$, l'indice $\xi_B(B^\top x)$ peut être estimé efficacement par des techniques à noyau, comme nous le proposons dans la section 3 de l'article.

En résumé, les principales contributions de l'article sont :

- la définition d'un nouveau sous-espace de dimension réduite (CTI subspace) adapté à l'estimation de l'indice de queue conditionnel ;
- la construction d'un estimateur de ce sous-espace et la preuve de sa consistance ;
- la construction et l'étude théorique d'un estimateur de l'indice de queue construit sur l'espace de dimension réduite ;
- la proposition d'une méthode d'estimation de la dimension de l'espace CTI ;
- La comparaison avec d'autres approches de réduction de dimension comme celles de Gardes [33] et TIREX [1].

Chapitre 3 : Comparaison de méthodes récentes pour l'estimation de quantiles extrêmes conditionnels

L'article présenté dans ce chapitre a été soumis à la revue *International Statistical Review* :

L. Gardes and A. Podgorny (2025). Extreme Conditional Quantile Estimation in High Dimensions : A Comparative Study, *Preprint*.

Ce troisième article propose une étude comparative approfondie de méthodes récentes permettant l'estimation de quantiles extrêmes conditionnels dans des contextes de grande dimension. Nous nous intéressons en particulier à l'estimation des paramètres de queue $\gamma(x)$ et $a(\alpha_n^{-1}|x)$. L'estimation du quantile intermédiaire $Q(\alpha_n|x)$ est elle faite par forêts aléatoires généralisées (Generalized Random Forests), en raison de son efficacité prouvée pour estimer des quantiles conditionnels non-extrêmes dans des contextes non paramétriques de grande dimension.

Plusieurs méthodes de machine learning et de réduction de dimension sont considérées :

- **Méthodes de machine learning :**
 - *Extremal Random Forest (ERF)* [38], qui repose sur une estimation pondérée par forêts aléatoires de la densité de Pareto généralisée des excès.
 - *Extremal Neural Network (EQRN)* [55], qui emploie un réseau de neurones pour estimer directement les paramètres de queue via une fonction de perte donnée par l'approximation GPD.
 - Pour chaque méthode, une version dédiée au cas à queue lourde ($\gamma(x) > 0$) est également étudiée, en se fondant sur la convergence des log-excès vers une loi exponentielle.
- **Méthodes à réduction de dimension spécifique aux extrêmes :**
 - *TIREX* [1].
 - *CTI* [36], que nous avons introduite dans le deuxième article.

La contribution principale de cet article est une étude comparative extensive par simulation de ces approches, dans une grande variété de modèles. Ce travail apporte un éclairage nouveau et rigoureux sur les avantages comparés des approches modernes d'estimation de quantiles extrêmes conditionnels. Nous évaluons notamment :

- la sensibilité des méthodes au choix du niveau intermédiaire α_n ;

- leur robustesse à la dimension p de l'espace des covariables ;
- l'impact du comportement de second ordre de la loi conditionnelle (paramètre $\rho(x)$) ;
- la complexité structurelle de la fonction $x \mapsto \gamma(x)$ (linéaire, non linéaire).

Les résultats des simulations menées dans cet article montrent que les méthodes évaluées sont globalement robustes à la dimension p des covariables, ce qui les rend adaptées aux contextes de grande dimension. Toutefois, leur sensibilité au choix du niveau intermédiaire α_n varie selon les méthodes et les modèles, ce paramètre doit donc être sélectionné avec précaution. Le comportement de second ordre, caractérisé par le paramètre $\rho(x)$, a également un impact significatif sur la performance des estimateurs. Enfin, les méthodes par réduction de dimension sont plus efficaces lorsque la structure de $\gamma(x)$ est linéaire. En présence d'une structure non linéaire simple, les approches issues du machine learning, peuvent offrir de meilleures performances. En revanche, lorsque la structure devient trop complexe ou peu régulée, aucune méthode ne parvient à produire une estimation fiable.

Chapitre 1

Asymptotic confidence intervals for extreme quantiles in a maximum domain of attraction

Gardes, L., Maistre, S., and Podgorny, A. (2025). Asymptotic confidence intervals for extreme quantiles in a maximum domain of attraction. *Electronic Journal of Statistics*, 19(1), 1103-1132.

Les quantiles extrêmes sont couramment utilisés pour évaluer les risques rares tels que les crues majeures, les températures extrêmes ou les crises financières, entre autres. Lorsque la loi sous-jacente appartient à un domaine d'attraction maximal, l'estimation des quantiles extrêmes a été largement étudiée dans la littérature. Cependant, de manière surprenante, peu de travaux ont été consacrés à la construction d'intervalles de confiance asymptotiques pour ces quantiles. C'est précisément la question abordée dans cet article, dans lequel nous proposons une méthode de construction d'un intervalle de confiance pour un quantile extrême. Cette méthode est applicable à un grand nombre de lois appartenant à un domaine d'attraction. La convergence de la probabilité de couverture vers le niveau fixé est démontrée, et les performances en échantillon fini sont étudiées à travers une étude de simulation. Une application aux durées de vie des supercentenaires français est également présentée.

1.1 Introduction

Let X be a random variable defined on an atomless probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We denote by $F_X(\cdot) := \mathbb{P}(X \leq \cdot)$ its cumulative distribution function and by $\bar{F}_X = 1 - F_X$ its survival function. Let $\mathcal{S} \subset [0, \infty)$ be the support of the distribution. The associated quantile is given for $\beta \in (0, 1)$ by

$$Q_X(\beta) := F_X^{\leftarrow}(1 - \beta) = \inf\{x \in \mathcal{S} | F_X(x) \geq 1 - \beta\}.$$

Throughout this paper, we assume that F_X is a continuous and strictly increasing function so that $F_X^{\leftarrow}$ coincides with the classical inverse. In various fields of application, large quantiles are commonly used to assess the risk of extreme events such as large floods, extreme temperature, financial crisis among many others. More details can be found in the monographs of Pisarenko and Rodkin [57] and Embrechts et al. [29]. If one observes n independent copies of X , a quantile $Q_X(\beta_n)$ is said to be large if $\beta_n \rightarrow 0$ as $n \rightarrow \infty$. When β_n does not converge too fast to 0, i.e., $n\beta_n \rightarrow \infty$, the quantile can be simply estimated by an order statistic. A more challenging question arises when $n\beta_n \rightarrow c \geq 0$, i.e., when $Q_X(\beta_n)$ is an extreme quantile. In this situation a precise knowledge of the shape of the tail distribution is required. In this paper, we assume that F_X belongs to a maximum domain of attraction (MDA) that is to say there exist sequences $(a_n) > 0$ and (b_n) and a real γ such that

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{\max(X_1, \dots, X_n) - b_n}{a_n} \leq x \right) = \exp \left(-(1 + \gamma x)^{-1/\gamma} \right),$$

for all x satisfying $1 + \gamma x > 0$. This convergence in distribution can be reformulated in terms of the quantile function Q_X . More specifically, F_X belongs to an MDA if and only if there exist a parameter $\gamma \in \mathbb{R}$ and an auxiliary function $a : [1, \infty) \rightarrow (0, \infty)$ such that for all $t > 0$,

$$\lim_{\alpha \rightarrow 0_+} \frac{Q_X(t\alpha) - Q_X(\alpha)}{a(\alpha^{-1})} = K_\gamma(t^{-1}) := \int_1^{t^{-1}} s^{\gamma-1} ds, \quad (1.1)$$

see for instance de Haan and Ferreira [39, Theorem 1.1.6]. The parameter γ , called the extreme value index, plays an important role in modeling extreme events since it controls the tail heaviness of the distribution. More precisely, if $\gamma < 0$, the right endpoint $Q_X(0)$ is finite and $\bar{F}_X$ decreases at a polynomial

rate. When $\gamma = 0$, the survival function decreases at an exponential rate and when $\gamma > 0$, the distribution is said to be heavy-tailed i.e., for all $t > 0$,

$$\lim_{x \rightarrow \infty} \frac{\overline{F}_X(tx)}{\overline{F}_X(x)} = t^{-1/\gamma}.$$

In other words, the distribution is heavy-tailed if and only if the function $\overline{F}_X$ is a regularly varying function with index $-1/\gamma$ or equivalently, $\overline{F}_X(x) = x^{-1/\gamma}L(x)$ where L is a slowly varying function that is to say, for all $t > 0$, the ratio $L(tx)/L(x)$ converges to 1 as $x \rightarrow \infty$.

The estimation of extreme quantiles has been widely studied in the literature. In the case of heavy-tailed distributions, a widely used estimator was proposed by Weissman [74]. When no assumption is made on the sign of the tail-index, one can cite the estimator introduced in Dekkers and de Haan [26] and studied in de Haan and Rootzén [41] or the one defined in Daouia et al. [25]. Approximation of high quantiles has been considered in de Valk [68] and a bias reduced quantile estimator was introduced in de Haan et al. [40]. One can also cite the estimator proposed by El Methni et al. [28] which focuses on heavy and light-tailed distributions.

An alternative to point estimation of extreme quantiles is the construction of two-sided confidence intervals. More specifically, for a preselected level $\nu \in (0, 1)$, we are searching for positive random variables $L_{n,\nu}$ and $R_{n,\nu}$ such that

$$\lim_{n \rightarrow \infty} \mathbb{P}(L_{n,\nu} \leq Q_X(\beta_n)) = \lim_{n \rightarrow \infty} \mathbb{P}(R_{n,\nu} \geq Q_X(\beta_n)) = \frac{1 + \nu}{2}.$$

If $\mathbb{P}(R_{n,\nu} < L_{n,\nu}) \rightarrow 0$ as $n \rightarrow \infty$, the random interval $[L_{n,\nu}, R_{n,\nu}]$ is then an asymptotic two-sided confidence interval of level ν for $Q_X(\beta_n)$.

For non extreme quantile, i.e., when $\beta_n = \beta$ is fixed, several approaches can be considered for the construction of a confidence interval for quantiles. A common procedure is to use the asymptotic distribution of a given quantile estimator, see for instance Fan and Liu [30]. Other standard methods to construct confidence intervals are profile likelihood based intervals, see Cox and Snell [24] and resampling confidence intervals, see Kocherginsky et al. [49] and Parzen et al. [54] among others.

Despite of its interest, little work has been dedicated to the construction of an asymptotic confidence interval when $\beta_n \rightarrow 0$. This is precisely the question addressed in this paper. A first approach is to use an estimator of $Q_X(\beta_n)$ as a pivotal statistic. A general estimator of extreme quantiles can be found in [39, Section 4.1]. It is given by

$$\hat{Q}_{X,n}^{(HF)}(\beta_n) := \hat{b}_n(k_n) + \hat{a}_n(n/k_n) \frac{(k_n/(n\beta_n))^{\hat{\gamma}_n} - 1}{\hat{\gamma}_n}, \quad (1.2)$$

where, for a sequence of integers (k_n) such that $k_n \rightarrow \infty$ and $n/k_n \rightarrow \infty$ as $n \rightarrow \infty$, $\hat{b}_n(k_n)$ is an estimator of the quantile $Q_X(k_n/n)$, $\hat{a}_n(n/k_n)$ an estimator of the auxiliary function a appearing in (1.1) evaluated at point n/k_n and $\hat{\gamma}_n$ an estimator of the extreme value index γ . Under additional assumptions, it can be shown (see [39, Theorem 4.3.1]) that

$$\frac{k_n^{1/2}}{\hat{a}_n(n/k_n) \int_1^{k_n/(n\beta_n)} s^{\hat{\gamma}_n-1} \ln(s) ds} \left(\hat{Q}_{X,n}^{(HF)}(\beta_n) - Q_X(\beta_n) \right),$$

converges to a centered normal distribution with variance $v^2(\gamma)$ where v is a known function. We derive from this asymptotic result the confidence interval of level ν for $Q_X(\beta_n)$ given by

$$\text{CI}^{(HF)}(\nu, \beta_n) := \left[\hat{Q}_{X,n}^{(HF)}(\beta_n) \pm k_n^{-1/2} \text{q}_N \left(\frac{1+\nu}{2}; 0, v^2(\hat{\gamma}_n) \right) \hat{v}_n(\beta_n) \right], \quad (1.3)$$

where $\text{q}_N(u; \mu, \sigma^2)$ is the quantile function of a normal distribution with mean μ and variance σ^2 and $\hat{v}_n(\beta_n) = \hat{a}_n(n/k_n) \int_1^{k_n/(n\beta_n)} s^{\hat{\gamma}_n-1} \ln(s) ds$. A similar approach is used in Buitendag et al. [16] to construct an asymptotic confidence interval for extreme quantiles in the particular case where the underlying distribution is heavy-tailed. Although this approach is natural, it has some drawbacks. First, the asymptotic variance is unknown and needs to be estimated and, more importantly, the finite sample distribution of the pivotal statistic can be quite different from the asymptotic one.

To overcome these drawbacks, a different methodology, not based on the asymptotic distribution of a quantile estimator, was proposed in Gardes [34] and Gardes and Maistre [35]. To be self contained, let us briefly recall the procedure used in these works. Let $X_1, \dots, X_n$ be a sample of n independent replications of X such that (1.1) holds for the quantile function Q_X . Let

$X_{1,n} \leq \dots \leq X_{n,n}$ be the sample arranged in ascending order. Since the probability space is atomless, there exist independent standard uniform random variables $U_1, \dots, U_n$ such that the random vector $(X_{n-j,n}, j = 0, \dots, n-1)$ is distributed as the random vector $(Q_X(U_{j+1,n}), j = 0, \dots, n-1)$. Hence, for any integers $j \in \{0, \dots, n-1\}$, it is readily seen that

$$\mathbb{P}(X_{n-j,n} \geq Q_X(\beta_n)) = \mathbb{P}(U_{j+1,n} \leq \beta_n) = F_{\text{Beta}}^{(j+1, n-j)}(\beta_n), \quad (1.4)$$

where $F_{\text{Beta}}^{(r,s)}$ is the cumulative distribution function of a Beta distribution with parameters $r > 0$ and $s > 0$.

When β_n does not converge too fast to 0, i.e., when $n\beta_n \rightarrow \infty$, it is shown in Gardes [34] that one can find integers $j_R < j_L \in \{0, \dots, n-1\}$ (that may depend on n) such that

$$\mathbb{P}(X_{n-j_R,n} \geq Q_X(\beta_n)) \rightarrow \frac{1+\nu}{2} \text{ and } \mathbb{P}(X_{n-j_L,n} \leq Q_X(\beta_n)) \rightarrow \frac{1+\nu}{2},$$

so that $[X_{n-j_L,n}, X_{n-j_R,n}]$ is a two-sided confidence interval for $Q_X(\beta_n)$.

Unfortunately, for an extreme quantile ($n\beta_n \rightarrow c \geq 0$) and for a given probability ν , the existence of such integers j_L and j_R is not guaranteed. To illustrate this point, let us consider for instance the case $\beta_n = 1/(2n)$. We prove in Lemma 1 that for all $j \in \{0, \dots, n-1\}$

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_{n-j,n} \geq Q_X(\beta_n)) = \lim_{n \rightarrow \infty} F_{\text{Beta}}^{(j+1, n-j)}(\beta_n) = F_{\text{Gamma}}^{(j+1, 1)}(1/2),$$

where $F_{\text{Gamma}}^{(r,s)}$ is the cumulative distribution function of a Gamma distribution with parameters $r > 0$ and $s > 0$. Since $F_{\text{Gamma}}^{(j+1, 1)}(1/2) \leq F_{\text{Gamma}}^{(1, 1)}(1/2) \approx 0.39 < (1+\nu)/2$ for all $\nu \in (0, 1)$, there is no j_R such that $\mathbb{P}(X_{n-j_R,n} \geq Q_X(\beta_n)) \rightarrow (1+\nu)/2$. To obtain a confidence interval in the extreme case, a solution consists in finding positive extrapolation factors $t_{n,L}$ and $t_{n,R}$ in order to have

$$\mathbb{P}(t_{n,R}X_{n-j_R,n} \geq Q_X(\beta_n)) \approx \mathbb{P}(t_{n,L}X_{n-j_L,n} \leq Q_X(\beta_n)) \approx (1+\nu)/2,$$

for some j_L and j_R in $\{0, \dots, n-1\}$. Plugging estimators $\widehat{t_{n,L}}$ and $\widehat{t_{n,R}}$ of these extrapolation factors leads to the confidence interval $[\widehat{t_{n,L}}X_{n-j_L,n}, \widehat{t_{n,R}}X_{n-j_R,n}]$. This procedure is used in Gardes and Maistre [35] with $j_L = j_R = j_n \rightarrow \infty$ to construct a confidence interval for extreme quantiles. The coverage probability of this interval converges to the nominal level only under the restrictive condition that the underlying distribution belongs to a subspace of

heavy-tailed distributions for which $\bar{F}_X(x) = x^{-1/\gamma}L(x)$ with $\gamma > 0$ and $L(x) \rightarrow d > 0$ as $x \rightarrow \infty$.

The novelty of this work lies in overcoming this restriction by proposing different extrapolation factors that can be used for both heavy-tailed distributions and distributions satisfying (1.1) with a non positive extreme value index γ . Of course, as shown in the main result of this paper, the required conditions to obtain a consistent confidence interval strongly depend on γ . In particular, when $\gamma < -1/2$, it seems quite difficult to construct a consistent confidence interval. Although the confidence intervals proposed in Gardes and Maistre [35] and in this work are constructed using a similar methodology, they are nevertheless very different since a different technique is required to estimate the extrapolation factors. More importantly, the interval proposed in this paper has a much broader scope, allowing us to study data whose distribution is not necessarily heavy-tailed as it is the case for the dataset studied in Section 1.5. A more detailed comparison of these two intervals is made in Section 1.2.

The paper is organized as follows. The definition of the proposed confidence interval is given in Section 1.2 and the convergence of its coverage probability to ν is established under general conditions. The estimation of the extrapolation factors is considered in Section 1.3. First, a general estimation procedure is presented in Section 1.3.1 where sufficient conditions for the convergence of the coverage probability to the nominal one are given. An application of this procedure is then proposed in Section 1.3.2. The finite sample behavior of the confidence interval is investigated on simulated data in Section 1.4 and on a real dataset in Section 1.5. All the proofs are gathered in Section 1.6.

1.2 Construction of the confidence interval

Let us consider a random variable X such that (1.1) holds for the quantile function Q_X . Let $X_1, \dots, X_n$ be a sample of n independent replications of X and let $X_{1,n} \leq \dots \leq X_{n,n}$ be the sample arranged in ascending order. As explained in the introduction, for given integers $0 \leq j_R \leq j_L \leq n-1$, we are searching for extrapolation factors $t_{n,L}$ and $t_{n,R}$ in order to have

$$\mathbb{P}(t_{n,R}X_{n-j_R,n} \geq Q_X(\beta_n)) \approx \mathbb{P}(t_{n,L}X_{n-j_L,n} \leq Q_X(\beta_n)) \approx (1 + \nu)/2.$$

Note that the integers j_L and j_R do not depend on n . Let us explain the procedure used to find the right extrapolation factor $t_{n,R}$. The same procedure is applied to find $t_{n,L}$. Using (1.4), we start by remarking that

$$\mathbb{P}(t_{n,R}X_{n-j_R,n} \geq Q_X(\beta_n)) = F_{\text{Beta}}^{(j_R+1,n-j_R)}(\bar{F}_X(t_{n,R}^{-1}Q_X(\beta_n))).$$

A good candidate is then the sequence $t_{n,R} := Q_X(\beta_n)/Q_X(\mathfrak{a}_{n,R}(\nu))$ where, denoting by $q_{\text{Beta}}^{(r,s)}$ the inverse of $F_{\text{Beta}}^{(r,s)}$,

$$\mathfrak{a}_{n,R}(\nu) := q_{\text{Beta}}^{(j_R+1,n-j_R)}\left(\frac{1+\nu}{2}\right).$$

Of course, the factor $t_{n,R}$ is unknown and needs to be estimated by any consistent estimator $\widehat{t_{n,R}}$, see Section 1.3. In the same way, we propose to take for the left extrapolation factor, the sequence $t_{n,L} := Q_X(\beta_n)/Q_X(\mathfrak{a}_{n,L}(\nu))$ where

$$\mathfrak{a}_{n,L}(\nu) := q_{\text{Beta}}^{(j_L+1,n-j_L)}\left(\frac{1-\nu}{2}\right).$$

Denoting by $\widehat{t_{n,L}}$ any consistent estimator of $t_{n,L}$, we propose the interval $[L_{n,\nu}, R_{n,\nu}]$ with $R_{n,\nu} = \widehat{t_{n,R}}X_{n-j_R,n}$ and $L_{n,\nu} = \widehat{t_{n,L}}X_{n-j_L,n}$ as an asymptotic confidence interval for $Q_X(\beta_n)$. It remains to check that $L_{n,\nu} < R_{n,\nu}$ with a probability as large as we want. Note that the fact that X is a positive random variable ensures that the extrapolation factors $t_{n,L}$ and $t_{n,R}$ are positive. From Lemma 1, we know that

$$\lim_{n \rightarrow \infty} \frac{\mathfrak{a}_{n,L}(\nu)}{\mathfrak{a}_{n,R}(\nu)} = \frac{q_{\text{Gam}}^{(j_L+1,1)}((1-\nu)/2)}{q_{\text{Gam}}^{(j_R+1,1)}((1+\nu)/2)},$$

where $q_{\text{Gam}}^{(r,s)}$ is the quantile function of a Gamma distribution with parameters $r > 0$ and $s > 0$. Hence, if the couple $(j_L, j_R) \in \{0, \dots, n-1\}^2$ is such that

$$\frac{q_{\text{Gam}}^{(j_L+1,1)}((1-\nu)/2)}{q_{\text{Gam}}^{(j_R+1,1)}((1+\nu)/2)} < 1 \quad \text{and} \quad j_L \geq j_R, \quad (1.5)$$

then, for n large enough, $t_{n,L}/t_{n,R} < 1 - \eta$ for some $\eta \in (0, 1)$. This entails that $\mathbb{P}(R_{n,\nu} < L_{n,\nu}) \rightarrow 0$ as $n \rightarrow \infty$. To sum up, we give below the definition of the two-sided asymptotic confidence interval for $Q_X(\beta_n)$ that is considered in all what follows.

Definition 1. Let X be a positive random variable such that (1.1) holds for the quantile function Q_X . Let $X_1, \dots, X_n$ be n independent replications of X . For a preselected probability $\nu \in (0, 1)$, a sequence $\beta_n \in (0, 1)$ and a couple $(j_L, j_R) \in \{0, \dots, n-1\}^2$ satisfying (1.5), the two-sided confidence interval for the extreme quantile $Q_X(\beta_n)$ is

$$\text{CI}(\nu, \beta_n) := \left[\widehat{t_{n,L}} X_{n-j_L,n}, \widehat{t_{n,R}} X_{n-j_R,n} \right],$$

where $\widehat{t_{n,L}}$ and $\widehat{t_{n,R}}$ are any consistent estimators of the ratios $t_{n,L}$ and $t_{n,R}$ with

$$t_{n,\bullet} := \frac{Q_X(\beta_n)}{Q_X(\mathbf{a}_{n,\bullet}(\nu))},$$

the symbol $\bullet$ being either L or R .

Comments

1. As mentioned before, condition (1.5) on the couple (j_L, j_R) ensures that $\text{CI}(\nu, \beta_n)$ is a proper interval. Note that this condition is satisfied for instance when $j_L = j_R$. The choice of the two integers j_L and j_R is left to the user. A possible choice is given in the simulation study.
2. In Gardes and Maistre [35], under the assumption that $x^{1/\gamma} \overline{F}_X(x) \rightarrow d > 0$ as $x \rightarrow \infty$ (or equivalently $\alpha^\gamma Q_X(\alpha) \rightarrow d^\gamma$ as $\alpha \rightarrow 0$), the following confidence interval for $Q_X(\beta_n)$ has been introduced :

$$\left[\check{t}_{n,-} X_{n-j_n,n}; \check{t}_{n,+} X_{n-j_n,n} \right], \quad (1.6)$$

where (j_n) is a sequence of integers such that $j_n \rightarrow \infty$ and $n/j_n \rightarrow \infty$ and, as in Definition 1, $\check{t}_{n,\pm}$ is an estimator of the extrapolation factor $t_{n,\pm} := Q_X(\beta_n)/Q_X(\tilde{\mathbf{a}}_{n,\pm}(\nu))$ with $\tilde{\mathbf{a}}_{n,\pm}(\nu) := q_{\text{Beta}}^{(j_n+1, n-j_n)}((1 \pm \nu)/2)$. Under the assumption that $\alpha^\gamma Q_X(\alpha) \rightarrow d^\gamma$ as $\alpha \rightarrow 0$, it is readily seen that $t_{n,\pm} \sim (\tilde{\mathbf{a}}_{n,\pm}(\nu)/\beta_n)^\gamma$. Naturally, for some consistent estimator $\hat{\gamma}_n$ of γ , Gardes and Maistre [35] propose to use $\check{t}_{n,\pm} := (\tilde{\mathbf{a}}_{n,\pm}(\nu)/\beta_n)^{\hat{\gamma}_n}$ as an estimator of the extrapolation factor. This estimator relies heavily on the assumption made on the underlying distribution and can no longer be used when (1.1) holds for some non positive extreme value index. A complete different technique must be used to estimate the extrapolation factors $t_{n,L}$ and $t_{n,R}$, see Section 1.3. The confidence interval (1.6) differs from $\text{CI}(\nu, \beta_n)$ in several other

points. First, the same large order statistics $X_{n-j_n,n}$ is used for the left and right bounds. This is not necessarily the case for the confidence interval proposed in this paper allowing for greater flexibility. Next, in (1.6), the sequence of integers (j_n) must converges to infinity while j_L and j_R are fixed (independent of n) in $\text{CI}(\nu, \beta_n)$. In fact, in Gardes and Maistre [35], the authors proposed to take $j_n = \max(3, \lfloor (\ln k_n)^{0.85} \rfloor)$ which converges to infinity at a very slow rate (for instance, for $n = 2,000$ with $k_n = n^{1/2}$, we have $j_n = 3$). Assuming that j_L and j_R are fixed seems more reasonable.

In the next result, we provide sufficient conditions on the estimators $\widehat{t_{n,L}}$ and $\widehat{t_{n,R}}$ ensuring that the asymptotic coverage probability of $\text{CI}(\nu, \beta_n)$ is equal to ν . Before that, let us remind that a sequence of real random variables Z_n is $\mathcal{O}_{\mathbb{P}}(1)$ if for all $\epsilon > 0$, there exist $N_\epsilon \in \mathbb{N}$ and $c_\epsilon > 0$ such that for all $n \geq N_\epsilon$, $\mathbb{P}(Z_n \in [-c_\epsilon, c_\epsilon]) \geq 1 - \epsilon$. In particular, if Z_n converges in distribution, then $Z_n = \mathcal{O}_{\mathbb{P}}(1)$.

Theorem 1. Assume that (1.1) holds for Q_X . If the extrapolation factors $\widehat{t_{n,L}}$ and $\widehat{t_{n,R}}$ in Definition 1 are such that $\sigma_n^{-1}(t_{n,\bullet}/\widehat{t_{n,\bullet}} - 1) = \mathcal{O}_{\mathbb{P}}(1)$ where (σ_n) is a positive sequence satisfying

$$\lim_{n \rightarrow \infty} \sigma_n \frac{Q_X(1/n)}{a(n)} = 0, \quad (1.7)$$

then the probabilities

$$\mathbb{P}\left(\widehat{t_{n,L}} X_{n-j_L,n} \leq Q_X(\beta_n)\right) \quad \text{and} \quad \mathbb{P}\left(\widehat{t_{n,R}} X_{n-j_R,n} \geq Q_X(\beta_n)\right),$$

both converge to $(1 + \nu)/2$ as $n \rightarrow \infty$.

From de Haan and Ferreira [39, Lemma 1.2.9], we have

$$\lim_{n \rightarrow \infty} \frac{a(n)}{Q_X(1/n)} = \max(0, \gamma).$$

Hence, condition (1.7) entails that the sequence σ_n converges to 0 and thus that $\widehat{t_{n,\bullet}}$ is a consistent estimator of $t_{n,\bullet}$ in the sense that the ratio $\widehat{t_{n,\bullet}}/t_{n,\bullet}$ converges in probability to 1. More specifically, if $\gamma > 0$, condition (1.7) reduces to $\sigma_n \rightarrow 0$. When $\gamma < 0$, we have that $Q_X(1/n) \rightarrow Q_X(0) =: x^* < \infty$ and, according to [39, Lemma 1.2.9 and Corollary 1.2.10], $a(n) \sim -\gamma\{x^* -$

$Q_X(1/n)\} = n^\gamma L(n)$, where L is a slowly varying function. So, when $\gamma < 0$, condition (1.7) takes the form $\sigma_n n^{-\gamma}/L(n) \rightarrow 0$. In particular, when $\gamma < -1/2$, the ratio $\widehat{t}_{n,\bullet}/t_{n,\bullet}$ must converge to 1 faster than the classical parametric rate of convergence $n^{1/2}$. We are not aware of any such estimators. So the construction of an asymptotic confidence interval for $Q_X(\beta_n)$ in this situation is an open problem.

1.3 Estimation of the extrapolation factors

Starting from a sample $X_1, \dots, X_n$ of independent copies of X , this section is dedicated to the estimation of the extrapolation factors $t_{n,L}$ and $t_{n,R}$. Recall that, for some preselected probability ν , these extrapolation factors are given by

$$t_{n,\bullet} := \frac{Q_X(\beta_n)}{Q_X(\mathfrak{a}_{n,\bullet}(\nu))}.$$

As before, we assume that the quantile function Q_X satisfies (1.1). Note that $Q_X(\mathfrak{a}_{n,\bullet}(\nu))$ is an extreme quantile since from Lemma 1, $n\mathfrak{a}_{n,\bullet}(\nu)$ converges to a positive constant.

1.3.1 General estimation procedure

The natural way to estimate $t_{n,\bullet}$ is to plug any consistent estimator $\widehat{Q}_{X,n}$ of $Q_X(\cdot)$ in the expression of $t_{n,\bullet}$ leading to the class of estimators

$$\widehat{t}_{n,\bullet}^{(*)} := \frac{\widehat{Q}_{X,n}(\beta_n)}{\widehat{Q}_{X,n}(\mathfrak{a}_{n,\bullet}(\nu))}. \quad (1.8)$$

For this class of estimators, we have the following version of Theorem 1.

Corollary 1. *Assume that (1.1) holds for Q_X and let $(\beta_n) \in (0, 1)$ be a sequence such that $n\beta_n \rightarrow c \geq 0$. If there exists a positive sequence $(\vartheta_n(\beta_n))$ with*

$$\lim_{n \rightarrow \infty} \frac{\vartheta_n(\beta_n)}{a(n)} \frac{Q_X(1/n)}{Q_X(\beta_n)} = 0,$$

and

$$(\vartheta_n(\beta_n))^{-1} \left(\widehat{Q}_{X,n}(\beta_n) - Q_X(\beta_n) \right) = \mathcal{O}_{\mathbb{P}}(1),$$

and a positive sequence ϑ_n with $\vartheta_n/a(n) \rightarrow 0$ and

$$\vartheta_n^{-1} \left(\hat{Q}_{X,n}(u_n) - Q_X(u_n) \right) = \mathcal{O}_{\mathbb{P}}(1),$$

for any sequence (u_n) such that nu_n converges to a positive constant, then the conclusion of Theorem [1](#) holds for $\hat{t}_{n,\bullet}^{(*)}$ instead of $\hat{t}_{n,\bullet}$.

1.3.2 Application : Moment Estimator

As mentioned in the introduction, a classical estimator of the extreme quantile $Q_X(u_n)$ where the sequence $(u_n) \in (0, 1)$ satisfies $nu_n \rightarrow c \geq 0$ as $n \rightarrow \infty$ is provided by $\hat{Q}_{X,n}^{(HF)}(u_n)$ which is defined in [\(1.2\)](#). Recall that $\hat{Q}_{X,n}^{(HF)}$ depends on estimators of the three quantities $Q_X(k_n/n)$, γ and $a(n/k_n)$. Several ways to obtain these estimators can be found in the literature. For instance, one can use a maximum likelihood approach (see [\[39\]](#), Section 4.3.1), the Probability-Weighted Moment approach, see Hosking et al. [\[46\]](#), etc. In this paper, we propose to use estimators based on the moment statistics

$$M_n^{(j)} := \frac{1}{k_n} \sum_{i=1}^{k_n} (\ln X_{n-i+1,n} - \ln X_{n-k_n,n})^j, \quad j \in \{1, 2\},$$

see [\[39\]](#), Section 4.3.2]. Note that we can apply the logarithm to the order statistics since X is a positive random variable. The quantile $Q_X(k_n/n)$ is estimated by the order statistic $X_{n-k_n,n}$, the extreme value index γ by

$$\hat{\gamma}_n^{(M)} := M_n^{(1)} + 1 - \frac{1}{2} \left(1 - \frac{\{M_n^{(1)}\}^2}{M_n^{(2)}} \right)^{-1},$$

which is the estimator introduced in Dekkers et al. [\[27\]](#) and $a(n/k_n)$ by

$$\hat{a}_n^{(M)} \left(\frac{n}{k_n} \right) := X_{n-k_n,n} M_n^{(1)} (1 - \hat{\gamma}_n^{(M)} + M_n^{(1)}),$$

Plugging these estimators in [\(1.2\)](#) leads to the Moment Estimator $\hat{Q}_{X,n}^{(M)}$ given by

$$\hat{Q}_{X,n}^{(M)}(u_n) := X_{n-k_n,n} + \hat{a}_n^{(M)} \left(\frac{n}{k_n} \right) \frac{(k_n/(nu_n))^{\hat{\gamma}_n^{(M)}} - 1}{\hat{\gamma}_n^{(M)}}. \quad (1.9)$$

To deal with the asymptotic behavior of the Moment Estimator, the rate of convergence in [\(1.1\)](#) needs to be specified. The following second-order condition is then introduced.

(SOC) There exist $\rho < 0$, $\gamma \neq \rho$ and functions $a > 0$ and A with $A(x) \rightarrow 0$ as $x \rightarrow \infty$ such that for all $t > 0$,

$$\lim_{\alpha \rightarrow 0} \frac{1}{A(\alpha^{-1})} \left(\frac{Q_X(t\alpha) - Q_X(\alpha)}{a(\alpha^{-1})} - K_\gamma(t^{-1}) \right) = H_{\gamma, \rho}(t^{-1}),$$

where $H_{\gamma, \rho}(s) := \{K_{\gamma+\rho}(s) - K_\gamma(s)\}/\rho$.

A detailed discussion on this condition can be found in the monographs of Beirlant et al. [10, Section 3.3] and de Haan and Ferreira [39, Section 2.3]. The second-order parameter $\rho < 0$, controls the rate of convergence in (1.1) : the larger the value of ρ , the slower the convergence. Note that when there exists an auxiliary function a such that

$$\frac{Q_X(t\alpha) - Q_X(\alpha)}{a(\alpha^{-1})} = K_\gamma(t^{-1}),$$

the second-order parameter ρ is by convention equal to $-\infty$. From [39, Lemma B.3.16], if Q_X satisfies the second-order condition (SOC) then, there exist $\rho' \in [\rho, 0]$ and a function B with $B(x) \rightarrow 0$ as $x \rightarrow \infty$ such that

$$\lim_{\alpha \rightarrow 0} \frac{1}{B(\alpha^{-1})} \left(\frac{\ln Q_X(t\alpha) - \ln Q_X(\alpha)}{a(\alpha^{-1})/Q_X(\alpha)} - K_{\gamma_-}(t^{-1}) \right) = H_{\gamma_-, \rho'}(t^{-1}), \quad (1.10)$$

where $\gamma_- := \min(0, \gamma)$. Note that the function B is regularly varying with index ρ' and, as a consequence, $A(x) = \mathcal{O}(B(x))$.

The asymptotic behavior of $\hat{Q}_{X,n}^{(M)}(u_n)$ has been established in de Haan and Ferreira [39, Section 4.3.2]. Let us recall the obtained result.

Proposition 2. *Assume that the second-order condition (SOC) holds. For any intermediate sequence (k_n) satisfying $k_n^{1/2} B(n/k_n) \rightarrow \lambda \in \mathbb{R}$ where B satisfies (1.10), we have, for any sequence (β_n) with $n\beta_n = o(k_n)$ and $\ln(n\beta_n) = o(k_n^{1/2})$,*

$$\{\vartheta_n^{(M)}(\beta_n)\}^{-1} \left(\hat{Q}_{X,n}^{(M)}(\beta_n) - Q_X(\beta_n) \right) = \mathcal{O}_{\mathbb{P}}(1),$$

where

$$\vartheta_n^{(M)}(\beta_n) = k_n^{-1/2} a\left(\frac{n}{k_n}\right) \int_1^{k_n/(n\beta_n)} s^{\gamma-1} \ln(s) ds.$$

The corresponding asymptotic confidence interval as defined in Definition [1](#) is given by

$$\text{CI}^{(M)}(\nu, \beta_n) := \left[\hat{t}_{n,L}^{(M)} X_{n-j_{L,n}}, \hat{t}_{n,R}^{(M)} X_{n-j_{R,n}} \right], \quad (1.11)$$

with

$$\hat{t}_{n,\bullet}^{(M)} = \frac{\hat{Q}_{X,n}^{(M)}(\beta_n)}{\hat{Q}_{X,n}^{(M)}(\mathbf{a}_{n,\bullet}(\nu))}, \quad \bullet \in \{L, R\}.$$

This interval is referred to as the Moment confidence interval. Note that if $n\beta_n$ converges to a positive constant then $Q_X(\beta_n)$ is asymptotically proportional to $Q_X(1/n)$ and

$$\vartheta_n^{(M)}(\beta_n) \propto \vartheta_n^{(M)} := k_n^{-1/2} a\left(\frac{n}{k_n}\right) \int_1^{k_n} s^{\gamma-1} \ln(s) ds,$$

where the symbol $\propto$ means *asymptotically proportional to*. Hence, according to Corollary [1](#), the random interval $\text{CI}^{(M)}(\nu, \beta_n)$ is an asymptotic two-sided confidence interval for $Q_X(\beta_n)$ if the condition

$$\lim_{n \rightarrow \infty} \frac{\vartheta_n^{(M)}(\beta_n)}{a(n)} \frac{Q_X(1/n)}{Q_X(\beta_n)} = 0, \quad (1.12)$$

holds together with the conditions of Proposition [2](#). Note that we do not need to know the asymptotic distribution of $\hat{Q}_{X,n}^{(M)}$ to construct the confidence interval $\text{CI}^{(M)}(\nu, \beta_n)$.

In the following result, we provide conditions on the sequences (k_n) and (β_n) ensuring that $\text{CI}^{(M)}(\nu, \beta_n)$ is an asymptotic two-sided confidence interval for $Q_X(\beta_n)$.

Proposition 3. *Assume that the second-order condition (SOC) holds for the quantile function Q_X .*

- i) *If $\gamma > -1/2$, conditions of Proposition [2](#) and [\(1.12\)](#) hold by taking for (k_n) any sequence asymptotically proportional to n^δ with $\delta = -2(\eta + \rho')/(1 - 2\rho')$ where $\eta \in (0, -\rho')$ can be taken as close as we want to 0 and for (β_n) any sequence asymptotically proportional to n^{-b} with $b > 1 - \delta$.*
- ii) *If $\gamma < -1/2$, there does not exist a sequence (k_n) for which condition [\(1.12\)](#) is satisfied.*

A clear conclusion for the moment estimator cannot be given in the situation where $\gamma = -1/2$. The validity of the confidence interval (1.11) depends on the underlying distribution.

When $\gamma < -1/2$, the moment estimator cannot be used to construct a valid confidence interval. As mentioned before, in this case, we need to use an estimator of $t_{n,\bullet}$ that converges faster than the parametric rate $n^{1/2}$. Up to our knowledge, such an estimator is not available in the literature.

1.4 Simulation study

In this section, we are interested in the behavior of the confidence interval $\text{CI}^{(M)}(\nu, \beta_n)$ for $Q_X(\beta_n)$ when a sample of size $n \in \{200; 500; 1,000; 2,000\}$ is observed. For this study, we take $\nu = .95$ and $\beta_n \in \{1/n, 1/(n \ln(n))\}$. Several distributions are considered. The two distributions below correspond to the case $\gamma > 0$ i.e., to heavy-tailed quantiles.

- **Absolute student distribution** — The quantile function is given for all $u \in (0, 1)$ by $Q_X(u) = q_t(1 - u/2; \mu)$, where $q_t(\cdot; \mu)$ is the quantile function of a Student distribution with $\mu > 0$ degree of freedom. The extreme-value index is $\gamma = 1/\mu$ and the second-order parameter is $\rho = -2/\mu$.
- **Burr distribution** — For $\kappa > 0$ and $d > 0$, the quantile function is given for all $u \in (0, 1)$ by $Q_X(u) = \{u^{-1/\kappa} - 1\}^{1/d}$. For this distribution, $\gamma = 1/(d\kappa)$ and $\rho = -1/\kappa$.

The extreme value index of the next three distributions is $\gamma = 0$. In this case, the distribution is said to be light-tailed.

- **Exponential distribution** — For all values of the rate parameter $\lambda > 0$, the second-order parameter is $\rho = -\infty$.
- **Absolute Normal distribution** — For $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, the quantile function is given for all $u \in (0, 1)$ by $Q_X(u) = q_N(u/2; \mu, \sigma^2)$, where $q_N(\cdot; \mu, \sigma^2)$ is the quantile function of a normal distribution with mean μ and variance σ^2 . The second-order parameter is $\rho = 0$ whatever the mean and the standard deviation of the distribution.
- **Log-Normal distribution** — For $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, the quantile function is given for all $u \in (0, 1)$ by $Q_X(u) = \exp\{q_N(u; \mu, \sigma^2)\}$, where $q_N(\cdot; \mu, \sigma^2)$ is the quantile function of a normal distribution with mean μ and variance σ^2 . The second-order parameter is $\rho = 0$.

Finally, the following distributions belong to an MDA with a negative ex-

treme value index.

- **Beta distribution** — For $\alpha > 0$ and $\beta > 0$, the probability density function is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$. The extreme value index is $\gamma = -1/\beta$ and the second-order parameter $\rho = -1/\beta$ if $\alpha \neq 1$.
- **Positive reverse Burr distribution** — For $\kappa > 0$ and $d > 0$, the quantile function is given for all $u \in (0, 1)$ by $Q_X(u) = 1 - \{2u^{-1/\kappa} - 1\}^{-1/d}$. The extreme value index and the second-order parameter are $\gamma = -1/(d\kappa)$ and $\rho = -1/\kappa$.

As explained in the introduction, a confidence interval $\text{CI}^{(HF)}(\nu, \beta_n)$ for $Q_X(\beta_n)$ can be derived from the asymptotic distribution of the moment estimator $\hat{Q}_{X,n}^{(M)}$ defined in (1.9). Its expression is given in (1.3) by taking $\hat{Q}_{X,n}^{(HF)} = \hat{Q}_{X,n}^{(M)}$, $\hat{a}_n(n/k_n) = \hat{a}_n^{(M)}(n/k_n)$, $\hat{\gamma} = \hat{\gamma}^{(M)}$ and

$$v^2(\gamma) = \begin{cases} 1 + \gamma^2, & \gamma \geq 0 \\ \frac{(1 - \gamma)^2(1 - 3\gamma + 4\gamma^2)}{(1 - 2\gamma)(1 - 3\gamma)(1 - 4\gamma)}, & \gamma < 0. \end{cases}$$

A natural way to measure the accuracy of $\text{CI}^{(\bullet)}(\nu, \beta_n) =: [L_{n,\nu}^{(\bullet)}, R_{n,\nu}^{(\bullet)}]$ where $\bullet \in \{M, HF\}$ is to compare the coverage probability

$$\text{CP}^{(\bullet)}(k_n) := \mathbb{P}([L_{n,\nu}^{(\bullet)}, R_{n,\nu}^{(\bullet)}] \ni Q_X(\beta_n)),$$

to the nominal level ν , leading to the error $\tilde{E}^{(\bullet)}(k_n) := |\text{CP}^{(\bullet)}(k_n) - \nu|$. The accuracy can also be evaluated by the quantity

$$E^{(\bullet)}(k_n) := \frac{1}{2} \left\{ \left| \nu_L^{(\bullet)}(k_n) - \frac{1 + \nu}{2} \right| + \left| \nu_R^{(\bullet)}(k_n) - \frac{1 + \nu}{2} \right| \right\},$$

where

$$\nu_L^{(\bullet)}(k_n) := \mathbb{P}(L_{n,\nu}^{(\bullet)} \leq Q_X(\beta_n)) \quad \text{and} \quad \nu_R^{(\bullet)}(k_n) := \mathbb{P}(R_{n,\nu}^{(\bullet)} \geq Q_X(\beta_n)).$$

Since it takes into account the errors of both the left and right bounds, $E^{(\bullet)}(k_n)$ is more suitable to measure the accuracy of a two-sided confidence interval. Indeed, $\tilde{E}^{(\bullet)}(k_n)$ is close to 0% if $\nu_L^{(\bullet)}(k_n) \approx 1$ and $\nu_R^{(\bullet)}(k_n) \approx 0.95$ which is not what is expected for a two-sided confidence interval. Conversely, in this situation, the quantity $E^{(\bullet)}(k_n) \times 100$ is close to 2.5%. This kind of situation was encountered in our simulation study, see the commentary

in the last paragraph of this section. We can also verify that $\tilde{E}^{(\bullet)}(k_n) \leq 2E^{(\bullet)}(k_n)$ so that if the quantity $E^{(\bullet)}(k_n)$ is close to 0, so is the common error $\tilde{E}^{(\bullet)}(k_n)$ (though the reverse is not necessarily true). We thus choose to focus on the quantity $E^{(\bullet)}(k_n)$ in our simulation study. The coverage probability $\text{CP}^{(\bullet)}(k_n)$, which is perhaps more intuitive, is also given but its value should be interpreted with caution. Finally, the true value of $E^{(\bullet)}(k_n)$ is obviously unknown but it can be approximated with a desired precision by means of a Monte Carlo (MC) procedure. In this study, the MC procedure is done by generating $N = 25,000$ independent samples.

Selection of the hyper-parameters

For the interval $\text{CI}^{(M)}(\nu, \beta_n)$, two integers j_L and j_R must be chosen. We propose to select them in order to minimize the extrapolation factors $\hat{t}_{n,L}^{(M)}$ and $\hat{t}_{n,R}^{(M)}$ in $\text{CI}^{(M)}(\nu, \beta_n)$. In other words, we take j_L and j_R so that $\mathbf{a}_{n,L}(.95) = q_{\text{Beta}}^{(j_L+1, n-j_L)}(.025)$ and $\mathbf{a}_{n,R}(.95) = q_{\text{Beta}}^{(j_R+1, n-j_R)}(.975)$ are as near as possible to β_n that is to say

$$j_L = \arg \min_{j \in \mathbb{N}} \left| q_{\text{Beta}}^{(j+1, n-j)}(.025) - \beta_n \right| \text{ and } j_R = \arg \min_{j \in \mathbb{N}} \left| q_{\text{Beta}}^{(j+1, n-j)}(.975) - \beta_n \right|.$$

For all $n \in \{200; 500; 1,000; 2,000\}$, we have $(j_L, j_R) = (3, 0)$ when $\beta_n = 1/n$ and $(j_L, j_R) = (1, 0)$ when $\beta_n = 1/(n \ln(n))$. Note that for $\nu = .95$, these two couples satisfy condition [\(1.5\)](#) since $j_L > j_R$,

$$\frac{q_{\text{Gam}}^{(4,1)}((1-\nu)/2)}{q_{\text{Gam}}^{(1,1)}((1+\nu)/2)} \approx 0.295 < 1 \text{ and } \frac{q_{\text{Gam}}^{(2,1)}((1-\nu)/2)}{q_{\text{Gam}}^{(1,1)}((1+\nu)/2)} \approx 0.066 < 1.$$

To select the number of observations k_n used in intervals $\text{CI}^{(M)}(\nu, \beta_n)$ and $\text{CI}^{(HF)}(\nu, \beta_n)$, we consider a method based on the excesses over a random threshold that has been proposed by Hüsler et al. [\[47\]](#). Let us briefly describe this selection procedure. For a given intermediate sequence (k_n) , let $\{Z_i = X_{n-i+1,n} - X_{n-k_n,n}, i = 1, \dots, k_n\}$ be the k_n excesses over the random threshold $X_{n-k_n,n}$. From [\[39, Lemma 3.4.1\]](#), given $X_{n-k_n} = t$, the random vector $(Z_1, \dots, Z_{k_n})$ is distributed as the random vector of order statistics $(Z_{k_n, k_n}^*, \dots, Z_{1, k_n}^*)$ obtained from independent and identically distributed random variables $Z_1^*, \dots, Z_{k_n}^*$. Moreover, under [\(1.1\)](#), for t large enough, the

common distribution of the Z_i^* can be approximated by a generalized Pareto (GP) distribution with cumulative distribution function

$$H_\gamma\left(\frac{x}{f(t)}\right) := 1 - \left(1 + \gamma \frac{x}{f(t)}\right)^{-1/\gamma}$$

for all $x > 0$ such that $1 + \gamma x/f(t) > 0$, where $f(t) = a(1/(1 - F_X(t)))$. Hence, since the intermediate order statistic $X_{n-k_n, n}$ is a consistent estimator of $Q_X(k_n/n)$ we can reasonably assume that, for n large enough and for a well chosen sequence (k_n) , the random variables $Z_1^*, \dots, Z_{k_n}^*$ are independent with common distribution function $H_\gamma(\cdot/a(n/k_n))$ and consequently that the random vector $(H_\gamma(Z_1/a(n/k_n)), \dots, H_\gamma(Z_{k_n}/a(n/k_n)))$ is distributed as the order statistics $(U_{k_n, k_n}, \dots, U_{1, k_n})$ of a standard uniform distribution. Replacing γ and $a(n/k_n)$ by their moment estimators $\hat{\gamma}_n^{(M)}$ and $\hat{a}_n^{(M)}(n/k_n)$ yields the discrepancy measure defined by

$$D(k_n) := \frac{1}{k_n} \sum_{i=1}^{k_n} \left| H_{\hat{\gamma}_n^{(M)}}\left(\frac{Z_i}{\hat{a}_n^{(H)}(n/k_n)}\right) - \frac{k_n - i + 1}{k_n + 1} \right|.$$

In view of the previous discussion, if (k_n) is well chosen, the discrepancy measure $D(k_n)$ should be close to 0. The selected value of k_n is thus given by

$$\hat{k}_n := \arg \min_{k \in \{k_1, \dots, k_2\}} D(k),$$

where $\{k_1, \dots, k_2\} \subset \{1, \dots, n\}$. In this simulation study, we take $k_1 = \lfloor .02n \rfloor$ and $k_2 = \lfloor .8n \rfloor$.

Comments on the obtained results – The results are gathered in Tables [1.1](#) to [1.6](#). First, in order to check if the data-driven procedure to select k_n provides satisfactory results, we have compared the value $E^{(\bullet)}(\hat{k}_n)$ with the 'best' error $E^{(\bullet)}(k_{n, opt}^{(\bullet)})$ where

$$k_{n, opt}^{(\bullet)} := \arg \min_{k \in [k_1, k_2]} E^{(\bullet)}(k).$$

It appears that for 6 out of the 9 distributions considered in our simulation study, the confidence intervals $CI^{(\bullet)}(.95, \beta_n)$ with $\bullet \in \{M, HF\}$ performs better when $k_n = \hat{k}_n$ than when $k_n = k_{n, opt}^{(\bullet)}$, no matter the sample size is. Let

us emphasize that since the value of $\hat{k}_n$ is different for each replication, is it quite possible that the value of $E^{(\bullet)}(\hat{k}_n)$ is less than that of $E^{(\bullet)}(k_{n,opt}^{(\bullet)})$.

The accuracy measure $E^{(M)}(\hat{k}_n)$ decreases as the sample size increases, except for log-normal and Beta distributions which do not fulfilled the conditions of Proposition 2. According to Proposition 3, the accuracy of the Moment confidence interval mainly depends on ρ' . For the absolute student distribution, we have $\rho' = \rho = -2/\mu$ so that a small value of μ should corresponds to a good behavior of the confidence interval. This is in line with the simulation study, see Tables 1.1 and 1.4. The absolute normal and the log-normal distributions have a second-order parameter $\rho = 0$ so that the consistency of the Moment confidence interval is not theoretically ensured. This can explain the poor results observed for the absolute normal distribution. Bad coverage probabilities are also obtained for the Beta distribution. This is probably due to the fact that this distribution does not satisfy the condition $\rho \neq \gamma$. For the positive reverse Burr distribution, we have $\rho' = \max(\gamma, \rho) = \gamma$. As shown in Tables 1.3 and 1.6, best results are obtained when $d = 2$ i.e., $\gamma = \rho' = -1/2$.

Comparison with $CI^{(HF)}$ – When $\beta_n = 1/n$, the proposed confidence interval clearly outperforms the interval $CI^{(HF)}(\nu, 1/n)$ for all the distributions except the Beta distribution although, in term of coverage probability, the later seems sometimes preferable. This is the case for instance for the Burr distribution with parameters $\kappa = 1$ and $d = 2$ for $n = 200$ and $\beta_n = 1/n$ (see Table 1.1). The coverage probability of the confidence interval $CI^{(HF)}(\nu, 1/n)$ is 94.9% which is very close to 95% while for the proposed interval, this probability is 92.208%. However, if we look at each bound separately, we see that $\nu_L^{(HF)}(\hat{k}_n) = 99.968\%$ and $\nu_R^{(HF)}(\hat{k}_n) = 94.932\%$. The confidence interval $CI^{(HF)}(\nu, 1/n)$ is thus not well balanced and cannot be used as a two-sided interval (by replacing the left bound by $-\infty$, we obtain approximatively the same coverage probability). The use of the proposed interval is clearly preferable since we have $\nu_L^{(M)}(\hat{k}_n) = 97.248\%$ and $\nu_R^{(M)}(\hat{k}_n) = 94.96\%$.

In the case $\beta_n = 1/(n \ln(n))$, both intervals are not accurate. The Moment confidence interval is better for the student distribution with $\mu = 1$ and provides comparable results for the Burr distribution. In the Gumbel MDA, $CI^{(M)}(\nu, \beta_n)$ is more accurate for the log-normal distribution. Finally, in the Weibull MDA, both estimators perform very poorly. As a conclusion, the proposed confidence interval should be preferred to the interval $CI^{(HF)}$ when

β_n is not too small (i.e., when $n\beta_n$ converges to a positive constant). The case $n\beta_n \rightarrow 0$ is more challenging and neither confidence interval is satisfactory even in the favourable situation where $k_n = k_{n,opt}^{(\bullet)}$ (except for the exponential and absolute Log-Normal distributions).

1.5 Real data study

Jeanne Calment was a French supercentenarian (i.e., a person who has reached the age of 110 years) who was born on 21 February 1875 and died on 4 August 1997 at the age of 122 years and 164 days. This exceptional longevity makes her the longest-living human at least among those whose birth and death dates could be verified. To get an idea of this exceptional lifespan, the second longest human lifespan observed in France (and in Europe) was 118 years and 341 days.

We would like to check whether Jeanne Calment's lifespan is an outlier or not. For this purpose, we compare her with French super-centenarians born during a similar period. Our study is based on the International Database on Longevity (IDL) which is the product of an international collaboration of several statistical offices and a large number of experts. This database contains birth and death dates of super-centenarians for 13 countries, and semi-super-centenarians (persons with a lifespan of between 105 and 110 years old) for 10 countries.

The population chosen for our comparison is the French super-centenarians and semi-super-centenarians born between 1890 and 1899. The number of observations per year over this period is comparable ranging from 162 for the year 1890 to 310 for the year 1898. We denote by $x_1, \dots, x_n$ the lifespans in years of the $n = 2400$ individuals in the study. We assume that these observations are realizations of n independent replications $X_1, \dots, X_n$ of a random variable X . In terms of the years studied and the period covered by the IDL (1978-2018), none of the persons concerned is still alive. This allows us to state that the variable X is not right-censored because of omissions of super-centenarians not included in the sample due to their longevity. We can consider that the lifespan of Jeanne Calment is not an outlier if it is close to the $1/2400$ -quantile which is the lifespan that is expected to be observed every 2400 people over the age of 105. We thus check if Jeanne Calment's lifespan belongs to the observed 95% confidence interval for $Q_X(1/2400)$.

We proceed with the confidence interval based on the Moment Estimator

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Absolute student distribution with $\mu = 4$ ($\gamma = 1/4$ and $\rho = -1/2$)				
$E^{(M)}(\hat{k}_n)$	2.408 %	1.842 %	1.280 %	0.926 %
$\text{CP}^{(M)}(\hat{k}_n)$	90.184 %	91.316 %	92.440 %	93.148 %
$E^{(HF)}(\hat{k}_n)$	3.206 %	1.874 %	1.196 %	1.346 %
$\text{CP}^{(HF)}(\hat{k}_n)$	93.580 %	96.108 %	96.928 %	97.692 %
$E^{(M)}(k_{n,opt}^{(M)})$	2.746 %	2.076 %	1.592 %	1.204 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	89.509 %	90.849 %	92.009 %	92.713 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	5.028 %	3.138 %	2.488 %	1.866 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	89.944 %	93.692 %	94.528 %	95.428 %
Absolute student distribution with $\mu = 1$ ($\gamma = 1$ and $\rho = -2$)				
$E^{(M)}(\hat{k}_n)$	1.535 %	0.940 %	0.624 %	0.296 %
$\text{CP}^{(M)}(\hat{k}_n)$	91.928 %	93.120 %	93.976 %	94.624 %
$E^{(HF)}(\hat{k}_n)$	5.558 %	4.244 %	3.542 %	2.696 %
$\text{CP}^{(HF)}(\hat{k}_n)$	88.884 %	91.512 %	92.868 %	93.920 %
$E^{(M)}(k_{n,opt}^{(M)})$	0.620 %	0.144 %	0.016 %	0.026 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	93.760 %	94.712 %	95.032 %	95.052 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	1.652 %	0.944 %	0.914 %	0.698 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	96.696 %	96.888 %	96.796 %	96.396 %
Burr distribution with $\kappa = 1$ and $d = 2$ ($\gamma = 1/2$ and $\rho = -1$)				
$E^{(M)}(\hat{k}_n)$	2.128 %	1.396 %	0.946 %	0.560 %
$\text{CP}^{(M)}(\hat{k}_n)$	90.744 %	92.208 %	93.228 %	93.992 %
$E^{(HF)}(\hat{k}_n)$	3.76 %	2.518 %	1.862 %	1.102 %
$\text{CP}^{(HF)}(\hat{k}_n)$	92.480 %	94.900 %	95.596 %	96.308 %
$E^{(M)}(k_{n,opt}^{(M)})$	2.308 %	1.430 %	0.920 %	0.518 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	90.385 %	92.141 %	93.369 %	94.196 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	4.444 %	2.884 %	1.164 %	0.254 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	91.112 %	93.736 %	92.760 %	94.492 %

TABLE 1.1 – Frechet MDA. Values of the accuracy measures $E^{(M)}(k_n)$ and $E^{(HF)}(k_n)$ and of the coverage probabilities $\text{CP}^{(M)}(k_n)$ and $\text{CP}^{(HF)}(k_n)$ for $\beta_n = 1/n$ and $k_n \in \{\hat{k}_n; k_{n,opt}^{(\bullet)}\}$.

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Exponential distribution with $\lambda = 1/2$ ($\gamma = 0$ and $\rho = -\infty$)				
$E^{(M)}(\hat{k}_n)$	1.442 %	1.012 %	0.478 %	0.152 %
$\text{CP}^{(M)}(\hat{k}_n)$	92.116 %	92.976 %	94.044 %	94.696 %
$E^{(HF)}(\hat{k}_n)$	2.434 %	1.194 %	1.102 %	0.852 %
$\text{CP}^{(HF)}(\hat{k}_n)$	95.124 %	96.708 %	97.204 %	96.704 %
$E^{(M)}(k_{n,opt}^{(M)})$	0.592 %	0.230 %	0.070 %	0.028 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	93.816 %	94.540 %	94.932 %	94.944 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	3.594 %	0.170 %	0.638 %	0.910 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	92.812 %	94.764 %	96.276 %	96.812 %
Absolute Normal distribution with $\mu = 0$ and $\sigma^2 = 1$ ($\gamma = 0$ and $\rho = 0$)				
$E^{(M)}(\hat{k}_n)$	2.004 %	1.796 %	1.322 %	1.038 %
$\text{CP}^{(M)}(\hat{k}_n)$	91.296 %	91.644 %	92.464 %	93.016 %
$E^{(HF)}(\hat{k}_n)$	3.034 %	2.442 %	1.856 %	1.444 %
$\text{CP}^{(HF)}(\hat{k}_n)$	93.932 %	95.100 %	96.184 %	96.968 %
$E^{(M)}(k_{n,opt}^{(M)})$	2.712 %	2.186 %	1.826 %	1.472 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	89.577 %	90.629 %	91.477 %	92.257 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	6.126 %	4.510 %	3.656 %	3.118 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	87.748 %	90.980 %	92.688 %	93.764 %
Log-Normal distribution with $\mu = 0$ and $\sigma^2 = 1$ ($\gamma = 0$ and $\rho = 0$)				
$E^{(M)}(\hat{k}_n)$	1.028 %	0.474 %	0.312 %	0.558 %
$\text{CP}^{(M)}(\hat{k}_n)$	92.944 %	94.052 %	95.152 %	95.668 %
$E^{(HF)}(\hat{k}_n)$	1.772 %	0.926 %	2.632 %	6.07 %
$\text{CP}^{(HF)}(\hat{k}_n)$	96.456 %	96.852 %	93.016 %	87.012 %
$E^{(M)}(k_{n,opt}^{(M)})$	0.268 %	0.226 %	0.056 %	0.020 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	94.480 %	94.548 %	94.888 %	95.040 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	1.234 %	1.052 %	1.034 %	1.092 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	97.420 %	97.032 %	97.068 %	97.184 %

TABLE 1.2 – Gumbel MDA. Values of the accuracy measures $E^{(M)}(k_n)$ and $E^{(HF)}(k_n)$ and of the coverage probabilities $\text{CP}^{(M)}(k_n)$ and $\text{CP}^{(HF)}(k_n)$ for $\beta_n = 1/n$ and $k_n \in \{\hat{k}_n; k_{n,opt}^{(\bullet)}\}$.

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Beta distribution with $\alpha = 4$ and $\beta = 4$ ($\gamma = -1/4$ and $\rho = -1/4$)				
$E^{(M)}(\hat{k}_n)$	2.304 %	2.268 %	2.198 %	1.942 %
$\text{CP}^{(M)}(\hat{k}_n)$	90.392 %	90.464 %	90.604 %	91.116 %
$E^{(HF)}(\hat{k}_n)$	1.310 %	1.260 %	1.458 %	1.908 %
$\text{CP}^{(HF)}(\hat{k}_n)$	97.216 %	97.412 %	97.364 %	97.068 %
$E^{(M)}(k_{n,opt}^{(M)})$	3.160 %	2.686 %	2.282 %	1.898 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	88.681 %	89.629 %	90.685 %	91.429 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	5.136 %	4.102 %	3.634 %	3.124 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	89.728 %	91.796 %	92.732 %	93.752 %
Positive Reverse Burr distribution with $\kappa = 1$ and $d = 4$ ($\gamma = -1/4$ and $\rho = -1$)				
$E^{(M)}(\hat{k}_n)$	1.690 %	1.450 %	0.976 %	0.680 %
$\text{CP}^{(M)}(\hat{k}_n)$	91.936 %	92.224 %	93.172 %	93.720 %
$E^{(HF)}(\hat{k}_n)$	2.148 %	1.524 %	1.266 %	1.450 %
$\text{CP}^{(HF)}(\hat{k}_n)$	95.704 %	96.920 %	97.532 %	97.900 %
$E^{(M)}(k_{n,opt}^{(M)})$	2.666 %	2.002 %	1.452 %	0.982 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	89.669 %	90.997 %	92.177 %	93.213 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	5.758 %	4.264 %	3.360 %	2.728 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	88.484 %	91.472 %	93.280 %	94.544 %
%				
Positive Reverse Burr distribution with $\kappa = 1$ and $d = 2$ ($\gamma = -1/2$ and $\rho = -1$)				
$E^{(M)}(\hat{k}_n)$	0.956 %	0.952 %	0.786 %	0.550 %
$\text{CP}^{(M)}(\hat{k}_n)$	93.372 %	93.280 %	93.752 %	94.140 %
$E^{(HF)}(\hat{k}_n)$	1.922 %	2.246 %	2.362 %	2.424 %
$\text{CP}^{(HF)}(\hat{k}_n)$	98.844 %	99.492 %	99.724 %	99.848 %
$E^{(M)}(k_{n,opt}^{(M)})$	2.728 %	2.156 %	1.674 %	1.292 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	89.545 %	90.689 %	91.853 %	92.617 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	5.520 %	4.278 %	3.668 %	3.088 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	88.960 %	91.444 %	92.664 %	93.824 %

TABLE 1.3 – Weibull MDA. Values of the accuracy measures $E^{(M)}(k_n)$ and $E^{(HF)}(k_n)$ and of the coverage probabilities $\text{CP}^{(M)}(k_n)$ and $\text{CP}^{(HF)}(k_n)$ for $\beta_n = 1/n$ and $k_n \in \{\hat{k}_n; k_{n,opt}^{(\bullet)}\}$.

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Absolute student distribution with $\mu = 4$ ($\gamma = 1/4$ and $\rho = -1/2$)				
$E^{(M)}(\hat{k}_n)$	6.424 %	5.404 %	4.552 %	3.802 %
$\text{CP}^{(M)}(\hat{k}_n)$	82.152 %	84.192 %	85.896 %	87.396 %
$E^{(HF)}(\hat{k}_n)$	4.404 %	2.904 %	2.212 %	1.618 %
$\text{CP}^{(HF)}(\hat{k}_n)$	91.192 %	94.192 %	95.520 %	96.652 %
$E^{(M)}(k_{n,opt}^{(M)})$	8.008 %	6.852 %	5.858 %	5.024 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	78.985 %	81.297 %	83.285 %	84.953 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	7.322 %	5.372 %	4.594 %	3.892 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	85.357 %	89.256 %	90.812 %	92.216 %
Absolute student distribution with $\mu = 1$ ($\gamma = 1$ and $\rho = -2$)				
$E^{(M)}(\hat{k}_n)$	3.696 %	2.38 %	1.622 %	1.044 %
$\text{CP}^{(M)}(\hat{k}_n)$	87.608 %	90.240 %	91.756 %	92.912 %
$E^{(HF)}(\hat{k}_n)$	6.372 %	4.914 %	4.066 %	3.270 %
$\text{CP}^{(HF)}(\hat{k}_n)$	87.256 %	90.172 %	91.868 %	93.420 %
$E^{(M)}(k_{n,opt}^{(M)})$	1.365 %	0.247 %	0.215 %	0.189 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	92.268 %	94.504 %	94.568 %	94.620 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	2.043 %	1.259 %	1.226 %	1.079 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	95.912 %	97.480 %	97.452 %	97.152 %
Burr distribution with $\kappa = 1$ and $d = 2$ ($\gamma = 1/2$ and $\rho = -1$)				
$E^{(M)}(\hat{k}_n)$	5.26 %	3.822 %	2.916 %	2.024 %
$\text{CP}^{(M)}(\hat{k}_n)$	84.480 %	87.356 %	89.168 %	90.948 %
$E^{(HF)}(\hat{k}_n)$	4.834 %	3.252 %	2.650 %	1.906 %
$\text{CP}^{(HF)}(\hat{k}_n)$	90.332 %	93.496 %	94.676 %	95.980 %
$E^{(M)}(k_{n,opt}^{(M)})$	6.077 %	4.121 %	2.899 %	1.869 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	82.845 %	86.757 %	89.200 %	91.260 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	6.531 %	4.621 %	3.531 %	2.097 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	86.937 %	90.756 %	92.824 %	93.172 %

TABLE 1.4 – Frechet MDA. Values of the accuracy measures $E^{(M)}(k_n)$ and $E^{(HF)}(k_n)$ and of the coverage probabilities $\text{CP}^{(M)}(k_n)$ and $\text{CP}^{(HF)}(k_n)$ for $\beta_n = 1/(n \ln n)$ and $k_n \in \{\hat{k}_n; k_{n,opt}^{(\bullet)}\}$.

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Exponential distribution with $\lambda = 1/2$ ($\gamma = 0$ and $\rho = -\infty$)				
$E^{(M)}(\hat{k}_n)$	3.508 %	2.934 %	2.444 %	2.054 %
$\text{CP}^{(M)}(\hat{k}_n)$	87.984 %	89.132 %	90.112 %	90.892 %
$E^{(HF)}(\hat{k}_n)$	2.718 %	1.532 %	1.146 %	0.898 %
$\text{CP}^{(HF)}(\hat{k}_n)$	94.564 %	96.720 %	97.292 %	96.796 %
$E^{(M)}(k_{n,opt}^{(M)})$	2.048 %	0.970 %	0.970 %	0.966 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	90.904 %	93.060 %	93.060 %	93.068 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	3.948 %	0.136 %	0.556 %	0.856 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	92.104 %	94.776 %	96.112 %	96.712 %
Absolute Normal distribution with $\mu = 0$ and $\sigma^2 = 1$ ($\gamma = 0$ and $\rho = 0$)				
$E^{(M)}(\hat{k}_n)$	5.594 %	5.466 %	5.06 %	4.43 %
$\text{CP}^{(M)}(\hat{k}_n)$	83.812 %	84.068 %	84.880 %	86.136 %
$E^{(HF)}(\hat{k}_n)$	4.404 %	3.894 %	3.232 %	2.684 %
$\text{CP}^{(HF)}(\hat{k}_n)$	91.192 %	92.212 %	93.496 %	94.560 %
$E^{(M)}(k_{n,opt}^{(M)})$	8.064 %	7.272 %	6.822 %	6.188 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	78.873 %	80.457 %	81.357 %	82.625 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	8.648 %	6.764 %	5.862 %	5.152 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	82.705 %	86.473 %	88.276 %	89.696 %
Log-Normal distribution with $\mu = 0$ and $\sigma^2 = 1$ ($\gamma = 0$ and $\rho = 0$)				
$E^{(M)}(\hat{k}_n)$	2.18 %	1.384 %	1.378 %	1.982 %
$\text{CP}^{(M)}(\hat{k}_n)$	90.640 %	92.232 %	92.804 %	93.316 %
$E^{(HF)}(\hat{k}_n)$	1.574 %	1.742 %	1.942 %	6.390 %
$\text{CP}^{(HF)}(\hat{k}_n)$	96.852 %	98.484 %	94.892 %	86.660 %
$E^{(M)}(k_{n,opt}^{(M)})$	0.562 %	0.824 %	0.810 %	0.638 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	93.876 %	93.352 %	93.524 %	93.724 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	1.262 %	1.200 %	1.152 %	1.198 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	97.524 %	97.400 %	97.288 %	97.396 %

TABLE 1.5 – Gumbel MDA. Values of the accuracy measures $E^{(M)}(k_n)$ and $E^{(HF)}(k_n)$ and of the coverage probabilities $\text{CP}^{(M)}(k_n)$ and $\text{CP}^{(HF)}(k_n)$ for $\beta_n = 1/(n \ln n)$ and $k_n \in \{\hat{k}_n; k_{n,opt}^{(\bullet)}\}$.

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
Beta distribution with $\alpha = 4$ and $\beta = 4$ ($\gamma = -1/4$ and $\rho = -1/4$)				
$E^{(M)}(\hat{k}_n)$	6.346 %	6.716 %	6.808 %	6.698 %
$\text{CP}^{(M)}(\hat{k}_n)$	82.308 %	81.568 %	81.380 %	81.600 %
$E^{(HF)}(\hat{k}_n)$	2.720 %	3.306 %	3.994 %	5.090 %
$\text{CP}^{(HF)}(\hat{k}_n)$	94.924 %	94.620 %	93.940 %	93.008 %
$E^{(M)}(k_{n,opt}^{(M)})$	8.988 %	8.080 %	7.472 %	6.900 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	77.025 %	78.841 %	80.057 %	81.201 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	8.010 %	6.584 %	5.920 %	5.248 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	83.981 %	86.833 %	88.160 %	89.504 %
Positive Reverse Burr distribution with $\kappa = 1$ and $d = 4$ ($\gamma = -1/4$ and $\rho = -1$)				
$E^{(M)}(\hat{k}_n)$	4.302 %	4.118 %	3.65 %	2.98 %
$\text{CP}^{(M)}(\hat{k}_n)$	86.396 %	86.764 %	87.700 %	89.036 %
$E^{(HF)}(\hat{k}_n)$	2.73 %	2.24 %	1.714 %	1.442 %
$\text{CP}^{(HF)}(\hat{k}_n)$	94.540 %	95.512 %	96.428 %	96.724 %
$E^{(M)}(k_{n,opt}^{(M)})$	7.314 %	6.044 %	5.016 %	3.882 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	80.373 %	82.913 %	84.969 %	87.237 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	8.064 %	6.006 %	4.898 %	4.016 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	83.873 %	87.988 %	90.204 %	91.968 %
Positive Reverse Burr distribution with $\kappa = 1$ and $d = 2$ ($\gamma = -1/2$ and $\rho = -1$)				
$E^{(M)}(\hat{k}_n)$	2.362 %	2.382 %	2.334 %	1.984 %
$\text{CP}^{(M)}(\hat{k}_n)$	90.276 %	90.236 %	90.328 %	91.028 %
$E^{(HF)}(\hat{k}_n)$	1.976 %	2.200 %	2.338 %	2.384 %
$\text{CP}^{(HF)}(\hat{k}_n)$	98.952 %	99.400 %	99.676 %	99.768 %
$E^{(M)}(k_{n,opt}^{(M)})$	7.290 %	6.164 %	5.216 %	4.268 %
$\text{CP}^{(M)}(k_{n,opt}^{(M)})$	80.421 %	82.673 %	84.569 %	86.465 %
$E^{(HF)}(k_{n,opt}^{(HF)})$	8.048 %	6.224 %	5.328 %	4.524 %
$\text{CP}^{(HF)}(k_{n,opt}^{(HF)})$	83.905 %	87.552 %	89.344 %	90.952 %

TABLE 1.6 – Weibull MDA. Values of the accuracy measures $E^{(M)}(k_n)$ and $E^{(HF)}(k_n)$ and of the coverage probabilities $\text{CP}^{(M)}(k_n)$ and $\text{CP}^{(HF)}(k_n)$ for $\beta_n = 1/(n \ln n)$ and $k_n \in \{\hat{k}_n; k_{n,opt}^{(\bullet)}\}$.

(see Section 1.3.2). The different parameters are taken following the procedure described in section 1.4. The obtained parameters are $j_L = 3$, $j_R = 0$ and $\hat{k}_n = 1388$. The extreme value index is estimated by $\hat{\gamma}_X = -0.059$ and the observed confidence interval is $[113.8; 116.1]$. Clearly, the value 122.5 is much larger than the upper bound. It thus seems that Jeanne Calment's lifespan can be considered as an outlier.

To go further, for each year, the 95% confidence intervals for the 1/250-th quantile are represented in Figure 1.1.

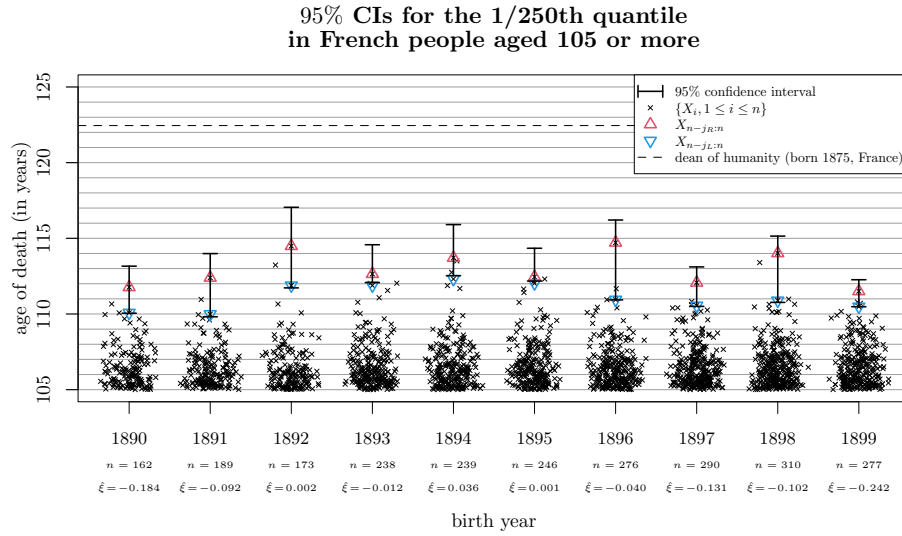


FIGURE 1.1 – Confidence intervals of level 95% for the lifespan of French super-centenarians for the years 1890 to 1899.

As expected, the lifespan of Jeanne Calment is clearly larger than all the upper bounds. We also note that the estimates of the extreme value index are always close to 0. This is in line with several studies such as the one of Rootzén and Zholud [60] and Alvarez et al. [5] where it is shown that the risk of dying appears constant over the age of 105. This surprising phenomenon is known as the "mortality plateau". In other words, it seems that X follows an exponential law (constant hazard rate). Note that Ferreira and Huang [31] conclude that the extreme value index is more likely to be negative.

1.6 Proofs

1.6.1 Preliminary results

Lemma 1. *For all non negative integer j and for all $\delta \in (0, 1)$,*

$$\lim_{n \rightarrow \infty} nq_{\text{Beta}}^{(j+1, n-j)}(\delta) = q_{\text{Gam}}^{(j+1, 1)}(\delta).$$

Proof – Let us first give some notations. Let $E_1, \dots$ be independent standard exponential random variables and, for $k \in \mathbb{N} \setminus \{0\}$, let $T_k := E_1 + \dots + E_k$. The distribution of T_k is then a Gamma distribution with parameters k and 1. From Rényi's representation of ordered standard uniform random variables, the random vector $(U_{j+1, n}, j = 0, \dots, n-1)$ is distributed as the random vector $(T_{j+1}/T_{n+1}, j = 0, \dots, n-1)$. Since $U_{j+1, n}$ is distributed according to a Beta distribution with parameters $j+1$ and $n-j$, we have for all $\delta \in (0, 1)$

$$\mathbb{P} \left(\frac{T_{j+1}}{T_{n+1}} \leq q_{\text{Beta}}^{(j+1, n-j)}(\delta) \right) = \mathbb{P} \left(\frac{nT_{j+1}}{T_{n+1}} \leq nq_{\text{Beta}}^{(j+1, n-j)}(\delta) \right) = \delta.$$

Moreover, from the law of large numbers, n/T_{n+1} converges almost surely to 1 and thus nT_{j+1}/T_{n+1} converges in distribution to a Gamma distribution with parameters $j+1$ and 1. Since the pointwise convergence of a sequence of increasing functions to a continuous function is uniform, we get

$$\lim_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} \left| F_{\text{Gam}}^{(j+1, 1)}(t) - \mathbb{P} \left(\frac{nT_{j+1}}{T_{n+1}} \leq t \right) \right| = 0.$$

Finally,

$$\begin{aligned} & \left| F_{\text{Gam}}^{(j+1, 1)}(nq_{\text{Beta}}^{(j+1, n-j)}(\delta)) - \mathbb{P} \left(\frac{nT_{j+1}}{T_{n+1}} \leq nq_{\text{Beta}}^{(j+1, n-j)}(\delta) \right) \right| \\ &= \left| F_{\text{Gam}}^{(j+1, 1)}(nq_{\text{Beta}}^{(j+1, n-j)}(\delta)) - \delta \right| \\ &\leq \sup_{t \in \mathbb{R}} \left| F_{\text{Gam}}^{(j+1, 1)}(t) - \mathbb{P} \left(\frac{nT_{j+1}}{T_{n+1}} \leq t \right) \right| \rightarrow 0, \end{aligned}$$

as $n \rightarrow \infty$. We conclude the proof by using the fact that $F_{\text{Gam}}^{(j+1, 1)}$ is bijective.

□

Lemma 2. Let $f_{\text{Beta}}^{(r,s)}$ be the density function of a Beta distribution with parameters $r > 0$ and $s > 0$. For any sequence $(z_n) \in (0, 1)$ such that $nz_n \rightarrow c > 0$ as $n \rightarrow \infty$ and for any positive integer j , we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} f_{\text{Beta}}^{(j+1, n-j)}(z_n) = \frac{c^j e^{-c}}{j!}.$$

Proof – The density of a Beta distribution with parameters $j+1$ and $n-j$ is given for $j < n$ by

$$f_{\text{Beta}}^{(j+1, n-j)}(z_n) = \frac{n!}{j!(n-j-1)!} z_n^j (1-z_n)^{n-j-1}.$$

Since $nz_n \rightarrow c > 0$, it is readily seen that $z_n^j \sim c^j n^{-j}$ and $(1-z_n)^{n-j-1} \rightarrow e^{-c}$ as $n \rightarrow \infty$. Hence,

$$f_{\text{Beta}}^{(j+1, n-j)}(z_n) \sim \frac{c^j e^{-c}}{j!} \frac{n!}{n^j (n-j-1)!}.$$

Stirling's formula ensures that, as $n \rightarrow \infty$, we have $n! \sim \sqrt{2\pi n} n^{n+1/2} e^{-n}$. Replacing in the above expression $n!$ and $(n-j-1)!$ by their equivalents leads to

$$\frac{n!}{n^j (n-j-1)!} \sim n,$$

and the proof is complete. $\square$

From [39, Theorem 1.1.6], condition (1.1) on the quantile function Q_X is equivalent to assume that for all t such that $1 + \gamma t > 0$ and for all sequence (x_n) converging to the right endpoint $Q_X(0)$,

$$\lim_{n \rightarrow \infty} \frac{\overline{F}_X(x_n + tg(x_n))}{\overline{F}_X(x_n)} = \frac{1}{K_\gamma^\leftarrow(t)}, \quad (1.13)$$

where $g(x) := a(1/\overline{F}_X(x))$. Recall that $K_\gamma(t) = (t^\gamma - 1)/\gamma$ if $\gamma \neq 0$ and $K_0(t) = \ln(t)$. Hence, $K_\gamma^\leftarrow(t) = (1 + \gamma t)^{1/\gamma}$, if $\gamma \neq 0$ and $K_0^\leftarrow(t) = \exp(t)$. A generalization of (1.13) when t is replaced by a sequence $t_n \rightarrow t_0$ was proved in [25, Lemma 2].

Lemma 3. Assume that (1.1) holds and let (x_n) be a sequence converging to the right endpoint $Q_X(0)$. For all sequences (t_n) such that $t_n \rightarrow t_0$, if there exists $\eta > 0$ for which $1 + \gamma t_0 \geq \eta$ then,

$$\lim_{n \rightarrow \infty} \frac{\overline{F}_X(x_n + t_n g(x_n))}{\overline{F}_X(x_n)} = \frac{1}{K_\gamma^\leftarrow(t_0)}.$$

1.6.2 Proofs of main results

Proof of Theorem 1

We need to prove that

$$\mathbb{P}\left(\widehat{t_{n,L}}X_{n-j_L,n} \leq Q_X(\beta_n)\right) \rightarrow \frac{1+\nu}{2}, \quad (1.14)$$

and

$$\mathbb{P}\left(\widehat{t_{n,R}}X_{n-j_R,n} \geq Q_X(\beta_n)\right) \rightarrow \frac{1+\nu}{2}. \quad (1.15)$$

We only prove (1.14), the proof of (1.15) being similar. Let $\epsilon > 0$. By assumption, there exist $c_\epsilon > 0$ and $N_\epsilon \in \mathbb{N}$ such that for all $n \geq N_\epsilon$, $\mathbb{P}(A_{n,\epsilon}) \geq 1 - \epsilon$, with

$$A_{n,\epsilon} := \left\{ \sigma_n^{-1} \left(\frac{t_{n,L}}{\widehat{t_{n,L}}} - 1 \right) \in [-c_\epsilon, c_\epsilon] \right\},$$

where we recall that $t_{n,L} := Q_X(\beta_n)/Q_X(\mathbf{a}_{n,L}(\nu))$. We have the decomposition

$$\begin{aligned} \mathbb{P}\left(\widehat{t_{n,L}}X_{n-j_L,n} \leq Q_X(\beta_n)\right) &= \mathbb{P}\left(\left\{\widehat{t_{n,L}}X_{n-j_L,n} \leq Q_X(\beta_n)\right\} \cap A_{n,\epsilon}\right) \\ &\quad + \mathbb{P}\left(\left\{\widehat{t_{n,L}}X_{n-j_L,n} \leq Q_X(\beta_n)\right\} \cap \Omega \setminus A_{n,\epsilon}\right) \\ &=: T_{1,n} + T_{2,n}. \end{aligned}$$

Obviously,

$$0 \leq T_{2,n} \leq \epsilon, \quad (1.16)$$

and thus $T_{2,n} \rightarrow 0$ as $n \rightarrow \infty$. Let us focus on the term $T_{1,n}$. We have

$$T_{1,n} = \mathbb{P}\left(\left\{\overline{F}_X(X_{n-j_L,n}) \geq \overline{F}_X\left(\frac{Q_X(\beta_n)}{\widehat{t_{n,L}}}\right)\right\} \cap A_{n,\epsilon}\right).$$

Let $x_{n,L} := Q_X(\mathbf{a}_{n,L}(\nu))$. When the event $A_{n,\epsilon}$ occurs, we have for all $n \geq N_\epsilon$,

$$\overline{F}_X(x_{n,L} + s_{n,\epsilon}g(x_{n,L})) \leq \overline{F}_X\left(\frac{Q_X(\beta_n)}{\widehat{t_{n,L}}}\right) \leq \overline{F}_X(x_{n,L} - s_{n,\epsilon}g(x_{n,L})),$$

where $g(y) = a(1/\overline{F}_X(y))$ and $s_{n,\epsilon} := c_\epsilon \sigma_n x_{n,L}/g(x_{n,L})$. Recall that $\mathbf{a}_{n,\bullet}(\nu) = q_{\text{Beta}}^{(j_\bullet+1, n-j_\bullet)}((1+\nu)/2)$. From Lemma 1, we have $n\mathbf{a}_{n,\bullet}(\nu) \rightarrow c^* := q_{\text{Gam}}^{(j_\bullet+1)}((1+\nu)/2)$.

$\nu)/2$). Hence $\mathbf{a}_{n,L}(\nu)$ is asymptotically proportional to $1/n$. Using the fact that the functions $Q_X(1/\cdot)$ and a are regularly varying functions, respectively with indices $\max(0, \gamma)$ and γ (see [39, Lemma 1.2.9 and Corollary 1.2.10]) and remarking that, since $\overline{F}_X$ is asymptotically bijective, $g(x_{n,L}) = a(\mathbf{a}_{n,L}^{-1})$, we get that $x_{n,L}/g(x_{n,L})$ is asymptotically proportional to $Q_X(1/n)/a(n)$. Hence, by assumption, $s_{n,\epsilon} \rightarrow 0$ as $n \rightarrow \infty$. Furthermore, $x_{n,L}$ converges to the right endpoint $Q_X(0)$ since $\mathbf{a}_{n,L}(\nu)$ converges to 0. We can thus apply Lemma 3 leading to

$$\overline{F}_X(x_{n,L} \pm s_{n,\epsilon}g(x_{n,L})) \sim \overline{F}_X(x_{n,L}) = \mathbf{a}_{n,L}(\nu).$$

As a consequence, there exists $N_\epsilon^{(*)} \in \mathbb{N}$ such that for all $n \geq \max(N_\epsilon, N_\epsilon^{(*)})$,

$$\mathbf{a}_{n,L}(\nu)(1 - \epsilon) \leq \overline{F}_X\left(\frac{Q_X(\beta_n)}{\widehat{t_{n,L}}}\right) \leq \mathbf{a}_{n,L}(\nu)(1 + \epsilon).$$

Note that $\overline{F}_X(X_{n-j_L,n})$ is distributed as a Beta distribution with parameters $j_L + 1$ and $n - j_L$. We obtain for $n \geq \max(N_\epsilon, N_\epsilon^{(*)})$,

$$\begin{aligned} T_{1,n} &\leq \mathbb{P}(\{\overline{F}_X(X_{n-j_L,n}) \geq \mathbf{a}_{n,L}(\nu)(1 - \epsilon)\} \cap A_{n,\epsilon}) \\ &\leq \mathbb{P}(\overline{F}_X(X_{n-j_L,n}) \geq \mathbf{a}_{n,L}(\nu)(1 - \epsilon)) = \overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 - \epsilon)). \end{aligned}$$

Moreover, still for $n \geq \max(N_\epsilon, N_\epsilon^{(*)})$,

$$\begin{aligned} T_{1,n} &\geq \mathbb{P}(\{\overline{F}_X(X_{n-j_L,n}) \geq \mathbf{a}_{n,L}(\nu)(1 + \epsilon)\} \cap A_{n,\epsilon}) \\ &= \mathbb{P}(\overline{F}_X(X_{n-j_L,n}) \geq \mathbf{a}_{n,L}(\nu)(1 + \epsilon)) \\ &\quad - \mathbb{P}(\{\overline{F}_X(X_{n-j_L,n}) \geq \mathbf{a}_{n,L}(\nu)(1 + \epsilon)\} \cap \Omega \setminus A_{n,\epsilon}) \\ &\geq \overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 + \epsilon)) - \epsilon. \end{aligned}$$

To sum up, we have proved that for all $\epsilon > 0$ and for all $n \geq \max(N_\epsilon, N_\epsilon^{(*)})$,

$$\overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 + \epsilon)) - \epsilon \leq T_{1,n} \leq \overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 - \epsilon)).$$

To prove that $T_{1,n} \rightarrow (1 + \nu)/2$ and to conclude the proof, it remains to show that

$$\lim_{n \rightarrow \infty} \overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 \pm \epsilon)) = \frac{1 + \nu}{2}. \quad (1.17)$$

Recall that $\overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)) = (1+\nu)/2$. The mean value theorem entails that there exists $\kappa \in (0, 1)$ such that

$$\overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 \pm \epsilon)) = \frac{1+\nu}{2} \mp \epsilon \mathbf{a}_{n,L}(\nu) f_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 \pm \kappa\epsilon)).$$

From Lemma [1](#), there exists $c_1 = c_1(\nu, j_L) > 0$ such that $\mathbf{a}_{n,L}(\nu) \sim c_1/n$. Hence, from Lemma [2](#)

$$f_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 \pm \kappa\epsilon)) \sim \frac{n}{j_L!} [c_1(1 \pm \kappa\epsilon)]^{j_L} \exp[-c_1(1 \pm \kappa\epsilon)] =: nc_2.$$

It is then readily seen that there exists positive constants c_3 and c_4 such that for all $v > 0$ and for n large enough,

$$\frac{1+\nu}{2} - c_3v \leq \overline{F}_{\text{Beta}}^{(j_L+1, n-j_L)}(\mathbf{a}_{n,L}(\nu)(1 \pm v)) \leq \frac{1+\nu}{2} + c_4v,$$

proving [\(1.17\)](#). $\square$

1.6.3 Proof of Corollary [1](#)

Using the assumption on the asymptotic behavior of the extreme quantile estimator $\hat{Q}_{X,n}(\beta_n)$, we get

$$\frac{\hat{Q}_{X,n}(\beta_n)}{Q_X(\beta_n)} = 1 + \mathcal{O}_{\mathbb{P}}\left(\frac{\vartheta_n(\beta_n)}{Q_X(\beta_n)}\right). \quad (1.18)$$

Since from [\[39\]](#), Lemma 1.2.9 and Corollary 1.2.10], we have that $a(x)/Q_X(x^{-1}) \rightarrow \gamma_+ := \max(\gamma, 0)$ as $x \rightarrow \infty$, the assumption

$$\lim_{n \rightarrow \infty} \frac{\vartheta_n(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} = 0,$$

entails that

$$\lim_{n \rightarrow \infty} \frac{\vartheta_n(\beta_n)}{Q_X(\beta_n)} = 0.$$

Moreover, recall that $\mathbf{a}_{n,\bullet}(\nu) = q_{\text{Beta}}^{(j_{\bullet}+1, n-j_{\bullet})}((1+\nu)/2)$. From Lemma [1](#), we have $n\mathbf{a}_{n,\bullet}(\nu) \rightarrow c^* := q_{\text{Gam}}^{(j_{\bullet}+1)}((1+\nu)/2)$. Hence, by assumption,

$$\frac{\hat{Q}_{X,n}(\mathbf{a}_{n,\bullet}(\nu))}{Q_X(\mathbf{a}_{n,\bullet}(\nu))} = 1 + \mathcal{O}_{\mathbb{P}}\left(\frac{\vartheta_n}{Q_X(\mathbf{a}_{n,\bullet}(\nu))}\right).$$

From [39, Lemma 1.2.9 and Corollary 1.2.10], the function $Q_X(1/\cdot)$ is regularly varying with index $\gamma_+ := \max(0, \gamma)$. Hence, as $n \rightarrow \infty$, $Q_X(\mathfrak{a}_{n,\bullet}(\nu)) \sim \{c^*\}^{-\gamma_+} Q_X(1/n)$ leading to

$$\frac{\hat{Q}_{X,n}(\mathfrak{a}_{n,\bullet}(\nu))}{Q_X(\mathfrak{a}_{n,\bullet}(\nu))} = 1 + \mathcal{O}_{\mathbb{P}}\left(\frac{\vartheta_n}{Q_X(1/n)}\right). \quad (1.19)$$

Since $a(n)/Q_X(1/n) \rightarrow \gamma_+$ as $n \rightarrow \infty$, it is readily seen that assumption $\vartheta_n/a(n) \rightarrow 0$ entails that $\vartheta_n/Q_X(1/n) \rightarrow 0$. We deduce from the equations (1.18) and (1.19) that

$$\frac{\hat{t}_{n,\bullet}^{(*)}}{t_{n,\bullet}} - 1 = \mathcal{O}_{\mathbb{P}}\left(\frac{\vartheta_n(\beta_n)}{Q_X(\beta_n)}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{\vartheta_n}{Q_X(1/n)}\right).$$

An application of Theorem 1 with

$$\sigma_n = \max\left(\frac{\vartheta_n(\beta_n)}{Q_X(\beta_n)}, \frac{\vartheta_n}{Q_X(1/n)}\right),$$

concludes the proof. $\square$

1.6.4 Proof of Proposition 3

Proof of part i) : First, assume that $\gamma > 0$. From [39, Remark 4.3.3] and since $k_n \rightarrow \infty$, we have as $n \rightarrow \infty$

$$\int_1^{k_n/(n\beta_n)} s^{\gamma-1} \ln(s) ds \sim \frac{1}{\gamma} \left(\frac{k_n}{n\beta_n}\right)^{\gamma} \ln\left(\frac{k_n}{n\beta_n}\right).$$

Moreover, the auxiliary function a is regularly varying with index $\gamma > 0$ (see [39, Lemma 1.2.9 and Corollary 1.2.10]). Hence, there exists a slowly varying function L such that

$$\frac{a(n/k_n)}{a(n)} = k_n^{-\gamma} \frac{L(n/k_n)}{L(n)}.$$

We then get that as $n \rightarrow \infty$,

$$\frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} \sim \frac{1}{\gamma} k_n^{-1/2} (n\beta_n)^{-\gamma} \frac{L(n/k_n)}{L(n)} \ln\left(\frac{k_n}{n\beta_n}\right) \frac{Q_X(1/n)}{Q_X(\beta_n)}.$$

Using the fact that the function $Q_X(1/\cdot)$ is regularly varying with index γ and since k_n and β_n are asymptotically proportional to a power of n ,

$$\frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} \propto n^{-\delta/2} L^*(n) \rightarrow 0,$$

where L^* is a slowly varying function. It is finally readily seen that conditions $k_n \rightarrow \infty$, $n/k_n \rightarrow \infty$ and $k_n^{1/2} B(n/k_n) \rightarrow \lambda \in \mathbb{R}$ are checked if $\delta = -2(\eta + \rho')/(1 - 2\rho')$ where $\eta \in (0, -\rho')$ can be taken as close as we want to 0.

Assume now that $\gamma = 0$. We have from [39, Remark 4.3.3] that, as $n \rightarrow \infty$,

$$\int_1^{k_n/(n\beta_n)} s^{-1} \ln(s) ds \sim \frac{1}{2} \ln^2 \left(\frac{k_n}{n\beta_n} \right).$$

Since a and $Q_X(1/\cdot)$ are in this case slowly varying functions, it is easy to check that

$$\lim_{n \rightarrow \infty} \frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} = 0,$$

for any sequence (k_n) which is asymptotically equivalent to n^δ with $\delta > 0$.

We next conclude as in the case $\gamma > 0$.

Finally, if $\gamma \in (-1/2, 0)$, we have from [39, Remark 4.3.3] that,

$$\lim_{n \rightarrow \infty} \int_1^{k_n/(n\beta_n)} s^{\gamma-1} \ln(s) ds = \frac{1}{\gamma^2}.$$

and thus, as $n \rightarrow \infty$,

$$\frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} \sim \frac{1}{\gamma^2} k_n^{-1/2} \frac{a(n/k_n)}{a(n)} \frac{Q_X(1/n)}{Q_X(\beta_n)}. \quad (1.20)$$

Since $Q_X(1/\cdot)$ is a slowly varying function and a is regularly varying with index γ , we have

$$\frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} \propto n^{-\delta(1/2+\gamma)} L^{**}(n),$$

where L^{**} is a slowly varying function. We then conclude as in the case $\gamma = 0$.

Proof of part ii) : We start with (1.20). It appears that

$$\frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} \sim \frac{U(n/k_n)}{U(n)},$$

where U is a regularly varying function with index $\gamma + 1/2 < 0$. Using [13, Theorem 1.5.2], since $k_n \rightarrow \infty$ and $n/k_n \rightarrow \infty$, we have for any $d > 0$ and for n large enough,

$$\left| \frac{U(n)}{U(n/k_n)} - k_n^{\gamma+1/2} \right| \leq \sup_{\lambda \geq d} \left| \frac{U(\lambda n/k_n)}{U(n/k_n)} - \lambda^{\gamma+1/2} \right| \rightarrow 0.$$

Hence, since $k_n^{\gamma+1/2} \rightarrow 0$, we have $U(n)/U(n/k_n) \rightarrow 0$ proving that, for any sequence (k_n) satisfying $k_n \rightarrow \infty$ and $n/k_n \rightarrow \infty$,

$$\frac{\vartheta_n^{(M)}(\beta_n)}{Q_X(\beta_n)} \frac{Q_X(1/n)}{a(n)} \rightarrow \infty.$$

□

Chapitre 2

Dimension reduction for the estimation of the conditional tail index

Gardes, L., and Podgorny, A. (2025). Dimension reduction for the estimation of the conditional tail-index, *Scandinavian Journal of Statistics*, to appear.

Nous nous intéressons à la relation entre les grandes valeurs d'une variable aléatoire réelle et ses covariables multidimensionnelles, dans un contexte où la distribution conditionnelle présente une queue lourde. L'estimation de l'indice de queue conditionnel positif d'une telle loi est une étape essentielle pour l'inférence statistique, mais cette tâche devient d'autant plus difficile que la dimension des covariables augmente. Dans ce travail, nous supposons l'existence d'un sous-espace linéaire de plus faible dimension tel que l'indice de queue conditionnel ne dépend des covariables qu'à travers leur projection sur ce sous-espace. Nous proposons une méthode pour estimer ce sous-espace de réduction de dimension et en établissons la consistance. Nous introduisons également un estimateur de l'indice de queue conditionnel exploitant cette réduction de dimension, dont nous démontrons également la consistance. Les bénéfices de cette approche sont illustrés à l'aide de simulations et d'une application à des données réelles.

2.1 Introduction

Let (X, Y) be a random pair, where Y is a real-valued response variable associated with a covariate X taking values in the set $\mathcal{X} \subset \mathbb{R}^p$, with $p \in \mathbb{N} \setminus \{0\}$. To better understand the relationship between Y and X , we seek to find a link function $g : \mathcal{X} \rightarrow \mathbb{R}$ such that $Y \approx g(X)$. For instance, if Y is square-integrable, a natural choice for the link function is the measurable function $g(\cdot) = \mathbb{E}(Y|X = \cdot)$ that minimizes the expected squared error $\mathbb{E}[(Y - g(X))^2]$. The conditional expectation $\mathbb{E}(Y|X = \cdot)$ is also referred to as the regression function. Given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of independent copies of (X, Y) and a point $x \in \mathcal{X}$, the primary goal of nonparametric regression is to construct an estimator $\hat{g}_n$ of g without assuming a specific form for the link function. Many approaches to nonparametric regression estimation can be found in the literature. Without being exhaustive, we mention some of them : local estimates can be obtained using the kernel method, introduced independently by [53] and [73], or the nearest-neighbor approach (see, for instance, [23]). A global estimation can be achieved using the smoothing splines method (see [64] and [7] for a comparison with the kernel method), where the regression function is estimated by a piecewise polynomial function. One can also mention machine learning methods such as neural networks [65] and random forests [15].

However, it is well known that, for a given level of accuracy of one of these estimators, the number n of required observations grows exponentially with respect to the dimension p . This phenomenon is often referred to as the curse of dimensionality (see [8]). A constraint can be added to the general regression model by assuming that all the information contained in the covariate X about the response is captured by a lower-dimensional projection of the predictor vector, $B^\top X \in \mathbb{R}^q$, with $q \ll p$. More precisely, Cook [20] introduced the notion of Sufficient Dimension Reduction (SDR) subspace : a q -dimensional linear subspace $\mathcal{S}$ is an SDR subspace if the distribution of Y given $B^\top X$ is the same as that of Y given X , where $B \in \mathbb{R}^{p \times q}$ is a basis of $\mathcal{S}$. In other words, denoting by $S(y, x) := \mathbb{P}(Y > y|X = x)$ the conditional survival function, we have $S(\cdot, x) = S(\cdot, B^\top x)$ for all $x \in \mathcal{X}$. The estimation of an SDR subspace has received considerable attention in the literature. It can be based on the conditional distribution of Y given X (*direct* approach) ; see, for instance, the MAVE approach [77] and the Sliced Regression (SR) estima-

tor [72]. The information provided by the moments of the inverse conditional distribution of X given Y can also be used to estimate the SDR subspace (*inverse* approach). As examples, we can cite the Sliced Inverse Regression method introduced by [51] and the SAVE estimator [22]. If one is only interested in a specific characteristic of the conditional distribution, the relevant dimension reduction subspace is not necessarily the SDR subspace, which can be much larger than required. For instance, when the characteristic of interest is the regression function $\mathbb{E}(Y|X)$, Cook and Li [21] assumed the existence of a smallest subspace $\mathcal{S}$, called the Mean Dimension Reduction (MDR) subspace, which is a q -dimensional linear subspace such that, for a basis B of $\mathcal{S}$, we have $\mathbb{E}(Y|X) = \mathbb{E}(Y|B^\top X)$ almost surely. If the MDR subspace is known (or at least can be estimated), inference on $\mathbb{E}(Y|X)$ can be performed by replacing X with the lower-dimensional covariate $B^\top X$, where $q \leq p$. When q is much smaller than p , nonparametric approaches provide more accurate estimates. Note that MAVE approach [77] provides an estimator of a MDR subspace with basis B_0 minimizing the expectation $\mathbb{E}[\{Y - \mathbb{E}(Y|B^\top X)\}^2]$.

In some practical cases, special attention needs to be paid to the right tail of the distribution rather than its central part. This is the case in finance (see, e.g., [59]) and insurance (see, for instance, [14, 58]). In these fields of application, it is often assumed that the conditional distribution of Y given X is heavy-tailed. More specifically, the conditional distribution is said to be heavy-tailed when there exists a positive function $\gamma(\cdot)$ such that, for all $t > 0$ and $x \in \mathcal{X}$,

$$\lim_{y \rightarrow \infty} \frac{S(ty, x)}{S(y, x)} = t^{-1/\gamma(x)}. \quad (2.1)$$

Condition (2.1) can be restated by saying that, for all $x \in \mathcal{X}$, the function $S(\cdot, x)$ is regularly varying with index $-1/\gamma(x)$. We thus have $S(y, x) = y^{-1/\gamma(x)} \mathcal{L}(y, x)$, where $\mathcal{L}(\cdot, x)$ is a positive, slowly varying function, meaning that $\mathcal{L}(ty, x)/\mathcal{L}(y, x) \rightarrow 1$ as $y \rightarrow \infty$ for all $t > 0$ and $x \in \mathcal{X}$ (see [13]). Condition (2.1) is also equivalent to assuming that the conditional distribution belongs to the maximum domain of attraction of Fréchet (see, for instance, [39, Theorem 1.2.1]). The positive function $\gamma(\cdot)$ is referred to as the conditional tail index. It determines the tail behavior of the conditional distribution.

In this paper, we focus on heavy-tailed conditional distributions (i.e., when condition (2.1) holds) and are interested in the behavior of the survival function $S(y, x)$ as y goes to infinity. As in the classical regression context presented above, we still face the curse of dimensionality, all the more so because, in order to analyze the tail of the distribution, only the largest observations are retained in the estimation process. In the literature, dimension reduction when the distribution of X given Y is elliptical (including certain types of heavy-tailed distributions) has been studied in [17] and [19]. However, despite its practical interest, little work has been done on dimension reduction for the conditional tail distribution of Y given X , in particular when the latter is heavy-tailed. A first attempt can be found in [33], where it is assumed that there exists a q -dimensional subspace $\mathcal{S}$, referred to as the Tail Dimension Reduction (TDR) subspace, for which

$$\lim_{y \rightarrow \infty} \sup_{x \in \mathcal{X}} \left| \frac{S(y, x)}{S_B(y, B^\top x)} - 1 \right| = 0, \quad (2.2)$$

where $B \in \mathbb{R}^{p \times q}$ is a basis of $\mathcal{S}$, and $S_B(\cdot, B^\top x)$ is the conditional survival function of Y given $B^\top X = B^\top x$. If the TDR subspace $\mathcal{S}$ exists, inference on the conditional tail distribution, such as the estimation of extreme quantiles, can be performed by replacing the covariate X with its reduced version $B^\top X$. Roughly speaking, under (2.2), we have $S(y|B^\top x) \approx S(y|x)$ for sufficiently large y , and thus the TDR subspace can be seen as a natural adaptation of the SDR subspace for the right-tail distribution. Recently, a different approach to reduce the dimensionality in the tail of the conditional distribution was introduced in [1]. Their method can be seen as an adaptation of the Sliced Inverse Regression technique introduced by [51] and differs significantly from that of [33]. In particular, condition (2.2) and the specific condition used in [1] are not equivalent. Moreover, the ultimate objective in [33] is the estimation of extreme quantiles, whereas in [1], it is to predict the occurrence of tail events. Another attempt to reduce the dimensionality can be found in [12]. The authors propose a specific inverse regression model, and the dimension reduction is achieved using a tail-adapted version of the partial least squares method.

Instead of the tail conditional distribution, one may be interested in one of its particular characteristics. An important example is, of course, the conditional tail index $\gamma(\cdot)$, whose estimation is a crucial step in making inferences about the conditional heavy-tailed distribution. Unfortunately, similarly to

what happens in the classical regression context, the dimension reduction subspaces from the aforementioned papers can be too large for this purpose, as they do not focus on the finer structure of the conditional tail index. Let us consider a simple example. Assume that for all $x \in \mathcal{X}$, there exists $B \in \mathbb{R}^{p \times q}$ with $q < p$ such that (2.1) holds with $\gamma(x) = g(B^\top x)$ and $\mathcal{L}(y, x) = c(x)$, where $g(\cdot)$ and $c(\cdot)$ are positive functions. Here, the TDR subspace defined in [33] and the SDR subspace are both equal to the set $\mathbb{R}^p$. However, it is intuitively clear that as soon as we are interested in the tail index $\gamma(\cdot)$, the appropriate dimension reduction subspace to consider is $\text{span}(B)$, the linear subspace of $\mathbb{R}^p$ spanned by the columns of B .

The main contribution of this work is the introduction of a dimension reduction subspace that captures the specific structure of the conditional tail index. This subspace is referred to as the Tail Index Dimension Reduction (TIDR) subspace. Roughly speaking, the TIDR subspace can be seen as the analogue of the MDR subspace when the characteristic of interest is the conditional tail index rather than the regression function. The framework and the main definitions, in particular the definition of the TIDR subspace $\mathcal{S}$, are given in Section 2.2. Section 2.3 is dedicated to the estimation of a basis of $\mathcal{S}$, which is then used to estimate the conditional tail index. The convergence in probability of the proposed estimators towards their target is established under certain mild conditions. Section 2.4 presents the results of a simulation study, followed by an application to a real dataset in Section 2.5. A procedure to estimate the dimension of the TIDR subspace is also presented and illustrated in Section 2.4. The proofs are provided in Section 2.6, except for the proof of Theorem 4 (see Section 2.3.3) provided in appendix.

2.2 Tail Index Dimension Reduction subspace

2.2.1 Framework

Throughout this paper, we consider a $(\mathbb{R}^p \times \mathbb{R})$ -valued random couple (X, Y) defined on a given probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For any $q \leq p$, $B \in \mathbb{R}^{p \times q}$ and $y \in \mathbb{R}$, let

$$S_B(y, B^\top X) := \mathbb{P}(Y > y | B^\top X)$$

be the conditional survival function of Y given $B^\top X$. When B is the identity matrix (i.e., $q = p$), the conditional survival function of Y given X is deno-

ted $S(y, X)$. For $\omega \in \Omega$ such that $B^\top X(\omega) = z$, we write $S_B(y, B^\top X)(\omega) = S_B(y, z)$. The couple (X, Y) follows the distribution described by the following model.

- (M) The support $\mathcal{X}$ of X is compact with a non-empty interior and the support of Y is confined to the positive half-line. Moreover, for all $x \in \mathcal{X}$, the conditional distribution of Y given $X = x$ is heavy-tailed with tail index $\gamma(x)$ i.e., $S(y, x) = y^{-1/\gamma(x)} \mathcal{L}(y, x)$, for some slowly varying function $\mathcal{L}(\cdot, x)$.

In Section [2.2.2](#) we give sufficient conditions ensuring that if the distribution of Y given X is heavy-tailed so is the conditional distribution of Y given $B^\top X$ for any $B \in \mathbb{R}^{p \times q}$. The definition of the tail index dimension reduction subspace is given in Section [2.2.3](#)

2.2.2 Preliminary result

For a given matrix $B \in \mathbb{R}^{p \times q}$ with $1 \leq q < p$, we can ask ourselves what happens if the covariate X is replaced by its projection $X_B := B(B^\top B)^{-1} B^\top X$ onto the linear subspace $\text{span}(B)$. Obviously, the knowledge of the conditional distribution of Y given X_B is equivalent to the one of Y given $B^\top X$. Since $S_B(y, B^\top X) = \mathbb{E}[S(y, X) | B^\top X]$, the conditional survival function of Y given $B^\top X$ can be seen as a mixture of the conditional survival function of Y given X . Hence, for all $x \in \mathcal{X}$, we are inclined to think that the decay of the conditional distribution of Y given $B^\top X = B^\top x$ is driven by the largest tail index involved in the mixture i.e., by

$$\xi_B(B^\top x) := \max_{z: B^\top z = B^\top x} \gamma(z). \quad (2.3)$$

Note that $\xi_B(\cdot)$ is a measurable function corresponding to the right endpoint of the conditional distribution of $\gamma(X)$ given $B^\top X = \cdot$. Under additional assumptions on the functions $\mathcal{L}(\cdot, \cdot)$ and $\gamma(\cdot)$, we prove hereafter that, as soon as model (M) holds, the conditional distribution of Y given $B^\top X$ is also heavy-tailed with tail index $\xi_B(B^\top X)$. These additional conditions are given below. First, we assume that the conditional tail index in model (M) is bounded from below and above by positive constants.

- (C.1) The distribution of the random variable $\gamma(X)$ is absolutely continuous with respect to the Lebesgue measure. Moreover,

$$\underline{\gamma} := \inf_{x \in \mathcal{X}} \gamma(x) > 0 \text{ and } \bar{\gamma} := \sup_{x \in \mathcal{X}} \gamma(x) < \infty,$$

The second condition is related to the function $\mathcal{L}(\cdot, \cdot)$.

(C.2) There exists a positive measurable function $c(\cdot)$ for which

$$\lim_{y \rightarrow \infty} \sup_{x \in \mathcal{X}} \left| \frac{\mathcal{L}(y, x)}{c(x)} - 1 \right| = 0, \text{ with } \inf_{x \in \mathcal{X}} c(x) > 0 \text{ and } \sup_{x \in \mathcal{X}} c(x) < \infty,$$

This condition entails in particular that, for all $x \in \mathcal{X}$, $\mathcal{L}(y, x) \rightarrow c(x)$ as $y \rightarrow \infty$. Let us give an example of slowly varying function $\mathcal{L}(\cdot, x)$ satisfying (C.2). Assume that

$$\mathcal{L}(y, x) = c(x) + \gamma(x)\beta(x)y^{-\tau(x)} + \varepsilon(y, x)y^{-\tau(x)}, \quad (2.4)$$

for some functions $c(\cdot) > 0$, $\tau(\cdot) > 0$ and $\beta(\cdot) \in \mathbb{R}$ and with $\varepsilon(y, x) \rightarrow 0$ as $y \rightarrow \infty$. This expression is a direct adaptation to the conditional case of the well known Hall and Welsh model [42] for slowly varying functions. This model is satisfied by most of the usual heavy-tailed distributions such as Pareto, Fréchet, Burr, student, etc. It is easy to check that condition (C.2) holds as soon as the functions $c(\cdot)$, $\gamma(\cdot)$ and $\tau(\cdot)$ are bounded from below and above by positive constants, if $\sup_{x \in \mathcal{X}} |\beta(x)| \leq \beta_0$ for some $\beta_0 \geq 0$ and if $\sup_{x \in \mathcal{X}} |\varepsilon(y, x)| \leq \varepsilon_1(y) \rightarrow 0$.

We are now in a position to establish the following result.

Proposition 4. *Assume that model (M) holds with conditions (C.1) and (C.2). For each full rank matrix $B \in \mathbb{R}^{p \times q}$ with $1 \leq q < p$, the conditional distribution of Y given $B^\top X$ is heavy-tailed with tail index $\xi_B(B^\top X)$.*

Note that under (C.1), the tail index of the conditional distribution of Y given $B^\top X$ belongs to the interval $[\underline{\gamma}, \bar{\gamma}] \subset (0, \infty)$.

2.2.3 Main definitions

We can now give the definition of the dimension reduction subspace that will be referred hereafter to as the TIDR subspace. From now on, we assume that model (M) holds together with conditions (C.1) and (C.2). Proposition 4 ensures that the conditional distribution of Y given $B^\top X$ is heavy-tailed for all full rank matrices B .

Definition 2. *Let $B \in \mathbb{R}^{p \times q}$ be a matrix with q linearly independent columns. The linear subspace $\mathcal{S} = \text{span}(B)$ is a Tail Index Dimension Reduction*

(TIDR) subspace if the conditional distributions of Y given X and Y given $B^\top X$ share the same conditional tail index, that is to say if $\xi_B(B^\top X) = \gamma(X)$ almost surely.

The basis $B \in \mathbb{R}^{p \times q}$ such that $\mathcal{S} = \text{span}(B)$ is of course not unique. In all what follows, we use the canonical basis of $\mathcal{S}$ which is defined hereafter.

Definition 3. Let $\tilde{B} \in \mathbb{R}^{p \times q}$ be any basis of $\mathcal{S}$ and let $\check{B}^\top$ be the matrix $\tilde{B}^\top$ in its reduced row echelon form. The canonical basis B of $\mathcal{S}$ is the orthogonal basis obtained by applying the Gram-Schmidt process on the columns of $\check{B}$.

It is well known that the reduced row echelon form of a given matrix is unique and thus so is the canonical basis. We denote by $\mathcal{B}_q$ the closure of the set that gathers the canonical bases of all linear subspaces of dimension q . If $B \in \mathcal{B}_q$ then $B^\top B = I_q$, the identity matrix of dimension q . When $q = 1$, the elements of the set $\mathcal{B}_1$ are the vectors $\beta \in \mathbb{R}^p$ with a positive first nonzero entry for which $\beta^\top \beta = 1$. Note also that $\mathcal{B}_q$ is a compact subset of $\mathbb{R}^{p \times q}$.

When $B = I_p$, it is readily seen that $\xi_B(B^\top x) = \gamma(x)$ for all $x \in \mathcal{X}$ and thus that $\mathcal{S} = \mathbb{R}^p$ is always a TIDR subspace. Naturally, the smallest subspace is of primary interest, which leads us to the definition of the central tail index subspace.

Definition 4. A linear subspace $\mathcal{S}_0$ is the Central Tail Index (CTI) subspace if $\mathcal{S}_0$ is a TIDR subspace such that $\mathcal{S}_0 \subset \mathcal{S}$ for all TIDR subspace $\mathcal{S}$.

The CTI subspace is a straightforward adaptation of the central mean subspace as defined in [21]. The CTI subspace does not necessarily exist but if it is the case, it is unique since it is the intersection of all the TIDR subspaces. We assume in the remainder of this paper that the random couple (X, Y) is such that the CTI subspace $\mathcal{S}_0$ exists. We also suppose that the dimension q of the CTI subspace is known.

In what follows, we denote by $B_0 \in \mathcal{B}_q$ the canonical basis of the CTI subspace $\mathcal{S}_0$. As a consequence of Proposition 4, the conditional distribution of Y given $B_0^\top X$ (which is not necessarily equal to the one of Y given X) is heavy-tailed with tail index $\xi_{B_0}(B_0^\top X) = \gamma(X)$. Hence, assuming that B_0 is known, the conditional tail index can be estimated using a sample of the random couple $(B_0^\top X, Y) \in \mathbb{R}^q \times \mathbb{R}$ instead of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$. When $q \ll p$, this permits us to construct a more efficient estimator of $\gamma(\cdot)$. Of course in practice, B_0 is unknown and its estimation is investigated in the next section.

2.3 Estimation of the CTI subspace and the conditional tail index

Throughout this section, we assume that the CTI subspace $\mathcal{S}_0$ exists and that its dimension, q , is known. As previously mentioned, the canonical basis of $\mathcal{S}_0$ is denoted by B_0 . Naturally, knowing the CTI dimension is unrealistic in practice; therefore, its estimation is addressed in the simulation study (see Section 2.4). Furthermore, as with most dimension reduction methods in the literature, verifying the existence of $\mathcal{S}_0$ is not straightforward. In Section 2.5, we present a graphical tool that can help assess whether this condition is satisfied.

This section is organized as follows. In Section 2.3.1, given an arbitrary estimator $\hat{\xi}_{n,B}(B^\top x)$ of $\xi_B(B^\top x)$, we propose an estimator $\hat{B}_n$ of the canonical basis of the CTI subspace $\mathcal{S}_0$. We next estimate the conditional tail index $\gamma(x) = \xi_{B_0}(B_0^\top x)$ by

$$\hat{\gamma}_n(x, \hat{B}_n) := \hat{\xi}_{n, \hat{B}_n}(\hat{B}_n^\top x).$$

In Section 2.3.2, we establish the consistency of the CTI subspace estimator. More specifically, we prove that

$$\|\hat{B}_n - B_0\| \xrightarrow{\mathbb{P}} 0,$$

for any norm $\|\cdot\|$ in $\mathbb{R}^{p \times q}$. Next, for a given compact subset $\mathcal{X}_0$ in the interior of $\mathcal{X}$, we establish the uniform consistency over the set $\mathcal{X}_0$ of the conditional tail index estimator :

$$\sup_{x \in \mathcal{X}_0} \left| \hat{\xi}_{n, \hat{B}_n}(\hat{B}_n^\top x) - \gamma(x) \right| \xrightarrow{\mathbb{P}} 0.$$

This last convergence is obtained under the following condition on the subspace $\mathcal{X}_0$.

(C.3) If there exists $B \in \mathcal{B}_q$ such that $\xi_B(B^\top x) = \gamma(x)$ for all $x \in \mathcal{X}_0$ then $B = B_0$.

This condition prevents, for instance, the pathological case where $\gamma(\cdot)$ is constant over the set $\mathcal{X}_0$, a scenario that can arise even if the CTI subspace exists.

2.3.1 Construction and definition of the estimators

To estimate the canonical basis $B_0 \in \mathcal{B}_q$ of the CTI subspace, we start by remarking that for all $x \in \mathcal{X}$ and $B \in \mathcal{B}_q$, one has $\gamma(x) = \xi_{B_0}(B_0^\top x) \leq \xi_B(B^\top x)$, since $\gamma(x)$ belongs to the support of the conditional distribution of $\gamma(X)$ given $B^\top X = B^\top x$ with right endpoint $\xi_B(B^\top x)$. If condition (C.3) holds for a compact subset $\mathcal{X}_0$ in the interior of $\mathcal{X}$ then

$$\arg \min_{B \in \mathcal{B}_q} \mathbb{E} [\xi_B(B^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}] = \{B_0\}.$$

Given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of independent copies of the random couple (X, Y) , the natural idea to estimate the canonical basis of the CTI subspace is to minimize over $\mathcal{B}_q$ an estimation of the expectation

$$\Psi(B, \mathcal{X}_0) := \mathbb{E} [\xi_B(B^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}]. \quad (2.5)$$

Let us first present the estimation procedure of the quantity $\Psi(B, \mathcal{X}_0)$. Assume that for all $B \in \mathcal{B}_q$, an estimator $\hat{\xi}_{n,B}(B^\top x)$ of $\xi_B(B^\top x)$ is available. When the dimension p of the covariate is not too large, numerous estimators of the conditional tail index can be found in the literature, see for instance [25, 32, 37]. An example of such an estimator will be given in Section 2.3.3. The empirical counterpart of $\Psi(B, \mathcal{X}_0)$ is given by

$$\tilde{\Psi}_n(B, \mathcal{X}_0) := \frac{1}{n} \sum_{i=1}^n \xi_B(B^\top X_i) \mathbb{I}\{X_i \in \mathcal{X}_0\}.$$

It can be shown that under model (M) where the tail index satisfies condition (C.1) and is such that

(C.4) for all $x \in \mathcal{X}$, the function $B \mapsto \xi_B(B^\top x)$ is Lipschitz in $B \in \mathcal{B}_q$, that is to say there exists a measurable function $M : \mathcal{X} \rightarrow [0, \infty)$ such that $\sup_{x \in \mathcal{X}} M(x) = \overline{M} < \infty$ and, for all $B_1 \in \mathcal{B}_q$, $B_2 \in \mathcal{B}_q$ and $x \in \mathcal{X}$,

$$|\xi_{B_1}(B_1^\top x) - \xi_{B_2}(B_2^\top x)| \leq M(x) \|B_1 - B_2\|,$$

then,

$$\sup_{B \in \mathcal{B}_q} |\tilde{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0)| \xrightarrow{\mathbb{P}} 0,$$

see the proof of Lemma 5. We propose to estimate $\Psi(B, \mathcal{X}_0)$ for all $B \in \mathcal{B}_q$ by the plug-in estimate

$$\hat{\Psi}_n(B, \mathcal{X}_0) := \frac{1}{n} \sum_{i=1}^n \hat{\xi}_{n,B}(B^\top X_i) \mathbb{I}\{X_i \in \mathcal{X}_0\}. \quad (2.6)$$

The definition of the proposed estimator of B_0 is given below.

Definition 5. Under model (M), for a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of independent copies of the random couple (X, Y) , the estimator $\hat{B}_n$ of the canonical basis of the CTI subspace minimizes the function $B \mapsto \hat{\Psi}_n(B, \mathcal{X}_0)$.

Our estimator belongs to the class of M-estimators. Note that, in a complete different setting, [48] also consider an M-estimator for the estimation of the direction of a single-index model. As mentioned in the introduction of this section, we estimate the conditional tail index $\gamma(x)$ by

$$\hat{\gamma}_n(x, \hat{B}_n) := \hat{\xi}_{n, \hat{B}_n}(\hat{B}_n^\top x).$$

The theoretical study of $\hat{B}_n$ and $\hat{\gamma}_n(x, \hat{B}_n)$ is done in the next section.

2.3.2 Main results

The first main result of this section establishes that if the conditional tail index estimator is uniformly consistent i.e.,

$$(C.5) \quad \sup_{B \in \mathcal{B}_q} \sup_{x \in \mathcal{X}_0} \left| \hat{\xi}_{n, B}(B^\top x) - \xi_B(B^\top x) \right| \xrightarrow{\mathbb{P}} 0,$$

then the estimator $\hat{B}_n$ in Definition 5 is consistent. Such an estimator is given in Section 2.3.3.

Theorem 2. Assume that model (M) holds and let $\mathcal{X}_0$ be a compact subset in the interior of $\mathcal{X}$ satisfying (C.3). Under conditions (C.1), (C.2) and (C.4), if the conditional tail index estimator appearing in $\hat{\Psi}_n(\cdot, \mathcal{X}_0)$ satisfies (C.5) then $\|\hat{B}_n - B_0\| \xrightarrow{\mathbb{P}} 0$ for any norm $\|\cdot\|$ in $\mathbb{R}^{p \times q}$.

The uniform consistency on $\mathcal{X}_0$ of the conditional tail index estimator $\hat{\gamma}_n(\cdot, \hat{B}_n)$ can be deduced from that of $\hat{B}_n$.

Theorem 3. Assume that model (M) holds and let $\mathcal{X}_0$ be a compact subset in the interior of $\mathcal{X}$ satisfying (C.3). Under conditions (C.1), (C.2) and (C.4), if the conditional tail index estimator $\hat{\xi}_{n, B}(B^\top x)$ satisfies (C.5) then

$$\sup_{x \in \mathcal{X}_0} \left| \hat{\gamma}_n(x, \hat{B}_n) - \gamma(x) \right| \xrightarrow{\mathbb{P}} 0.$$

The next section is dedicated to the definition and the theoretical study of a conditional tail index estimator satisfying condition (C.5).

2.3.3 A local-Hill estimator

For $B \in \mathcal{B}_q$ and $x \in \mathcal{X}$, let us now propose an appropriate estimator of the conditional tail index $\xi_B(B^\top x)$ of the conditional distribution of Y given $B^\top X = B^\top x$ (see Proposition 4) leading to a consistent estimator of B_0 . To motivate the definition of our estimator, let us start with the unrealistic situation where we have at our disposal a sample $\{\check{Y}_i(B, x), i = 1, \dots, m\}$ of $m \in \mathbb{N} \setminus \{0\}$ independent random variables with common distribution the conditional distribution of Y given $B^\top X = B^\top x$. The associated order statistics are denoted $\check{Y}_{(1)}(B, x) < \dots < \check{Y}_{(m)}(B, x)$. The best known estimator of the tail index is the Hill estimator which is given by

$$\check{\xi}_B^{(H)}(B^\top x) := \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} \ln \frac{\check{Y}_{(m-i+1)}(B, x)}{\check{Y}_{(m-[\alpha m])}(B, x)},$$

for $\alpha \in (1/m, 1)$. Of course, this is not a proper estimator since the random variables $\{\check{Y}_i(B, x), i = 1, \dots, m\}$ are not observed. We therefore propose to replace them in the expression of $\check{\xi}_B^{(H)}(B^\top x)$ by a suitable set of observed random variables. More precisely, let us introduce the set,

$$\mathcal{T}(B, x, h) := \{z \in \mathcal{X} \mid \|B^\top z - B^\top x\| \leq h\},$$

where $h = h_n > 0$ and let

$$M^\star = M(B, x, h) := \sum_{i=1}^n \mathbb{I}\{X_i \in \mathcal{T}(B, x, h)\},$$

be the random number of covariates in the set $\mathcal{T}(B, x, h)$. We denote by $\{X_i^\star = X_i(B, x, h), i = 1, \dots, M^\star\}$ the set of covariates that belong to $\mathcal{T}(B, x, h)$. The associated response variables are denoted $\{Y_i^\star = Y_i(B, x, h), i = 1, \dots, M^\star\}$. In the rest of that paper, the star ($\star$) serves as a reminder of the dependence on B, x and h . Intuitively, when h is close enough to zero, the random variable $Y_i^\star$ is approximatively distributed as $\check{Y}_i(B, x)$. For some sequence $(\alpha_n) = (\alpha) \in (0, 1)$, this leads us to introduce the estimator of the conditional tail index defined below.

Definition 6. *Under model (M), for all $(B, x) \in \mathcal{B}_q \times \mathcal{X}$, the local-Hill estimator of the conditional tail index $\xi_B(B^\top x)$ is*

$$\hat{\xi}_B^{(H)}(B^\top x) = \hat{\xi}_B^{(H)}(B^\top x, \alpha, h) := \frac{1}{[\alpha M^\star]} \sum_{i=1}^{[\alpha M^\star]} \ln \frac{Y_{(M^\star-i+1)}^\star}{Y_{(M^\star-[\alpha M^\star])}^\star},$$

if $\alpha M^\star > 1$ and $\hat{\xi}_B^{(H)}(B^\top x, \alpha, h) = 0$ otherwise.

Note that this estimator is quite similar to the one introduced in [37]. Under additional assumptions given below, we show that this estimator is uniformly consistent over the set $\Theta := \mathcal{B}_q \times \mathcal{X}_0$ where $\mathcal{X}_0$ is any compact subset in the interior of $\mathcal{X}$. This set is endowed with the product distance δ defined for all $\theta = (B, x) \in \Theta$ and $\theta' = (B', x') \in \Theta$ by $\delta(\theta, \theta') := \|B - B'\| + \|x - x'\|$. Below, we give the assumptions required to establish the uniform consistency. The first assumption is classically used in a dimension reduction context, see for instance [20].

(H.1) The distribution of X is absolutely continuous with respect to the Lebesgue measure.

Note that since $B \in \mathcal{B}_q$ is a full rank matrix, condition (H.1) entails that for all $B \in \mathcal{B}_q$, the distribution of $B^\top X$ is also absolutely continuous. Let us denote by $f_B(\cdot)$ the associated density function. The following regularity condition is required.

(H.2) For all $B \in \mathcal{B}_q$ and $x \in \mathcal{X}_0$, $f_B(B^\top x) > 0$. Moreover, there exists a positive constant K_f such that for all $(B, B') \in \mathcal{B}_q^2$ and $(x, x') \in \mathcal{X}^2$,

$$|f_B(B^\top x) - f_{B'}([B']^\top x')| \leq K_f (\|B - B'\| + \|x - x'\|).$$

Since $\Theta = \mathcal{B}_q \times \mathcal{X}_0$ is compact, this condition entails that

$$\sup_{(B,x) \in \Theta} f_B(B^\top x) < C_2 < \infty \quad \text{and} \quad \inf_{(B,x) \in \Theta} f_B(B^\top x) > C_1 > 0.$$

The following assumptions are commonly used to establish a uniform consistency result in an extreme value theory context, see [37]. The first one is a regularity condition on the conditional tail index of Y given $B^\top X$.

(H.3) The function $\mathcal{B}_q \times \mathcal{X} \ni (B, x) \mapsto \xi_B(B^\top x)$ is continuous.

For all $B \in \mathcal{B}_q$ and $z \in B^\top \mathcal{X}$, let $Q_B(\cdot, z)$ be the inverse function of the conditional survival function $y \mapsto S_B(y, z)$. From Proposition 4, we have for almost all $z \in B^\top \mathcal{X}$ and $u \in (0, 1)$ that

$$Q_B(u, z) = u^{-\xi_B(z)} \ell_B(u^{-1}, z),$$

where for all $B \in \mathcal{B}_q$ and $z \in B^\top \mathcal{X}$, $\ell_B(\cdot, z)$ is a slowly varying function. The following condition ensures that the convergence of $\ell_B(ty, B^\top x)/\ell_B(y, B^\top x)$ to 1 as $y \rightarrow \infty$ holds uniformly. More specifically, with the notation

$$\Delta_{B,x}(y) := \sup_{t>1} \ln \left(\frac{\ell_B(ty, B^\top x)}{\ell_B(y, B^\top x)} \right),$$

we assume that

$$(H.4) \quad \lim_{y \rightarrow \infty} \sup_{B \in \mathcal{B}_q} \sup_{x \in \mathcal{X}} \Delta_{B,x}(y) = 0.$$

Let us mention that if the slowly varying function $\mathcal{L}(\cdot, x)$ in model (M) is given as in (2.4), one can check that for all $B \in \mathcal{B}_q$ and $x \in \mathcal{X}$, the slowly varying function $\ell_B(y, B^\top x)$ converges to $c_B(B^\top x) := \{\mathbb{E}[c(X)|B^\top x]\}^{\xi_B(B^\top x)} > 0$. As a consequence,

$$\lim_{y \rightarrow \infty} \Delta_{B,x}(y) = 0.$$

Condition (H.4) ensures that the above convergence holds uniformly on B and x .

Finally, for all $B \in \mathcal{B}_q$, we measure the largest oscillation of the log-quantile function with respect to its second variable by

$$\omega(u, v, B, x, h) = \omega^*(u, v) := \sup_{(\beta, x') \in \mathcal{D}_{u,v}^*} \left| \ln \left(\frac{Q_B(\beta, B^\top x)}{Q_B(\beta, B^\top x')} \right) \right|$$

where

$$\mathcal{D}_{u,v}^* = \mathcal{D}_{u,v}(B, x, h) = \{(b, y) \in (0, 1) \times \mathcal{X}; b \in [u, v] \text{ and } \|B^\top(x - y)\| \leq h\}.$$

This oscillation is widely used in the extreme value theory literature (see for instance [25], [37]). It acts as a measure of how close conditional distributions are for two neighboring values of the covariate. It is controlled by the following assumption.

(H.5) There exists $\kappa > 0$ such that

$$\lim_{n \rightarrow \infty} \sup_{B \in \mathcal{B}_q} \sup_{x \in \mathcal{X}} \omega^*(n^{-(1+\kappa)}, 1 - n^{-(1+\kappa)}) = 0.$$

We establish below the uniform consistency of the local-Hill estimator given in Definition 6.

Theorem 4. *Assume that model (M) holds and let $\mathcal{X}_0$ be a compact subset in the interior of $\mathcal{X}$ satisfying (C.3). Under conditions (C.1), (H.1)-(H.5), if the sequences (α) and (h) are such that $nh^q \alpha / \ln(n) \rightarrow \infty$ and $\alpha \ln(nh^q) \rightarrow 0$ as $n \rightarrow \infty$ then,*

$$\sup_{B \in \mathcal{B}_q} \sup_{x \in \mathcal{X}_0} \left| \hat{\xi}_B^{(H)}(B^\top x) - \xi_B(B^\top x) \right| \xrightarrow{\mathbb{P}} 0.$$

The proof of Theorem 4 is quite long and technical. It is available in the appendix. The scheme of the proof is similar to the one of [37, Theorem 1] with however some additional difficulties due to the necessity to prove uniformity on $\mathcal{B}_q$. Using the local-Hill estimator of the conditional tail index we define the local-Hill CTI estimator of B_0 and the estimator of $\gamma(x)$ as follows.

Definition 7. *Under model (M), for a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of independent copies of the random couple (X, Y) , the local-Hill estimator $\hat{B}_n^{(H)}$ of the canonical basis of the CTI subspace minimizes the function*

$$B \mapsto \hat{\Psi}_n^{(H)}(B, \mathcal{X}_0) := \frac{1}{n} \sum_{i=1}^n \hat{\xi}_B^{(H)}(B^\top X_i) \mathbb{I}\{X_i \in \mathcal{X}_0\}.$$

The associated conditional tail index estimator is given for $x \in \mathcal{X}_0$ by

$$\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(H)}) := \hat{\xi}_{\hat{B}_n^{(H)}}^{(H)}([\hat{B}_n^{(H)}]^\top x).$$

The consistencies of the estimators $\hat{B}_n^{(H)}$ and $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(H)})$ follow directly from Theorems 2, 3 and 4.

Corollary 2. *Assume that model (M) holds and let $\mathcal{X}_0$ be a compact subset in the interior of $\mathcal{X}$ satisfying (C.3). Under conditions (C.1) and (H.1)-(H.5), if the sequences (α) and (h) are such that $nh^q \alpha / \ln(n) \rightarrow \infty$ and $\alpha \ln(nh^q) \rightarrow 0$ as $n \rightarrow \infty$ then,*

$$\|\hat{B}_n^{(H)} - B_0\| \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \sup_{x \in \mathcal{X}_0} \left| \hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(H)}) - \gamma(x) \right| \xrightarrow{\mathbb{P}} 0.$$

2.4 Finite sample size behavior

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent copies of a random couple (X, Y) for which there exists a CTI subspace $\mathcal{S}_0$ with canonical basis B_0 . In this section, we are interested in the finite sample behavior of the estimator $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(H)})$, where $\hat{B}_n^{(H)}$ denotes the local-Hill CTI estimator of B_0 , see Definition 7. We start by assuming that the dimension q of $\mathcal{S}_0$ is known, corresponding to the situation where the estimator is consistent. Its estimation is investigated in Section 2.4.5. This Section is organized as follows. A short

discussion about the choice of the tuning parameters required to compute these estimators is provided in Section 2.4.1. We explain in Section 2.4.2 the procedure we adopted to find the solution to the optimization problem involved in the definition of the local-Hill CTI estimator. Different models for the distribution of (X, Y) are given in Section 2.4.3 and the simulation results are commented in Section 2.4.4. Finally, a way to estimate the dimension q of the CTI subspace is presented in Section 2.4.5. The **R** source codes used in the simulation study as well as for the analysis of real data are available on a GitHub page¹.

2.4.1 Tuning parameters

The local-Hill estimator of the conditional tail index $\hat{\gamma}_n^{(H)}(\cdot, \hat{B}_n^{(H)})$ depends on two tuning parameters. The first one is the bandwidth h used to select the observations closest to the point of interest. We propose to take h in such a way that the number of points in the ball of center $[\hat{B}_n^{(H)}]^\top x$ and radius h is proportional to n^{1-b} for some $b \in (0, 1)$. In view of Lemma 3, this number of points is proportional to $\vartheta_q n h^q$ where ϑ_q is the volume of the unit q -ball leading to $h = (n^{-b}/\vartheta_q)^{1/q}$. The selection of h thus reduces to the one of $b \in [0, 1)$. In all what follows, we use the uniform norm $\|\cdot\|_\infty$ for which $\vartheta_q = 2^q$.

The second tuning parameter is the probability $\alpha \in (0, 1)$ controlling the number of largest observations. We take $\alpha = n^{-a}$ for some $a \in (0, 1)$.

Proposing a method to select the parameters a and b is beyond the scope of this paper and left for future work. In Section 2.4.4, we consider a grid of values for these two tuning parameters and keep the couple (a, b) leading to the best estimation, see Section 2.4.4 for more details.

A third tuning parameter $n_0 \ll n$ is considered in order to reduce the computational burden associated with the evaluation of $\hat{\Psi}_n^{(H)}(B, \mathcal{X}_0)$. More precisely, we randomly select n_0 observations $\tilde{X}_1, \dots, \tilde{X}_{n_0}$ in the original sample $\{X_1, \dots, X_n\}$ and replace the evaluation of $\hat{\Psi}_n^{(H)}(B, \mathcal{X}_0)$ by the one of

$$\tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0) := \frac{1}{n_0} \sum_{i=1}^{n_0} \hat{\xi}_B^{(H)}(B^\top \tilde{X}_i) \mathbb{I}\{\tilde{X}_i \in \mathcal{X}_0\}.$$

1. <https://github.com/Alex-Podgorny/Dimension-Reduction-tail-index/tree/master>

This approach significantly decreases the computation time without noticeably affecting the accuracy of the results, this is also used in [33]. If $n_0 \sim cn$ for some constant $c \in (0, 1)$, the theoretical results presented above remain unchanged. We set $n_0 = \lfloor n/20 \rfloor$, which provides a good trade-off between computational efficiency and precision.

2.4.2 Finding the solution of the optimization problem

The estimator $\hat{B}_n^{(H)}$ is solution of the constrained optimization problem

$$(P) \min_{B \in \mathbb{R}^{p \times q}} \tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0) \quad \text{subject to} \quad B \in \mathcal{B}_q.$$

Searching for the solution by moving along the feasible set is rather difficult. In particular, the solution strongly depends on the initialization of the algorithm. We thus decided to relax the constraint by considering the problem

$$(P') \min_{B \in \mathbb{R}^{p \times q}} \tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0) \quad \text{subject to} \quad B^\top B = I_q.$$

Since $\mathcal{B}_q$ is the set of all canonical q -dimensional basis, each element of $\mathcal{B}_q$ corresponds to a different subspace of $\mathbb{R}^p$. Of course, this is no longer the case if we only impose B to be an orthogonal matrix but (P') is easier to solve than (P). Moreover, if B (resp. B') is a solution of (P) (resp. (P')), then

$$\tilde{\Psi}_{n_0}^{(H)}(B', \mathcal{X}_0) \leq \tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0).$$

To solve the constrained optimization problem (P'), we apply row by row the Simulated Annealing (SANN) algorithm (which is available in the `optim` function in R). The algorithm is described below :

1. Initialization : We generate 100 points from a Halton sequence (cf. [43]) in the hypercube $[-1, 1]^{pq}$. The function $\tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0)$ is evaluated for each $p \times q$ matrix obtained with these points, and the matrix corresponding to the lowest value of $\tilde{\Psi}_{n_0}^{(H)}$ is selected as the initial matrix $B^{(0)}$.
2. For $i \geq 0$, until convergence :
 - Let $B^{(i,0)} = B^{(i)}$.
 - For $j = 0, \dots, p-1$:
 - $\hookrightarrow$ optimize the function $B \mapsto \tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0)$ over the $(j+1)$ th row of $B^{(i,j)}$ (all the other rows are fixed). We use the SANN algorithm with 20 iterations ;

$\hookrightarrow$ apply a QR decomposition on the optimized matrix to obtain an orthonormalized version denoted by $B^{(i,j+1)}$.

- Let $B^{(i+1)} = B^{(i,p)}$

The convergence is achieved when

$$\|P_{B^{(i+1)}} - P_{B^{(i)}}\|_2 < 10^{-2},$$

where $P_B = BB^\top$ is the projector on the subspace spanned by the orthogonal basis B and $\|\cdot\|_2$ denotes the spectral norm. The choice of 10^{-2} was made as a trade-off between the accuracy of the method and computational efficiency.

2.4.3 Models

Four models are considered for the distribution of the random couple (X, Y) . The random vector X is uniformly distributed on the space $\mathcal{X} := [0, 1]^p$ with $p \in \{4, 30\}$. Let $Q(\cdot, x)$ be the inverse of the function $y \mapsto S(y, x)$. The random variable Y is given by $Y := U^{-\gamma(x)}\ell(U^{-1}, x)$ where U is a standard uniform random variable. Two different slowly varying functions are used in our simulation study :

$$\ell^{(1)}(u^{-1}, x) := [1 + \exp\{B_1^\top x - u^{-1}\}]^{-1} \text{ and } \ell^{(2)}(u^{-1}, x) := \exp(-u/2)B_1^\top x,$$

with $B_1 = (0, 0, 5, 5, 0, \dots, 0)^\top \in \mathbb{R}^p$. For the first three models, the dimension of the CTI subspace is $q = 1$ with canonical basis $B_0 = (2, 1, 0, \dots, 0)^\top / \sqrt{5} \in \mathbb{R}^p$. Two functions $\gamma(\cdot) = \xi_{B_0}(B_0^\top \cdot)$ are considered : for all $z \in \mathbb{R}$,

$$\xi_{B_0}^{(1)}(z) := \frac{1}{10} + \frac{9}{10} \left(\frac{\sqrt{5}z}{3} \right)^4 \text{ and } \xi_{B_0}^{(2)}(z) := \frac{1}{10} + \frac{9}{10} |\cos(2z)|.$$

— **Model 1** - For all $x \in \mathcal{X}$, $\gamma(x) = \xi_{B_0}^{(1)}(B_0^\top x)$ and $\ell(u^{-1}, x) = \ell^{(1)}(u^{-1}, x)$. For this model, since $\ell^{(1)}(u^{-1}, x) \rightarrow 1$ as $u \rightarrow 0$, the basis B_0 also satisfies condition [\(2.2\)](#) that is to say that the CTI subspace corresponds to the dimension reduction subspace introduced in [\[33\]](#). For the next model, we consider the slowly varying function $\ell^{(2)}(u^{-1}, x)$ which converges to $B_1^\top x$ as $u \rightarrow 0$. As a consequence, the CTI subspace is still $\text{span}(B_0)$ while the dimension reduction subspace in [\[33\]](#) is $\text{span}(B_0, B_1)$.

— **Model 2** - For all $x \in \mathcal{X}$, $\gamma(x) = \xi_{B_0}^{(1)}(B_0^\top x)$ and $\ell(u^{-1}, x) = \ell^{(2)}(u^{-1}, x)$.

In Models 1 and 2, the conditional tail index $\xi_{B_0}^{(1)}(\cdot)$ is strictly monotone. The function $\xi_{B_0}^{(2)}(\cdot)$, which is symmetric with respect to the vertical line $y = \pi/4$, is used in Model 3.

— **Model 3** - For all $x \in \mathcal{X}$, $\gamma(x) = \xi_{B_0}^{(2)}(B_0^\top x)$ and $\ell(u^{-1}, x) = \ell^{(1)}(u^{-1}, x)$. For the last model, the dimension of the CTI linear subspace is $q = 2$ with canonical basis $B_0 = (e_1, e_2) \in \mathbb{R}^{p \times 2}$ where $e_1 = (1, 0, \dots, 0)^\top$ and $e_2 = (0, 1, 0, \dots, 0)^\top$.

— **Model 4** - For all $x \in \mathcal{X}$, $\ell(u^{-1}, x) := \ell^{(1)}(u^{-1}, x)$ and

$$\gamma(x) = \xi_{B_0}(B_0^\top x) := \frac{1}{10} + \frac{9}{5} [(e_1^\top x - 0.5)^2 + (e_2^\top x - 0.5)^2].$$

2.4.4 Simulation results

Our aim is to compare the finite sample performance of a local-Hill type estimator $\hat{\gamma}_n^{(H)}(x, \check{B}_n)$ of the conditional tail index $\gamma(x)$ with different estimated dimension reduction subspaces $\text{span}(\check{B}_n)$. More specifically, we consider 7 different estimated bases :

- $\check{B}_n = \hat{B}_n^{(H)}$, the proposed local-Hill CTI estimator, see Definition [7](#);
- $\check{B}_n = \hat{B}_n^{(G)}$, the TDR estimator proposed by Gardes [33](#);
- $\check{B}_n = \hat{B}_n^{(T_1)}$ and $\check{B}_n = \hat{B}_n^{(T_2)}$ which are the two estimators proposed by Aghbalou et al. [11](#) (respectively, TIREX1 and TIREX2 as named by the authors);
- $\check{B}_n = \hat{B}_n^{(L)}$ corresponding to the vector of dimension p obtained by minimizing the $(1 - \alpha)$ th quantile loss function with a LASSO penalization, see [75](#). This approach is referred to as the LASSO Quantile Regression (QR) method;
- $\check{B}_n = B_0$ corresponding to the ideal situation where the true basis B_0 is known;
- $\check{B}_n = I_p$ corresponding to the case where no dimension reduction is used.

Note that $\hat{B}_n^{(H)}$ is consistent only if $\gamma(x) > 0$ for all $x \in \mathcal{X}$ while $\hat{B}_n^{(G)}$, $\hat{B}_n^{(T_1)}$ and $\hat{B}_n^{(T_2)}$ are consistent regardless the sign of the conditional tail index. The TDR estimator in [33](#) is also the solution of a constrained optimization problem which is solved with the procedure described in Section [2.4.2](#). Moreover, because the execution time increases exponentially with the dimension p , the

estimator $\hat{B}_n^{(G)}$ is only considered in the case $p = 4$. Finally, we use the function `rq.pen.cv()` of the **R** package `rqPen` to compute the estimator $\hat{B}_n^{(L)}$. A cross-validation procedure is used to select the regularization parameter. Note that the dimension reduction subspace obtained with the LASSO-QR method is of dimension $q = 1$.

For each model, we generate $N = 100$ samples of size $n = 2000$ for $p = 4$ and $n = 4000$ for $p = 30$. The accuracy of the estimator $\hat{\gamma}_n^{(H)}(x, \check{B}_n)$ for the conditional tail index $\gamma(x)$ is evaluated on the compact $\mathcal{X}_0 := [\varepsilon_p, 1 - \varepsilon_p]^p$. We take $\varepsilon_p = (1 - 0.9^p)/2$ so that approximatively 90% of observations belong to $\mathcal{X}_0$.

As mentioned in Section [2.4.1](#), the tuning parameters are given by $h = n^{-b/q}/2$ and $\alpha = n^{-a}$ for $(a, b) \in (0, 1)^2$. The couple (a, b) is taken in the set

$$\mathcal{G} := \{0.25, 0.3, 0.35, 0.4, 0.45, 0.5\} \times \{0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}.$$

For $r \in \{1, \dots, N\}$, let $\hat{\gamma}_n^{(H,r)}(x, \check{B}_n)$ be the local-Hill estimator of $\gamma(x)$ evaluated on the r th replication. For each $r \in \{1, \dots, N\}$ and each $(a, b) \in \mathcal{G}$, we compute the error

$$E^{(r)}(\check{B}_n, a, b) := \frac{1}{M} \sum_{i=1}^M \left(\hat{\gamma}_n^{(H,r)}(x_i, \check{B}_n) - \gamma(x_i) \right)^2,$$

where $M = 10.000$ and $\{x_i; i = 1, \dots, M\}$ are the points of a Halton sequence on the compact $\mathcal{X}_0$. Note that $E^{(r)}(\check{B}_n, a, b)$ is an approximation of the integral

$$\int_{\mathcal{X}_0} \left(\hat{\gamma}_n^{(H,r)}(x, \check{B}_n) - \gamma(x) \right)^2 dx.$$

For $r \in \{1, \dots, N\}$, let $E^{(r, opt)}(\check{B}_n) := E^{(r)}(\check{B}_n, \hat{a}_n^{(opt)}, \hat{b}_n^{(opt)})$ be the optimal error of the r -th replication where

$$(\hat{a}_n^{(opt)}, \hat{b}_n^{(opt)}) := \arg \min_{(a,b) \in \mathcal{G}} \widehat{\text{MISE}}(\check{B}_n, a, b) := \arg \min_{(a,b) \in \mathcal{G}} \frac{1}{N} \sum_{r=1}^N E^{(r)}(\check{B}_n, a, b),$$

minimizes an estimation of the Mean Integrated Squared Error

$$\mathbb{E} \left[\int_{\mathcal{X}_0} \left(\hat{\gamma}_n^{(H,r)}(x, \check{B}_n) - \gamma(x) \right)^2 dx \right].$$

For each estimator $\check{B}_n$, the boxplot of the values $\{E^{(r,opt)}(\check{B}_n); r = 1, \dots, N\}$ are given in Figures 2.1 and 2.2. Note that the different methods are compared in the best possible scenario i.e., by using the true dimension q of the CTI subspace and the optimal hyper parameters. The proposed CTI estimator consistently achieves lower errors compared to the alternatives, indicating its effectiveness in estimating the tail index $\gamma(x)$ when dimension reduction is properly applied. As expected, the ideal scenario where $\check{B}_n = B_0$ achieves slightly better results, but this serves only as a benchmark, as the true subspace is generally unknown in practice. In contrast, using $\check{B}_n = I_p$ leads to significantly higher errors, underscoring the importance of reducing dimensionality for accurate tail index estimation. For Model 2, where the CTI subspace is included in the TDR subspace introduced in [33], the estimation of B_0 by $\hat{B}_n^{(G)}$ is inaccurate. This is in line with the theory. For Model 3 where the function $\xi_{B_0}(\cdot)$ is symmetric with respect to the vertical line $y = \pi/4$, $\hat{B}_n^{(G)}$, $\hat{B}_n^{(T_1)}$ and $\hat{B}_n^{(T_2)}$ are clearly less efficient. This phenomenon is also observed with the SIR method which fails to recover the true direction in presence of a symmetric relationship. Note that Gardes' estimator and TIREX1 failed to estimate the dimension reduction subspace in Model 4 for which $q = 2$. Finally, the LASSO-QR method only provides satisfactory results under Model 1 where the slowly varying function converges to a constant. In addition to the boxplots, heatmaps are used to display the estimated MISE values over the entire grid of tuning parameters (a, b) . These heatmaps provide a detailed visualization of the sensitivity of each method to the choice of parameters. All the methods demonstrate robust performance across a wide range of tuning parameters. Moreover it seems that the choice of the optimal parameters does not strongly depend of the model. The choice $(a, b) = (0.3, 0.2)$ give good results in all cases for our estimation procedure, this is why we recommend it. Figure 2.3 shows the heatmaps for our method on the different models. The heatmaps for the other methods are available in the Github page mentioned in the introduction of Section 2.4.

2.4.5 Estimation of the dimension of the CTI subspace

In the above simulations, the true dimension q of the CTI subspace was used. In practice, this dimension is, in most of the cases, unknown and need to be estimated. Our procedure of estimation lies on the following result. For all $d \in \{1, \dots, p\}$, let $B_{0,d} \in \mathcal{B}_d$ be a matrix such that $\Psi(B_{0,d}, \mathcal{X}_0) \leq \Psi(B, \mathcal{X}_0)$

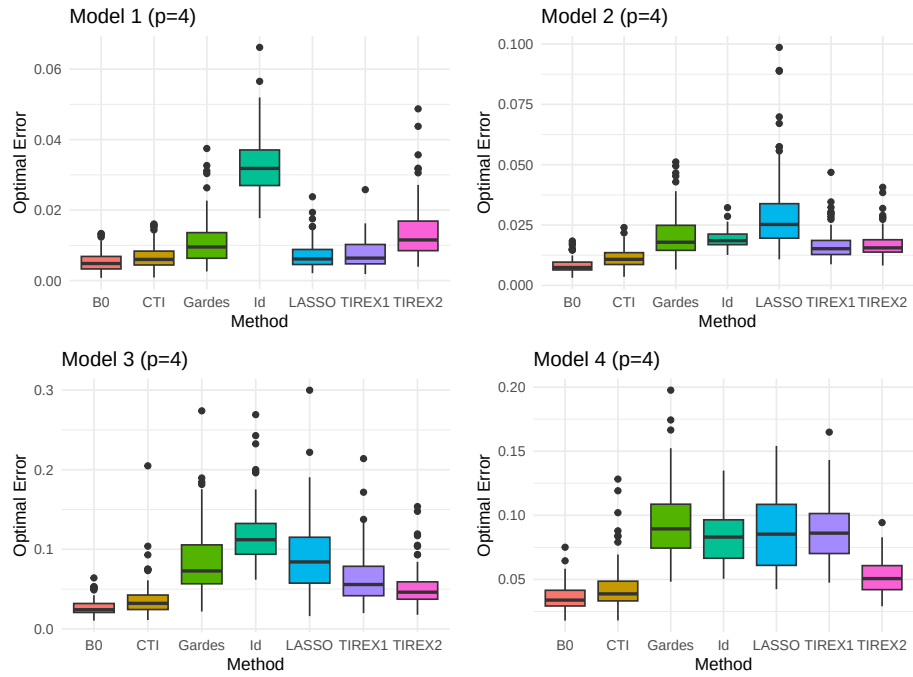


FIGURE 2.1 – Boxplots of the values $\{E^{(r,opt)}(\check{B}_n); r = 1, \dots, N\}$ for different estimated bases $\check{B}_n$ on $N = 100$ replications of a sample of size $n = 2000$ from Models 1 to 4 with $p = 4$.

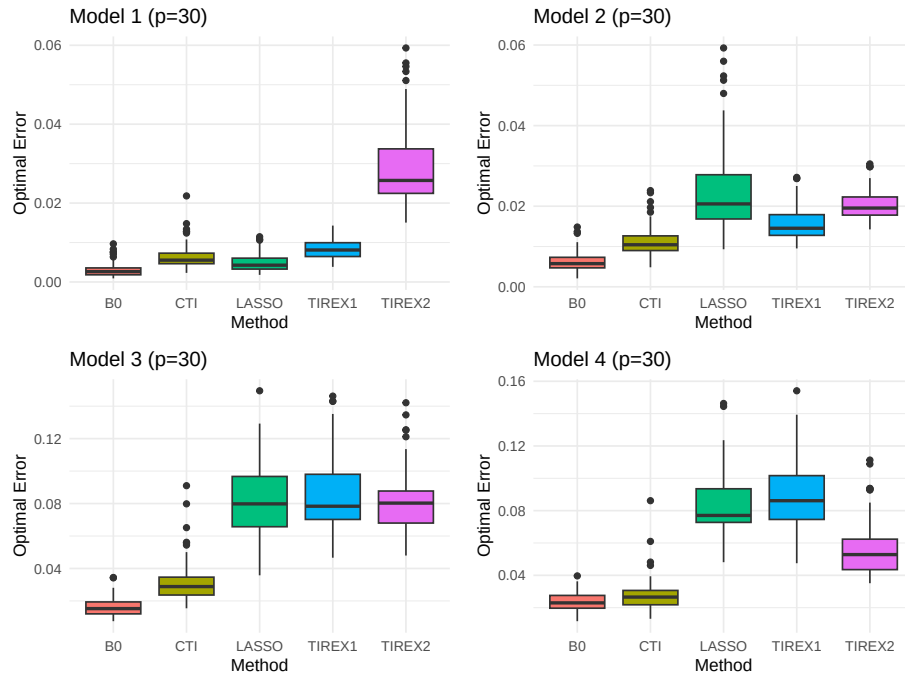


FIGURE 2.2 – Boxplots of the values $\{E^{(r,opt)}(\check{B}_n); r = 1, \dots, N\}$ for different estimated bases $\check{B}_n$ on $N = 100$ replications of a sample of size $n = 4000$ from Models 1 to 4 with $p = 30$. For the case $\check{B}_n = I_p$ the errors are very high and are not represented (For instance, medians for the four models are respectively 5.64, 0.33, 6.25, 5.91).

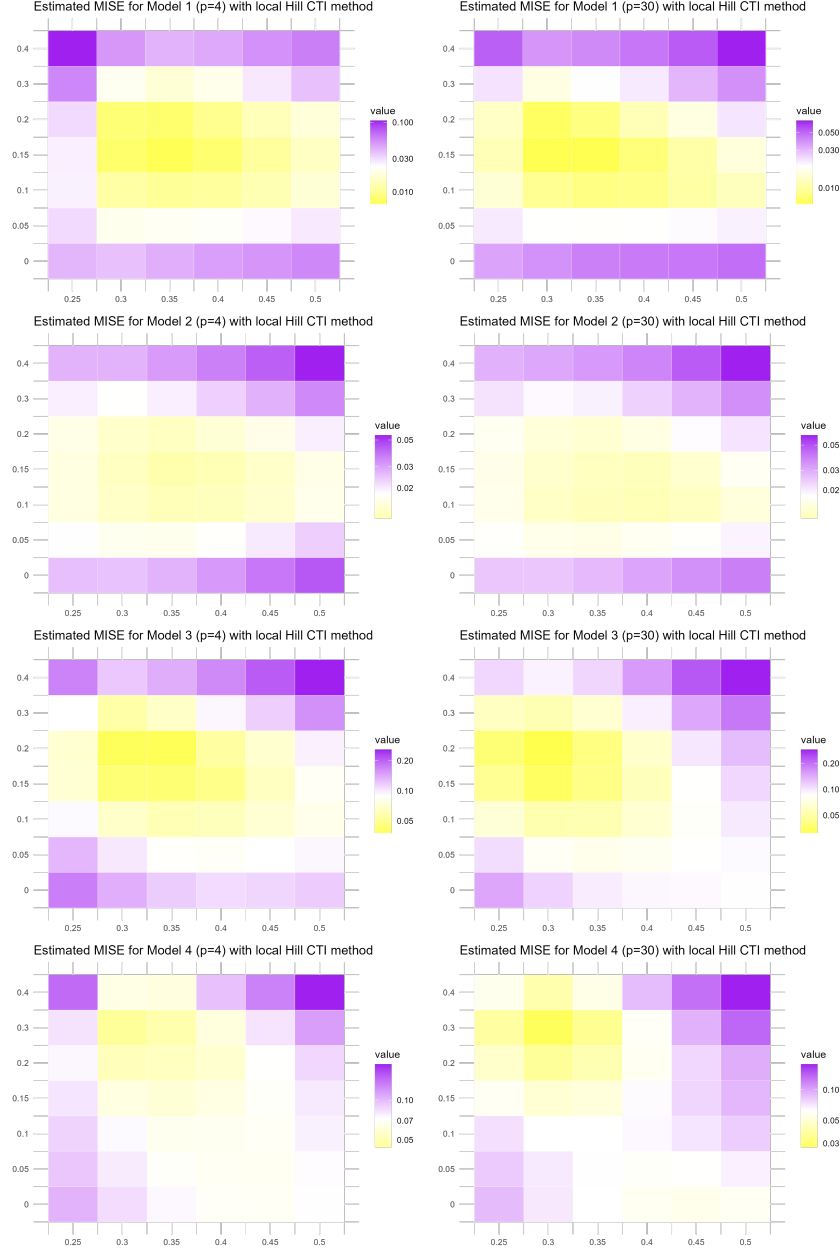


FIGURE 2.3 – Estimated MISE for the proposed CTI estimator over the entire grid of tuning parameters (a, b) (a on x-axis, b on y-axis).

for all $B \in \mathcal{B}_d$ where the function $\Psi(\cdot, \mathcal{X}_0)$ is defined in (2.5). Let also us introduce the notation,

$$c(d, \mathcal{X}_0) := \Psi(B_{0,d}, \mathcal{X}_0).$$

Proposition 5. *Assume that (C.3) holds. If for all $d \in \{1, \dots, p-1\}$ there exists a matrix $C_{d+1} \in \mathcal{B}_{d+1}$ with $\text{span}(B_{0,d}) \subset \text{span}(C_{d+1})$ and such that*

$$\mathbb{P} [\{\xi_{C_{d+1}}(C_{d+1}^\top X) \neq \xi_{B_{0,d}}(B_{0,d}^\top X)\} \cap \{X \in \mathcal{X}_0\}] > 0,$$

then, for all $d \in \{1, \dots, q-1\}$, one has $c(d+1, \mathcal{X}_0) < c(d, \mathcal{X}_0)$ and for all $d \in \{q, \dots, p\}$, $c(d, \mathcal{X}_0) = \mathbb{E}[\gamma(X)\mathbb{I}\{X \in \mathcal{X}_0\}]$.

As a consequence, we get that $q = \min\{d \in \{1, \dots, p\}; c(d, \mathcal{X}_0) \leq c(d+1, \mathcal{X}_0)\}$. We thus naturally propose to estimate the unknown dimension by

$$\hat{q}_n := \min\{d \in \{1, \dots, p\}; \hat{c}_n(d, \mathcal{X}_0) \leq \hat{c}_n(d+1, \mathcal{X}_0)\},$$

where

$$\hat{c}_n(d, \mathcal{X}_0) = \min_{B \in \mathcal{B}_d} \tilde{\Psi}_{n_0}^{(H)}(B, \mathcal{X}_0).$$

As recommended in the previous section, we compute the estimator $\hat{c}_n(d, \mathcal{X}_0)$ with the tuning parameters $h = n^{-b/d}/2$ and $\alpha = n^{-a}$ with $a = 0.3$ and $b = 0.2$. To assess the validity of our estimation procedure, we conduct a simulation study in which the dimension is estimated for the above models with $p = 4$ and $p = 30$. The obtained results are provided in the left part of Tables 2.1 and 2.2. The true dimension is selected in at least 90% of the cases for $q = 1$ and in at least 80% of the cases for $q = 2$. The worst result is obtained for Model 4 with $p = 4$, where our estimator failed to recover the true dimension in 19% of the replications. We also note that our method tends to select smaller dimensions q rather than larger ones.

For $d \in \{1, 2, 3, \hat{q}_n\}$, we also look at the finite sample behavior of the estimator $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(d,H)})$ where

$$\hat{B}_n^{(d,H)} = \arg \min_{B \in \mathcal{B}_d} \hat{\Psi}_n^{(H)}(B, \mathcal{X}_0).$$

The estimated MISE are given in the right part of Tables 2.1 and 2.2. As expected, the smallest value of the estimated MISE is observed when $d = q$ i.e., for the true dimension. However, for all the consider models, using the

Estimation of q				Estimated MISE			
Model	$\hat{q}_n = 1$	$\hat{q}_n = 2$	$\hat{q}_n = 3$	$d = 1$	$d = 2$	$d = 3$	$d = \hat{q}_n$
1 ($q = 1$)	100 %	0 %	0 %	0.012	0.015	0.026	0.012
2 ($q = 1$)	96 %	4 %	0 %	0.014	0.014	0.016	0.014
3 ($q = 1$)	99 %	1 %	0 %	0.066	0.075	0.107	0.065
4 ($q = 2$)	18 %	81 %	1 %	0.085	0.063	0.075	0.065

TABLE 2.1 – Left part of the table : estimation of the CTI subspace dimension over $N = 100$ replications of a sample of size $n = 2000$ generated from Models 1 to 4 with $p = 4$. Right part of the table : Estimated MISE of the local-Hill tail index estimator $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(d,H)})$ for different values of d .

Estimation of q				Estimated MISE			
Model	$\hat{q}_n = 1$	$\hat{q}_n = 2$	$\hat{q}_n = 3$	$d = 1$	$d = 2$	$d = 3$	$d = \hat{q}_n$
1 ($q = 1$)	99 %	1 %	0 %	0.010	0.013	0.020	0.010
2 ($q = 1$)	93 %	6 %	1 %	0.014	0.016	0.018	0.013
3 ($q = 1$)	98 %	2 %	0 %	0.053	0.066	0.090	0.052
4 ($q = 2$)	12 %	88 %	0 %	0.073	0.050	0.062	0.052

TABLE 2.2 – Left part of the table : estimation of the CTI subspace dimension over $N = 100$ replications of a sample of size $n = 4000$ generated from Models 1 to 4 with $p = 30$. Right part of the table : Estimated MISE of the local-Hill tail index estimator for different values of d .

estimated dimension $\hat{q}_n$ instead of the true dimension provides very similar results.

Let us now focus on the finite sample behavior of the CTI estimator $\hat{B}_n^{(H)}$, computed with the estimated dimension $\hat{q}_n$ and the recommended choice of hyper parameters, i.e., $(\alpha_n, h_n) = (n^{-0.3}, n^{-0.2/\hat{q}_n}/2)$. We follow the procedure described in Section 2.4.4 to compare the errors obtained for the CTI estimator with the optimal errors $\{E^{(r,opt)}(\hat{B}_n); r = 1, \dots, N\}$ for $\hat{B}_n \in \{\hat{B}_n^{(G)}, \hat{B}_n^{(T_1)}, \hat{B}_n^{(T_2)}, \hat{B}_n^{(L)}, B_0, I_n\}$. Let us highlight again that, except for the CTI estimator, all estimators are obtained in the best possible scenario, i.e., with the true dimension q and the optimal hyper parameters. The corresponding boxplots are provided in Figures 2.4 and 2.5. It appears that, even though this study does not favor our estimator, it remains superior to or at least equivalent to its competitors.

Finally, we are interested in robustness to an incorrect choice of the dimension q . We focus on the three estimators CTI, TDR, and TIREX2, since the other approaches (TIREX1 and LASSO-QR) are consistent only when $q = 1$. For all models with $p \in \{4, 30\}$, we compute the three estimated bases $\hat{B}_n^{(d,H)}$, $\hat{B}_n^{(d,G)}$, and $\hat{B}_n^{(d,T_2)}$ when the dimension reduction subspace is assumed to be of dimension $d \in \{1, 2, 3\}$. For each method, the hyper parameters are set to $(\alpha_n, h_n) = (n^{-0.3}, n^{-b/d}/2)$. We then obtain the estimated MISE of the local-Hill tail index estimators $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(d,H)})$, $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(d,G)})$, and $\hat{\gamma}_n^{(H)}(x, \hat{B}_n^{(d,T_2)})$. The obtained results are available in Tables 2.3 and 2.4. It appears that the influence of the choice of d , the dimension of the estimated reduction subspace, strongly depends on the model and the estimation procedure. For instance, under Model 2 with $p \in \{4, 30\}$, the CTI and TIREX2 approaches are quite robust to an incorrect choice of the dimension. Conversely, under Model 3 with $p = 4$, the ratio between the estimated MISE obtained with the incorrect dimension $d = 3$ and the estimated MISE with the true dimension ($q = 1$) is approximately 2.1 for TIREX2 and 1.7 for CTI. These results confirm the necessity of estimating the dimension q . Let us highlight again that the estimator $\hat{q}_n$ proposed in this work always leads to an error very close to the lowest one.

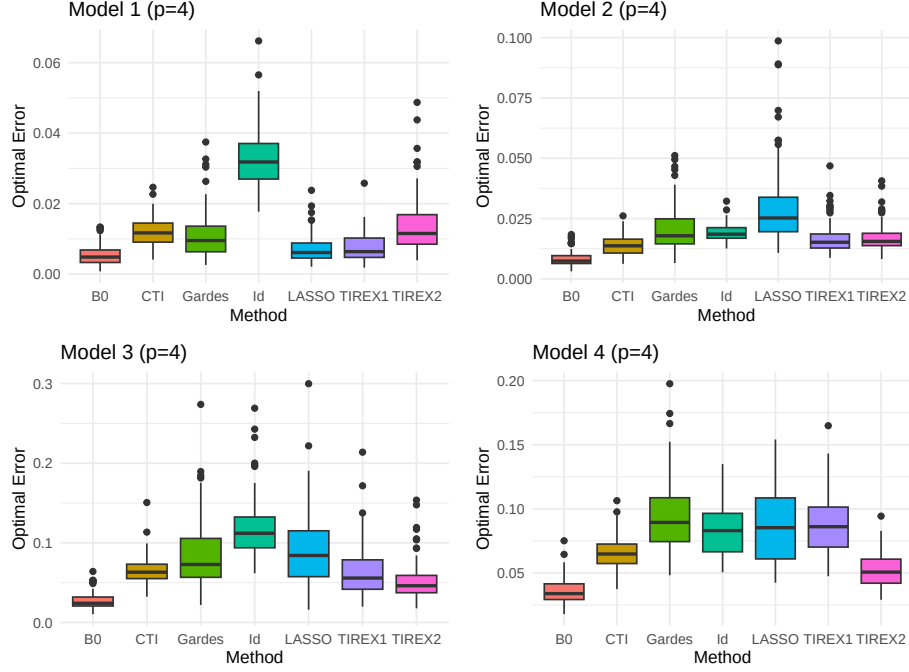


FIGURE 2.4 – Boxplots of the errors for different choices for $\check{B}_n$ on $N = 100$ replications of a sample of size $n = 2000$ from Models 1 to 4 with $p = 4$. The optimal error is considered for all the estimators with the true dimension q except for the CTI estimator which is computed with the estimated dimension $\hat{q}_n$ and the hyper parameters $(\alpha_n, h_n) = (n^{-0.3}, n^{-0.2/\hat{q}_n}/2)$.

	$\check{B}_n^{(d)} = \hat{B}_n^{(d,H)}$			$\check{B}_n^{(d)} = \hat{B}_n^{(d,G)}$			$\check{B}_n^{(d)} = \hat{B}_n^{(d,T_2)}$		
Model	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$
1 ($q = 1$)	0.012	0.015	0.026	0.033	0.069	0.069	0.034	0.043	0.0470
2 ($q = 1$)	0.014	0.014	0.016	0.028	0.041	0.036	0.019	0.018	0.019
3 ($q = 1$)	0.066	0.075	0.107	0.133	0.159	0.158	0.057	0.084	0.122
4 ($q = 2$)	0.086	0.063	0.076	0.148	0.118	0.116	0.088	0.056	0.087

TABLE 2.3 – For samples of size $n = 2000$ from Models 1 to 4 with $p = 4$, values of the estimated MISE of the local-Hill tail index estimator for different estimated dimension reduction subspaces $\check{B}_n^{(d)} \in \{\hat{B}_n^{(d,H)}, \hat{B}_n^{(d,G)}, \hat{B}_n^{(d,T_2)}\}$ with different values of $d \in \{1, 2, 3\}$.

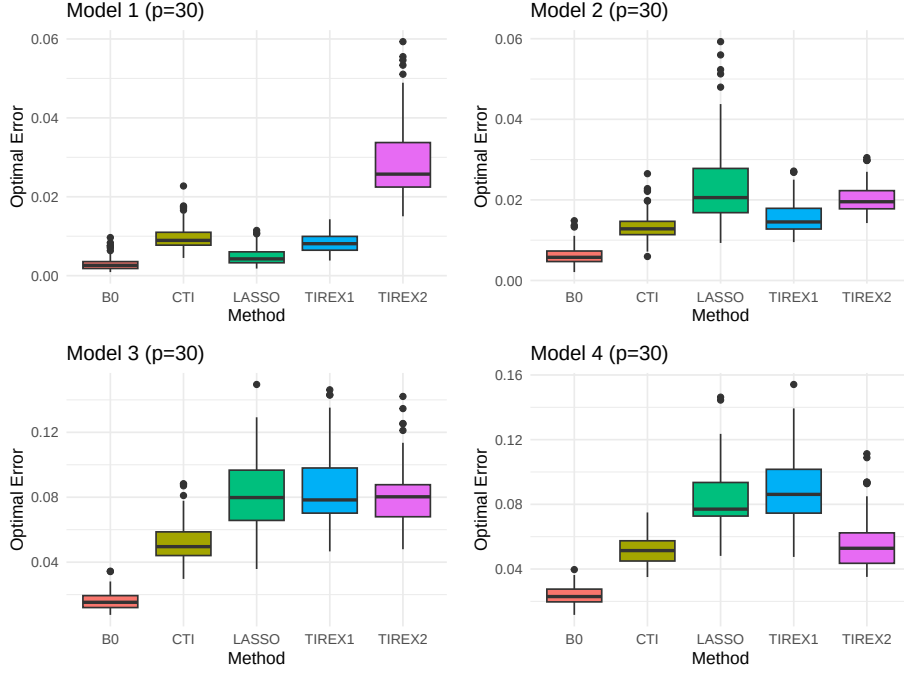


FIGURE 2.5 – Boxplots of the errors for different choices for $\check{B}_n$ on $N = 100$ replications of a sample of size $n = 4000$ from Models 1 to 4 with $p = 30$. The optimal error is considered for all the estimators with the true dimension q except for the CTI estimator which is computed with the estimated dimension $\hat{q}_n$ and the hyper parameters $(\alpha_n, h_n) = (n^{-0.3}, n^{-0.2/\hat{q}_n}/2)$. For the case $\check{B}_n = I_p$ the errors are very high and are not represented (For instance, medians for the four models are respectively 5.64, 0.33, 6.25, 5.91).

	$\check{B}_n^{(d)} = \hat{B}_n^{(d,H)}$			$\check{B}_n^{(d)} = \hat{B}_n^{(d,T_2)}$		
Model	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$
1 ($q = 1$)	0.010	0.013	0.020	0.044	0.041	0.043
2 ($q = 1$)	0.014	0.016	0.018	0.021	0.023	0.024
3 ($q = 1$)	0.053	0.066	0.090	0.088	0.097	0.107
4 ($q = 2$)	0.073	0.050	0.062	0.086	0.074	0.074

TABLE 2.4 – For samples of size $n = 4000$ from Models 1 to 4 with $p = 30$, values of the estimated MISE of the local-Hill tail index estimator for different estimated dimension reduction subspaces $\check{B}_n^{(d)} \in \{\hat{B}_n^{(d,H)}, \hat{B}_n^{(d,T_2)}\}$ with different values of $d \in \{1, 2, 3\}$.

2.5 Application to a real data set

As the world's economies and industries develop, people are increasingly exposed to pollution, especially in urban areas. Some pollutants, such as ozone, can cause serious health problems. It forms under certain meteorological conditions (temperature, humidity, etc.) from primary pollutants emitted directly into the atmosphere by human activities. The aim here is to identify the primary pollutants that can lead to extreme ozone concentrations.

Our study is based on the record of $n = 4841$ daily concentration pollutants in Chicago from 1987 to 2000. These data were obtained few years ago from the Internet-based Health and Air Pollution Surveillance System (iHAPSS) website but unfortunately they are no longer available. The data set provides n observations of the maximum daily concentration of ozone (O_3) and of several primary pollutants including the maximum daily concentration of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), particulate matter with a diameter smaller than 10 micrometers (PM_{10}) and carbon monoxide (CO). This data set was considered by many authors to illustrate the effect of dimension reduction (see, for example, [33, 62, 76]).

The response variable is given by $Y = \exp(O_3/\sigma_{O_3})$ where $\sigma_{O_3}^2$ is the variance of O_3 . In practice this variance is replaced by its empirical counterpart. The considered covariate is the random vector $X = (X_{NO_2}, X_{SO_2}, X_{PM_{10}}, X_{CO})^\top \in [0, 1]^p$ of dimension $p = 4$ with $X_\bullet = [\bullet - \max(\bullet)]/[\min(\bullet) - \max(\bullet)]$. It is a normalized version of the random vector $(NO_2, SO_2, PM_{10}, CO)$. The exponential function is applied to the maximum daily concentration of ozone in order to obtain a heavy-tailed distributed random variable. By looking at the Pareto quantile plot $\{(\ln(k/i), \ln(Y_{n-i,n}/Y_{n-k,n})); i \in \{1, \dots, k = 200\}\}$, see Figure 2.6, we can see that the points are approximately located on a straight line which is what it is expected for a heavy-tailed distribution. It should be noted that we should check that the conditional distribution of Y given $X = x$ is heavy-tailed for all x in the support of X . This task cannot be done in practice, but the fact that Y is heavy-tailed allows us to hope that the same is true for conditional distributions.

We apply the estimation procedure proposed in this paper to this data set with $\mathcal{X}_0 := [0.02, 0.4]^p$. This compact contains around 88% of the observations. Following the recommendation made in the simulation study, the tuning parameters are set to $a = 0.3$ and $b = 0.2$. Applying the procedure described in Section 2.4.5, we find that $\hat{q}_n = 1$ (with $\hat{c}_n(1, \mathcal{X}_0) = 0.531$ and

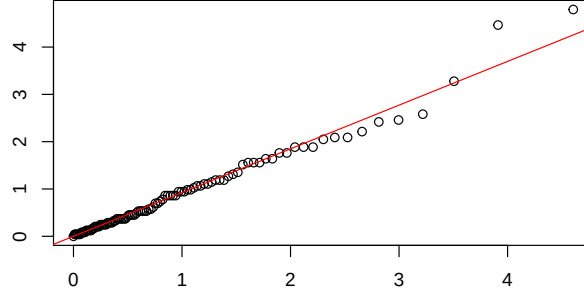


FIGURE 2.6 – Pareto quantile plot (x-axis : $\ln(k/i)$, y-axis : $\ln(Y_{n-i,n}/Y_{n-k,n})$ for $i \in \{0, \dots, k = 200\}$. The fitted line is depicted in red.)

$\hat{c}_n(2, \mathcal{X}_0) = 0.614$). By applying the minimization procedure described in Section 2.4.2, the estimated basis of the CTI subspace is given by

$$\hat{B}_n^{(H)} = (0.831, 0.095, 0.500, 0.223)^\top.$$

Note that the two largest components of $\hat{B}_n^{(H)}$ correspond to NO_2 and PM_{10} . In Figure 2.7, the $(100 \times \alpha)\%$ largest observations of O_3 are represented versus the projection of the covariate unto the estimated CTI subspace. It appears that an increase in the extreme values of ozone is associated with increasing values of the 1-dimensional reduced covariate that is to say with increasing values of NO_2 and PM_{10} .

We can now estimate the conditional tail index function. The estimated values $\hat{\gamma}_n^{(H)}(x_i, \hat{B}_n^{(H)})$ for $x_i \in \mathcal{X}_0$ as a function of $[\hat{B}_n^{(H)}]^\top x_i$ are depicted in Figure 2.8. This figure shows a positive correlation between the tail index and the CTI direction. As a conclusion, it seems that the magnitude of extreme ozone concentrations increases with the concentrations of NO_2 and PM_{10} . Note that a study of this data set was also conducted by [62]. The author concluded that ozone concentration (not necessarily extreme) mostly depends on the level of NO_2 . Our study shows that PM_{10} becomes another important factor for extreme ozone concentration.

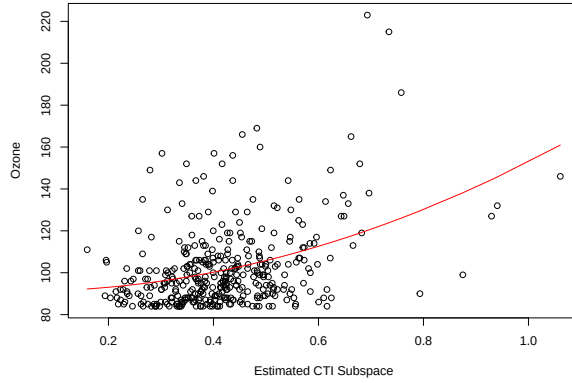


FIGURE 2.7 – Largest observations of O_3 versus the projected covariates on the linear subspace spanned by $\hat{B}_n^{(H)}$. The smooth curve is obtained by fitting a polynomial of order 2.

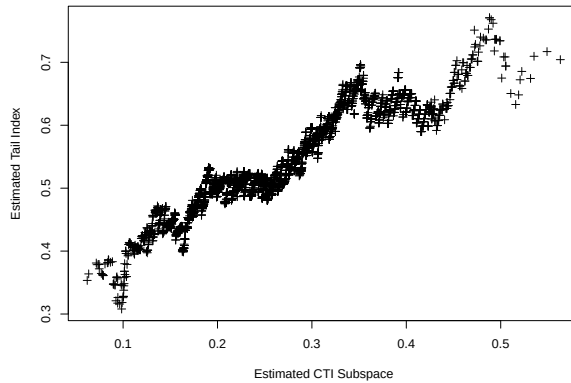


FIGURE 2.8 – Estimated tail index versus the projected covariates on the linear subspace spanned by $\hat{B}_n^{(H)}$.

Conclusion

When the tail index function is positive and defined on a lower-dimensional linear subspace (referred to as the CTI subspace), this paper proposes an estimation procedure for a basis of this subspace, as well as an estimation of

the conditional tail index. The consistency of these estimators is established, and a simulation study demonstrates that our method outperforms those of our main competitors. Additionally, we propose a method for estimating the unknown dimension of the dimension reduction subspace, which yields satisfactory results on synthetic data. The implementation of the estimation procedure described in this paper in an **R** package is currently in progress; however, all the codes are available on a GitHub page.

2.6 Proofs

2.6.1 Preliminary results

Lemma 4. *Assume that model (M) holds with the conditions (C.1) and (C.2). For $B \in \mathcal{B}_q$, if the conditional tail index $\gamma(\cdot)$ and the positive measurable function $c(\cdot)$ involved in (C.2) are such that the function*

$$y \in (0, \infty) \mapsto \mathbb{E} [c(X)y^{-1/\gamma(X)} | B^\top X],$$

is almost surely regularly varying with index $-1/\xi_B(B^\top X) < 0$, then the conditional distribution of Y given $B^\top X$ is heavy-tailed with a tail index $\xi_B(B^\top X)$.

Proof – We have to prove that the conditional survival function $S_B(\cdot, B^\top X)$ is almost surely regularly varying with tail index $-1/\xi_B(B^\top X)$. In the rest of the proof, all the equalities and inequalities are assumed to hold almost surely. We start with the following equalities.

$$\begin{aligned} S_B(y, B^\top X) &= \mathbb{E} [S(y, X) | B^\top X] = \int_0^1 \mathbb{P} (S(y, X) > z | B^\top X) dz \\ &= \int_0^1 \mathbb{P} \left(\frac{1}{\gamma(X)} < \frac{\ln(\mathcal{L}(y, X))}{\ln(y)} - \frac{\ln(z)}{\ln(y)} \middle| B^\top X \right) dz. \end{aligned}$$

Set $s = -\ln(z)/\ln(y)$. We get

$$S_B(y | B^\top X) = \ln(y) \int_0^\infty y^{-s} \mathbb{P} \left(\frac{1}{\gamma(X)} < s + \frac{\ln(\mathcal{L}(y, X))}{\ln(y)} \middle| B^\top X \right) ds$$

By assumption (C.2), for all $\varepsilon > 0$, there exists $\mathcal{Y}_0 > 0$ such that for all $y \geq \mathcal{Y}_0$,

$$S_B(y | B^\top X) \leq \ln(y) \int_0^\infty y^{-s} \mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < s + \frac{\varepsilon}{\ln(y)} \middle| B^\top X \right) ds.$$

Hence, let $t = s + \varepsilon/\ln(y)$,

$$S_B(y|B^\top X) \leq e^\varepsilon \ln(y) \int_{\varepsilon/\ln(y)}^{\infty} y^{-t} \mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < t \middle| B^\top X \right) dt.$$

Since for y large enough,

$$\sup_{x \in \mathcal{X}} \left(\frac{\varepsilon}{\ln(y)} + \frac{\ln(c(x))}{\ln(y)} \right) < \inf_{x \in \mathcal{X}} \frac{1}{\gamma(x)},$$

we have, for y large enough, that

$$\mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < \frac{\varepsilon}{\ln(y)} \middle| B^\top X \right) = 0.$$

As a consequence,

$$\int_0^{\varepsilon/\ln(y)} y^{-s} \mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < s \middle| B^\top X \right) ds = 0,$$

leading to

$$S_B(y|B^\top X) \leq e^\varepsilon \ln(y) \int_0^{\infty} y^{-t} \mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < t \middle| B^\top X \right) dt.$$

In the same way, we have for all $\varepsilon > 0$ that

$$S_B(y|B^\top X) \geq e^{-\varepsilon} \ln(y) \int_0^{\infty} y^{-t} \mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < t \middle| B^\top X \right) dt.$$

Remarking that

$$\ln(y) \int_0^{\infty} y^{-t} \mathbb{P} \left(\frac{1}{\gamma(X)} - \frac{\ln(c(X))}{\ln(y)} < t \middle| B^\top X \right) dt = \mathbb{E} [c(X) y^{-1/\gamma(X)} | B^\top X],$$

we have shown that for all $\varepsilon > 0$, we have for y large enough, that

$$e^{-\varepsilon} \mathbb{E} [c(X) y^{-1/\gamma(X)} | B^\top X] \leq S(y|B^\top X) \leq e^\varepsilon \mathbb{E} [c(X) y^{-1/\gamma(X)} | B^\top X],$$

which is the desired result. ■

The next result establishes the uniform consistency of the general estimator $\widehat{\Psi}_n(B, \mathcal{X}_0)$ of $\Psi(B, \mathcal{X}_0)$.

Lemma 5. *Assume that model (M) holds together with conditions (C.4) and (C.5). Then,*

$$\sup_{B \in \mathcal{B}_q} \left| \widehat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right| \xrightarrow{\mathbb{P}} 0.$$

Proof – We start by the inequality

$$\sup_{B \in \mathcal{B}_q} \left| \widehat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right| \leq T_{n,1} + T_{n,2},$$

$$T_{n,1} := \sup_{B \in \mathcal{B}_q} \left| \widetilde{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right|$$

and $T_{n,2} := \sup_{B \in \mathcal{B}_q} \sup_{x \in \mathcal{X}_0} \left| \widehat{\xi}_{n,B}(B^\top x) - \xi_B(B^\top x) \right|.$

Condition (C.5) ensures that $T_{n,2} \xrightarrow{\mathbb{P}} 0$. For the term $T_{n,1}$, we must prove a uniform law of large numbers for the set of parametric functions $\mathcal{F}_q := \{x \in \mathcal{X} \mapsto \xi_B(B^\top x) \mathbb{I}\{X \in \mathcal{X}_0\}; B \in \mathcal{B}_q\}$. This result is established by compiling different results that can be found for instance in the monograph of [69]. Note that under (C.1), $\mathcal{F}_q \subset L_1(X)$ where $L_1(X)$ is the set of all functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $\mathbb{E}(|f(X)|) < \infty$. Let $\|\cdot\|_1$ be the $L_1(X)$ -norm defined by $\|f\|_1 = \mathbb{E}(|f(X)|)$. According to [69, Theorem 2.4.1], the result of Lemma 5 is true if the bracketing number $N_{[]}(\varepsilon, \mathcal{F}_q, \|\cdot\|_1)$ is finite for all $\varepsilon > 0$ (see [69, Definition 2.1.6] for a definition of bracketing numbers). We now use [69, Theorem 2.7.11] ensuring that under (C.4), one has for all $\varepsilon > 0$,

$$N_{[]}(\varepsilon, \mathcal{F}_q, \|\cdot\|_1) \leq N\left(\frac{\varepsilon}{2\mathbb{E}(M(X))}, \mathcal{B}_q, \|\cdot\|\right),$$

where for all $\eta > 0$ and $\|\cdot\|$ any norm in $\mathbb{R}^{p \times q}$, $N(\eta, \mathcal{B}_q, \|\cdot\|)$ is the covering number of the set $\mathcal{B}_q$ (see [69, Definition 2.1.5]). We conclude the proof by using the compacity of $\mathcal{B}_q$ and the fact that the covering number of a compact subset is finite. $\blacksquare$

2.6.2 Proofs of main results

Proof of Proposition 4 – Thanks to Lemma 4, we only need to prove that for all $B \in \mathcal{B}_q$, the function

$$y \in (0, \infty) \mapsto \mathbb{E} [c(X)y^{1/\gamma(X)} | B^\top X], \quad (2.7)$$

is almost surely regularly varying with index $-1/\xi_B(B^\top X) < 0$. According to [39, Theorem 1.2.1 and Remark 1.2.3], for any positive random variable Z , the limit

$$\lim_{y \rightarrow \infty} \frac{\mathbb{E}[\ln(Z/y)\mathbb{I}\{Z > y\} | B^\top X]}{\mathbb{P}(Z > y | B^\top X)} = \xi_B(B^\top X), \quad (2.8)$$

holds almost surely if and only if the function

$$y \in (0, \infty) \mapsto \mathbb{P}(Z > y | B^\top X) = \mathbb{E} [\mathbb{P}(Z > y | X) | B^\top X],$$

is almost surely regularly varying with index $-1/\xi_B(B^\top X) < 0$. Taking for Z a positive random variable such that for all $x \in \mathcal{X}$ and $y > [c(x)]^{\gamma(x)}$,

$$\mathbb{P}(Z > y | X = x) = c(x)y^{-1/\gamma(x)},$$

proving (2.7) is thus equivalent to prove (2.8). In the rest of the proof, all equalities, inequalities, limits, etc. are assumed to hold almost surely. Using the tower property of the conditional expectation, we get

$$\mathbb{E}[\ln(Z/y)\mathbb{I}\{Z > y\} | B^\top X] = \mathbb{E} [\mathbb{E}[\ln(Z/y)\mathbb{I}\{Z > y\} | X] | B^\top X].$$

Moreover

$$\begin{aligned} \mathbb{E}[\ln(Z/y)\mathbb{I}\{Z > y\} | X] &= \int_0^\infty \mathbb{P}(Z > ye^z | X) dz \\ &= \gamma(X)c(X)y^{-1/\gamma(X)} = \gamma(X)\mathbb{P}(Z > y | X). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}[\ln(Z/y)\mathbb{I}\{Z > y\} | B^\top X] &= \mathbb{E} [\gamma(X)\mathbb{P}(Z > y | X) | B^\top X] \\ &= \mathbb{E} [\mathbb{E}[\gamma(X)\mathbb{I}\{Z > y\} | X] | B^\top X] = \mathbb{E}[\gamma(X)\mathbb{I}\{Z > y\} | B^\top X] \\ &= \int_0^{\xi_B(B^\top X)} \mathbb{P}(\{\gamma(X) > t\} \cap \{Z > y\} | B^\top X) dt. \end{aligned}$$

As a consequence, prove (2.8) is equivalent to prove that

$$\lim_{y \rightarrow \infty} \int_0^{\xi_B(B^\top X)} \frac{\mathbb{P}(\{\gamma(X) > t\} \cap \{Z > y\} | B^\top X)}{\mathbb{P}(Z > y | B^\top X)} dt = \xi_B(B^\top X).$$

Since the integrand is positive and smaller than 1, one can use the dominated convergence theorem to exchange limit and integral. It thus finally remains to prove that for all $t \in (0, \xi_B(B^\top X))$,

$$\lim_{y \rightarrow \infty} \frac{\mathbb{P}(\{\gamma(X) > t\} \cap \{Z > y\} | B^\top X)}{\mathbb{P}(Z > y | B^\top X)} = 1.$$

Under (C.1), the distribution of the random variable $\gamma(X)$ is absolutely continuous and thus $\{\gamma(X) > t\} = \{\gamma(X) \geq t\}$. This leads to the equality

$$\frac{\mathbb{P}(\{\gamma(X) > t\} \cap \{Z > y\} | B^\top X)}{\mathbb{P}(Z > y | B^\top X)} = 1 - \frac{\mathbb{P}(\{\gamma(X) < t\} \cap \{Z > y\} | B^\top X)}{\mathbb{P}(Z > y | B^\top X)}.$$

The rest of the proof consists in showing that for all $t \in (0, \xi_B(B^\top X))$,

$$\lim_{y \rightarrow \infty} y^{1/t} \mathbb{P}(\{\gamma(X) < t\} \cap \{Z > y\} | B^\top X) = 0, \quad (2.9)$$

and

$$\lim_{y \rightarrow \infty} y^{1/t} \mathbb{P}(Z > y | B^\top X) = \infty. \quad (2.10)$$

Let us start by proving (2.9). We have

$$\begin{aligned} y^{1/t} \mathbb{P}(\{\gamma(X) > t\} \cap \{Z > y\} | B^\top X) &= \mathbb{E}[y^{1/t} \mathbb{I}\{\gamma(X) < t\} \mathbb{P}(Z > y | X) | B^\top X] \\ &= \mathbb{E}[c(X) y^{1/t-1/\gamma(X)} \mathbb{I}\{\gamma(X) < t\} | B^\top X]. \end{aligned}$$

Since for $y > 1$, one has $c(X) y^{1/t-1/\gamma(X)} \mathbb{I}\{\gamma(X) < t\} \leq c(X)$ and since by assumption $c(\cdot)$ is bounded, we can apply the dominated convergence theorem to get

$$\begin{aligned} &\lim_{y \rightarrow \infty} y^{1/t} \mathbb{P}(\{\gamma(X) > t\} \cap \{Z > y\} | B^\top X) \\ &= \mathbb{E}\left[c(X) \mathbb{I}\{\gamma(X) < t\} \lim_{y \rightarrow \infty} y^{1/t-1/\gamma(X)} \middle| B^\top X\right] = 0, \end{aligned}$$

proving (2.9). Let us now turn to the proof of (2.10). First, note that

$$\mathbb{P}(Z > y | B^\top X) = \mathbb{E}[\mathbb{P}(Z > y | X) | B^\top X] = \mathbb{E}[c(X) y^{-1/\gamma(X)} | B^\top X].$$

Moreover, for $t \in (0, \xi_B(B^\top X))$, let $\varepsilon_t(B^\top X) := (t^{-1} - \xi_B^{-1}(B^\top X))/2$. Since $\xi_B(B^\top X)$ is the right endpoint of the conditional distribution of $\gamma(X)$ given $B^\top X$, we have

$$\mathbb{P}(\gamma^{-1}(X) < t^{-1} - \varepsilon_t(B^\top X) | B^\top X) > 0.$$

Hence, for all $y > 0$,

$$\begin{aligned} y^{1/t} \mathbb{P}(Z > y | B^\top X) &\geq \mathbb{E} \left[c(X) y^{t^{-1} - \gamma^{-1}(X)} \mathbb{I}\{\gamma^{-1}(X) \leq t^{-1} - \varepsilon_t(B^\top X)\} \middle| B^\top X \right] \\ &\geq \inf_{x \in \mathcal{X}} c(x) \mathbb{P}(\gamma^{-1}(X) < t^{-1} - \varepsilon_t(B^\top X) | B^\top X) y^{\varepsilon_t(B^\top X)}, \end{aligned}$$

which converges to infinity as $y \rightarrow \infty$ since under (C.2), $\inf_{x \in \mathcal{X}} c(x) > 0$. ■

Proof of Theorem 2 – The scheme of this proof is similar to the one used in the proof of [48, Theorem 5.1]. From the definition (5) of $\hat{B}_n$, we have

$$\mathbb{P}(\hat{\Psi}_n(\hat{B}_n, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)) = 1.$$

For a norm $\|\cdot\|$ in $\mathbb{R}^{p \times q}$, let $U(B_0, \rho)$ be the open ball in $\mathbb{R}^{p \times q}$ of center B_0 and radius $\rho > 0$. For all $\rho > 0$,

$$\begin{aligned} 1 &= \mathbb{P}(\hat{\Psi}_n(\hat{B}_n, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)) \\ &= \mathbb{P}\left(\left\{\hat{\Psi}_n(\hat{B}_n, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)\right\} \cap \{\hat{B}_n \in U(B_0, \rho)\}\right) \\ &\quad + \mathbb{P}\left(\left\{\hat{\Psi}_n(\hat{B}_n, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)\right\} \cap \{\hat{B}_n \notin U(B_0, \rho)\}\right). \end{aligned}$$

Hence, since

$$\mathbb{P}\left(\left\{\hat{\Psi}_n(\hat{B}_n, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)\right\} \cap \{\hat{B}_n \in U(B_0, \rho)\}\right) \leq \mathbb{P}(\hat{B}_n \in U(B_0, \rho)),$$

and

$$\begin{aligned} &\mathbb{P}\left(\left\{\hat{\Psi}_n(\hat{B}_n, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)\right\} \cap \{\hat{B}_n \notin U(B_0, \rho)\}\right) \\ &\leq \mathbb{P}\left(\inf_{B \notin U(B_0, \rho)} \hat{\Psi}_n(B, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)\right), \end{aligned}$$

we have that

$$\mathbb{P}(\hat{B}_n \in U(B_0, \rho)) + \mathbb{P}\left(\inf_{B \notin U(B_0, \rho)} \hat{\Psi}_n(B, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0)\right) \geq 1.$$

As a consequence, if we show that

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\inf_{B \notin U(B_0, \rho)} \hat{\Psi}_n(B, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0) \right) = 0, \quad (2.11)$$

then, for all $\rho > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\hat{B}_n \in U(B_0, \rho) \right) = 1,$$

which is the desired result. To prove (2.11), let us remark first that

$$\inf_{B \notin U(B_0, \rho)} \left(\hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right) + \inf_{B \notin U(B_0, \rho)} \Psi(B, \mathcal{X}_0) \leq \inf_{B \notin U(B_0, \rho)} \hat{\Psi}_n(B, \mathcal{X}_0).$$

Hence,

$$\begin{aligned} p_n &:= \mathbb{P} \left(\inf_{B \notin U(B_0, \rho)} \hat{\Psi}_n(B, \mathcal{X}_0) \leq \hat{\Psi}_n(B_0, \mathcal{X}_0) \right) \\ &\leq \mathbb{P} \left(\inf_{B \notin U(B_0, \rho)} \left(\hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right) \right. \\ &\quad \left. + \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \leq \Psi(B_0, \mathcal{X}_0) - \inf_{B \notin U(B_0, \rho)} \Psi(B, \mathcal{X}_0) \right) \end{aligned}$$

Now, since for all $(x, y) \in \mathbb{R}^2$, if $x \leq y$ then $|x| \geq -y$, we have

$$\begin{aligned} p_n &\leq \mathbb{P} \left(\left| \inf_{B \notin U(B_0, \rho)} \left(\hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right) \right. \right. \\ &\quad \left. \left. + \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \right| \geq \inf_{B \notin U(B_0, \rho)} \Psi(B, \mathcal{X}_0) - \Psi(B_0, \mathcal{X}_0) \right) \end{aligned}$$

Under condition (C.3), for all $\rho > 0$, there exists $\varepsilon > 0$ such that

$$\inf_{B \notin U(B_0, \rho)} \Psi(B, \mathcal{X}_0) - \Psi(B_0, \mathcal{X}_0) > \varepsilon.$$

Furthermore,

$$\begin{aligned} &\left| \inf_{B \notin U(B_0, \rho)} \left(\hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right) + \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \right| \\ &\leq \left| \inf_{B \notin U(B_0, \rho)} \left(\hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right) \right| + \left| \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \right| \\ &\leq \sup_{B \notin U(B_0, \rho)} \left| \hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right| + \left| \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \right| \\ &\leq \sup_{B \in \mathcal{B}_q} \left| \hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right| + \left| \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \right|. \end{aligned}$$

Hence,

$$p_n \leq \mathbb{P} \left(\sup_{B \in \mathcal{B}_q} \left| \hat{\Psi}_n(B, \mathcal{X}_0) - \Psi(B, \mathcal{X}_0) \right| + \left| \Psi(B_0, \mathcal{X}_0) - \hat{\Psi}_n(B_0, \mathcal{X}_0) \right| \geq \varepsilon \right).$$

We conclude the proof by using Lemma 5. ■

Proof of Theorem 3 – Let $\varepsilon > 0$. We need to prove that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{P} \left[\sup_{x \in \mathcal{X}_0} \left| \hat{\gamma}_n(x, \hat{B}_n) - \gamma(x) \right| > \varepsilon \right] \\ &= \lim_{n \rightarrow \infty} \mathbb{P} \left[\sup_{x \in \mathcal{X}_0} \left| \hat{\xi}_{n, \hat{B}_n}([\hat{B}_n]^\top x) - \xi_{B_0}(B_0^\top x) \right| > \varepsilon \right] \\ &= 0. \end{aligned}$$

We start with the following decomposition : let $\eta \in]0, \varepsilon/(2\bar{M})[$,

$$\mathbb{P} \left[\sup_{x \in \mathcal{X}_0} \left| \hat{\gamma}_n(x, \hat{B}_n) - \gamma(x) \right| > \varepsilon \right] =: T_{1,n} + T_{2,n},$$

where

$$\begin{aligned} T_{1,n} &= \mathbb{P} \left[\left\{ \sup_{x \in \mathcal{X}_0} \left| \hat{\gamma}_n(x, \hat{B}_n) - \gamma(x) \right| > \varepsilon \right\} \cap \left\{ \|\hat{B}_n - B_0\| < \eta \right\} \right] \\ \text{and } T_{2,n} &= \mathbb{P} \left[\left\{ \sup_{x \in \mathcal{X}_0} \left| \hat{\gamma}_n(x, \hat{B}_n) - \gamma(x) \right| > \varepsilon \right\} \cap \left\{ \|\hat{B}_n - B_0\| \geq \eta \right\} \right] \end{aligned}$$

Of course, from Theorem 2,

$$T_{2,n} \leq \mathbb{P} \left(\|\hat{B}_n - B_0\| \geq \eta \right) \rightarrow 0.$$

Next,

$$\begin{aligned} T_{1,n} &\leq \mathbb{P} \left[\sup_{B: \|B - B_0\| < \eta} \sup_{x \in \mathcal{X}_0} \left| \hat{\xi}_{n,B}(B^\top x) - \xi_{B_0}(B_0^\top x) \right| > \varepsilon \right] \\ &\leq \mathbb{P} \left[\sup_{B: \|B - B_0\| < \eta} \sup_{x \in \mathcal{X}_0} \left| \hat{\xi}_{n,B}(B^\top x) - \xi_B(B^\top x) \right| > \varepsilon/2 \right] \\ &\quad + \mathbb{I} \left\{ \sup_{B: \|B - B_0\| < \eta} \sup_{x \in \mathcal{X}_0} \left| \xi_B(B^\top x) - \xi_{B_0}(B_0^\top x) \right| > \varepsilon/2 \right\}. \end{aligned}$$

The first term converges to 0 according to (C.5) and the second term is equal to 0 since from (C.4),

$$\sup_{B: \|B-B_0\| < \eta} \sup_{x \in \mathcal{X}_0} |\xi_B(B^\top x) - \xi_{B_0}(B_0^\top x)| \leq \overline{M}\eta < \varepsilon/2.$$

The proof is then complete. $\blacksquare$

Proof of Proposition 5 – Let us start with some notations. For all matrix $B \in \mathcal{B}^d$ with $d \in \{1, \dots, p\}$ and for any $x \in \mathcal{X}$, let

$$\mathcal{A}_x(B) := \{z \in \mathbb{R}^p; B^\top z = B^\top x\}$$

. Note that, according to (2.3), $\xi_B(B^\top x) = \max\{\gamma(z); z \in \mathcal{A}_x(B)\}$. Let $C_{d+1} \in \mathcal{B}_{d+1}$ be a matrix satisfying the condition of Proposition 5. Since $\mathcal{A}_x(C_{d+1}) \subset \mathcal{A}_x(B_{0,d})$, for all $x \in \mathcal{X}$, we have $\xi_{C_{d+1}}(C_{d+1}^\top x) \leq \xi_{B_{0,d}}(B_{0,d}^\top x)$ and thus

$$\begin{aligned} \min_{B \in \mathcal{B}_{d+1}} \Psi(B, \mathcal{X}_0) &= \min_{B \in \mathcal{B}_{d+1}} \mathbb{E} [\xi_B(B^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}] \leq \mathbb{E} [\xi_{C_{d+1}}(C_{d+1}^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}] \\ &\leq \mathbb{E} [\xi_{B_{0,d}}(B_{0,d}^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}] = \min_{B \in \mathcal{B}_d} \Psi(B, \mathcal{X}_0), \end{aligned} \quad (2.12)$$

for all $d \in \{1, \dots, p\}$. In particular, the assumption on the matrix C_{d+1} ensures that, for all $d \in \{1, \dots, q-1\}$,

$$\min_{B \in \mathcal{B}_{d+1}} \Psi(B, \mathcal{X}_0) < \min_{B \in \mathcal{B}_d} \Psi(B, \mathcal{X}_0).$$

Moreover, since for all $B \in \mathcal{B}_d$ and $x \in \mathcal{X}$, $\xi_B(B^\top x) = \max\{\gamma(z); B^\top z = B^\top x\} \geq \gamma(x)$, we have, for all $d \in \{1, \dots, p\}$,

$$\mathbb{E} [\gamma(X) \mathbb{I}\{X \in \mathcal{X}_0\}] \leq \min_{B \in \mathcal{B}_d} \Psi(B, \mathcal{X}_0). \quad (2.13)$$

From (2.12), we also have that for all $d \geq q$,

$$\min_{B \in \mathcal{B}_d} \Psi(B, \mathcal{X}_0) \leq \min_{B \in \mathcal{B}_q} \Psi(B, \mathcal{X}_0) = \mathbb{E} [\gamma(X) \mathbb{I}\{X \in \mathcal{X}_0\}]. \quad (2.14)$$

Gathering (2.13) and (2.14) entails that for all $d \geq q$,

$$\min_{B \in \mathcal{B}_d} \Psi(B, \mathcal{X}_0) = \mathbb{E} [\gamma(X) \mathbb{I}\{X \in \mathcal{X}_0\}],$$

and the proof is complete. $\blacksquare$

Appendix – Proof of Theorem 4

We shall prove that the estimator $\hat{\xi}_B^{(H)}(B^\top x; \alpha, h) = \hat{\xi}_B^{(H)}(B^\top x)$ satisfies condition (C.5) i.e., that for all $\varepsilon > 0$, the probability

$$p_n := \mathbb{P} \left(\sup_{(B,x) \in \Theta} \left| \hat{\xi}_B^{(H)}(B^\top x) - \xi_B(B^\top x) \right| > \varepsilon \right),$$

converges to 0 as n goes to infinity. The proof consists in showing the uniform consistency on a sequence of “sufficiently large” subsets Θ_n of Θ and deal with the oscillation of the estimator. We use the same scheme as the one used for the proof of [37, Theorem 1] but here we have to deal with the uniformity on $\mathcal{B}_q \times \mathcal{X}$ and not only on $\mathcal{X}$. This gives rise to technical difficulties.

Since Θ is a compact subset, for a fixed $\eta > q$ and every $n \in \mathbb{N} \setminus \{0\}$, there exists a finite subset Θ_n of Θ such that for all $\theta := (B, x) \in \Theta$, one can find $\tilde{\theta}_n(\theta) = (\tilde{B}_n(\theta), \tilde{x}_n(\theta)) \in \Theta_n$ satisfying $\delta(\tilde{\theta}_n(\theta), \theta) < n^{-\eta}$. Note that Θ_n can be chosen such that for some $c = c(\eta) > 0$, $\text{card}(\Theta_n) = \mathcal{O}(n^c)$ for all $n \in \mathbb{N} \setminus \{0\}$, see [44, Proof of Lemma 1]. The triangular inequality yields :

$$\begin{aligned} p_n &\leq \mathbb{I} \left\{ \sup_{\theta=(B,x) \in \Theta} \left| \xi_B(B^\top x) - \xi_{\tilde{B}_n(\theta)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta) \right) \right| > \varepsilon/3 \right\} \\ &\quad + \mathbb{P} \left(\sup_{\tilde{\theta}=(\tilde{B},\tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)} \left(\tilde{B}^\top \tilde{x} \right) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon/3 \right) \\ &\quad + \mathbb{P} \left(\sup_{\theta=(B,x) \in \Theta} \left| \hat{\xi}_{\tilde{B}_n(\theta)}^{(H)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta) \right) - \hat{\xi}_B^{(H)}(B^\top x) \right| > \varepsilon/3 \right). \end{aligned}$$

We must prove that the three terms in the above inequality converge to 0 as n goes to infinity. This is carried out in Propositions 6, 7 and 8 given below.

Proposition 6. *Under condition (H.3), for n large enough and for all $\varepsilon > 0$,*

$$\sup_{\theta=(B,x) \in \Theta} \left| \xi_B(B^\top x) - \xi_{\tilde{B}_n(\theta)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta) \right) \right| > \varepsilon.$$

Proposition 7. *Under conditions (C.1), (H.1), (H.2), (H.4) and (H.5), if the sequences (α) and (h) are such that $nh^q\alpha/\ln(n) \rightarrow \infty$ and $\alpha \ln(nh^q) \rightarrow 0$ as $n \rightarrow \infty$ then, for all $\varepsilon > 0$,*

$$\mathbb{P} \left(\sup_{\tilde{\theta}=(\tilde{B},\tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)} \left(\tilde{B}^\top \tilde{x} \right) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right) \rightarrow 0.$$

Proposition 8. *Under conditions (C.1), (H.1), (H.2), (H.4) and (H.5), if the sequences (α) and (h) are such that $nh^q\alpha/\ln(n) \rightarrow \infty$ and $\alpha \ln(nh^q) \rightarrow 0$ as $n \rightarrow \infty$ then, for all $\varepsilon > 0$,*

$$\mathbb{P} \left(\sup_{\theta=(B,x) \in \Theta} \left| \hat{\xi}_{\tilde{B}_n(\theta)}^{(H)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta) \right) - \hat{\xi}_B^{(H)} (B^\top x) \right| > \varepsilon \right) \rightarrow 0.$$

Proof of Proposition 6

From (H.3), the function $(B, x) \in \Theta \mapsto \xi_B(B^\top x)$ is continuous. Since Θ is compact, it is also uniformly continuous which shows the result since $\delta(\tilde{\theta}_n(\theta), \theta) < n^{-\eta} \rightarrow 0$ as n goes to infinity. ■

Proof of Proposition 7

Notations

For all $i \in \{1, \dots, n\}$ and $B \in \mathcal{B}_q$, let

$$V_i(B) := S_B(Y_i, B^\top X_i) \quad \text{and} \quad U_i^\star = U_i(B, x, h) = S_B(Z_i^\star, B^\top W_i^\star), \quad (2.15)$$

where the random variables $Z_i^\star$ and $W_i^\star$ have been introduced in Section 2.3.3. Recall that the star ($\star$) is here to recall the dependence on B , x and h .

Let $\mathbf{m} := (m(\tilde{B}, \tilde{x}); (\tilde{B}, \tilde{x}) \in \Theta_n)$ be a list of $\text{card}(\Theta_n)$ integers. We introduce the event

$$\mathfrak{B}_n(\mathbf{m}) := \bigcap_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left\{ M(\tilde{B}, \tilde{x}, h) = m(\tilde{B}, \tilde{x}) \right\}.$$

Finally, denoting by ϑ_q the volume of the unit q -ball and by

$$\mathbf{L}_n = \left\{ \mathbf{m}; m(\tilde{B}, \tilde{x}) \in \left\{ \left\lceil \frac{C_1 \vartheta_q}{2} n h^q \right\rceil, \dots, \left\lceil \frac{3C_2 \vartheta_q}{2} n h^q \right\rceil \right\} \text{ for all } (\tilde{B}, \tilde{x}) \in \Theta_n \right\},$$

we introduce the event

$$\mathfrak{A}_n := \bigcup_{\mathbf{m} \in \mathbf{L}_n} \mathfrak{B}_n(\mathbf{m}).$$

Note that the events $\{\mathfrak{B}_n(\mathbf{m}); \mathbf{m} \in \mathbf{L}_n\}$ are disjoint.

Preliminary results

The following result of real analysis will be useful in the proof of Lemma 9. A proof can be found in [37, Lemma 1].

Lemma 6. *Let $\{a_1, \dots, a_n\}$ and $\{b_1, \dots, b_n\}$ be two sets of distincts real numbers such that for all $i = 1, \dots, n$, $a_i \leq b_i$. Denoting by $\{a_{(1)}, \dots, a_{(n)}\}$ and $\{b_{(1)}, \dots, b_{(n)}\}$ the ordered sets, we have for all $i = 1, \dots, n$ that $a_{(i)} \leq b_{(i)}$.*

The next result is dedicated to the random variables $V_i(B)$ and U_i^* .

Lemma 7. *i) The random variables $V_1(B), \dots, V_n(B)$ are independent standard uniform random variables which are independent from $X_1, \dots, X_n$.
ii) For all $(\tilde{B}, \tilde{x}) \in \Theta_n$ and conditionally to $\mathfrak{B}_n(\mathbf{m})$ with $\mathbf{m} := (m(\tilde{B}, \tilde{x}); (\tilde{B}, \tilde{x}) \in \Theta_n)$, the random variables $U_1^*, \dots, U_{m(\tilde{B}, \tilde{x})}^*$ are independent standard uniform random variables.
iii) For all $(\tilde{B}, \tilde{x}) \in \Theta_n$ and conditionally to $\mathfrak{B}_n(\mathbf{m})$, the random variables*

$$\left\{ E_i(\tilde{B}, \tilde{x}, h) = E_i^* := i \ln \left(\frac{U_{(i)}^*}{U_{(i+1)}^*} \right), i = 1, \dots, m(\tilde{B}, \tilde{x}) \right\},$$

are independent standard exponential random variables.

Proof – Points i) and ii) are proved in [37, Lemma 4]. For the point iii), let $m = m(\tilde{B}, \tilde{x})$ and for all $i = 1, \dots, m$, let $F_i^* := -\ln(U_i^*)$. Using the result of ii), we get that conditionally to $\mathfrak{B}_n(\mathbf{m})$, $F_1^*, \dots, F_m^*$ are independent standard exponential random variables. Hence, if $F_{(1)} < \dots < F_{(m)}$ are the associated order statistics,

$$\{E_i^*, i = 1, \dots, m\} = \{i (F_{(m-i+1)}^* - F_{(m-i)}^*), i = 1, \dots, m\}.$$

We conclude the proof by using Rényi representation theorem. ■

The next result is dedicated to the number of observations in the set $\mathcal{T}(B, x, h)$.

Lemma 8. *Under conditions (H.1) and (H.2), if $nh^q/\ln(n) \rightarrow \infty$ as $n \rightarrow \infty$ then,*

$$\frac{1}{nh^q} \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| M(\tilde{B}, \tilde{x}, h) - \vartheta_q nh^q f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| \xrightarrow{\mathbb{P}} 0.$$

Lemma 8 means that, uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$,

$$\frac{M(\tilde{B}, \tilde{x}, h)}{nh^q} \xrightarrow{\mathbb{P}} \vartheta_q f_{\tilde{B}}(\tilde{B}^\top \tilde{x}).$$

As a consequence, since

$$\mathfrak{A}_n := \bigcap_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left\{ \frac{M(\tilde{B}, \tilde{x}, h)}{nh^q} \in \left[\frac{C_1 \vartheta_q}{2}, \frac{3C_2 \vartheta_q}{2} \right] \right\},$$

and since, from (H.2), $0 < C_1 < f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) < C_2$ for all $(\tilde{B}, \tilde{x}) \in \Theta_n$, it is readily seen that $\mathbb{P}(\mathfrak{A}_n) \rightarrow 1$.

Proof Lemma 8 – For all $(\tilde{B}, \tilde{x}) \in \Theta_n$, we have the inequality

$$\left| M(\tilde{B}, \tilde{x}, h) - \vartheta_q nh^q f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| \leq T_{1,n}(\tilde{B}, \tilde{x}, h) + T_{2,n}(\tilde{B}, \tilde{x}, h),$$

where

$$T_{1,n}(\tilde{B}, \tilde{x}, h) := \left| n\mathbb{P}\left(X \in \mathcal{T}(\tilde{B}, \tilde{x}, h)\right) - \vartheta_q nh^q f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right|,$$

and

$$T_{2,n}(\tilde{B}, \tilde{x}, h) := \left| M(\tilde{B}, \tilde{x}, h) - n\mathbb{P}\left(X \in \mathcal{T}(\tilde{B}, \tilde{x}, h)\right) \right|.$$

We thus have to prove that, as $n \rightarrow \infty$,

$$\frac{1}{nh^q} \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} T_{1,n}(\tilde{B}, \tilde{x}, h) \rightarrow 0 \quad \text{and} \quad \frac{1}{nh^q} \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} T_{2,n}(\tilde{B}, \tilde{x}, h) \xrightarrow{\mathbb{P}} 0. \quad (2.16)$$

Denoting by $\mathcal{U}_q$ the unit q -ball, we get

$$\begin{aligned} \mathbb{P}\left(X \in \mathcal{T}(\tilde{B}, \tilde{x}, h)\right) &= h^q \int_{\mathcal{U}_q} f_{\tilde{B}}(\tilde{B}^\top \tilde{x} + hu) du \\ &= \vartheta_q h^q f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) + h^q \int_{\mathcal{U}_q} \left[f_{\tilde{B}}(\tilde{B}^\top \tilde{x} + hu) - f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right] du. \end{aligned} \quad (2.17)$$

Since $\tilde{B} \in \mathcal{B}_q$ is an orthogonal matrix, we can write

$$f_{\tilde{B}}(\tilde{B}^\top \tilde{x} + hu) = f_{\tilde{B}}\left(\tilde{B}^\top \left(\tilde{x} + h\tilde{B}(\tilde{B}^\top \tilde{B})^{-1}u\right)\right).$$

By construction of the set $\mathcal{B}_q$, we have $\|\tilde{B}(\tilde{B}^\top \tilde{B})^{-1}\| = \|\tilde{B}\| < \infty$. Hence, for all $u \in \mathcal{U}_q$, the vector $\tilde{x} + h\tilde{B}(\tilde{B}^\top \tilde{B})^{-1}u$ belongs to the ball of center $\tilde{x}$ and radius $h\|\tilde{B}\|$. Since $h \rightarrow 0$ as $n \rightarrow \infty$, using [37, Lemma 2] we get that for n large enough and uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$, the vector $\tilde{x} + h\tilde{B}u$ belongs to $\mathcal{X}$. Hence, from condition (H.2), for n large enough, we have that for all $u \in \mathcal{U}_q$, there exists $K_f > 0$ such that

$$\left| f_{\tilde{B}}(\tilde{B}^\top \tilde{x} + hu) - f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| \leq K_f h \|\tilde{B}u\| \leq Kh. \quad (2.18)$$

Gathering (2.17) and (2.18), we obtain that for all $(\tilde{B}, \tilde{x}) \in \Theta_n$, $T_{1,n}(\tilde{B}, \tilde{x}, h) \leq K_f \vartheta_q n h^{q+1}$, proving the first part of (2.16). Let us now focus on the second part of (2.16). For all $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P} \left(\frac{1}{nh^q} \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} T_{2,n}(\tilde{B}, \tilde{x}, h) > \varepsilon \right) &= \mathbb{P} \left(\bigcup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left\{ \frac{1}{nh^q} T_{2,n}(\tilde{B}, \tilde{x}, h) > \varepsilon \right\} \right) \\ &\leq \text{card}(\Theta_n) \times \mathbb{P} \left(\frac{1}{nh^q} T_{2,n}(\tilde{B}, \tilde{x}, h) > \varepsilon \right) \end{aligned}$$

We now use Bernstein's inequality to get

$$\mathbb{P} \left(\frac{1}{nh^q} T_{2,n}(\tilde{B}, \tilde{x}, h) > \varepsilon \right) \leq 2 \exp \left(- \frac{\varepsilon n h^q}{2[\lambda_n(\tilde{B}, \tilde{x}, h) - 1/3]} \right),$$

where

$$\lambda_n(\tilde{B}, \tilde{x}, h) := \frac{1}{\varepsilon h^q} \mathbb{P} \left(X \in \mathcal{T}(\tilde{B}, \tilde{x}, h) \right) \mathbb{P} \left(X \notin \mathcal{T}(\tilde{B}, \tilde{x}, h) \right).$$

Using (2.17) and (2.18), we have

$$\begin{aligned} \mathbb{P} \left(X \in \mathcal{T}(\tilde{B}, \tilde{x}, h) \right) \mathbb{P} \left(X \notin \mathcal{T}(\tilde{B}, \tilde{x}, h) \right) &\leq \mathbb{P} \left(X \in \mathcal{T}(\tilde{B}, \tilde{x}, h) \right) \\ &\leq \vartheta_q h^q \left(f_{\tilde{B}}(\tilde{B}^\top \tilde{x}) + R_n \right), \end{aligned}$$

where $|R_n| \leq K \vartheta_q h \rightarrow 0$. Hence, from (H.2), there exists $C_2 > 0$ such that $\lambda_n(\tilde{B}, \tilde{x}, h) \leq 2\vartheta_q C_2 / \varepsilon$. Recall that $\text{card}(\Theta_n) = \mathcal{O}(n^c)$, we finally get that

$$\mathbb{P} \left(\frac{1}{nh^q} \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} T_{2,n}(\tilde{B}, \tilde{x}, h) > \varepsilon \right) = \mathcal{O} \left(n^c \exp \left(- \frac{\varepsilon n h^q}{2[2\vartheta_q C_2 / \varepsilon - 1/3]} \right) \right),$$

which converges to 0 since $nh^q / \ln(n) \rightarrow \infty$. ■

Let us now provide an expansion of the conditional tail-index estimator. We use the notations introduced in Lemma 7.

Lemma 9. *Assume that $nh^q\alpha \rightarrow \infty$ as $n \rightarrow \infty$. Uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$, for n large enough, we have conditionally to $\mathfrak{B}_n(\mathbf{m})$ with $\mathbf{m} := (m = m(\tilde{B}, \tilde{x}); (\tilde{B}, \tilde{x}) \in \Theta_n)$, that*

$$\left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} E_i^* \right| \leq 2\omega^* \left(V_{(1)}(\tilde{B}), V_{(n)}(\tilde{B}) \right) + \Delta_{\tilde{B}, \tilde{x}} \left(1/U_{([\alpha M^*]+1)}^* \right).$$

Proof – Since $nh^q\alpha \rightarrow \infty$, we have conditionally to $\mathfrak{B}_n(\mathbf{m})$ and uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$ that for n large enough,

$$\hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) = \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} \ln \left(\frac{Z_{(m-i+1)}^*}{Z_{(m-[\alpha m])}^*} \right),$$

where we use the shortcuts $m = m(\tilde{B}, \tilde{x})$ and $Z_i^* = Z_i(\tilde{B}, \tilde{x}, h)$. For all $i \in \{1, \dots, [\alpha m]\}$, we have

$$\ln Z_i^* = \ln \left(Q_{\tilde{B}}(U_i^*, \tilde{B}^\top W_i^*) \right) \leq \ln \left(Q_{\tilde{B}}(U_i^*, \tilde{B}^\top \tilde{x}) \right) + \omega^* \left(U_{(1)}^*, U_{(m)}^* \right),$$

where $U_i^* = U_i(\tilde{B}, \tilde{x}, h)$ and $\omega^*(u, v) = \omega(u, v, \tilde{B}, \tilde{x}, h)$. Using Lemma 6, we get that for all $i = 1, \dots, m$

$$\ln Z_{(m-i+1)}^* \leq \ln \left(Q_{\tilde{B}}(U_{(i)}^*, \tilde{B}^\top \tilde{x}) \right) + \omega^* \left(U_{(1)}^*, U_{(m)}^* \right).$$

Similarly,

$$\ln Z_{(m-i+1)}^* \geq \ln \left(Q_{\tilde{B}}(U_{(i)}^*, \tilde{B}^\top \tilde{x}) \right) - \omega^* \left(U_{(1)}^*, U_{(m)}^* \right).$$

We deduce from these two last inequalities that

$$\left| \ln \left(\frac{Z_{(m-i+1)}^*}{Z_{(m-[\alpha m])}^*} \right) - \ln \left(\frac{Q_{\tilde{B}}(U_{(i)}^*, \tilde{B}^\top \tilde{x})}{Q_{\tilde{B}}(U_{([\alpha m]+1)}^*, \tilde{B}^\top \tilde{x})} \right) \right| \leq 2\omega^* \left(V_{(1)}(\tilde{B}), V_{(m)}(\tilde{B}) \right),$$

since $V_{(1)}(\tilde{B}) \leq U_{(1)}^\star$ and $V_{(m)}(\tilde{B}) \geq U_{(m)}^\star$. As a consequence, conditionally to $\mathfrak{B}_n(\mathbf{m})$ and uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$ we have for n large enough,

$$\left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} \ln \left(\frac{Q_{\tilde{B}}(U_{(i)}^\star, \tilde{B}^\top \tilde{x})}{Q_{\tilde{B}}(U_{([\alpha m]+1)}^\star, \tilde{B}^\top \tilde{x})} \right) \right| \leq 2\omega^\star \left(V_{(1)}(\tilde{B}), V_{(m)}(\tilde{B}) \right).$$

Now remark that from Lemma 7 iii) and since $\ln(Q_{\tilde{B}}(u, \tilde{B}^\top \tilde{x})) = \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \ln(u^{-1}) + \ln(\ell_{\tilde{B}}(u^{-1}, \tilde{B}^\top \tilde{x}))$,

$$\begin{aligned} \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} \ln \left(\frac{Q_{\tilde{B}}(U_{(i)}^\star, \tilde{B}^\top \tilde{x})}{Q_{\tilde{B}}(U_{([\alpha m]+1)}^\star, \tilde{B}^\top \tilde{x})} \right) &= \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} i \ln \left(\frac{Q_{\tilde{B}}(U_{(i)}^\star, \tilde{B}^\top \tilde{x})}{Q_{\tilde{B}}(U_{(i+1)}^\star, \tilde{B}^\top \tilde{x})} \right) \\ &= \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} E_i^\star + \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} i \ln \left(\frac{\ell_{\tilde{B}}(1/U_{(i)}^\star, \tilde{B}^\top \tilde{x})}{\ell_{\tilde{B}}(1/U_{(i+1)}^\star, \tilde{B}^\top \tilde{x})} \right). \end{aligned}$$

We conclude the proof by remarking that

$$\begin{aligned} &\left| \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} i \ln \left(\frac{\ell_{\tilde{B}}(1/U_{(i)}^\star, \tilde{B}^\top \tilde{x})}{\ell_{\tilde{B}}(1/U_{(i+1)}^\star, \tilde{B}^\top \tilde{x})} \right) \right| \\ &= \left| \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} \ln \left(\frac{\ell_{\tilde{B}}(1/U_{(i)}^\star, \tilde{B}^\top \tilde{x})}{\ell_{\tilde{B}}(1/U_{([\alpha m]+1)}^\star, \tilde{B}^\top \tilde{x})} \right) \right| \leq \Delta_{\tilde{B}, \tilde{x}}(1/U_{([\alpha m]+1)}^\star). \end{aligned}$$

■

We are now in position to prove Proposition 7

Proof of Proposition 7

For all $\varepsilon > 0$, we have

$$\begin{aligned} &\mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right) \\ &\leq [1 - \mathbb{P}(\mathfrak{A}_n)] + \mathbb{P} \left(\left\{ \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right\} \cap \mathfrak{A}_n \right). \end{aligned}$$

Since Lemma 8 entails that $\mathbb{P}(\mathfrak{A}_n) \rightarrow 1$, we only need to focus on the second term. We have

$$\begin{aligned} & \mathbb{P} \left(\left\{ \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right\} \cap \mathfrak{A}_n \right) \\ & \leq \sum_{\mathbf{m} \in \mathbf{L}_n} \mathbb{P} \left(\left\{ \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right\} \middle| \mathfrak{B}_n(\mathbf{m}) \right) \mathbb{P}(\mathfrak{B}_n(\mathbf{m})). \end{aligned}$$

Since the events $\{\mathfrak{B}_n(\mathbf{m}), \mathbf{m} \in \mathbf{L}_n\}$ are disjoint, we obtain

$$\begin{aligned} & \mathbb{P} \left(\left\{ \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right\} \cap \mathfrak{A}_n \right) \\ & \leq \sup_{\mathbf{m} \in \mathbf{L}_n} \mathbb{P} \left(\left\{ \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \hat{\xi}_{\tilde{B}}^{(H)}(\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \right| > \varepsilon \right\} \middle| \mathfrak{B}_n(\mathbf{m}) \right) =: \sup_{\mathbf{m} \in \mathbf{L}_n} T(\mathbf{m}). \end{aligned}$$

Using Lemma 9 and keeping the same notations, we get that $T(\mathbf{m}) \leq T_1(\mathbf{m}) + T_2(\mathbf{m}) + T_3(\mathbf{m})$, where

$$\begin{aligned} T_1(\mathbf{m}) &:= \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \omega^\star \left(V_{(1)}(\tilde{B}), V_{(n)}(\tilde{B}) \right) > \varepsilon/8 \middle| \mathfrak{B}_n(\mathbf{m}) \right), \\ T_2(\mathbf{m}) &:= \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \xi_{\tilde{B}}(\tilde{B}^\top \tilde{x}) \left| \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} E_i^\star - 1 \right| > \varepsilon/2 \middle| \mathfrak{B}_n(\mathbf{m}) \right), \end{aligned}$$

and

$$T_3(\mathbf{m}) := \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \Delta_{\tilde{B}, \tilde{x}}(1/U_{([\alpha m]+1)}^\star) > \varepsilon/4 \middle| \mathfrak{B}_n(\mathbf{m}) \right).$$

Let us first focus on the term $T_1(\mathbf{m})$. Recall that there exists some $c > 0$ such that $\text{card}(\Theta_n) = \mathcal{O}(n^c)$. Let $\kappa > c$ and let us introduce the set

$$\mathcal{G}_n := \bigcap_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left\{ V_{(1)}(\tilde{B}) > n^{-(1+\kappa)} \right\} \cap \left\{ V_{(n)}(\tilde{B}) < 1 - n^{-(1+\kappa)} \right\}.$$

We have,

$$\begin{aligned} 1 - \mathbb{P}(\mathcal{G}_n) &= \mathbb{P} \left(\bigcup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left\{ V_{(1)}(\tilde{B}) \leq n^{-(1+\kappa)} \right\} \cup \left\{ V_{(n)}(\tilde{B}) \geq 1 - n^{-(1+\kappa)} \right\} \right) \\ &\leq 2 \text{card}(\Theta_n) (1 - [1 - n^{-(1+\kappa)}]^n) = \mathcal{O}(n^{c-\kappa}) \rightarrow 0. \end{aligned}$$

Hence, for all $\mathbf{m} \in \mathbf{L}_n$, we have

$$T_1(\mathbf{m}) \leq \mathbb{I} \left\{ \omega^\star (n^{-(1+\kappa)}, 1 - n^{-(1+\kappa)}) > \varepsilon/8 \right\} + [1 - \mathbb{P}(\mathcal{G}_n | \mathfrak{B}_n(\mathbf{m}))].$$

From Lemma [7](#) i), we know that the events $\mathcal{G}_n$ and $\mathfrak{B}_n(\mathbf{m})$ are independent. Hence, $\mathbb{P}(G_n | \mathfrak{B}_n(\mathbf{m})) = \mathbb{P}(\mathcal{G}_n) \rightarrow 1$ and using condition (H.5),

$$\sup_{\mathbf{m} \in \mathbf{L}_n} T_1(\mathbf{m}) \leq \mathbb{I} \left\{ \omega^\star (n^{-(1+\kappa)}, 1 - n^{-(1+\kappa)}) > \varepsilon/8 \right\} + [1 - \mathbb{P}(\mathcal{G}_n)] \rightarrow 0. \quad (2.19)$$

For the term $T_2(\mathbf{m})$, we have under (C.1)

$$T_2(\mathbf{m}) \leq \text{card}(\Theta_n) \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(\left| \frac{1}{[\alpha m]} \sum_{i=1}^{[\alpha m]} E_i^\star - 1 \right| > \frac{\varepsilon}{2\gamma} \middle| \mathfrak{B}_n(\mathbf{m}) \right).$$

Using Chernoff's inequality for exponential random variables, there exists a constant $C_3 > 0$ such that for all $\mathbf{m} \in \mathbf{L}_n$,

$$T_2(\mathbf{m}) \leq \text{card}(\Theta_n) \exp(-C_3[\alpha m]) \leq \text{card}(\Theta_n) \exp \left(-\frac{3C_2 C_3 \vartheta_q}{2} n h^q \alpha \right).$$

Since $\text{card}(\Theta_n) = \mathcal{O}(n^c)$ and $n h^q \alpha / \ln(n) \rightarrow \infty$, we get that

$$\sup_{\mathbf{m} \in \mathbf{L}_n} T_2(\mathbf{m}) \rightarrow 0. \quad (2.20)$$

Finally, for the term $T_3(\mathbf{m})$, we start with the inequality

$$T_3(\mathbf{m}) \leq \text{card}(\Theta_n) \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(\Delta_{\tilde{B}, \tilde{x}} (1/U_{([\alpha m]+1)}^\star) > \varepsilon/4 \middle| \mathfrak{B}_n(\mathbf{m}) \right).$$

For all sequence (ε_n) such that $\varepsilon_n \rightarrow 0$, condition (H.4) entails that for n large enough and uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$, if the event $\{U_{([\alpha m]+1)}^\star \leq \varepsilon_n\}$ holds then $\Delta_{\tilde{B}, \tilde{x}} (1/U_{([\alpha m]+1)}^\star) \leq \varepsilon/4$. As a consequence,

$$T_3(\mathbf{m}) \leq \text{card}(\Theta_n) \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(U_{([\alpha m]+1)}^\star > \varepsilon_n \middle| \mathfrak{B}_n(\mathbf{m}) \right).$$

From the second part of Lemma [7](#), conditionally to $\mathfrak{B}_n(\mathbf{m})$, the random variable $U_{([\alpha m]+1)}^\star$ follows a Beta distribution with parameters $[\alpha m] + 1$ and

$m - \lfloor \alpha m \rfloor$. Hence,

$$\begin{aligned} & \mathbb{P} \left(U_{(\lfloor \alpha m \rfloor + 1)}^* > \varepsilon_n \mid \mathfrak{B}_n(\mathbf{m}) \right) \\ &= \frac{m!}{\lfloor \alpha m \rfloor! (m - \lfloor \alpha m \rfloor - 1)!} \int_{\varepsilon_n}^1 x^{\lfloor \alpha m \rfloor} (1 - x)^{m - \lfloor \alpha m \rfloor - 1} dx \\ &\leq m^{\lfloor \alpha m \rfloor + 1} (1 - \varepsilon_n)^{m - \lfloor \alpha m \rfloor - 1} \end{aligned}$$

Since $\ln(1 - u) \leq -u$ for all $u \in (0, 1)$, we have for all $\mathbf{m} \in \mathbf{L}_n$,

$$\begin{aligned} & \mathbb{P} \left(U_{(\lfloor \alpha m \rfloor + 1)}^* > \varepsilon_n \mid \mathfrak{B}_n(\mathbf{m}) \right) \\ &\leq \exp \left[-m\varepsilon_n \left(\frac{m - \lfloor \alpha m \rfloor + 1}{2m} - (\lfloor \alpha m \rfloor + 1) \frac{\ln(m)}{m\varepsilon_n} \right) \right]. \end{aligned}$$

Since $m = \mathcal{O}(nh^q)$ uniformly on $(\tilde{B}, \tilde{x}) \in \Theta_n$, it is readily seen that for all sequence ε_n such that $\alpha \ln(nh^q)/\varepsilon_n \rightarrow 0$, for n large enough, there exists a constant $C_4 > 0$ such that

$$\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(U_{(\lfloor \alpha m \rfloor + 1)}^* > \varepsilon_n \mid \mathfrak{B}_n(\mathbf{m}) \right) \leq \exp(-C_4 nh^q \varepsilon_n).$$

As a consequence, since $\text{card}(\Theta_n) = \mathcal{O}(n^c)$, for a sequence (ε_n) such that $\alpha \ln(nh^q)/\varepsilon_n \rightarrow 0$ and $nh^q \varepsilon_n / \ln(n) \rightarrow \infty$, we get that

$$\sup_{\mathbf{m} \in \mathbf{L}_n} T_3(\mathbf{m}) \rightarrow 0. \quad (2.21)$$

For instance, one can take $\varepsilon_n = (\alpha \ln(nh^q))^{1/2}$. Since by assumption, $\alpha \ln(nh^q) \rightarrow 0$, we have $\alpha \ln(nh^q)/\varepsilon_n \rightarrow 0$. Moreover,

$$\frac{nh^q \varepsilon_n}{\ln(n)} = \frac{nh^q \alpha}{\ln(n)} \left(\frac{\ln(nh^q)}{\alpha} \right)^{1/2} \rightarrow \infty,$$

since by assumption $nh^q \alpha / \ln(n) \rightarrow \infty$.

The conclusion of the proof is then straightforward by gathering (2.19), (2.20) and (2.21). $\blacksquare$

Proof of Proposition 8

Notations

Recall that for all $\theta = (B, x) \in \Theta$ we have $\delta(\tilde{\theta}_n(\theta), \theta) < n^{-\eta}$ with $\eta > q$. For all $\tilde{C} > 0$ and $(\tilde{B}, \tilde{x}) \in \Theta_n$, let

$$\begin{aligned}\tilde{\mathcal{T}}_{n, \tilde{C}}(\tilde{B}, \tilde{x}, h) &:= \left\{ y \in \mathcal{X}; \ h - \tilde{C}n^{-\eta} < \left\| \tilde{B}^\top(\tilde{x} - y) \right\| \leq h + \tilde{C}n^{-\eta} \right\} \\ &= \mathcal{T}(\tilde{B}, \tilde{x}, h + \tilde{C}n^{-\eta}) \setminus \mathcal{T}(\tilde{B}, \tilde{x}, h - \tilde{C}n^{-\eta})\end{aligned}$$

Note that since $\eta > q$, for n large enough, $\tilde{C}n^{-\eta} < h$. Let also

$$\tilde{D}_n(\tilde{B}, \tilde{x}) := \sum_{i=1}^n \mathbb{I} \left\{ X_i \in \tilde{\mathcal{T}}_{n, \tilde{C}}(\tilde{B}, \tilde{x}, h) \right\}.$$

Preliminary results

Lemma 10. *There exists a constant $\tilde{C} > 0$ such that for all $\theta = (B, x) \in \Theta$,*

$$\mathcal{T}(\tilde{B}_n(\theta), \tilde{x}_n(\theta), h - \tilde{C}n^{-\eta}) \subset \mathcal{T}(B, x, h) \subset \mathcal{T}(\tilde{B}_n(\theta), \tilde{x}_n(\theta), h + \tilde{C}n^{-\eta})$$

Proof of Lemma 10 – From the reverse triangular inequality, $\|B^\top(y - x)\| \geq |T_{n,1}(\theta) - T_{n,2}(\theta)|$, where

$$\begin{aligned}T_{n,1}(y, \theta) &:= \left\| [\tilde{B}_n(\theta)]^\top(\tilde{x}_n(\theta) - y) \right\| \\ \text{and } T_{n,2}(y, \theta) &:= \left\| (B - \tilde{B}_n(\theta))^\top(y - x) + [\tilde{B}_n(\theta)]^\top(x - \tilde{x}_n(\theta)) \right\|.\end{aligned}$$

Now, remark that there exists a constant $K_0 > 0$ such that $\|(B - \tilde{B}_n(\theta))^\top\| \leq K_0\|B - \tilde{B}_n(\theta)\|$. In particular, $\|[\tilde{B}_n(\theta)]^\top\| \leq K_0$.

Furthermore, since x and y belong to the compact $\mathcal{X}$, we have $\|x - y\| \leq 2N(\mathcal{X})$ where $N(\mathcal{X}) = \sup_{x \in \mathcal{X}} \|x\| < \infty$. As a consequence, letting

$\tilde{C} := K_0 \max(2N(\mathcal{X}), 1)$, we get

$$T_{n,2}(y, \theta) \leq \tilde{C}\delta(\tilde{\theta}_n(\theta), \theta) \leq \tilde{C}n^{-\eta}.$$

Hence, if $T_{n,1}(y, \theta) > h + \tilde{C}n^{-\eta}$ then $\|B^\top(y - x)\| > h$. In other word, if $y \notin \mathcal{T}(\tilde{B}_n(\theta), \tilde{x}_n(\theta), h + \tilde{C}n^{-\eta})$ then $y \notin \mathcal{T}(B, x, h)$ proving the second inclusion. Next, if $T_{n,1}(y, \theta) \leq h - \tilde{C}n^{-\eta}$ then $\|B^\top(y - x)\| \leq \min(h - \tilde{C}n^{-\eta}, \tilde{C}n^{-\eta}) < h$. This entails that if $y \in \mathcal{T}(\tilde{B}_n(\theta), \tilde{x}_n(\theta), h - \tilde{C}n^{-\eta})$ then $y \in \mathcal{T}(B, x, h)$ proving the first inclusion and completing the proof. ■

Lemma 11. *Assume that conditions (H.1) and (H.2) holds. If $nh^q/\ln(n) \rightarrow \infty$ as $n \rightarrow \infty$ then for all integer K such that $K > \lfloor c/(\eta - 1) \rfloor$,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \tilde{D}_n(\tilde{B}, \tilde{x}) \geq K \right) = 0.$$

Proof of Lemma 11 – It suffices to prove that for all $K > c/(\eta - 1)$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \tilde{D}_n(\tilde{B}, \tilde{x}) \geq K \right) &\leq \lim_{n \rightarrow \infty} \sum_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(\tilde{D}_n(\tilde{B}, \tilde{x}) \geq K \right) \\ &\leq \lim_{n \rightarrow \infty} \text{card}(\Theta_n) \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(\tilde{D}_n(\tilde{B}, \tilde{x}) \geq K \right) = 0. \end{aligned}$$

Let $h_{\pm} := h \pm \tilde{C}n^{-\eta}$ and let I be the set of all list $\mathbf{i} := \{i_1, \dots, i_K\}$ of K distinct indices in $\{1, \dots, n\}^K$.

$$\begin{aligned} \mathbb{P} \left(\tilde{D}_n(\tilde{B}, \tilde{x}) \geq K \right) &= \mathbb{P} \left(\bigcup_{\mathbf{i} \in I} \bigcap_{s=1}^K \left\{ h_- < \|\tilde{B}(X_{i_s} - \tilde{x})\| \leq h_+ \right\} \right) \\ &= \binom{n}{K} \left[\mathbb{P} \left(h_- < \|\tilde{B}(X - \tilde{x})\| \leq h_+ \right) \right]^K. \end{aligned}$$

From (2.17) in the proof of Lemma 8,

$$\begin{aligned} \mathbb{P} \left(h_- < \|\tilde{B}(X - \tilde{x})\| \leq h_+ \right) &= \mathbb{P} \left(\|\tilde{B}(X - \tilde{x})\| \leq h_+ \right) - \mathbb{P} \left(\|\tilde{B}(X - \tilde{x})\| \leq h_- \right) \\ &= \vartheta_q f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} \right) \left[(h_+)^q - (h_-)^q \right] + (h_+)^q R_+ - (h_-)^q R_-, \end{aligned}$$

where

$$R_{\pm} := \int_{\mathcal{U}_q} \left[f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} + h_{\pm} u \right) - f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} \right) \right] du.$$

First remark that

$$(h_+)^q - (h_-)^q = h^q \left[\left(1 + \tilde{C} \frac{n^{-\eta}}{h} \right)^q - \left(1 - \tilde{C} \frac{n^{-\eta}}{h} \right)^q \right].$$

Since $\eta > q$, we know that $n^{-\eta}/h \rightarrow 0$ and thus,

$$(h_+)^q - (h_-)^q = \mathcal{O} \left(h^{q-1} n^{-\eta} \right). \quad (2.22)$$

Furthermore, $(h_+)^q R_+ - (h_-)^q R_- = (h_+)^q R_{1,+} - (h_-)^q R_{1,-} + [(h_+)^q - (h_-)^q] R_2$, where

$$R_{1,\pm} := \int_{\mathcal{U}_q} \left[f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} + h_\pm u \right) - f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} + hu \right) \right] du$$

and $R_2 := \int_{\mathcal{U}_q} \left[f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} + hu \right) - f_{\tilde{B}} \left(\tilde{B}^\top \tilde{x} \right) \right] du.$

Mimicking the proof of (2.18) in the proof of Lemma 8, we have

$$|R_{1,\pm}| \leq \vartheta_q K_f \tilde{C} n^{-\eta} \quad \text{and} \quad |R_2| \leq \vartheta_q K_f h.$$

From (2.22), $(h_+)^q R_+ - (h_-)^q R_- = \mathcal{O}(h^q n^{-\eta})$ and, since from (H.2) $|f_B(B^\top x)|$ is uniformly bounded on $(B, x) \in \Theta$,

$$\mathbb{P} \left(h_- \leq \|\tilde{B}(X - \tilde{x})\| \leq h_+ \right) = \mathcal{O}(h^{q-1} n^{-\eta}) + \mathcal{O}(h^q n^{-\eta}) = \mathcal{O}(n^{-\eta}). \quad (2.23)$$

Since $\binom{n}{K} = \mathcal{O}(n^K)$ and $\text{card}(\Theta_n) = \mathcal{O}(n^c)$, we get

$$\text{card}(\Theta_n) \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \mathbb{P} \left(\tilde{D}_n(\tilde{B}, \tilde{x}) \geq K \right) = \mathcal{O} \left(n^{c+K(1-\eta)} \right).$$

We conclude the proof by remarking that the condition $K > \lfloor c/(\eta - 1) \rfloor$ entails that $n^{c+K(1-\eta)} \rightarrow 0$. $\blacksquare$

Proof of Proposition 8

We keep the notations used for the definition of the conditional tail-index estimator in Section 2.3.3. In particular, $M^\star = M(B, x, h)$ and $Z_i^\star = Z_i(B, x, h)$. In addition, for all $\theta = (B, x) \in \Theta$, let $\tilde{M}_n^\pm := M(\tilde{B}_n(\theta), \tilde{x}_n(\theta), h_\pm)$ and for $i = 1, \dots, \tilde{M}_n^\pm$, let $\tilde{Z}_i^{\star, \pm} := Z_{(\tilde{M}_n^\pm - i + 1)}(\tilde{B}_n(\theta), \tilde{x}_n(\theta), h_\pm)$. Moreover, for $K > \lfloor c/(\eta - 1) \rfloor$, let

$$\mathfrak{E}_n(K) := \left\{ \sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \tilde{D}_n(\tilde{B}, \tilde{x}) < K \right\} \cap \mathfrak{A}_n^+ \cap \mathfrak{A}_n^-,$$

where

$$\mathfrak{A}_n^\pm := \bigcap_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left\{ \frac{M(\tilde{B}, \tilde{x}, h_\pm)}{n(h_\pm)^q} \in \left[\frac{C_1 \vartheta_q}{2}, \frac{3C_2 \vartheta_q}{2} \right] \right\},$$

From Lemma 8 and Lemma 11, we know that $\mathbb{P}(\mathfrak{C}_n(K)) \rightarrow 1$. In what follows, we assume that the event $\mathfrak{C}_n(K)$ holds. We have the following inequalities : for all $(B, x) \in \Theta$,

$$\begin{aligned} M^\star - K + 1 &\leq \widetilde{M}_n^- \leq M^\star, \quad M^\star \leq \widetilde{M}_n^+ \leq M^\star + K - 1, \\ \text{and} \quad \frac{C_1 \vartheta_q}{2} n(h_-)^q &\leq M^\star \leq \frac{3C_2 \vartheta_q}{2} n(h_+)^q. \end{aligned} \quad (2.24)$$

As a straightforward consequence, for n large enough and uniformly on $(B, x) \in \Theta$, we have $\lfloor \alpha M^\star \rfloor < \widetilde{M}_n^- < \widetilde{M}_n^+$.

Now, Lemma 10 ensures that for all $i \in \{1, \dots, \lfloor \alpha M^\star \rfloor\}$,

$$\widetilde{Z}_{(\widetilde{M}_n^- - i + 1)}^{\star, -} \leq Z_{(M^\star - i + 1)}^\star \leq \widetilde{Z}_{(\widetilde{M}_n^+ - i + 1)}^{\star, +}. \quad (2.25)$$

Moreover, for all $k_+ \geq \lfloor \alpha M^\star \rfloor + K$, we have $Z_{(M^\star - \lfloor \alpha M^\star \rfloor)}^\star \geq \widetilde{Z}_{(\widetilde{M}_n^+ - k_+)}^{\star, +}$. Indeed, the largest value of $\widetilde{Z}_{(\widetilde{M}_n^+ - k_+)}^{\star, +}$ is obtained when $\widetilde{M}_n^+ = M^\star + K - 1$ with the additional $(K - 1)$ observations larger than the maximal observation $Z_{(M^\star)}^\star$. In the same way, for all $k_- \leq \lfloor \alpha M^\star \rfloor - K + 1$, we have $Z_{(M^\star - \lfloor \alpha M^\star \rfloor)}^\star \leq \widetilde{Z}_{(\widetilde{M}_n^- - k_-)}^{\star, -}$. Note that for n large enough and uniformly on $(B, x) \in \Theta$, $\lfloor \alpha M^\star \rfloor > K - 1$. To sum up, for $k_+ \geq \lfloor \alpha M^\star \rfloor + K$, and $k_- \leq \lfloor \alpha M^\star \rfloor - K + 1$, we have for n large enough, that

$$\widetilde{Z}_{(\widetilde{M}_n^- - k_-)}^{\star, -} \geq Z_{(M^\star - \lfloor \alpha M^\star \rfloor)}^\star \geq \widetilde{Z}_{(\widetilde{M}_n^+ - k_+)}^{\star, +}. \quad (2.26)$$

Taking $\alpha_+ := \alpha + 2(K + 2)/[C_1 \vartheta_q n(h_+)^q]$, we have

$$\lfloor \alpha_+ \widetilde{M}_n^+ \rfloor \geq \alpha \widetilde{M}_n^+ + \frac{2(K + 2)}{C_1 \vartheta_q n(h_+)^q} \widetilde{M}_n^+ + 1 \geq \alpha \widetilde{M}_n^+ + K + 1 \geq \lfloor \alpha M^\star \rfloor + K.$$

Similarly, we show that for $\alpha_- := \alpha - 2(K + 1)/[C_1 \vartheta_q n(h_-)^q]$,

$$\lfloor \alpha_- \widetilde{M}_n^- \rfloor \leq \lfloor \alpha M^\star \rfloor - K + 1.$$

We can thus use inequality (2.25) and inequality (2.26) with $k_- = \lfloor \alpha_- \widetilde{M}_n^- \rfloor$ and $k_+ = \lfloor \alpha_+ \widetilde{M}_n^+ \rfloor$ to obtain for n large enough and uniformly on $\theta = (B, x) \in \Theta$,

$$\begin{aligned} \frac{\lfloor \alpha_- \widetilde{M}_n^- \rfloor}{\lfloor \alpha M^\star \rfloor} \widehat{\xi}_{\widetilde{B}_n(\theta)}^{(H)} \left([\widetilde{B}_n(\theta)]^\top \widetilde{x}_n(\theta); \alpha_-, h_- \right) &\leq \widehat{\xi}_B^{(H)} (B^\top x; \alpha, h) \\ &\leq \frac{\lfloor \alpha_+ \widetilde{M}_n^+ \rfloor}{\lfloor \alpha M^\star \rfloor} \widehat{\xi}_{\widetilde{B}_n(\theta)}^{(H)} \left([\widetilde{B}_n(\theta)]^\top \widetilde{x}_n(\theta); \alpha_+, h_+ \right). \end{aligned}$$

Note from (2.24), it easy to check that there exist constants $C_- > 0$ and $C_+ > 0$ such that for all $(B, x) \in \Theta$

$$1 - \frac{C_-}{nh^q} \leq \frac{[\alpha_- \widetilde{M}_n^-]}{[\alpha M^\star]} \leq 1 \text{ and } 1 \leq \frac{[\alpha_+ \widetilde{M}_n^+]}{[\alpha M^\star]} \leq 1 + \frac{C_+}{nh^q}$$

Hence, uniformly on $(B, x) \in \Theta$,

$$\begin{aligned} \left(1 - \frac{C_-}{nh^q}\right) \widehat{\xi}_{\tilde{B}_n(\theta)}^{(H)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta); \alpha_-, h_-\right) &\leq \widehat{\xi}_B^{(H)} (B^\top x; \alpha, h) \\ &\leq \left(1 + \frac{C_+}{nh^q}\right) \widehat{\xi}_{\tilde{B}_n(\theta)}^{(H)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta); \alpha_+, h_+\right). \end{aligned} \quad (2.27)$$

Recall that all the previous results hold under the event $\mathfrak{C}_n(K)$. We thus write for all $\varepsilon > 0$,

$$\mathbb{P} \left(\sup_{\theta=(B,x) \in \Theta} \left| \widehat{\xi}_{\tilde{B}_n(\theta)}^{(H)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta)\right) - \widehat{\xi}_B^{(H)} (B^\top x) \right| > \varepsilon \right) \leq T_n + (1 - \mathbb{P}(\mathfrak{C}_n(K))),$$

where

$$T_n := \mathbb{P} \left(\left\{ \sup_{\theta=(B,x) \in \Theta} \left| \widehat{\xi}_{\tilde{B}_n(\theta)}^{(H)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta)\right) - \widehat{\xi}_B^{(H)} (B^\top x) \right| > \varepsilon \right\} \cap \mathfrak{C}_n(K) \right).$$

Since $\mathbb{P}(\mathfrak{C}_n(K)) \rightarrow 1$, we only need to focus on T_n . We have $T_n \leq T_{n,1} + T_{n,2}$ where

$$\begin{aligned} T_{n,1} &= \mathbb{P} \left(\left\{ \sup_{\theta=(B,x) \in \Theta} \left| \widehat{\xi}_B^{(H)} (B^\top x) - \xi_{\tilde{B}_n(\theta)} \left([\tilde{B}_n(\theta)]^\top \tilde{x}_n(\theta)\right) \right| > \varepsilon/2 \right\} \cap \mathfrak{C}_n(K) \right) \\ \text{and } T_{n,2} &= \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \widehat{\xi}_{\tilde{B}}^{(H)} (\tilde{B}^\top \tilde{x}) - \xi_{\tilde{B}} (\tilde{B}^\top \tilde{x}) \right| > \varepsilon/2 \right). \end{aligned}$$

From Proposition 7, we have $T_{n,2} \rightarrow 0$. It thus remain to show that $T_{n,1} \rightarrow 0$. Using (2.27), $T_{n,1} \leq T_{n,3}^+ + T_{n,3}^-$ where

$$T_{n,3}^\pm = \mathbb{P} \left(\sup_{(\tilde{B}, \tilde{x}) \in \Theta_n} \left| \left(1 \pm \frac{C_\pm}{nh^q}\right) \widehat{\xi}_{\tilde{B}}^{(H)} (\tilde{B}^\top \tilde{x}; \alpha_\pm, h_\pm) - \xi_{\tilde{B}} (\tilde{B}^\top \tilde{x}) \right| > \varepsilon/2 \right).$$

Since $h_\pm \sim h$, $\alpha_\pm \sim \alpha$ and $nh^q \rightarrow \infty$, we can use Proposition 7 to conclude the proof. $\blacksquare$

Chapitre 3

Extreme Conditional Quantile Estimation in High Dimensions : A Comparative Study

Gardes L. and Podgorny A. (2025). Extreme Conditional Quantile Estimation in High Dimensions : A Comparative Study, *Preprint*.

Cet article traite de la problématique de l'estimation des quantiles extrêmes conditionnels dans des contextes de grande dimension. Nous considérons des avancées récentes fondées sur des techniques d'apprentissage automatique, telles que les réseaux de neurones et les forêts aléatoires, ainsi que des approches de réduction de dimension spécifiquement conçues pour le traitement des valeurs extrêmes. Nous introduisons un cadre unifié englobant ces différentes méthodes, en nous concentrant sur les modèles à queue lourde. Une étude de simulation approfondie évalue leurs performances dans une grande variété de scénarios, en analysant l'influence de la dimension des covariables, du comportement de second ordre de la queue, de la complexité de la relation entre les covariables et l'indice de queue, ainsi que du choix du niveau intermédiaire de quantile. Nos résultats permettent de dégager des recommandations pratiques pour le choix des méthodes adaptées et mettent en lumière les atouts et les limites de chaque approche.

3.1 Introduction

Let Y be a real-valued response variable and X its associated random covariate of support $\mathcal{X} \subset \mathbb{R}^p$. Given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of n independent replications of (X, Y) , we are interested in the estimation of the conditional quantile of level $\alpha \in (0, 1)$ at a point $x \in \mathbb{R}^p$ given by

$$Q(\alpha|x) := \inf \{t \in \mathbb{R}; \overline{F}(t|x) \geq \alpha\},$$

where $\overline{F}(t|x) := \mathbb{P}(Y > t|X = x)$ is the conditional survival function of Y given $X = x$.

Numerous methods exist for conditional quantile estimation [50, 52, 79] but most of them are not reliable in a high-dimensional setting i.e., for a large dimension p of the covariate. For instance kernel regression [9, 61] or nearest-neighbor methods [11] suffer severely from the curse of dimensionality [8], leading to poor estimation performance and computational inefficiency as the number of covariates increases.

To tackle the challenges of high-dimensional quantile estimation, several strategies have been proposed. In this paper, we focus on two broad families of methods. The first includes machine learning approaches such as neural networks, random forests, gradient boosting or penalized regression models. These methods aim to estimate the mapping $x \mapsto Q(\alpha|x)$ directly, leveraging the structure of the data through flexible, data-driven models. The second family of methods relies on dimension reduction techniques, such as Sliced Inverse Regression (SIR) [51] or Sliced Average Variance Estimation (SAVE) [22]. These approaches seek to identify a q -dimensional linear subspace $\mathcal{S} \subset \mathbb{R}^p$ with $q \ll p$ that retains all the information in X relevant for predicting Y . In this framework, the conditional quantile $Q(\alpha|x)$ is assumed to depend only on a projection $B^\top x$, for some basis $B \in \mathbb{R}^{p \times q}$ of $\mathcal{S}$, reducing the effective dimensionality of the problem and making the classical nonparametric techniques usable in the reduced space.

In the remainder of this paper, we shift our attention from classical conditional quantiles to the more delicate problem of estimating an extreme conditional quantile $Q(\beta_n|x)$, where the level β_n converges to 0 as the sample size increases. These quantiles play a crucial role in various domains, including

finance, environmental science, and risk management, and their estimation presents additional challenges due to the scarcity of observations in the tail of the distribution.

To ensure reliable estimation in such cases, we assume that for all $x \in \mathcal{X}$, the conditional distribution of Y given $X = x$ belongs to a maximum domain of attraction providing a theoretical framework for extrapolating beyond observed data. Specifically, we consider the following model :

- (M) For all $x \in \mathcal{X} \subset \mathbb{R}^p$, the cumulative distribution function $F(\cdot|x)$ of Y given $X = x$ is continuous, strictly increasing and such that there exist functions $\gamma : \mathcal{X} \rightarrow \mathbb{R}$ and $a : [1, \infty) \times \mathcal{X} \rightarrow (0, \infty)$ such that for any $t > 0$,

$$\lim_{\alpha \rightarrow 0_+} \frac{Q(t\alpha|x) - Q(\alpha|x)}{a(\alpha^{-1}|x)} = \frac{t^{-\gamma(x)} - 1}{\gamma(x)} =: K_{\gamma(x)}(t^{-1}), \quad (3.1)$$

where the right-hand side is interpreted as $\log(t^{-1})$ when $\gamma(x) = 0$.

In order to estimate extreme conditional quantiles, we exploit the asymptotic behavior given by model (M). More specifically, for a probability level β_n close to zero, we apply the limit in (3.1) with $t = \beta_n/\alpha_n$ to get

$$Q(\beta_n|x) \approx Q(\alpha_n|x) + \frac{a(\alpha_n^{-1}|x)}{\gamma(x)} \left[\left(\frac{\alpha_n}{\beta_n} \right)^{\gamma(x)} - 1 \right], \quad (3.2)$$

where α_n is a sufficiently ‘large’ level in order that the conditional quantile $Q(\alpha_n|x)$ can be estimated without using the convergence (3.1) in model (M). In particular, in [38], the authors assume that $\alpha_n > \beta_n$ with $\alpha_n \rightarrow 0$ and $n\alpha_n \rightarrow \infty$. Such a quantile is referred to as a classical conditional quantile. Naturally, the faster the convergence in (3.1), the better the approximation (3.2). The rate of convergence in (3.1) is usually controlled by the following condition.

- (SOC) For all $x \in \mathcal{X}$, there exist functions $\gamma : \mathcal{X} \rightarrow \mathbb{R}$, $\rho : \mathcal{X} \rightarrow (-\infty, 0]$, $a : [1, \infty) \times \mathcal{X} \rightarrow (0, \infty)$ and $A : [1, \infty) \times \mathcal{X} \rightarrow (0, \infty)$ with $A(y|x) \rightarrow 0$ as $y \rightarrow \infty$ such that for any $t > 0$,

$$\lim_{\alpha \rightarrow 0_+} \frac{1}{A(\alpha^{-1}|x)} \left(\frac{Q(t\alpha|x) - Q(\alpha|x)}{a(\alpha^{-1}|x)} - K_{\gamma(x)}(t^{-1}) \right) = H_{\gamma(x), \rho(x)}(t^{-1}), \quad (3.3)$$

with

$$H_{\gamma(x), \rho(x)}(y) := \frac{1}{\rho(x)} (K_{\gamma(x)+\rho(x)}(y) - K_{\gamma(x)}(y)).$$

This condition is referred to as the second-order condition (see, for example, de Haan and Ferreira [39, Section 2.3]). The function $|A(\cdot|x)|$ is regularly varying of index $\rho(x) \leq 0$ that is to say that for all $t > 0$,

$$\lim_{y \rightarrow \infty} \left| \frac{A(ty|x)}{A(y|x)} \right| = t^{\rho(x)}.$$

As a consequence, the so-called second-order parameter $\rho(x)$ drives the rate of convergence in (3.1); a value of $\rho(x)$ close to 0 is associated to a slow convergence and thus to a poor approximation (3.2).

To sum up, estimating $Q(\beta_n|x)$ requires :

- the estimation of a classical conditional quantile $Q(\alpha_n|x)$;
- the estimation of the scale parameter $a_n(x) := a(\alpha_n^{-1}|x)$ and of the tail index $\gamma(x)$. These parameters govern the tail behavior and enable the extrapolation beyond classical conditional quantiles.

In the specific case where $\gamma(x) > 0$ for all $x \in \mathcal{X}$, we have that $a(\alpha|x) \sim \gamma(x)Q(\alpha|x)$ as $\alpha \rightarrow 0$ (see for instance [39, Lemma 1.2.9]). In this case, convergence (3.1) rewrites

$$\lim_{\alpha \rightarrow 0+} \frac{Q(t\alpha|x)}{Q(\alpha|x)} = t^{-\gamma(x)}. \quad (3.4)$$

When $\gamma(x) > 0$, the conditional distribution is said to be heavy-tailed and γ is referred to as the conditional tail-index. This is equivalent to say that the survival function $\bar{F}(\cdot|x)$ is regularly varying of index $-1/\gamma(x)$ that is to say, for all $t > 0$,

$$\lim_{y \rightarrow \infty} \frac{\bar{F}(ty|x)}{\bar{F}(y|x)} = t^{-1/\gamma(x)}.$$

Equation (3.4) leads to a simplified estimation strategy for the estimation of the extreme conditional quantile since only the estimation of the conditional tail index γ is required. The approximated extreme quantile is then given by :

$$Q(\beta_n|x) \approx Q(\alpha_n|x) \left(\frac{\alpha_n}{\beta_n} \right)^{\gamma(x)}. \quad (3.5)$$

We pay particular attention to the estimation of extreme conditional quantiles in the heavy-tailed case $\gamma(x) > 0$, as it is the most relevant and challenging setting in many real-world applications. Heavy-tailed distributions

frequently arise in finance, insurance, environmental sciences, and various risk assessment problems, where accurate tail estimation is crucial.

In this paper, we focus on four recently proposed methods adapted to this setting. The first two are machine learning-based : the extremal random forest method developed by Gnecco, Terefe, and Engelke [38] and the extremal neural network approach introduced by Pasche and Engelke [55]. These methods can deal with the general case $\gamma(x) \in \mathbb{R}$ in model (M). In [38], the authors propose a variant that is specifically dedicated to heavy-tailed distributions ($\gamma(x) > 0$). In this paper, we also propose an adaption of the extremal neural network approach [55] to the case $\gamma(x) > 0$. The other two methods are based on dimension reduction techniques specifically designed for extremes : the TIREX method of Aghbalou et al. [1], and the CTI method proposed by Gardes and Podgorny [36]. As these methods have only been introduced recently, few numerical studies have evaluated their relative performance. In particular, the use of dimension reduction techniques developed specifically for extreme value analysis has not yet been systematically assessed for conditional extreme quantile estimation. The main contribution of this paper is to conduct a comprehensive numerical study, evaluating the performance of these techniques across a variety of models and settings. This analysis provides new insights into the strengths and limitations of each method, assessing their robustness to key factors such as the dimension of the covariate space, the choice of the classical quantile level α_n , and the underlying model structure.

The structure of the paper is as follows. Section 3.2 is dedicated to the estimation of classical conditional quantiles using the generalized random forest algorithm. As explained above, this estimation is the first step in the estimation process of an extreme conditional quantile. The following methods for the estimation of an extreme quantile :

1. extremal random forest and its variant for heavy-tailed distribution [38],
2. extremal neural network [55] and its adaptation to heavy-tailed distribution,
3. dimension reduction with TIREX [1],
4. dimension reduction with the CTI method [36],

are presented in detail in Section 3.3. Finally, these methods are compared through a simulation study in Section 3.4.

3.2 Classical conditional quantile estimation

As mentioned in the introduction, the first step in the estimation of an extreme conditional quantile is the estimation of a classical conditional quantile $Q(\alpha_n|x)$ where the order α_n converges to 0 not too fast. Estimation of classical conditional quantiles has been the subject of extensive research. Various effective methods exist, but in this work, we systematically estimate it using the generalized random forest method of Athey, Tibshirani, and Wager [6], which has demonstrated strong performance in nonparametric conditional quantile estimation in high-dimensional settings.

The random forest algorithm was initially introduced by Breiman [15] for estimating the conditional mean $\mu(x) := \mathbb{E}(Y|X = x)$. This method was later generalized by the generalized random forest algorithm, see Athey, Tibshirani, and Wager [6], allowing the use of other loss functions. As a consequence, this method can be used to estimate various quantities beyond the conditional mean, such as conditional quantiles. Hereafter, we present the generalized random forest algorithm for the estimation of a conditional quantile.

For a level $\alpha > 0$, the quantile $Q(\alpha|x)$ minimizes the conditional expectation of the quantile loss function $\rho_\alpha(c) := c(\mathbb{I}\{c > 0\} - \alpha)$:

$$Q(\alpha|x) = \arg \min_{q \in \mathcal{Y}} \mathbb{E}(\rho_\alpha(Y - q)|X = x),$$

where $\mathcal{Y}$ is the support of the distribution of the response variable Y . A natural way to estimate $Q(\alpha|x)$ is to replace the conditional expectation by its empirical counterpart :

$$Q(\alpha|x) := \arg \min_{q \in \mathcal{Y}} \sum_{i=1}^n W(x, X_i) \rho_\alpha(Y_i - q),$$

where $W(x, X_i)$ are positive weights summing to 1 and measuring the similarity between $Q(\alpha|x)$ and $Q(\alpha|X_i)$. The value of $W(x, X_i)$ is expected to be close to 0 if $Q(\alpha|x)$ and $Q(\alpha|X_i)$ are very different and close to 1 if $Q(\alpha|x) \approx Q(\alpha|X_i)$. One method among many others to estimate these similarity weights is the generalized random forest algorithm. More specifically, as for the classical random forest algorithm, generalized random forest aggregate B decision trees where each tree is drawn with replacement from the original dataset. The algorithm is described below. Let $s_n(\alpha) \in \{1, \dots, n\}$ and $m_{RF}(\alpha) \in \{1, \dots, s_n(\alpha)\}$.

1. For $b \in \{1, \dots, B\}$, select a random subsample with replacement of size $s_n(\alpha)$ of the training set. Next, randomly select $k := \min(\lfloor \sqrt{p} \rfloor + 20, \lfloor \sqrt{p} \rfloor)$ components of the covariate. We denote by $P^{(j,b)} := \{X_i^{(b)}; i = 1, \dots, s_n(\alpha)\}$ the obtained random vectors of dimension k and by $\{Y_i^{(b)}; i = 1, \dots, s_n(\alpha)\}$ the associated responses variables.
2. As long as $|P^{(j,b)}| \geq m_{RF}(\alpha)$, split the node $P^{(j,b)}$ into two axis-aligned child nodes C_1 and $C_2 = P^{(j,b)} \setminus C_1$ and compute the criterion

$$\Delta^{(j,b)}(C_1, C_2) = \sum_{\ell=1}^2 \frac{1}{|C_\ell|} \left(\sum_{X_i^{(b)} \in C_\ell} \mathbb{I}\{Y_i^{(b)} > \hat{Q}(\alpha|P^{(j,b)})\} \right)^2 \quad (3.6)$$

where, for a given subset C , $|C|$ is a number of elements in the subset C and

$$\hat{Q}(\alpha|P^{(j,b)}) := \arg \min_{q \in \mathcal{Y}} \sum_{i=1}^n \rho_\alpha(Y_i^{(b)} - q) \mathbb{I}\{X_i^{(b)} \in P^{(j,b)}\}. \quad (3.7)$$

3. Denote by $(P^{(2j,b)}, P^{(2j+1,b)})$ the couple which maximizes (3.6).
4. Repeat steps 2. and 3. until it is no longer possible. The obtained nodes are called the final leafs.

Let us emphasize that the couple that maximizes the criterion $\Delta^{(j,b)}(\cdot, \cdot)$ over all possible splits is the couple of sets where one contains all the $Y_i^{(b)}$'s exceeding the estimated conditional quantile $\hat{Q}(\alpha|P^{(j,b)})$ and the other all the $Y_i^{(b)}$'s located below $\hat{Q}(\alpha|P^{(j,b)})$. As a consequence, for the covariates belonging to the same final leaf, the corresponding conditional quantiles are expected to be close.

The last step of the generalized random forest algorithm is the computation of the estimated similarity weights. More specifically, let $L_b(x)$ denote the terminal leaf of tree b containing x and let

$$w^{(b)}(x, X_i) := \begin{cases} 1/|L_b(x)| & \text{if } X_i \in L_b(x), \\ 0 & \text{otherwise.} \end{cases}$$

The estimated similarity weights are given by

$$W_\alpha^{(RF)}(x, X_i) := \frac{1}{B} \sum_{b=1}^B w^{(b)}(x, X_i).$$

Hyper-parameter	Description	Selection
B	number of trees	$B = 2000$
$s_n(\alpha)$	size of the subsamples	$s_n(\alpha) = \lfloor n/2 \rfloor$
$m_{RF}(\alpha)$	minimal number of points if a node	$m_{RF}(\alpha) = 5$

TABLE 3.1 – Hyper-parameters for the estimation of a classical conditional quantile.

Averaging over the B trees smooths out variability across individual trees, ensuring a capacity to capture complex conditional structures. The classical conditional quantile estimator is finally given by

$$\check{Q}^{(\text{RF})}(\alpha|x) := \arg \min_{q \in \mathcal{Y}} \sum_{i=1}^n W_{\alpha}^{(\text{RF})}(x, X_i) \rho_{\alpha}(Y_i - q). \quad (3.8)$$

This method can be extended to estimate multiple quantiles simultaneously at different levels $\alpha_1 > \dots > \alpha_K > \alpha_{K+1} := 0$. For a given node P , instead of partitioning the data into two groups based on a single conditional quantile estimate, observations are divided into $K + 1$ intervals

$$I_k(P) := \left[\hat{Q}(\alpha_k|P), \hat{Q}(\alpha_{k+1}|P) \right), \quad k = 1, \dots, K.$$

For the b -th tree, the optimal split $(C_1, C_2 = P \setminus C_1)$ of node P maximizes the multiple quantiles criterion :

$$\sum_{k=1}^K \sum_{\ell=1}^2 \frac{1}{|C_{\ell}|} \left(\sum_{X_i^{(b)} \in C_{\ell}} \mathbb{I}\{Y_i^{(b)} \in I_k(P)\} \right)^2. \quad (3.9)$$

We end this section by giving the hyper-parameters required to perform this estimation procedure of $Q(\alpha_n|x)$ together with a proposition of selection.

3.3 Tail parameters estimation

After estimating the classical conditional quantile, the second step consists in estimating the tail parameters. We present here three main approaches :

extreme random forest, extreme neural network and dimension reduction techniques. Gradient boosting has also been proposed for extreme conditional quantile estimation by Velthoen, Dombry, Cai, and Engelke [70] and Yougmann uses generalized additive models for the parameters of a GPD distribution [78]. However, recent comparisons by Pasche and Engelke [55] and Gnecco, Terefe, and Engelke [38] have shown that they are slightly less effective than neural networks and random forests in capturing extreme quantiles. Given these findings, we focus on the most competitive methods, which have demonstrated superior performance in high-dimensional settings.

3.3.1 Extreme random forest

The extreme random forest (ERF) estimator was proposed by Gnecco, Terefe, and Engelke [38]. It is based on a result of Pickands [56], which states that under model (M), the excess distribution above a high threshold converges to a Generalized Pareto Distribution (GPD) : for all $z > 0$,

$$\lim_{\alpha \rightarrow 0_+} \mathbb{P} \left(\frac{Y - Q(\alpha|x)}{a(\alpha^{-1}|x)} < z \middle| Y > Q(\alpha|x), X = x \right) = G(z, 1, \gamma(x)), \quad (3.10)$$

where $G(z, a, \gamma) := 1 - (1 + \gamma z/a)^{-1/\gamma}$ is the cumulative distribution function of a GPD with parameters (a, γ) . The corresponding negative log-density function is given for $z > 0$ by

$$g(z, a, \gamma) := -\log \frac{\partial G(z, a, \gamma)}{\partial z} = \log(a) + \left(1 + \frac{1}{\gamma}\right) \log \left(1 + \frac{\gamma}{a} z\right). \quad (3.11)$$

The idea behind the ERF procedure is presented hereafter. For a given classical conditional quantile level α_n , let

$$\{\hat{Y}_i(\alpha_n); i = 1, \dots, M_n(\alpha_n)\},$$

be the random variables Y_i 's larger than the estimated classical conditional quantile $\check{Q}^{(\text{RF})}(\alpha_n|X_i)$ which is defined in (3.8). The random variables $\hat{Y}_i(\alpha_n) - \check{Q}^{(\text{RF})}(\alpha_n|X_i)$ are referred to as the empirical excesses. If a large weight is assigned to the empirical excesses for which the associated covariate X_i is such that the conditional distribution of Y given $X = X_i$ is similar to the one of Y given $X = x$, we get a sample that, according to (3.10), follows

approximatively a GPD with parameters $(a_n(x), \gamma(x))$. The difference between the two conditional distributions is measured by the similarity weights

$$\left\{ W_{\vec{\alpha}}^{(\text{RF})}(x, X_i); i = 1, \dots, n \right\},$$

which are obtained with the generalized random forest algorithm described in Section 3.2 by using the multiple quantiles criterion (3.9) with $\vec{\alpha} = (0.9, 0.5, 0.1)$. This leads us to the definition of the ERF estimator of an extreme conditional quantile as introduced in [38]. Let us first give some notations.

For $i \in \{1, \dots, n\}$, let $U_i := (Y_i - \check{Q}^{(\text{RF})}(\alpha_n | X_i))_+$ where $(\cdot)_+$ is the positive part function. Let also introduced the estimators

$$(\hat{a}_n, \hat{\gamma}_n) = \arg \min_{(a, \gamma)} \sum_{i=1}^n g(Y_i - Q_Y(\alpha_n), a, \gamma) \mathbb{I}\{Y_i > Q_Y(\alpha_n)\}, \quad (3.12)$$

where

$$Q_Y(\alpha_n) := \arg \min_{q \in \mathcal{Y}} \sum_{i=1}^n \rho_{\alpha_n}(Y_i - q). \quad (3.13)$$

Note that $\hat{a}_n$ and $\hat{\gamma}_n$ are the estimators of the tail-parameters obtained in the unconditional case that is to say when no covariate X is recorded with the response variable Y .

Definition 8. Let $x \in \mathcal{X}$ and let α_n be a classical conditional quantile level. For $\beta_n < \alpha_n$, the ERF estimator of the extreme conditional quantile $Q(\beta_n | x)$ is given by

$$\hat{Q}^{(\text{RF})}(\beta_n | x) := \check{Q}^{(\text{RF})}(\alpha_n | x) + \frac{\hat{a}_n^{(\text{RF})}(x)}{\hat{\gamma}_n^{(\text{RF})}(x)} \left[\left(\frac{\alpha_n}{\beta_n} \right)^{\hat{\gamma}_n^{(\text{RF})}(x)} - 1 \right],$$

where the estimated tail-parameters $\hat{a}_n^{(\text{RF})}(x)$ and $\hat{\gamma}_n^{(\text{RF})}(x)$ minimize the penalized negative log-likelihood

$$\sum_{i=1}^n W_{\vec{\alpha}}^{(\text{RF})}(x, X_i) g(U_i, a, \gamma) \mathbb{I}\{U_i > 0\} + \lambda(\gamma - \hat{\gamma}_n)^2,$$

for some regularization parameter $\lambda > 0$ and with $\hat{\gamma}_n$ defined in (3.12).

The penalization is introduced to prevent excessive fluctuations the conditional tail index estimator across different values of x . Large variations can lead to erratic behavior in extreme quantile estimates, particularly in regions where data is sparse.

A heavy-tailed variant with the log-exceedances is also proposed by Gnecco, Terefe, and Engelke [38]. It is based on the asymptotic result (see [39, Theorem 3.2.2]) :

$$\lim_{\alpha \rightarrow 0_+} \mathbb{P} \left(\log \frac{Y}{Q(\alpha|x)} < z \mid Y > Q(\alpha|x), X = x \right) = \mathcal{E}(1/\gamma(x)),$$

where $\mathcal{E}(1/\gamma(x))$ denotes the cumulative distribution function of an exponential distribution with mean $\gamma(x)$. This suggests an alternative estimation strategy by approximating the weighted expectation of the log-exceedances $\log \hat{Y}_i(\alpha_n) - \log \check{Q}^{(RF)}(\alpha_n|X_i)$ by $\gamma(x)$. More specifically, introducing the random variables

$$V_i := \left(\log \frac{Y_i}{\check{Q}^{(RF)}(\alpha_n|X_i)} \right)_+$$

the following estimator was introduced in [38].

Definition 9. Let $x \in \mathcal{X}$ and let α_n be a classical conditional quantile level. For $\beta_n < \alpha_n$, the heavy-tailed extreme random forest (HT-ERF) estimator of the extreme conditional quantile $Q(\beta_n|x)$ is given by

$$\hat{Q}^{(HRF)}(\beta_n|x) = \check{Q}^{(RF)}(\alpha_n|x) \left(\frac{\alpha_n}{\beta_n} \right)^{\hat{\gamma}_n^{(HRF)}(x)},$$

where the estimated tail index parameter is given by :

$$\hat{\gamma}_n^{(HRF)}(x) \in \arg \min_{\gamma > 0} \sum_{i=1}^n W_{\vec{\alpha}}^{(RF)}(x, X_i) (V_i - \gamma)^2 \mathbb{I}\{V_i > 0\} + \lambda(\gamma - \hat{\gamma}_n)^2,$$

for some regularization parameter $\lambda > 0$ and with $\hat{\gamma}_n$ defined in (3.12).

In the following table, we give all the parameters required for the extreme random forest algorithm.

For the estimator $\tilde{Q}^{(\text{RF})}(\alpha_n x)$		
Hyper-parameter	Description	Selection
α_n	classical quantile level	Choice left to the user
B	number of trees	$B = 2000$
$s_n(\alpha_n)$	size of the subsamples	$s_n(\alpha_n) = \lfloor n/2 \rfloor$
$m_{RF}(\alpha_n)$	minimal number of points if a node	$m_{RF}(\alpha_n) = 5$
For the similarity weights $W_{\vec{\alpha}}^{(\text{RF})}(x, X_i)$		
Hyper-parameter	Description	Selection
B	number of trees	$B = 2000$
$s_n(\vec{\alpha})$	size of the subsamples	$s_n(\vec{\alpha}) = \lfloor n/2 \rfloor$
$m_{RF}(\vec{\alpha})$	minimal number of points if a node	cross-validation
Estimation of the tail parameter(s)		
Hyper-parameter	Description	Selection
λ	regularization parameter	cross-validation

TABLE 3.2 – Hyper-parameters for the extreme random forest algorithm.

The Random Forest approach is advantageous for extreme quantile estimation due to its theoretical guarantees and computational efficiency. Gnecco, Terefe, and Engelke [38] established the consistency under reasonable assumptions, ensuring reliable estimation as the sample size grows. Additionally, the method is computationally efficient, making it well-suited for large datasets.

3.3.2 Extreme neural network

Neural networks are powerful models for capturing complex, non-linear relationships in data, and their use has become increasingly popular across a range of fields, including extreme value theory in recent years [3, 4]. In this section, we focus on the Pacsche and Engelke approach’s [55]. The field of neural networks is vast, and many additional methods exist that are beyond the scope of this work. For a more comprehensive introduction to neural networks, we refer the reader to classic works such as [2].

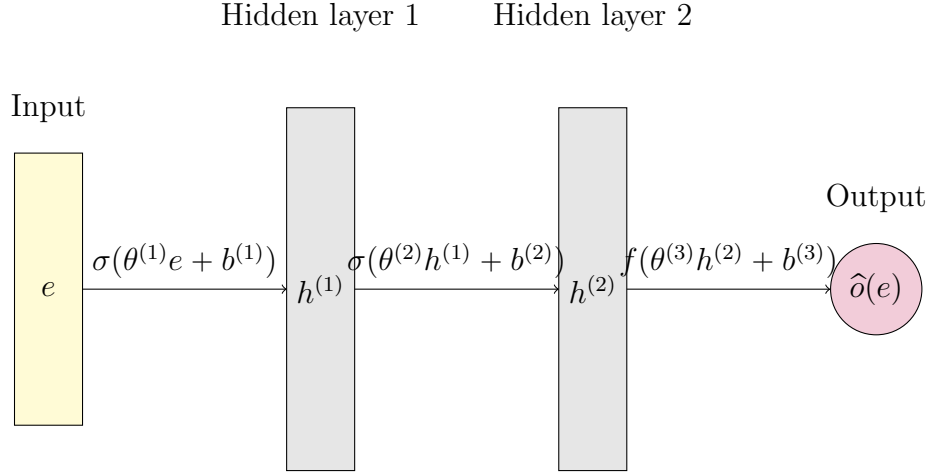


FIGURE 3.1 – Schematic representation of an FNN with two hidden layers. The operation between each layer is denoted above the arrows.

Introduction to feedforward neural networks

A feedforward neural network (FNN) maps an input $e \in \mathbb{R}^p$ to an output $\hat{o}(e) \in \mathbb{R}^{p_o}$. It consists of $L - 1$ hidden layers and 1 output layer, each performing a linear transformation followed by a non-linear activation function. Formally, for each hidden layer $l \in \{1, \dots, L - 1\}$, the transformation can be written as :

$$h^{(l)} = \sigma(\theta^{(l)}h^{(l-1)} + b^{(l)}) \in \mathbb{R}^{d_l},$$

where $\theta^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ is the weight matrix (with $d_0 = p$), $b^{(l)} \in \mathbb{R}^{d_l}$ is the bias vector, and σ is the activation function. Note that σ applies to a vector element-wise i.e., $\sigma((x_1, \dots, x_d)^\top) = (\sigma(x_1), \dots, \sigma(x_d))^\top$. The input for the first hidden layer is $h^{(0)} = e$. The output layer produces the final prediction of dimension p_o :

$$\hat{o}(e) = f(\theta^{(L)}h^{(L-1)} + b^{(L)}) \in \mathbb{R}^{p_o}.$$

where f is a arbitrary function, $\theta^{(L)} \in \mathbb{R}^{p_o \times d_{L-1}}$ and $b^{(L)} \in \mathbb{R}^{p_o}$. Figure 3.1 illustrates the architecture of a simple FNN with two hidden layers ($L = 3$).

The output function $\hat{o}(\cdot) = \hat{o}(\cdot|\mathcal{W})$ is thus a quite complex function depending on a vector $\mathcal{W}$ of parameters. This vector contains the coefficients of the

matrices $\{\theta^{(l)}, l = 1, \dots, L\}$ and the components of the vectors $\{b^{(l)}, l = 1, \dots, L\}$ so that the dimension of W is

$$D_{FNN} := \sum_{l=1}^L (d_{l-1} + 1) d_l,$$

with $d_L = p_o$. Given a training set $\{(e_1, o_1), \dots, (e_n, o_n)\}$, the parameter vector of the output function is denoted by $\widehat{W}_n$ and is obtained by maximizing the average loss function :

$$\mathcal{AL}(\mathcal{W}) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\widehat{o}(e_i|W), o_i),$$

where $\mathcal{L}(\cdot, \cdot)$ is a given loss function that depends on the specific context in which the FNN is used (e.g., regression, classification, etc.).

This optimization problem is solved by using a gradient descent algorithm : for a given learning rate η , the parameter vector is updated by using the rule

$$\mathcal{W} \leftarrow \mathcal{W} - \eta \nabla_{\mathcal{W}}(\mathcal{AL}(\mathcal{W})). \quad (3.14)$$

To be computationally efficient, the gradient needs to be computed iteratively, layer by layer, starting from the last layer and propagating back to the first (backpropagation algorithm). This update rule is repeated until convergence (with respect to a criterion choose by the user) or until a given number of iterations is reached. In practice, instead of computing the gradient for each point of the training set, which is computationally demanding, we randomly divide the training set into n/s_B mini-batches $\{B_j, j = 1, \dots, n/s_B\}$ of size $s_B \in \{1, \dots, n\}$. We then perform iteratively from B_1 to B_{n/s_B} the update rule (3.14) by replacing the gradient $\nabla_{\mathcal{W}}(\mathcal{AL}(\mathcal{W}))$ by

$$\frac{1}{s_B} \sum_{i \in B_j} \nabla_{\mathcal{W}}(\mathcal{L}(\widehat{o}(e_i|\mathcal{W}), o_i)).$$

A complete pass through all mini-batches $\{B_j, j = 1, \dots, n/s_B\}$ is referred to as an epoch. In general, several epochs (say $n_e > 1$) are required to ensure convergence. Let us emphasize that the choice of the mini-batch size s_B plays an important role in training dynamics. Typically, values of s_B range from 32 to 512, depending on the dataset size n and computational constraints.

Estimation of an extreme conditional quantile

When observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$ are drawn from an independent sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, Pasche and Engelke [55] propose to estimate the tail parameters $a_n(x)$ and $\gamma(x)$ by the output of an FNN. Their method, referred to as the Extreme Quantile Regression Network (EQRN), was initially tailored for flood risk forecasting in the Swiss Aare river catchment. For $\zeta = (\nu, \gamma) \in (0, \infty) \times (-1/2, \infty)$, the FNN is trained by using the loss function

$$\mathcal{L}^{(\text{GPD})}(\zeta, u) := \left[\left(1 + \frac{1}{\gamma}\right) \log \left(1 + \gamma \frac{u}{\nu}\right) + \log(\nu) - \log(1 + \gamma) \right] \mathbb{I}\{u > 0\}.$$

Note that $\mathcal{L}^{(\text{GPD})}(\zeta, u)$ is a reparameterization of the negative log-density of a GPD with parameters a and γ (see (3.11)) where $\nu = a(1 + \gamma)$. The authors have observed that this new parameterization leads to a more stable estimation of the GPD parameters. The training set is $\{(e_1, o_1), \dots, (e_n, o_n)\}$ with $e_i = x_i$ and $o_i = u_i$ where

$$u_i := \left(y_i - \check{Q}^{(\text{RF})}(\alpha_n | x_i) \right)_+,$$

with $\check{Q}^{(\text{RF})}(\alpha_n | x_i)$ is defined in (3.8) for some classical conditional quantile level α_n . For an input x , denoting by $\mathcal{W}$ the parameter vector of the FNN, the output is the two-dimensional vector

$$\hat{\zeta}(x | \mathcal{W}) = (\hat{\nu}(x | \mathcal{W}), \hat{\gamma}(x | \mathcal{W})).$$

Following the recommendation of Pasche and Engelke [55], this output is obtained by applying the activation function f given for $x \in \mathbb{R}$ by

$$f(x) = (\exp(x), 0.6 \tanh(x) + 0.1),$$

so that the scale-related parameter $\hat{\nu}(x) \in (0, +\infty)$, and the tail index estimate $\hat{\gamma}(x) \in (-0.5, 0.7)$. Using the procedure described in Section 3.3.2, the parameter vector of the trained FNN is the vector $\widehat{\mathcal{W}}_n^{(\text{GPD})}$ minimizing the penalized average loss

$$\mathcal{AL}^{(\text{GPD})}(\mathcal{W}, \lambda) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}^{(\text{GPD})} \left(\hat{\zeta}(x_i | \mathcal{W}), u_i \right) + \lambda \|\mathcal{W}\|_2, \quad (3.15)$$

for some regularization parameter $\lambda > 0$ and where $\|\cdot\|_2$ is the L^2 -norm. The penalty term has been added to improve the generalization power of the FNN. We are now in a position to give the definition of the extreme conditional quantile $Q(\beta_n|x)$.

Definition 10. *Let $x \in \mathcal{X}$ and let α_n be a classical conditional quantile level. For $\beta_n < \alpha_n$, the Extreme Quantile Regression Network (EQRN) estimator of the extreme conditional quantile $Q(\beta_n|x)$ is given by*

$$\hat{Q}^{(NN)}(\beta_n|x) = \check{Q}^{(RF)}(\alpha_n|x) + \frac{\hat{a}_n^{(NN)}(x)}{\hat{\gamma}_n^{(NN)}(x)} \left[\left(\frac{\alpha_n}{\beta_n} \right)^{\hat{\gamma}_n^{(NN)}(x)} - 1 \right],$$

where the estimated parameters are given by

$$\hat{a}_n^{(NN)}(x) = \frac{\hat{v}(x|\widehat{\mathcal{W}}_n^{(GPD)})}{1 + \hat{\gamma}(x|\widehat{\mathcal{W}}_n^{(GPD)})}, \quad \text{and} \quad \hat{\gamma}_n^{(NN)}(x) = \hat{\gamma}(x|\widehat{\mathcal{W}}_n^{(GPD)}),$$

where $\widehat{\mathcal{W}}_n^{(GPD)}$ minimizes (3.15) for some regularization parameter $\lambda > 0$.

We finally propose to adapt the above estimator to the situation where the conditional tail index is positive, that is, for heavy-tailed conditional distributions. Note that this adaptation was not done in the paper by Pasche and Engelke [55]. In this case, the neural network is trained using the loss function

$$\mathcal{L}^{(HT)}(\gamma, v) = (v - \gamma)^2 \mathbb{I}\{v > 0\},$$

with the training set $\{(e_1, o_1), \dots, (e_n, o_n)\}$ where $e_i = x_i$,

$$o_i =: v_i := \log \left(\frac{y_i}{\check{Q}^{(RF)}(\alpha_n|x_i)} \right),$$

and where α_n is a classical conditional quantile level. For an input x , we denote by $\hat{\gamma}(x|\mathcal{W})$ the output of the FNN calculated with the parameter vector $\mathcal{W}$. Here, the activation function for the output layer is $f(x) = x$. As before, the trained parameter vector is $\widehat{\mathcal{W}}_n^{(HT)}$ which minimizes the penalized average loss

$$\mathcal{AL}^{(HT)}(\mathcal{W}, \lambda) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}^{(HT)}(\hat{\gamma}(x_i|\mathcal{W}), v_i) + \lambda \|\mathcal{W}\|_2, \quad (3.16)$$

for some positive regularization parameter λ . The estimator of the extreme conditional quantile for heavy-tailed distributions is given hereafter.

Definition 11. Let $x \in \mathcal{X}$ and let α_n be a classical conditional quantile level. For $\beta_n < \alpha_n$, the Heavy-Tailed Extreme Quantile Regression Network (HT-EQRN) estimator is given by :

$$\hat{Q}^{(HNN)}(\beta_n|x) = \hat{Q}^{(RF)}(\alpha_n|x) \left(\frac{\alpha_n}{\beta_n} \right)^{\hat{\gamma}_n^{(HNN)}(x)},$$

where the estimated tail index parameter is given by :

$$\hat{\gamma}_n^{(HNN)}(x) = \hat{\gamma}(x|\hat{\mathcal{W}}_n^{(HT)}),$$

where $\hat{\mathcal{W}}_n^{(HT)}$ minimizes (3.16) for some regularization parameter $\lambda > 0$.

Neural network-based methods offer high flexibility and do not require strong distributional assumptions. However, the choice of architecture and regularization is crucial to avoid overfitting ; see Table 3.3 where all the required hyperparameters are listed. Furthermore, they lack theoretical guarantees regarding consistency and convergence, which can make their results harder to interpret compared to classical statistical methods.

3.3.3 Dimension reduction procedures

While machine learning methods generally perform well in high-dimensional settings, reducing the dimensionality of the covariate space can sometimes give better results. The basic idea is to add a constraint to the model for the couple (X, Y) by assuming that all the information contained in the covariate X about the tail behavior of Y is captured by a lower-dimensional projection of the predictor vector, $B^\top X \in \mathbb{R}^{p \times q}$, with $q < p$. Hence, if B is known (or can be estimated), inference can be performed by replacing X with the lower-dimensional covariate $B^\top X$. When $q \ll p$, dimension reduction methods allows us to circumvent the curse of dimensionality.

To our knowledge, three dimension reduction methods have been developed for extreme values. The earliest, proposed by Gardes [33], is computationally demanding and its implementation is not easy when $q > 1$. Two more recent approaches will be presented in this section : the tail inverse regression proposed by Aghbalou et al. [1] and the method based on the central tail index subspace introduced in Gardes and Podgorny [36].

For the estimator $\check{Q}^{(\text{RF})}(\alpha_n x)$		
Hyper-parameter	Description	Selection
α_n	classical quantile level	Choice left to the user
B	number of trees	$B = 2000$
$s_n(\alpha_n)$	size of the subsamples	$s_n(\alpha_n) = \lfloor n/2 \rfloor$
$m_{RF}(\alpha_n)$	minimal number of points if a node	$m_{RF}(\alpha_n) = 5$
For the stochastic gradient descent algorithm		
Hyper-parameter	Description	Selection
η	learning rate	$\eta = 10^{-4}$
n_e	number of epochs	Early stopping*
s_B	size of mini-batches	128
Architecture of the neural network		
Hyper-parameter	Description	Selection
L	number of layers	cross-validation
σ	activation function for the hidden layers	$\sigma = \tanh$
$d_1, \dots, d_{L-1}$	dimensions of the weight matrices	cross-validation
Estimation of the tail parameter(s)		
Hyper-parameter	Description	Selection
λ	regularization parameter	cross-validation

TABLE 3.3 – Hyper-parameters for the extreme neural network algorithm.

* The early stopping procedure is described in Section [3.4](#)

Tail inverse regression

Aghbalou et al. [1] propose two dimension reduction methods, TIREX1 and TIREX2, which can be seen as adaptations of Sliced Inverse Regression (SIR) (Li [51]) and SAVE (Cook and Weisberg [22]), respectively. However, TIREX1 is theoretically limited to estimating subspaces of dimension at most 1. Hereafter, we chose to focus on TIREX2 method.

The TIREX2 method estimates a basis for the Extreme Sufficient Direction Reduction (Extreme SDR) subspace, defined as follows :

Definition 12. *Let $y^\star$ be the right endpoint of the distribution of Y . A linear subspace $\mathcal{S}_E$ spanned by a basis $B \in \mathbb{R}^{p \times q}$ is an Extreme SDR if*

$$\lim_{y \rightarrow y^\star} \mathbb{E} \left| \frac{\mathbb{P}(Y > y | B^\top X) - \mathbb{P}(Y > y | X)}{\mathbb{P}(Y > y)} \right| = 0. \quad (3.17)$$

Moreover, $\mathcal{S}_E$ is called a central Extreme SDR subspace if it is an Extreme SDR subspace containing all the Extreme SDR subspaces.

If B spans an Extreme SDR, the tails of the conditional distributions $Y|X = x$ and $Y|B^\top X = B^\top x$ are, roughly speaking, close to each other. However, Equation (3.17) does not imply that $Q(\alpha|x)/Q(\alpha|B^\top x) \rightarrow 1$ as $\alpha \rightarrow 0$. The authors' primary focus was not on extreme quantile estimation, but our simulations suggest that the method can be effectively used for this purpose in practice. Note also that the central Extreme SDR subspace does not necessarily exist but if it is the case, it is unique and given by the intersection of all Extreme SDR subspaces. In all what follows, we assume the the central Extreme SDR $\mathcal{S}_E$ exists.

Let us now give the main idea leading to the estimator of the linear subspace $\mathcal{S}_E$ proposed by Aghbalou et al. [1]. Let $Z := \Sigma^{-1/2}(X - m)$ be the normalized random vector of covariates with $m = \mathbb{E}(X)$ and $\Sigma = \mathbb{V}(X)$ and let P_E be the orthogonal projection matrix on the subspace $\mathcal{S}_E$. Under the linearity condition $\mathbb{E}(Z|P_E Z) = P_E Z$ and if $\mathbb{V}(Z|P_E Z)$ is almost surely constant, the author's establish under additional technical conditions (see [1, Theorem 1 and 2]) that the $p \times p$ matrix

$$M_T := \int_0^1 D_n(u) [D_n(u)]^\top du,$$

with, denoting by I_p the identity matrix and by Q_Y the quantile function of Y (see (3.13)),

$$D_n(u) := \frac{1}{\alpha_n} \mathbb{E} \left[(ZZ^\top - I_p) \mathbb{I}\{Y > Q_Y(\alpha_n)\} \right],$$

is such that $\text{span}(M_T) \subset \mathcal{S}_E$. The TIREX estimator of the central Extreme SDR subspace is the subspace spanned by the q largest eigenvalues of the estimated matrix :

$$\widehat{M}_n := \int_0^1 \widehat{D}_n(u) \left[\widehat{D}_n(u) \right]^\top du,$$

where

$$\widehat{D}_n(u) := \Sigma_n^{-1/2} \left(\frac{1}{[n\alpha_n]} \sum_{i=1}^n \overline{X}_i \overline{X}_i^\top \mathbb{I}\{Y_i > Y_{n-[n\alpha_n]+1,n}\} \right) \Sigma_n^{-1/2},$$

with $Y_{1,n} \leq \dots \leq Y_{n,n}$ the order statistics associated to the sample $\{Y_1, \dots, Y_n\}$,

$$\overline{X}_i := X_i - \frac{1}{n} \sum_{j=1}^n X_j \text{ and } \Sigma_n = \frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right) \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^\top.$$

Denoting by $\widehat{B}_n^{(T)}$ the $p \times q$ matrix of the q largest eigenvalues of $\widehat{M}_n$, we proposed the following kernel-based estimator of $Q(\beta_n|x)$ when the conditional distribution of Y given $X = x$ is heavy-tailed.

Definition 13. Let $x \in \mathcal{X}$ and let α_n be a classical conditional quantile level. For $\beta_n < \alpha_n$, the TIREX estimator is given by :

$$\widehat{Q}^{(T)}(\beta_n|x) = \check{Q}^{(RF)}(\alpha_n|x) \left(\frac{\alpha_n}{\beta_n} \right)^{\widehat{\gamma}_n^{(T)}(x)}, \quad (3.18)$$

where the estimated tail index is given for a non-negative function $\mathcal{K}$ on $\mathbb{R}^q$ and a bandwidth h_n by :

$$\widehat{\gamma}_n^{(T)}(x) \in \arg \min_{\gamma \in \mathbb{R}} \sum_{i=1}^n \mathcal{K} \left(\frac{[\widehat{B}_n^{(T)}]^\top (x - X_i)}{h_n} \right) (V_i - \gamma)^2 \mathbb{I}_{V_i > 0}.$$

For the estimator $\tilde{Q}^{(\text{RF})}(\alpha_n x)$		
Hyper-parameter	Description	Selection
α_n	classical quantile level	user choice
B	number of trees	$B = 2000$
$s_n(\alpha_n)$	size of the subsamples	$s_n(\alpha_n) = \lfloor n/2 \rfloor$
$m_{\text{RF}}(\alpha_n)$	minimal number of points if a node	$m_{\text{RF}}(\alpha_n) = 5$
Estimation of the basis $\hat{B}_n^{(T)}$		
Hyper-parameter	Description	Selection
q	dimension of the subspace	cross-validation
For the Kernel method		
Hyper-parameter	Description	Selection
$\mathcal{K}$	kernel density	$\mathcal{K}(\cdot) = \mathbb{I}\{\ \cdot\ _\infty < 1\}$
h_n	bandwidth parameter	$h_n = n^{-0.2/q}$

TABLE 3.4 – Hyper-parameters for the TIREX method.

Central Tail Index Subspace

Unlike TIREX, Gardes and Podgorny [36] propose a dimension reduction approach which is specifically designed for the estimation of the conditional tail index in the heavy-tailed case. In this section, we assume that for all $x \in \mathcal{X}$, the conditional distribution function of Y given $X = x$ is heavy-tailed with tail index $\gamma(x) > 0$. Under additional assumptions, the authors prove that for all $B \in \mathbb{R}^{p \times q}$, the conditional distribution of Y given $B^\top X = B^\top x$ remains heavy-tailed with tail index

$$\xi_B(B^\top x) := \max_{z: B^\top z = B^\top x} \gamma(z). \quad (3.19)$$

We then define the Tail-Index Dimension Reduction (TIDR) subspace as follows.

Definition 14. *A linear subspace $\mathcal{S}$ spanned by a basis $B \in \mathbb{R}^{p \times q}$ is a Tail-Index Dimension Reduction (TIDR) subspace if the conditional distributions of $Y|X$ and $Y|B^\top X$ share the same conditional tail index, that is to say if $\xi_B(B^\top X) = \gamma(X)$ almost surely. Moreover, $\mathcal{S}$ is called a central Tail Index (CTI) subspace if it is an TIDR subspace containing all the TIDR subspaces.*

In what follows, we assume that the CTI subspace exists. It is thus unique as the intersection of all the TIDR subspaces. We denote by $B_0 \in \mathcal{B}_q$ its canonical basis (see [36, Definition 2]). Here, $\mathcal{B}_q$ denotes the set that gathers the canonical bases of all linear subspaces of dimension q . Note that for all $x \in \mathcal{X}$ and $B \in \mathcal{B}_q$,

$$\gamma(x) = \xi_{B_0}(B_0^\top x) \leq \xi_B(B^\top x) = \max_{z: B^\top z = B^\top x} \gamma(z),$$

since $\gamma(x)$ belongs to the support of the conditional distribution of $\gamma(X)$ given $B^\top X = B^\top x$. Hence, for a compact subset $\mathcal{X}_0$ in the interior of $\mathcal{X}$ satisfying the condition

(C) For $B \in \mathcal{B}_q$, if $\xi_B(B^\top x) = \gamma(x)$ for all $x \in \mathcal{X}_0$ then $B = B_0$,

we get that

$$\arg \min_{B \in \mathcal{B}_q} \mathbb{E} [\xi_B(B^\top X) \mathbb{I}\{X \in \mathcal{X}_0\}] =: \arg \min_{B \in \mathcal{B}_q} \Psi(B, \mathcal{X}_0) = \{B_0\}.$$

Under condition (C), we propose to estimate B_0 by the canonical basis $\hat{B}_n^{(\text{CTI})}$ which minimizes the empirical version of $\Psi(B, \mathcal{X}_0)$ given by :

$$\hat{\Psi}_n(B, \mathcal{X}_0) := \frac{1}{n} \sum_{i=1}^n \hat{\xi}_{n,B}(B^\top X_i) \mathbb{I}\{X_i \in \mathcal{X}_0\}.$$

Similarly to TIREX, we define the following kernel-based estimator of an extreme conditional quantile.

Definition 15 (CTI). *Let $x \in \mathcal{X}$ and let α_n be a classical conditional quantile level. For $\beta_n < \alpha_n$, the CTI estimator is given by :*

$$\hat{Q}^{(\text{CTI})}(\beta_n | x) = \check{Q}^{(\text{RF})}(\alpha_n | x) \left(\frac{\alpha_n}{\beta_n} \right)^{\hat{\gamma}_n^{(\text{CTI})}(x)}, \quad (3.20)$$

where the estimated tail index is given for a non-negative $\mathcal{K}$ on $\mathbb{R}^q$ and a bandwidth h_n by :

$$\hat{\gamma}_n^{(\text{CTI})}(x) \in \arg \min_{\gamma \in \mathbb{R}} \sum_{i=1}^n \mathcal{K} \left(\frac{[\hat{B}_n^{(\text{CTI})}]^\top (x - X_i)}{h_n} \right) (V_i - \gamma)^2 \mathbb{I}_{V_i > 0}.$$

For the estimator $\tilde{Q}^{(\text{RF})}(\alpha_n x)$		
Hyper-parameter	Description	Selection
α_n	classical quantile level	Choice left to the user
B	number of trees	$B = 2000$
$s_n(\alpha_n)$	size of the subsamples	$s_n(\alpha_n) = \lfloor n/2 \rfloor$
$m_{\text{RF}}(\alpha_n)$	minimal number of points if a node	$m_{\text{RF}}(\alpha_n) = 5$
Estimation of the basis $\hat{B}_n^{(\text{CTI})}$		
Hyper-parameter	Description	Selection
q	dimension of the subspace	Procedure in [36]
For the Kernel method		
Hyper-parameter	Description	Selection
$\mathcal{K}$	kernel density	$\mathcal{K}(\cdot) = \mathbb{I}\{\ \cdot\ _\infty < 1\}$
h_n	bandwidth parameter	$h_n = n^{-0.2/q}$

TABLE 3.5 – Hyper-parameters for the CTI method.

Dimension reduction methods provide an efficient way to handle high-dimensional covariate spaces by projecting the data onto a lower-dimensional subspace while preserving relevant information for tail index estimation. Additionally, it can enhance interpretability by identifying key directions that influence extreme events. However, they rely on the correct identification of the relevant subspace, which may not always be theoretically justified.

3.4 Simulation Study

In this section, we evaluate the performance of the various methods presented in this paper for estimating extreme conditional quantiles in a high-dimensional setting. The goal is to assess their accuracy and robustness under controlled conditions where the true quantiles are known. The study is structured as follows. Section 4.1 describes the general setup for the data-generating process and the evaluation protocol. Section 4.2 discusses the selection of tuning parameters for each method. Section 4.3 presents and analyzes the results of the experiments across different scenarios.

3.4.1 General Setup

We generate $R = 100$ independent replications of samples $(Y_1, X_1), \dots, (Y_n, X_n)$, each of size $n = 2000$, where the covariates $X \in \mathbb{R}^p$ are drawn uniformly from the hypercube $[-1, 1]^p$. The conditional distribution of $Y|X = x$ is assumed to be heavy-tailed, with a conditional quantile function given for $x = (x_1, \dots, x_p)^\top$ by

$$Q(u|x) = q(u, \gamma(x)) \times (1 + 6\phi(x_1, x_2)),$$

where $q(\cdot, \gamma(x))$ is a heavy-tailed quantile function of a distribution with local tail index $\gamma(x) > 0$ and ϕ denotes the density of a standard bivariate normal distribution centered at $(0, 0)$ with correlation coefficient 0.9. Note that this conditional quantile function is also used by Pasche and Engelke [55] in their simulation study.

On each replication, we apply all the estimation methods presented in this paper to estimate the extreme conditional quantile of level $\beta_n = 2/n$. The performance of each method is evaluated by computing the estimation error on the set $\{x_i, i = 1, \dots, 1000\}$ of the first 1000 points of a Halton sequence in $[-1, 1]^p$. For each replication r , we evaluate the estimation accuracy using the symmetric relative error at each evaluation point x_i , defined as

$$E^{(r, \bullet)}(x_i) := \max \left(\left| \frac{\hat{Q}^{(\bullet)}(\beta_n|x_i)}{Q(\beta_n|x_i)} - 1 \right|, \left| \frac{Q(\beta_n|x_i)}{\hat{Q}^{(\bullet)}(\beta_n|x_i)} - 1 \right| \right),$$

with $\bullet \in \{\text{RF}, \text{HRF}, \text{NN}, \text{HNN}, \text{T}, \text{CTI}\}$. This criterion is scale-invariant and penalizes both underestimation and overestimation in a balanced way. The global error for the r -th replication is then defined as the median of the pointwise errors over the evaluation set :

$$\overline{E}^{(r, \bullet)} = \text{Med} \left(\{E^{(r, \bullet)}(x_i) : i = 1, \dots, 1000\} \right).$$

3.4.2 Choice of Tuning Parameters

All methods presented in this paper involve several hyperparameters that can significantly affect their performance. In this section, we describe the tuning procedures adopted for each approach. The hyper-parameter values used for

each method are summarized in the tables provided in their respective sections. Unless otherwise stated, these values follow the recommendations of the authors who introduced the corresponding methods. When parameters require tuning, we adopt a standard three-fold cross-validation procedure. More precisely, the training data is randomly partitioned into three approximately equally sized subsets. For each fold, two subsets are used to train the model, and the remaining subset is used to evaluate the performance. This process is repeated three times, rotating the validation subset at each iteration. The configuration minimizing the average of the three validation errors is selected.

For the random forest-based methods RF and HRF, we perform the cross-validation procedure over the grid of hyperparameters considered in [38]. Specifically, we are searching the regularization parameter λ in the set $\{0, 0.01, 0.1\}$ and the minimum leaf size $m_{RF}(\vec{\alpha})$ in the set $\{5, 40, 100\}$, resulting in a total of nine combinations. For a subset of indices $F \subset \{1, \dots, n\}$, the validation error for the method RF with the fold F is given by

$$CV^{(RF)}(\lambda, m_{RF}(\vec{\alpha})) := \sum_{i \in F} g(U_i, \hat{a}_n^{(RF)}(X_i), \hat{\gamma}_n^{(RF)}(X_i)) \mathbb{I}\{U_i > 0\},$$

where $\hat{a}_n^{(RF)}$ and $\hat{\gamma}_n^{(RF)}$ are obtained by the procedure described in Definition 8 with the observations $\{(X_i, Y_i); i \notin F\}$. For the method HRF, the validation error for the fold F is

$$CV^{(HRF)}(\lambda, m_{RF}(\vec{\alpha})) := \sum_{i \in F} (V_i - \hat{\gamma}_n^{(HRF)}(X_i))^2 \mathbb{I}\{V_i > 0\},$$

where the estimator $\hat{\gamma}_n^{(HRF)}$ is obtained as in Definition 9 by using the subsample $\{(X_i, Y_i); i \notin F\}$.

For the neural network-based approaches, we use a cross-validation grid over combinations of network architecture and L^2 regularization. As Pasche and Engelke [55], we consider three architectures : small networks with three hidden layers of width 10, intermediate networks with two layers of width 64, and bigger networks with three layers of width 128. For each architecture, we test three values of penalty : $\lambda \in \{0, 10^{-5}, 10^{-3}\}$, leading again to nine configurations. Here, denoting by L the number of layers, the validation error

for a fold F is

$$\text{CV}^{(\text{NN})}(\lambda, L) := \sum_{i \in F} \mathcal{L}^{(\text{GPD})} \left(\hat{\zeta}(X_i | \widehat{\mathcal{W}}_n^{(\text{GPD})}), U_i \right),$$

for the NN approach and

$$\text{CV}^{(\text{HNN})}(\lambda, L) := \sum_{i \in F} \mathcal{L}^{(\text{HT})} \left(\hat{\gamma}(X_i | \widehat{\mathcal{W}}_n^{(\text{HT})}), V_i \right),$$

where the parameters $\widehat{\mathcal{W}}_n^{(\text{GPD})}$ and $\widehat{\mathcal{W}}_n^{(\text{HT})}$ of the neural network are calculated by using the subsample $\{(X_i, Y_i); i \notin F\}$.

We also use an early stopping strategy to determine the number of epochs. More specifically, recall that in the three-fold cross-validation procedure, three validation errors are computed by rotating the fold dedicated to validation. The number of epochs used to obtain each validation error is selected as follows : the data belonging to the two training folds are randomly divided into a set E_T (80% of the data) and a set E_V . The model is trained on E_T , and the validation error is evaluated on E_V . This evaluation is performed over several epochs, and the training is stopped when no improvement in the validation error is observed after a fixed number of consecutive epochs (referred to as the patience threshold). The model weights corresponding to the best validation performance are retained. In our experiments, the patience threshold is set to 20 epochs. Once the best hyperparameters are selected through cross-validation, a final training is performed on the entire dataset, again applying early stopping with the same strategy.

For dimension reduction methods, the choice of the reduced dimension q plays a central role. In the case of TIREX, q is selected by cross-validation. For the CTI approach, q is estimated using the specific procedure proposed by Gardes and Podgorny [36]. For the compact $\mathcal{X}_0$, we take $\mathcal{X}_0 := [\varepsilon_p, 1 - \varepsilon_p]^p$ with $\varepsilon_p = (1 - 0.9^p)/2$ so that approximatively 90% of observations belong to $\mathcal{X}_0$.

Finally, the choice of the classical quantile level α_n cannot be tuned via cross-validation, since it directly affects the number of training examples (i.e., the number of observed excesses). As a consequence, α_n must be chosen by the user based on prior knowledge or practical constraints. We investigate the impact of this choice in Experiment 1.

3.4.3 Experimental Results

We now present the results of four experiments designed to evaluate the sensitivity and robustness of the different methods under various conditions. Each experiment investigates a specific factor influencing estimation accuracy :

- 1) the threshold level α_n ;
- 2) the dimension p of the covariate ;
- 3) the second order parameter function ρ ;
- 4) complexity of the conditional tail index function γ .

Experiment 1 We assess the sensitivity of the estimators to the choice of the classical quantile level α_n , which determines the number of excesses used in the tail parameter estimation. We consider three different values for this level, namely $\alpha_n \in \{0.05, 0.1, 0.2\}$. The dimension of the covariate is fixed to $p = 10$ and two different conditional distributions are considered :

- **Frechet** : $q(u, \gamma(x)) = (-\log(1 - u))^{-\gamma(x)}$,
- **Absolute student** : $q(u, \gamma(x)) = t_{1/\gamma(x)}(1 - u/2)$, where t_ν denotes the quantile function of a Student distribution with ν degrees of freedom.

As in Pasche and Engelke [55], the conditional tail-index is given for $x = (x_1, \dots, x_p)^\top$ by

$$\gamma(x) = \left(\frac{7}{1 + \exp(4x_1 + 1.2)} + 3 \right)^{-1} \in [0.1, 0.3]. \quad (3.21)$$

The results are reported in Figure 3.2. For the Fréchet conditional distribution, all methods are relatively robust to the choice of α_n , with the exception of HNN, which exhibits increased variability at small values of α_n . In this setting, heavy-tailed variants, HRF and HNN, generally outperform GPD-based methods RF and NN, except HNN for $\alpha_n = 0.05$. In particular, the random forest method RF is particularly unstable.

For the Student conditional distribution, the GPD-based machine learning approaches RF and NN outperform their heavy-tailed counterparts. The dimension reduction approaches (T and CTI), are highly sensitive to the choice of α_n . For these methods to perform well, a small value of α_n seems essential.

Experiment 2 We now evaluate the impact of the covariate dimension p . We take the setup of the first experiment but now with a fixed value for the

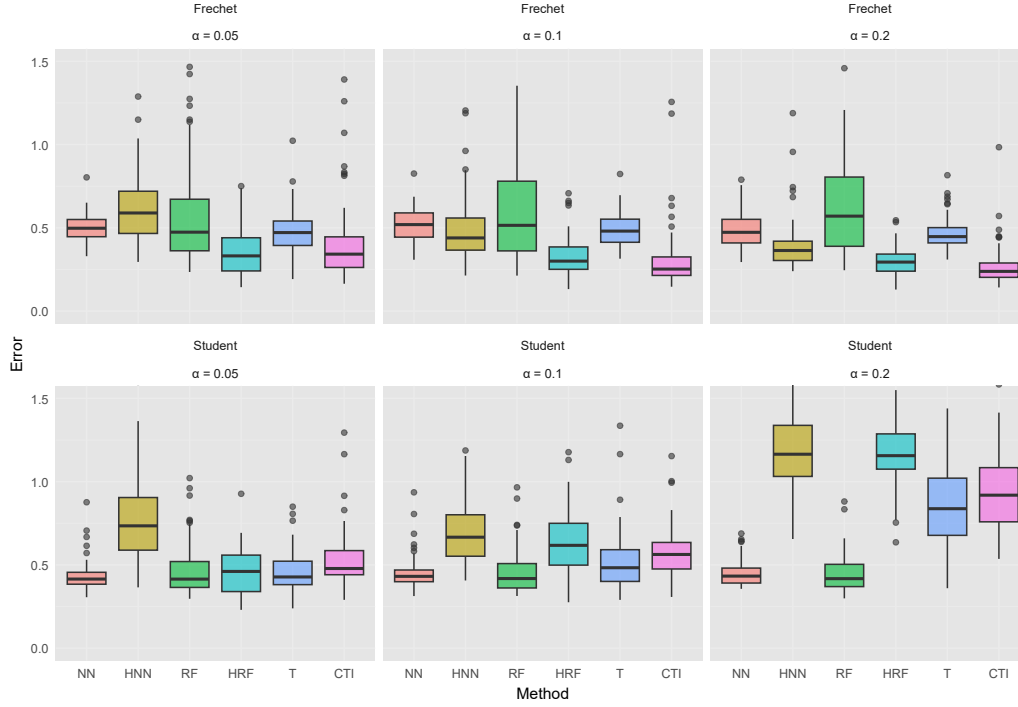


FIGURE 3.2 – **Experiment 1** – Boxplots of the symmetric relative error for the Fréchet and Absolute student distributions for different values of α_n

classical condition quantile level α_n (namely, $\alpha_n = 0.1$) and different values for $p \in \{5, 10, 30\}$. The results, depicted in Figure 3.3, indicate that, except for the HNN method, the dimension of the covariate does not strongly influence the performance of the conditional quantile estimator. The neural network approach HNN, which is designed for the heavy-tailed case, performs better than method NN when $p = 5$ but strongly deteriorates when $p = 30$. For the Fréchet conditional distribution, the best results are clearly obtained by the approaches HRF and CTI.

Experiment 3 As mentioned in the introduction, the second-order parameter function $\rho < 0$ quantifies the rate of convergence in the limit result of model (M). We examine its influence by considering, in addition to Fréchet

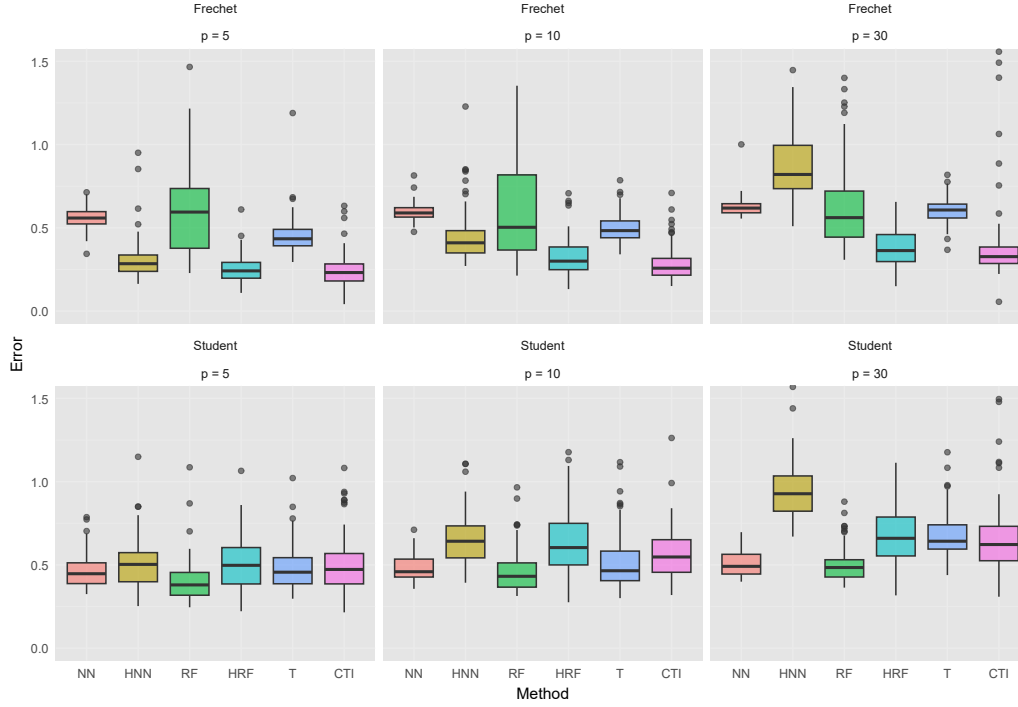


FIGURE 3.3 – **Experiment 2** – Boxplots of the symmetric relative error for the Fréchet and Absolute student distributions for different values of p .

and student distributions, the Burr conditional distribution :

$$q(u, \gamma(x)) = (u^\rho - 1)^{-\gamma(x)/\rho}, \text{ with } \rho \in \{-0.2, -0.5, -1, -\infty\}.$$

Here, the second-order parameter function is equal to ρ and does not depend on x . Note also that

$$\lim_{\rho \rightarrow -\infty} (u^\rho - 1)^{-\gamma(x)/\rho} = u^{-\gamma(x)}.$$

Hence, when $\rho = -\infty$, the Burr distribution coincides with a standard Pareto distribution. For the Fréchet distribution, $\rho = -1$ for all x , whereas for the student distribution, ρ depends on x and ranges approximately between -0.2 and $-2/3$. In this experiment, the classical quantile level is fixed to $\alpha_n = 0.1$ and the covariate dimension is $p = 10$.

The boxplots in Figure 3.4 highlight the effect of $\rho(x)$ on estimation accuracy. These results suggest that the second-order parameter plays a decisive role in determining the relative performance of the estimators. In general, when $\rho(x)$ is far from zero (i.e., fast convergence), heavy-tailed approaches tend to outperform GPD-based ones. Conversely, when $\rho(x)$ is close to zero (i.e., convergence is slower), the GPD methods yield more accurate estimates.

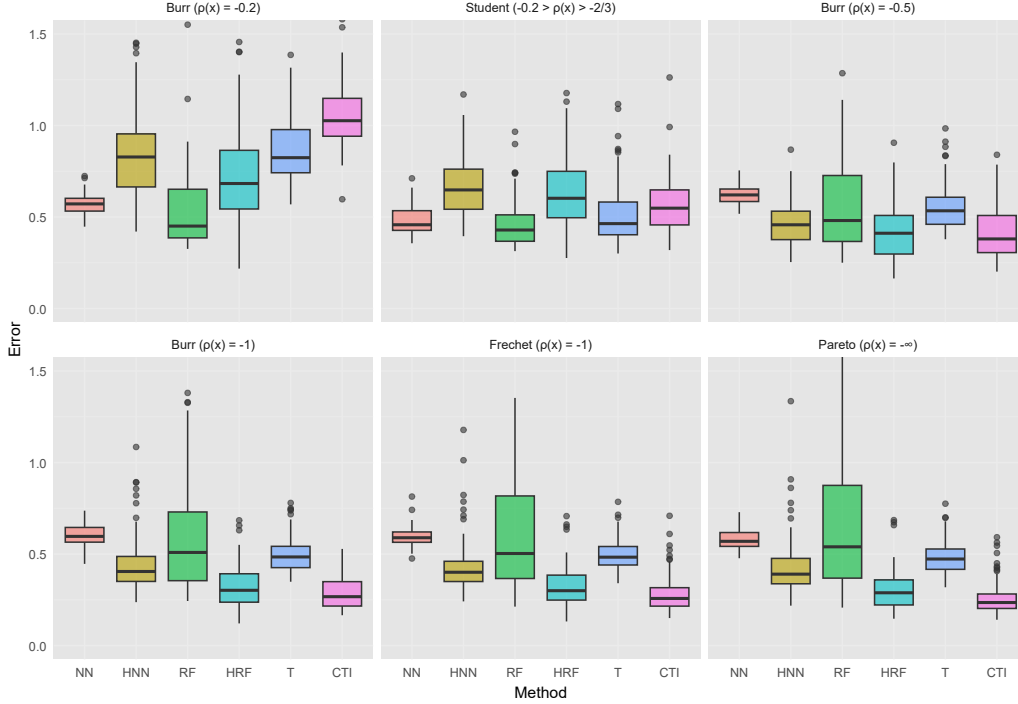


FIGURE 3.4 – **Experiment 3** – Boxplots of the symmetric relative error for the Fréchet, Absolute student and Burr distributions.

Experiment 4 In this final experiment, we investigate how the complexity of the conditional tail index γ affects the performance of the estimators. To this end, we take $\alpha_n = 0.1$, $p = 10$ and consider a Fréchet distribution with three different configurations for γ , each reflecting increasing levels of

structural complexity :

$$\begin{aligned}\gamma_1(x) &= 0.3 + 0.5 \mathbb{I}\{|x_1| > 1/2\}, \\ \gamma_2(x) &= 0.3 + 0.5 \mathbb{I}\{|x_1| + |x_2| > 1\}, \\ \gamma_3(x) &= 0.3 + 0.5 \mathbb{I}\{\|x\|_1 > p/2\}.\end{aligned}$$

The conditional tail index γ_1 only depends on the first component of the covariate. In this case, it can be shown that $\mathcal{S} = \text{span}(B_0)$ with $B_0 = (1, 0, \dots, 0)^\top$ is the CTI subspace as introduced in Definition 14. The function γ_2 depends on the first two components of the covariate so that the CTI subspace is spanned by the vectors $e_1 := (1, 0, \dots, 0)^\top$ and $e_2 := (0, 1, 0, \dots, 0)^\top$. Note that on the non linear set $\{(x, y); |x| + |y| > 1\}$, the function γ_2 is equal to 0.8 and to 0.3 on $\{(x, y); |x| + |y| \leq 1\}$. Finally, for the function γ_3 , the CTI is the set $\mathbb{R}^p$ and the function depends on the whole covariate.

The results are depicted in Figure 3.5. First remark that all methods perform poorly when the conditional tail index is γ_3 . This can be explained by the complexity of the pattern that cannot be effectively captured, even by flexible machine learning-based approaches. For the function γ_1 , the best results are obtained by the HRF and CTI methods. Finally, for γ_2 , HRF appears to benefit from its ability to detect nonlinear structures, in contrast to CTI, which relies on linear projections.

Conclusion Across the four experiments conducted in this study, we evaluated the performance of several recently proposed methods for extreme conditional quantile estimation under a variety of conditions. Our findings provide several practical insights for users seeking to apply these methods in high-dimensional settings.

First, HNN systematically underperforms and should generally be avoided. In contrast, NN exhibits moderate performance but proves robust across all scenarios, making it a reliable and safe choice when no prior information is available about the tail structure.

Among all the approaches, RF is the better choice when the second-order parameter function ρ is close to zero, corresponding to low convergence in the extreme value limit. However, it becomes highly unstable as ρ moves away from zero. On the other hand, HRF is more stable and outperforms RF in scenarios where $\rho(x) < -0.5$.

Regarding dimension reduction methods, CTI consistently dominates TIREX, which shows weaker performance in most settings. When the tail index

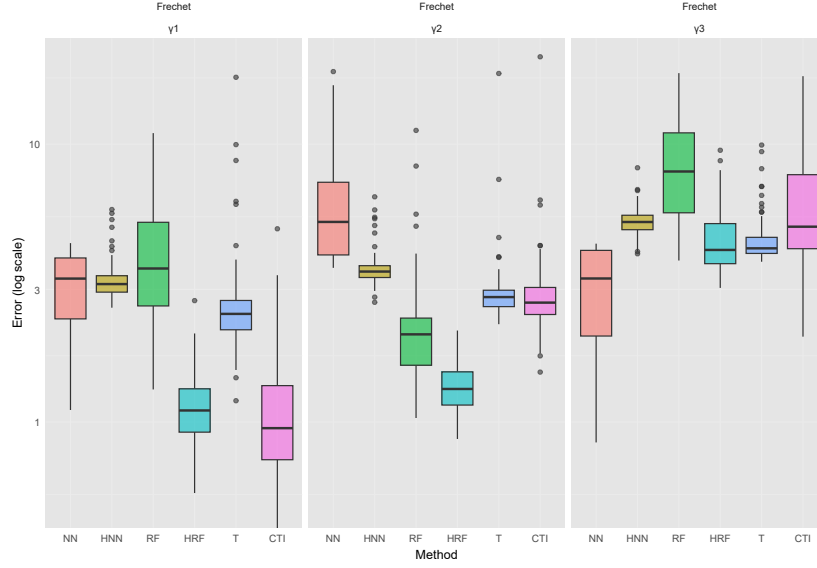


FIGURE 3.5 – **Experiment 4** – Boxplots of the symmetric relative error for the Fréchet distribution with three different tail index functions γ_1 , γ_2 , and γ_3 . The experiment illustrates the impact of structural complexity in the conditional tail index.

is defined on a lower dimensional linear subspace, it also outperforms all the tested methods as soon as $\rho(x) < -0.5$.

More broadly, our results show that GPD-based methods are generally less sensitive to the choice of the intermediate level α_n than their heavy-tailed counterparts. For heavy-tailed approaches, using a relatively small value such as $\alpha_n = 0.05$ or $\alpha_n = 0.1$ is recommended to ensure reliable estimation.

These findings offer practical guidance for selecting appropriate methods based on the underlying model characteristics and emphasize the importance of the unknown second-order function and of the structural complexity of the conditional tail index when estimating extreme quantiles.

Conclusion et perspectives

Ce manuscrit s'est attaché à étudier la problématique de l'estimation des quantiles extrêmes dans différents cadres, en s'appuyant sur les outils de la théorie des valeurs extrêmes. Ces quantités, essentielles pour modéliser le risque d'événements rares dans de nombreux domaines comme la finance, l'environnement ou l'assurance, posent des défis spécifiques en raison de la rareté des observations disponibles dans les queues de distribution.

Dans un premier temps, nous nous sommes intéressés à la construction d'intervalles de confiance pour les quantiles extrêmes dans le cas non conditionnel, en proposant une méthode valable quel que soit le domaine d'attraction de la loi sous-jacente. Cette approche, basée sur l'extrapolation à partir de statistiques d'ordre et non sur l'approximation normale, permet d'obtenir de meilleurs résultats.

Le deuxième axe de la thèse a porté sur le cadre conditionnel en grande dimension. Nous avons introduit une méthode innovante de réduction de dimension, spécifiquement conçue pour l'estimation de l'indice de queue conditionnel dans le cas de lois à queue lourde. Ce travail a permis de montrer qu'il est possible d'identifier un sous-espace de plus faible dimension conservant toute l'information pertinente pour modéliser le comportement extrême, et ainsi de contourner le fléau de la grande dimension.

Enfin, une étude comparative a permis d'évaluer la performance de différentes approches récentes, issues du machine learning ou de la réduction dimension. Cette analyse approfondie a mis en évidence la robustesse globale de ces méthodes à la dimension, tout en soulignant la sensibilité potentielle au choix du niveau intermédiaire α_n et au paramètre de second ordre $\rho(x)$. Elle a également montré que, si les méthodes de réduction de dimension sont plus

efficaces en présence de structures linéaires, les approches d'apprentissage automatique peuvent s'avérer plus performantes dans des cas non linéaires simples, bien qu'aucune méthode ne soit pleinement satisfaisante en cas de structure trop complexe.

Plusieurs perspectives de recherche émergent naturellement de ces travaux. Tout d'abord, la méthode proposée au chapitre 1 pour la construction d'intervalles de confiance dans le cadre non conditionnel pourrait être étendue de manière relativement naturelle au contexte conditionnel. Pour estimer un intervalle de confiance autour du quantile extrême conditionnel $Q(\beta_n|x)$, il serait envisageable de s'appuyer sur des statistiques d'ordre calculées à partir du sous-échantillon donné par des observations dont la covariable X_i est proche de x , par exemple contenues dans une boule de rayon h centrée en x . L'extrapolation serait alors effectuée de manière locale, en utilisant des estimateurs de l'indice de queue $\gamma(x)$ et de la fonction de normalisation $a(\alpha_n^{-1}|x)$, obtenus à partir des méthodes présentées au chapitre 3. Cette approche permettrait de construire des intervalles de confiance asymptotiques pour les quantiles extrêmes conditionnels.

Ensuite, la méthode de réduction de dimension linéaire CTI développée au chapitre 2 pourrait potentiellement être étendue à des structures non linéaires. Une idée serait d'enrichir le cadre en cherchant non plus un sous-espace linéaire mais une sous-variété non linéaire $\mathcal{S} \subset \mathbb{R}^p$ minimisant une quantité analogue à la fonctionnelle $\Psi(B, \mathcal{X}_0)$ définie dans le chapitre. Cette généralisation permettrait de capturer des dépendances plus complexes entre les covariables et le comportement extrême de la variable cible.

Par ailleurs, il serait pertinent d'explorer des combinaisons hybrides entre méthodes de réduction de dimension et approches d'apprentissage présentées dans le chapitre 3. Par exemple, les covariables pourraient être projetées sur un sous-espace appris par la méthode CTI du chapitre 2, avant d'être utilisées comme entrées d'un modèle de régression quantile basé sur une architecture neuronale ou sur des forêts aléatoires.

Enfin, la question du choix du niveau intermédiaire α_n demeure un enjeu important. Ce paramètre, souvent fixé de manière empirique, peut influencer fortement la qualité de l'estimation des quantiles extrêmes. Le développement de procédures automatiques et théoriquement fondées pour sélectionner α_n de façon optimale représente un axe de recherche important, en particulier dans le cas conditionnel.

Bibliographie

- [1] Aghbalou, A., Portier, F., Sabourin, A. and Zhou, C. (2024). Tail Inverse Regression : dimension reduction for prediction of extremes, *Bernoulli*, **30(1)**, 503–533.
- [2] Aggarwal, C. C. (2018). *Neural networks and deep learning : a textbook*. Springer.
- [3] Allouche, M., Girard, S., & Gobet, E. (2022). EV-GAN : Simulation of extreme events with ReLU neural networks, *Journal of Machine Learning Research*, **23(150)**, 1–39.
- [4] Allouche, M., Girard, S., & Gobet, E. (2024). Estimation of extreme quantiles from heavy-tailed distributions with neural networks, *Statistics and Computing*, **34(1)**, 12.
- [5] Alvarez, J. A., Villavicencio, F., Strozza, C. and Camarda, C.G. (2021). Regularities in human mortality after age 105, *PloS ONE*, **16(7)**, e0253940.
- [6] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests, *The Annals of Statistics*, **47(2)**, 1148–1178.
- [7] Aydin, D. (2007). A comparison of the nonparametric regression models using smoothing spline and kernel regression, *International Journal of Engineering, Mathematical and Physical Sciences*, **1(12)**, 575–579.
- [8] Bellman, R.E. (1961). *Adaptative Control Process : A Guided Tour*, Princeton University Press.
- [9] Berlinet, A., Gannoun, A., & Atzner-Løber, E. (2001). Asymptotic normality of convergent estimates of conditional quantiles, *Statistics*, **35(2)**, 139–169.

- [10] Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004). *Statistics of Extremes : Theory and Applications*, Wiley Series in Probability and Statistics.
- [11] Bhattacharya, P. K., & Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile, *The Annals of Statistics*, **18**(3)1400–1415.
- [12] Bousebata, M., Enjolras, G., and Girard, S. (2023). Extreme Partial Least-Squares regression, *Journal of Multivariate Analysis*, **194**, 105101.
- [13] Bingham, N.H., Goldie C.M. and Teugels, J.L. (1987). *Regular Variation*, Cambridge University Press.
- [14] Brazauskas, V., Jones, B.L., Puri, M.L. and Zitikis, R. (2008). Estimating conditional tail expectation with actuarial applications in view, *Journal of Statistical Planning and Inference*, **138**, 3590–3604.
- [15] Breiman, L. (2001). Random forests, *Machine Learning*, **45**(1), 5–32.
- [16] Buitendag, S., Beirlant, J. and de Wet, T. (2020). Confidence intervals for extreme Pareto-type quantiles, *Scandinavian Journal of Statistics*, **47**(1), 36–55.
- [17] Bura, E. and Forzani, L. (2015). Sufficient reductions in regressions with elliptically contoured inverse predictors, *Journal of the American Statistical Association*, **110**(509), 420–434.
- [18] Carreau, J., & Girard, S. (2011). Spatial extreme quantile estimation using a weighted log-likelihood approach, *Journal de la Société Française de Statistique*, **152**(3), 66–82.
- [19] Chen, X, Zhang, J. and Zhou, W. (2022). High-dimensional elliptical sliced inverse regression in non-gaussian distributions. *Journal of Business & Economic Statistics*, **40**(3), 1204–1215.
- [20] Cook, R.D. (1994). On the interpretation of regression plots, *Journal of the American Statistical Association*, **89**, 177–189.
- [21] Cook, R.D. and Li, B. (2002). Dimension reduction for conditional mean in regression, *Annals of Statistics*, **30**(2), 455–474.

- [22] Cook, R. D., & Weisberg, S. (1991). Sliced inverse regression for dimension reduction : Comment, *Journal of the American Statistical Association*, **86(414)**, 328–332.
- [23] Cover, T. (1968). Estimation by the nearest neighbor rule, *IEEE Transactions on Information Theory*, **14**, 50–55.
- [24] Cox, D.R., Snell, E.J. (1989). *Analysis of Binary Data*, Number 32 in Monographs on Statistics and Applied Probability, 2nd edn. Routledge, Boca Raton.
- [25] Daouia, A., Gardes, L. and Girard, S. (2013). On kernel smoothing for extremal quantile regression, *Bernoulli*, **19(5B)**, 2557–2589.
- [26] Dekkers, A.L.M. and de Haan, L (1989). On the estimation of the extreme-value index and large quantile estimation, *Annals of Statistics*, **17(4)**, 1795–1832.
- [27] Dekkers, A.L.M., Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics*, **17(4)**, 1833–1855.
- [28] El Methni, J., Gardes, L., Girard, S. and Guillou, A. (2012). Estimation of extreme quantiles from heavy and light tailed distributions, *Journal of Statistical Planning and Inference*, **142(10)**, 2735–2747.
- [29] Embrechts, P., Klüppelberg, C. and Mikosch, T. (2013). *Modelling extremal events for insurance and finance*, Springer Science & Business Media.
- [30] Fan, Y. and Liu, R. (2016). A direct approach to inference in nonparametric and semiparametric quantile models, *Journal of Econometrics*, **191(1)**, 196–216.
- [31] Ferreira, A. and Huang, F. (2018). Is human life limited or unlimited? (A discussion of the paper by Holger Rootzén and Dmitrii Zholud), *Extremes*, **21**, 373–382.
- [32] Gardes, L. (2015). A general estimator for the extreme value index : applications to conditional and heteroscedastic extremes, *Extremes*, **18(3)**, 479–510.

- [33] Gardes, L. (2018). Tail Dimension Reduction for extreme quantile estimation, *Extremes*, **21**(1), 57–95.
- [34] Gardes, L. (2020). Nonparametric confidence interval for conditional quantiles with large-dimensional covariates, *Electronic Journal of Statistics*, **14**(1), 661–701.
- [35] Gardes, L. and Maistre, S. (2023). Nonparametric asymptotic confidence intervals for extreme quantiles, *Scandinavian Journal of Statistics*, **50**(2), 825–841.
- [36] Gardes, L., and Podgorny, A. (2025). Dimension reduction for the estimation of the conditional tail-index, *Scandinavian Journal of Statistics*, to appear.
- [37] Gardes, L., & Stupfler, G. (2014). Estimation of the conditional tail index using a smoothed local Hill estimator, *Extremes*, **17**(1), 45–75.
- [38] Gnecco, N., Terefe, E. M., & Engelke, S. (2024). Extremal random forests, *Journal of the American Statistical Association*, **119**(548), 3059–3072.
- [39] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory : An introduction*, Springer Series in Operations Research and Financial Engineering, Springer.
- [40] de Haan, L., Mercadier, C. and Zhou, C. (2016). Adapting extreme value statistics to financial time series : dealing with bias and serial dependence, *Finance and Stochastics*, **20**, 321–354.
- [41] de Haan, L. and Rootzèn, H. (1993). On the estimation of high quantiles, *Journal of Statistical Planning and Inference*, **35**(1), 1–13.
- [42] Hall, P. and Welsh, A.H. (1985). Adaptative estimates of parameters of regular variation, *The Annals of Statistics*, **13**, 331–341.
- [43] Halton, J. H. (1964). Algorithm 247 : Radical-inverse quasi-random point sequence, *Communications of the ACM*, **7**(12), 701–702.
- [44] Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation, *The Annals of Statistics*, **13**(4), 1465–1481.

- [45] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.
- [46] Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985). Estimation of the generalized extreme value distribution by the method of probability-weighted moments, *Technometrics*, **27(3)**, 251–261.
- [47] Hüsler, J., Li, D. and Raschke, M. (2016). Extreme value index estimator using maximum likelihood and moment estimation, *Communications in Statistics-Theory and Methods*, **45(12)**, 3625–3636.
- [48] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics*, **58**, 71–120.
- [49] Kocherginsky, M., He, X. and Mu, Y. (2005). Practical confidence intervals for regression quantiles, *Journal of Computational and Graphical Statistics*, **14(1)**, 41–55.
- [50] Koenker, R. & Bassett Jr, G. (1978). Regression quantiles, *Econometrica*, **46(1)**, 33–50.
- [51] Li, K-C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86(414)**, 316–327.
- [52] Li, Q., & Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data, *Journal of Business and Economic Statistics*, **26(4)**, 423–434.
- [53] Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and its Application*, **9(1)**, 141–142.
- [54] Parzen, M.I., Wei, L.J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions, *Biometrika*, **81(2)**, 341–350.
- [55] Pasche, O. C., & Engelke, S. (2024). Neural networks for extreme quantile regression with an application to forecasting of flood risk, *The Annals of Applied Statistics*, **18(4)**, 2818–2839.
- [56] Pickands III, J. (1975). Statistical inference using extreme order statistics, *The Annals of Statistics*, **3(1)**, 119–131.

- [57] Pisarenko, V. F. and Rodkin, M. V. (2014). *Statistical analysis of natural disasters and related losses*, Springer.
- [58] Read, L.K. and Vogel, R.M. (2015). Reliability, return periods, and risk under nonstationarity, *Water Resources Research*, **51(8)**, 6381–6398.
- [59] Rockafellar, R.T. and Uryasev, S. (2002). Conditional Value-at-Risk for general loss distributions, *Journal of Banking and Finance*, **26**, 1443–1471.
- [60] Rootzén, H. and Zholud, D. (2017). Human life is unlimited – but short, *Extremes*, **20**, 713–728.
- [61] Samanta, M. (1989). Non-parametric estimation of conditional quantiles, *Statistics and Probability Letters*, **7(5)**, 407–412.
- [62] Scrucca, L. (2011). Model-based SIR for dimension reduction, *Computational Statistics and Data Analysis*, **55**, 3010–3026.
- [63] Shami, T.M., El-Saleh, A.A., Alswaitti, M., Al-Tashi, Q. and Mirjalili, S. (2022). Particle Swarm Optimization : A Comprehensive Survey, *IEEE Access*, **10**, 10031–10061.
- [64] Silverman, B.W. (1984). Spline smoothing : the equivalent variable kernel method, *The Annals of Statistics*, **12(3)**, 898–916.
- [65] Specht, D.F. (1991). A general regression neural network, *IEEE Transactions on Neural Networks*, **2(6)**, 568–576.
- [66] Thompson, W.R. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form, *Annals of Mathematical Statistics*, **7**, 122–128.
- [67] Tukey, J. W. (1975). Mathematics and the picturing of data, *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, **5(2)**, 523–531.
- [68] de Valk, C. (2016). Approximation of high quantiles from intermediate quantiles, *Extremes*, **19**, 661–686.
- [69] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer New York.

- [70] Velthoen, J., Dombry, C., Cai, J. J., & Engelke, S. (2023). Gradient boosting for extreme quantile regression, *Extremes*, **26**(4), 639–667.
- [71] Wang, D., Tan, D. and Liu, L. (2018). Particle swarm optimization algorithm : an overview, *Soft computing*, **22**, 387–408.
- [72] Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction, *Journal of the American Statistical Association*, **103**(482), 811–821.
- [73] Watson, G. S. (1964). Smooth regression analysis, *Sankhyā*, **26**(15), 175–184.
- [74] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *Journal of the American Statistical Association*, **73**, 812–815.
- [75] Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression, *Statistica Sinica*, **19**, 801–817.
- [76] Xia, Y. (2009). Model checking in regression via dimension reduction, *Biometrika*, **96**(1), 133–148.
- [77] Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **64**(3), 363–410.
- [78] Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for US wind gusts, *Journal of the American Statistical Association*, **114**(528), 1865–1879.
- [79] Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression, *Statistics and Probability Letters*, **54**(4), 437–447.