Théorie des valeurs extrêmes et applications en environnement

Théo Rietsch

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE

UMR 7501

Strasbourg

Thèse

présentée pour obtenir le grade de docteur de l'Université de Strasbourg
Spécialité MATHÉMATIQUES APPLIQUÉES

Théo Rietsch

Théorie des valeurs Extrêmes et applications en environnement

Soutenue le 14 Novembre 2013
devant la commission d'examen

Armelle Guillou, directeur de thèse
Philippe Naveau, co-directeur de thèse
Valérie Chavez-Demoulin, rapporteur
Anne-Catherine Fabre, rapporteur
Laurent Gardes, examinateur

www-irma.u-strasbg.fr

Remerciements

Je remercie l'ADEME pour avoir financé cette thèse. Je remercie les membres du jury : Laurent Gardes pour avoir accepté de faire partie du jury ; Valérie Chavez-Demoulin et Anne Catherine Fabre pour avoir pris le temps de lire ma thèse et de se déplacer. Je remercie également mes directeurs de thèse : Philippe Naveau pour les nombreuses discussions intéressantes que nous avons eues au labo et pour m'avoir proposé cette thèse ; Armelle Guillou pour son aide et son soutien constants. Sans elle je n'en serais pas là aujourd'hui. Merci à Yuri Goegebeur et Nicolas Gilardi pour leur collaboration sur deux de mes articles. Merci à Anne de m'avoir supporté.

Merci à ma famille, à Marie et Émilie. Sans vous je ne serais peut être pas allé au bout.

Merci aux 5 qui se reconnaîtront d'avoir été là sur la fin.

Les autres savent.

"Even things that are true can be proved." - O.W.

Table des matières

Publications et Conférences

Article accepté :

[1] A non-parametric entropy based approach to detect changes in climate extremes, *Journal of the Royal Statistical Society : Series B, Statistical Methodology*, (avec A. Guillou et P. Naveau, Chapitre 1).

Articles soumis :

[1] Network design for heavy rainfall analysis, (avec N. Gilardi, A. Guillou et P. Naveau, Chapitre 2).

[2] Robust conditional Weibull-type estimation, (avec Y. Goegebeur et A. Guillou, Chapitre 3).

Conférences :

[1] An entropy-based approach to detect changes in extreme time series, Workshop on Environmental Risk and Extreme Events, Ascona, 10-15 Juillet 2011.

[2] Query by Committee and Extreme Values, ERCIM, Oviedo, 1-3 Décembre 2012.

Séminaires :

[1] Plans d'expériences appliqués à la prévision des extrêmes climatiques régionaux, séminaire de l'équipe statistique, Strasbourg, Juin 2012.

[2] Optimization of a monitoring network in an extreme value context, séminaire du groupe de travail ESTIMR du LSCE, Gif-sur-Yvette, Mai 2013.

[3] Optimization of a monitoring network in an extreme value context, Météo France, Toulouse, Mai 2013.

Chapitre 1

Introduction

1.1 Différentes approches des extrêmes

La théorie des valeurs extrêmes est un sujet longuement étudié dans la littérature ces dernières années, tout particulièrement dans le cadre univarié. Dans ce contexte, deux approches sont utilisées pour caractériser l'appartenance à un domaine d'attraction. La première repose sur la loi limite du maximum d'un échantillon de variables aléatoires indépendantes et identiquement distribuées $X_1, ..., X_n$ de loi F. Plus spécifiquement, on dit que F est dans le domaine d'attraction D_γ si et seulement si il existe deux suites normalisantes (a_n), $a_n \in \mathbb{R}^+$ et (b_n), $b_n \in \mathbb{R}$ telles que $\forall x \in \mathbb{R}$:

$$\lim_{n\to\infty} \mathbb{P}\left(a_n^{-1}\left\{\max_{1\leq i\leq n} X_i - b_n\right\} \leq x\right) = H_\gamma(x) \tag{1.1}$$

où

$$H_\gamma(x) = \begin{cases} \exp\Big(-(1+\gamma x)^{-\frac{1}{\gamma}}\Big) & \text{pour tout } x \text{ tel que } 1+\gamma x > 0, \text{ si } \gamma \neq 0, \\ \exp\Big(-\exp(-x)\Big) & \text{pour tout } x \in \mathbb{R}, \text{ si } \gamma = 0, \end{cases}$$

s'appelle la distribution des valeurs extrêmes généralisées (distribution GEV).

De ce résultat asymptotique, il découle immédiatement que le comportement de la queue

de distribution d'une fonction est complètement caractérisé par un unique paramètre, noté γ ou ξ suivant le domaine d'applications, et appelé indice des valeurs extrêmes. Le signe de ce paramètre est un indicateur essentiel sur la forme de la queue de distribution.

Plus précisément, si on désigne par $\tau_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ le point terminal de F et si on définit une fonction à variations lentes ℓ comme une fonction positive satisfaisant

$$\lim_{x\to\infty} \frac{\ell(\lambda x)}{\ell(x)} = 1 \quad \text{pour } \lambda > 0, \tag{1.2}$$

on peut alors différencier trois domaines d'attraction :

• Le domaine de Fréchet correspondant à $\gamma > 0$, où les distributions sont de type Pareto, i.e. elles admettent une décroissance de type polynomiale au niveau de la queue : $1 - F(x) = x^{-1/\gamma}\ell_F(x)$ où ℓ_F est une fonction à variations lentes. On a alors $\tau_F = \infty$;

• Le domaine de Gumbel correspondant à $\gamma = 0$ où la décroissance est cette fois-ci de type exponentielle. Le point terminal τ_F peut alors être fini ou infini ;

• Le domaine de Weibull correspondant à $\gamma < 0$. On a alors $\tau_F < \infty$ et $1 - F(x) = (\tau_F - x)^{-1/\gamma}\ell_F((\tau_F - x)^{-1})$ où ℓ_F est également une fonction à variations lentes.

Cette approche basée sur la distribution GEV est appropriée si les données consistent en des maxima. Ceci dit, suivant le domaine d'applications, cela n'est pas toujours le cas et de plus il peut paraître très subjectif de réduire le jeu de données entier à une seule observation, la plus grande. Cette approche a donc été beaucoup controversée dans la littérature et une alternative a donc été proposée, faisant suite aux travaux de Balkema et de Haan (1974), ainsi qu'à ceux de Pickands (1975). Elle repose sur la loi des excès au-delà d'un seuil fixe u et est communément appelée l'approche « pics au-delà d'un seuil » (approche POT, « Peaks-Over-Threshold »). L'idée est la suivante : partant d'un échantillon $X_1, ..., X_n$, on se fixe un seuil u grand. On ne considère que les N_u observations dépassant ce seuil. On note Y_i, $i = 1, .., N_u$, les excès au-delà du seuil u, définis comme

l'écart entre l'observation et u. La fonction de répartition des excès au-delà de u est alors donnée par

$$F_u(y) = \mathbb{P}(Y \leq y | X > u) = \mathbb{P}(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}.$$

Théorème 1.1 *(Pickands, 1975) F appartient au domaine d'attraction D_γ si et seulement si il existe une fonction $\sigma(.)$ positive et un réel γ tels que la loi des excès F_u peut être uniformément approchée par une distribution de Pareto généralisée (GPD) notée $G_{\sigma,\gamma}$, i.e.*

$$\lim_{u \uparrow \tau_F} \sup_{x \in (0, \tau_F - u)} \left| F_u(x) - G_{\sigma(u),\gamma}(x) \right| = 0,$$

$$\text{où} \quad G_{\sigma,\gamma}(x) = \begin{cases} 1 - (1 + \frac{\gamma x}{\sigma})^{-\frac{1}{\gamma}} & si\ \gamma \neq 0,\ x \geq 0\ et\ x < -\frac{\sigma}{\gamma}\ si\ \gamma < 0 \\ \\ 1 - \exp(-\frac{x}{\sigma}) & si\ \gamma = 0,\ x \geq 0. \end{cases}$$

L'utilisation de l'une ou l'autre de ces approches requiert l'estimation de paramètres, γ pour la méthode GEV ou (γ, σ) pour l'approche POT. Pour cela différentes techniques d'estimation ont été proposées dans la littérature, parmi elles on peut citer les méthodes des moments, des moments pondérés (Greenwood *et al.*, 1979 ; Hosking et Wallis, 1987), des moments pondérés généralisés (Diebolt *et al.*, 2007), du maximum de vraisemblance (Smith, 1984)... Ces estimateurs sont en général basés sur les k plus grandes valeurs de l'échantillon et leurs consistances sont établies sous des hypothèses convenables sur ce paramètre. Par contre, pour établir leur normalité asymptotique, une hypothèse supplémentaire est nécessaire sur la fonction à variations lentes. Cette hypothèse est communément appelée hypothèse du second ordre et permet de spécifier la vitesse de convergence de (1.2) vers 1 :

Hypothèse $(\mathcal{R})$. *Il existe une constante $\rho < 0$ et une fonction $b(.)$ satisfaisant $b(x) \to 0$*

quand $x \to \infty$*, telles que pour tout* $\lambda \geq 1$*, on ait*

$$\ln\left(\frac{\ell(\lambda x)}{\ell(x)}\right) \sim b(x) \int_1^\lambda t^{\rho-1} dt$$

quand $x \to \infty$*.*

D'après Geluk et de Haan (1987), $(\mathcal{R})$ implique que $|b(.)|$ est à variations régulières d'indice ρ, i.e. est telle que $|b(\lambda x)|/|b(x)| \to \lambda^\rho$ quand $x \to \infty$ pour tout $\lambda > 0$. Par conséquent le paramètre ρ contrôle la vitesse de convergence du rapport $\ell(\lambda x)/\ell(x)$ vers 1. Si $|\rho|$ est petit, alors la convergence est lente et l'estimation des paramètres de queue est dans ce cas difficile (problème de biais).

1.2 Information de Kullback-Leibler

Dans un contexte de réchauffement climatique, une question naturelle qui intéresse à la fois les climatologues, les assureurs ou les modélisateurs de risque est de savoir si les températures élevées sur les dernières années, disons les 30 dernières années (durée habituelle en climatologie), diffèrent de façon significative de celles mesurées au cours des périodes antérieures. Plus particulièrement on cherchera dans le Chapitre 2 à déterminer si oui ou non il y a un changement dans la distribution des extrêmes. Pour cela, on ne fera pas d'hypothèses trop fortes sur la distribution, en particulier on n'imposera pas une densité paramétrique spécifique, mais on privilégiera plutôt une approche non-paramétrique qui puisse être utilisée pour de grands jeux de données.

Une approche classique en climat consiste à construire une série d'indicateurs dits indicateurs extrêmes météorologiques et à étudier leurs variabilités temporelles en termes de fréquence et d'intensité (Alexander *et al.*, 2006 ; Frich *et al.*, 2002). Le problème avec de tels indices est qu'ils se concentrent souvent sur les extrêmes dits « modérés » (inférieurs à 90%) mais pas sur les extrêmes dits « élevés » (au-delà de 95%). De plus leurs propriétés statistiques ont rarement été établies. Une orientation plus statistique consiste à

analyser les extrêmes « élevés » en ayant recours à la théorie des valeurs extrêmes. En particulier si l'on est prêt à supposer que les maxima de température pour un bloc donné (jours, mois, saisons, ...) sont approximativement de loi GEV $H_{\mu,\sigma,\xi}(.) = H_{\xi}(\frac{\cdot-\mu}{\sigma})$, où μ et σ sont respectivement des paramètres de position et d'échelle (cf. Fowler et Kilsby, 2003 ; Kharin *et al.*, 2007), alors il est possible d'étudier les changements dans les paramètres de la GEV. Cette méthode est attractive car elle tire pleinement profit de la théorie des valeurs extrêmes. Par exemple, Jarušková et Rencová (2008) ont étudié des tests statistiques pour détecter des changements dans des séries de maxima et minima annuels de températures mesurées sur 5 sites météorologiques : Bruxelles, Cadiz, Milan, St. Pétersbourg et Stockholm. Cette méthode souffre de trois limitations : c'est une approche basée sur le maximum, donc cela écarte toutes les données sauf la plus grande du bloc. Elle impose une forme GEV fixe ce qui est restrictif pour les petits blocs (la GEV étant une distribution limite). Trois paramètres doivent être étudiés pour détecter des changements dans une série chronologique. En ce qui concerne la première limitation une solution classique pour s'en affranchir est de travailler avec les excès au-delà d'un seuil élevé à la place d'un bloc de maxima. Cependant les deux autres limitations restent aussi d'actualité pour le modèle GPD (cf. Section 1.1). Dans le Chapitre 2, on n'imposera pas une forme paramétrique pour la densité et de ce fait, il sera impossible de suivre les changements dans les paramètres (e.g., Grigga et Tawn, 2012). Une autre stratégie doit donc être suivie pour comparer les différences de répartition entre les extrêmes. En théorie de l'information (Burnham et Anderson, 1998) il est coutumier de comparer les densités de probabilité de deux périodes en calculant l'entropie

$$I(f;g) = \mathbb{E}_f \left\{ \ln \left(\frac{f(\mathbf{X})}{g(\mathbf{X})} \right) \right\},$$

où $\mathbf{X}$ représente un vecteur aléatoire de densité f et g une autre densité. Bien qu'elle soit utilisée pour mesurer un écart, l'entropie n'est pas une distance au sens topologique du terme. En effet, l'inégalité triangulaire n'est pas vérifiée ainsi que la propriété de symétrie.

Ceci explique de fait l'importance du choix entre les fonctions f et g. Des modifications ont été proposées pour symétriser cette information, par exemple en utilisant

$$D(f;g) = I(f;g) + I(g;f).$$

On parle alors de divergence de Kullback-Leibler. Travailler avec cette quantité présente un certain nombre d'avantages. C'est une notion utilisée par de nombreuses communautés : physiciens, climatologues, statisticiens, informaticiens, ... qui résume l'information en une seule valeur. Par ailleurs, dans certains cas, elle peut être explicitement calculée, comme par exemple pour des distributions gaussiennes multivariées (cf. Penny et Roberts, 2000).

Dans le Chapitre 2, nous proposerons dans un premier temps une approximation de la divergence de façon à l'exprimer en fonction des queues de probabilité et non des densités. Nous construirons ensuite un estimateur de cette divergence approximée et nous étudierons ses propriétés asymptotiques. Son comportement à distance finie sera exhibé sur simulations et sur des données réelles climatiques concernant 24 stations météorologiques en Europe.

1.3 Réseaux de neurones et Query By Committee

Dans le contexte de changement climatique dans lequel nous nous trouvons et étant donné les différentes catastrophes naturelles survenues récemment, il est devenu crucial de bien comprendre comment se comportent les évènements extrêmes tels que les vagues de chaleur ou les fortes pluies. Une bonne connaissance de l'intensité de tels évènements peut permettre aux autorités de prendre des décisions préventives efficaces comme la construction de digues ou d'infrastructures adaptées dans les zones à risque. La principale information que nous possédons à propos de ces évènements provient des stations météorologiques si-

tuées sur le territoire. Celles-ci nous fournissent des observations en continu, par exemple sur la température ou les précipitations. Ces observations nous permettent d'estimer le comportement extrémal des quantités d'intérêt correspondantes. Malheureusement, même si le réseau est grand, ceci ne constitue qu'une information ponctuelle. En particulier, il n'y a pas nécessairement de stations aux endroits où l'on a besoin d'information et de ce fait il est nécessaire d'extrapoler le comportement des extrêmes climatiques à partir des données fournies par les stations du réseau. Une approche assez classique pour traiter ce problème est appelée l'analyse fréquentielle régionale. Elle consiste à supposer que la distribution des extrêmes suit une loi paramétrique (en l'occurence une GPD, cf Section 1.1) en chaque point du territoire, puis à estimer les paramètres de cette loi en chaque station avant de les interpoler sur tout le territoire. La première partie est basée sur la méthode POT, rappelée en Section 1.1. En ce qui concerne l'interpolation des paramètres sur tout le territoire, nous avons décidé d'utiliser une approche basée sur les réseaux de neurones pour la réaliser. En tirant profit des propriétés de ces réseaux, nous allons chercher à répondre à la question suivante : où faut-il ajouter (resp. retirer) des stations pour gagner (resp. perdre) le plus (resp. le moins) d'information sur le comportement des extrêmes ? Cette question est particulièrement importante car le contexte économique peut contraindre les autorités à devoir fermer des stations ou parfois leur permettre d'en ouvrir de nouvelles. Dans ce cas, la stratégie la plus naïve consiste à sélectionner les stations aléatoirement, ou à spéculer sur les endroits où elles seraient le plus utile. Ceci dit, la construction d'une nouvelle station est un processus coûteux, et le choix de l'endroit où l'on construit ou supprime une station ne peut être pris à la légère et doit être fait de la façon la plus réfléchie possible afin de disposer du maximum d'information sur le comportement des extrêmes. Pour cela un plan d'expériences doit être mis en oeuvre. Ce type d'approche a été initialement développé dans le cadre de modèles linéaires et étendu aux modèles non linéaires que nous utiliserons ici.

Nous avons choisi d'aborder cette question à l'aide du Query By Committee (QBC), une

approche basée sur les réseaux de neurones. Il s'agit d'un processus qui vise à ajouter progressivement des stations jusqu'à ce qu'un critère d'arrêt soit satisfait. Il est particulièrement utile lorsque les données sont coûteuses à obtenir, ce qui est typiquement le cas dans notre contexte.

Nous allons commencer par rappeler succinctement le principe des réseaux de neurones avant de présenter l'algorithme du QBC et la façon dont nous allons l'utiliser pour sélectionner les stations qui devraient être ajoutées/retirées du réseau.

1.3.1 Réseaux de neurones

Soit $\mathcal{Z} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ un échantillon où $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,q}) \in \mathcal{X} \subset \mathbb{R}^q$, l'ensemble des sites où les mesures sont faites, et y_i est l'observation de la quantité y au site i. On cherche à modéliser y par un réseau de neurone. Pour cela, on va interpoler sur tout le territoire le modèle suivant

$$y_i = g(\mathbf{x}_i, \boldsymbol{\beta}), \tag{1.3}$$

pour chaque site, $i = 1, ..., n$, où g sera une fonction non linéaire (ici un réseau de neurones) paramétrée par un vecteur $\boldsymbol{\beta}$ pour lequel un choix optimal sera proposé.

Nous allons commencer par définir en détail la notion de neurone, qui constitue l'unité élémentaire d'un réseau de neurones. Un neurone est simplement défini comme une fonction non linéaire, paramétrée et à valeurs bornées (dans certains cas, la construction d'un réseau nécessite l'emploi de fonctions linéaires, auquel cas nous parlerons de neurone linéaire). La sortie d'un neurone est une fonction non linéaire - appelée fonction d'activation - appliquée à une combinaison linéaire des variables d'entrée $\mathbf{x}$ plus une constante appelée biais. Un choix classique pour la fonction d'activation est la fonction tangente hyperbolique, que nous utiliserons dans la suite. La sortie des neurones que nous utilisons s'écrit

donc :

$$\begin{aligned} y_i &= \tanh\left[\beta_0 + \sum_{j=1}^{q} \beta_j x_{i,j}\right] \\ &= \tanh\left(\boldsymbol{\beta}' \cdot \widetilde{\mathbf{x}}_i\right), \end{aligned}$$

où $\widetilde{\mathbf{x}}_i = (1, x_{i,1}, \ldots, x_{i,q})' = (1, \mathbf{x}_i)$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)'$ et $\cdot$ désigne le produit scalaire dans $\mathbb{R}^{q+1}$. Notons que la dimension initiale q est augmentée ici d'une unité, en raison du terme constant de biais (β_0) qui est systématiquement rajouté.

Un réseau de neurone est tout simplement défini comme l'association de neurones en plusieurs couches où les entrées des neurones d'une couche correspondent à une combinaison linéaire des sorties des neurones de la couche précédente. La Figure 1.1 représente de manière schématique un réseau de neurones à une couche, mais il est possible d'en rajouter pour augmenter la complexité du réseau. La complexité d'un réseau correspond au nombre de paramètres intervenant dans le modèle qui sont directement liés au nombre de couches de neurones dits "cachés" ainsi qu'au nombre de neurones présents dans chaque couche. Ces deux nombres définissent ce qu'on appellera l'architecture du réseau. Les réseaux que nous allons utiliser sont des réseaux à une couche de N_c neurones cachés et à un neurone de sortie linéaire. L'expression développée de la sortie d'un tel réseau est

$$\begin{aligned} g(\widetilde{\mathbf{x}}_i, \mathbf{B}) &= \beta_0^{(N_c+1)} + \sum_{\ell=1}^{N_c} \left[\beta_\ell^{(N_c+1)} \tanh\left(\beta_0^{(\ell)} + \sum_{j=1}^{q} \beta_j^{(\ell)} x_{i,j}\right)\right] \\ &= \mathbf{B}^{(2)'} \cdot \mathbf{f}(\mathbf{B}^{(1)} \widetilde{\mathbf{x}}_i), \end{aligned} \tag{1.4}$$

avec $\widetilde{\mathbf{x}}_i \in \mathbb{R}^{q+1}$ le vecteur des variables (augmenté du terme constant de biais), $\mathbf{B}^{(1)}$ une matrice de dimension $N_c \times (q+1)$, de terme générique $\beta_j^{(\ell)}$, contenant les poids attribués aux variables en entrée de chaque neurone cachés, $\mathbf{B}^{(2)} = (\beta_0^{(N_c+1)}, ..., \beta_{N_c}^{(N_c+1)})'$ le vecteur des paramètres correspondant au neurone de sortie linéaire et le vecteur $\mathbf{f} \in \mathbb{R}^{N_c+1}$ constitué d'un biais (qui vaut 1) et de toutes les fonctions non linéaires correspondant aux neurones cachés. En l'occurence on a $\mathbf{f}_\ell(\mathbf{B}^{(1)} \widetilde{\mathbf{x}}_i) = \tanh\left(\sum_{j=0}^{q} \beta_j^{(\ell)} x_{i,j}\right)$ avec la convention $x_{i,0} := 1$ pour tout i. Le vecteur $\mathbf{B}$ du membre de gauche de (1.4) résulte de la concaténation de

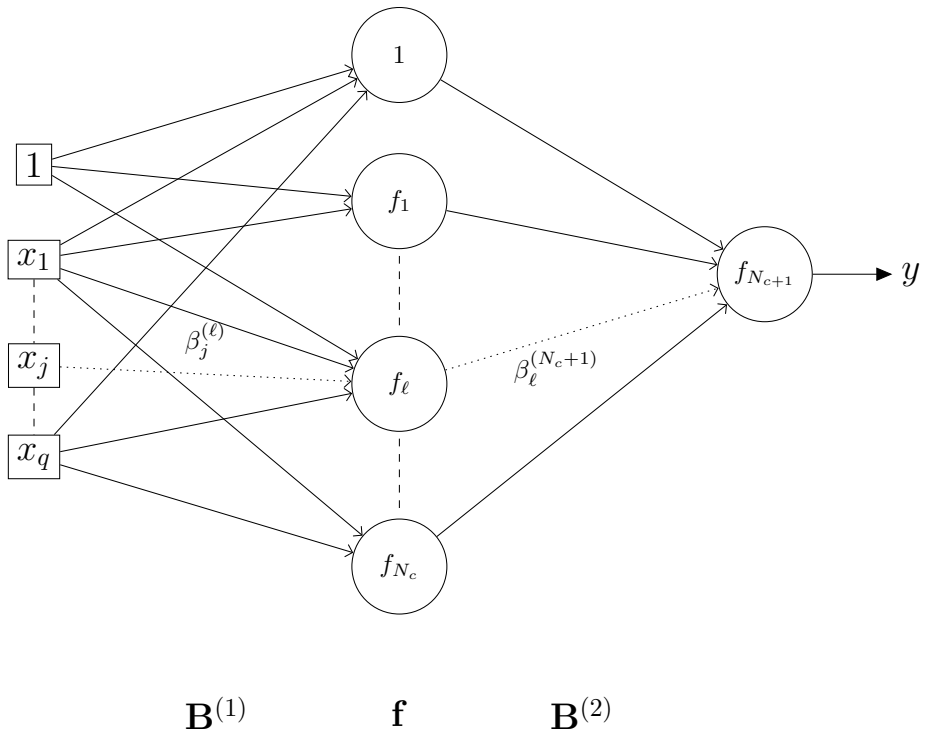

FIGURE 1.1 – Représentation schématique d'un réseau de neurones à une couche avec un neurone de sortie linéaire

toutes les composantes de $\mathbf{B}^{(1)}$ et $\mathbf{B}^{(2)}$. Il a donc une dimension $N_c(q+2)+1$.

Les réseaux de ce type possèdent une propriété remarquable, à l'origine de leur intérêt pratique, qui est celle d'être des approximateurs universels parcimonieux. Par approximateur, nous entendons que toute fonction bornée suffisamment régulière peut être approchée avec autant de précision que l'on veut, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire. Mais il ne s'agit pas d'une particularité des réseaux de neurones car d'autres familles de fonctions paramétrées possèdent également cette propriété, cependant sa spécificité provient de son caractère « parcimonieux » : à précision égale, les réseaux de neurones nécessitent moins

de paramètres ajustables que les autres approximateurs connus. Dans notre contexte cependant cette propriété fondamentale ne sera pas très utile compte tenu du fait que notre problème ne comporte que deux variables (latitude et longitude).

En revanche, nous allons décrire maintenant une autre propriété fondamentale, qui elle par contre sera utilisée dans notre contexte. Les paramètres optimaux d'un réseau de neurones ayant une architecture fixée sont obtenus en minimisant la fonction moindres carrés

$$LS(\mathbf{B}) = \sum_{i=1}^{n} \|y_i - g(\widetilde{\mathbf{x}}_i, \mathbf{B})\|^2.$$

Il existe différentes façons de minimiser cette quantité en $\mathbf{B}$. L'algorithme que nous utiliserons est appelé algorithme de rétropropagation du Gradient (nous renvoyons à LeCun *et al.*, 1998, pour plus de précisions). Cette optimisation est aussi appelée entraînement du réseau. Sa mise en oeuvre requiert la connaissance a priori du nombre de neurones cachés (sachant que dans notre cas il n'y a qu'une seule couche). Afin de déterminer le nombre optimal de neurones cachés, nous employons une méthode classique de validation croisée. Plus précisément, nous utiliserons l'approche "m-fold cross-validation", qui consiste à diviser l'échantillon en m sous-échantillons, puis à sélectionner un des m échantillons comme ensemble de validation et les $m-1$ autres comme ensemble d'apprentissage. Dans notre contexte, nous choisirons pour commencer un nombre maximal de neurones cachés dans la couche. Pour chaque architecture (i.e. pour chaque nombre de neurones cachés inférieur ou égal à ce maximum) nous décomposerons l'échantillon $\mathcal{Z}$ en m parties $(\mathcal{Z}_1, \ldots, \mathcal{Z}_m)$. Puis, pour tout $i \in \{1 \ldots m\}$, nous optimiserons le réseau en utilisant pour données d'entraînement $(\mathcal{Z}_1, \ldots, \mathcal{Z}_{i-1}, \mathcal{Z}_{i+1}, \ldots, \mathcal{Z}_m)$ et nous calculerons l'erreur de généralisation du modèle sur l'échantillon $\mathcal{Z}_i$ comme suit

$$\sum_{j:\mathbf{x}_j \in \mathcal{Z}_i} \|y_j - g(\widetilde{\mathbf{x}}_j, \mathbf{B}_i)\|^2,$$

où $\mathbf{B}_i$ est le vecteur des paramètres optimaux obtenu par minimisation de LS sur les données d'entraînement $(\mathcal{Z}_1, \ldots, \mathcal{Z}_{i-1}, \mathcal{Z}_{i+1}, \ldots, \mathcal{Z}_m)$. Finalement, nous sommerons ces

erreurs sur i et nous choisirons l'architecture pour laquelle l'erreur finale est minimale. Une fois l'architecture optimale trouvée, nous l'utiliserons et nous réentraînerons le réseau à l'aide des données de $\mathcal{Z}$, ce qui nous donnera le vecteur optimal final $\mathbf{B}$.

La propriété des réseaux de neurones dont nous allons tirer profit dans la prochaine section provient du fait que l'algorithme d'optimisation de LS converge vers des minima locaux. De ce fait, contrairement aux modèles linéaires en leurs paramètres, nous pouvons créer différents réseaux de neurones avec les mêmes données d'entraînement en changeant l'initialisation des paramètres dans le processus d'optimisation. Nous allons maintenant montrer l'utilité de cette propriété pour le Query By Committee.

1.3.2 Query By Committee

Nous allons dans cette section présenter l'algorithme QBC. Quelques figures illustreront, sur un exemple à une dimension, les différentes étapes de l'algorithme schématisées dans la Figure 1.2. Supposons que l'on veuille estimer une fonction

$$h : \mathbb{R}^q \to \mathbb{R},$$

par exemple celle de la Figure 1.3, avec un réseau de neurones tel que nous l'avons défini dans la section précédente (un réseau à une couche de neurones cachés et un neurone de sortie linéraire). On suppose connues uniquement les valeurs de la fonction en n points, i.e. on dispose d'un échantillon

$$\mathcal{Z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}.$$

Pour améliorer la connaissance de h, on pourrait ajouter aléatoirement de nouvelles observations à $\mathcal{Z}$. Cependant, si l'acquisition de ces nouvelles observations est coûteuse, en temps et/ou en argent, on aimerait être à même de choisir les observations qui nous apporteraient le plus d'information sur h. C'est pour répondre à ce problème que le QBC a été initialement introduit par Seung *et al.* (1992) dans un contexte de classification et

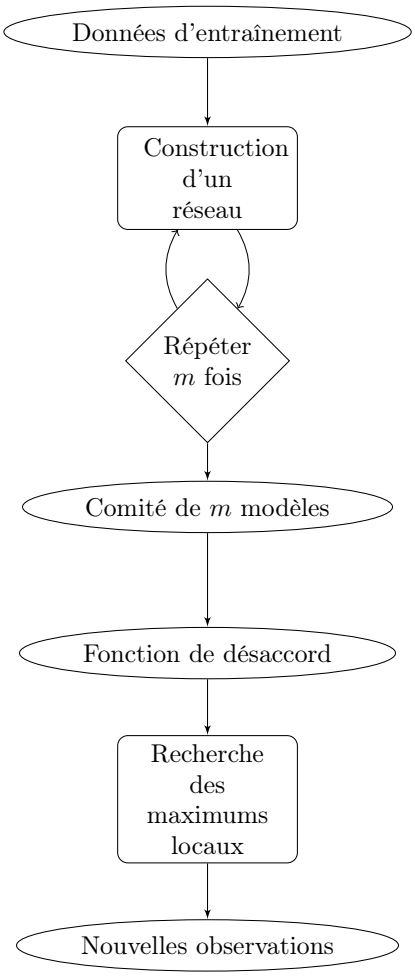

FIGURE 1.2 – Représentation schématique de l'algorithme.

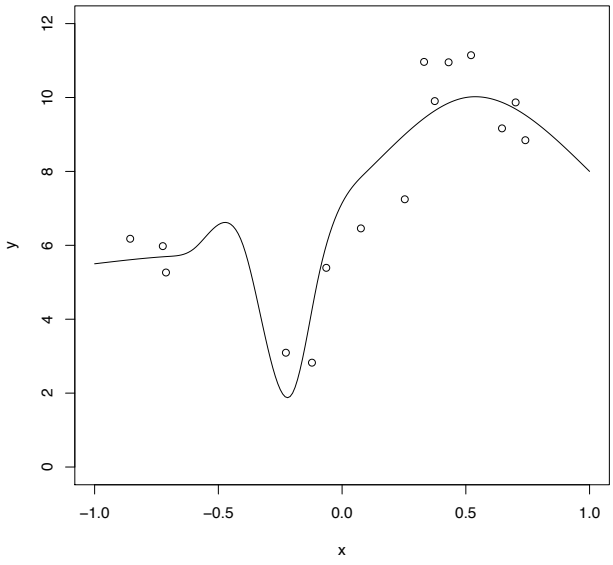

FIGURE 1.3 – Un exemple de fonction h et d'un échantillon initial $\mathcal{Z}$.

étendu au contexte de la régression dans Krogh et Vedelsby (1995). Son principe est de sélectionner les nouvelles données les plus informatives afin d'obtenir l'estimation la plus précise possible de h avec le moins de données possibles. L'idée du QBC est de comparer les résultats d'un comité de modèles (appelés aussi experts) en chaque point du territoire. On suppose donc avoir créé un ensemble de m modèles $\widetilde{h}^{(1)}, \ldots, \widetilde{h}^{(m)}$ (cf. Figure 1.4 avec $m = 5$). Nous discuterons ultérieurement de la façon de construire ces réseaux. On définit la moyenne pondérée

$$\overline{h}(\mathbf{x}) = \sum_{i=1}^{m} p_i \widetilde{h}^{(i)}(\mathbf{x}).$$

Ce type de moyenne est usuel en réseaux de neurones et permet d'appréhender le problème biais/variance (cf. e.g., Hansen et Salomon, 1990 ; Wolpert, 1992 ; Perrone et Cooper, 1993). Cette fonction sera considérée comme la sortie de notre modèle. Les poids p_i représentent l'importance accordée au réseau $\widetilde{h}^{(i)}$ dans l'estimation de h. Naturellement les p_i sont positifs et de somme égale à 1. Nous ne nous intéresserons pas au choix des p_i et les prendrons tous égaux à $1/m$. Nous définissons maintenant la fonction d'ambiguité

d'un élément comme

$$a^{(i)}(\mathbf{x}) = (\widetilde{h}^{(i)}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2$$

et le désaccord de l'ensemble comme

$$d(\mathbf{x}) := \overline{a}(\mathbf{x}) = \sum_i p_i a^{(i)}(\mathbf{x}) = \sum_i p_i (\widetilde{h}^{(i)}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2.$$

Nous définissons également l'erreur de prédiction des réseaux et des moyennes

$$\epsilon^{(i)} = (h(\mathbf{x}) - \widetilde{h}^{(i)}(\mathbf{x}))^2$$

$$e(x) = (h(\mathbf{x}) - \overline{h}(\mathbf{x}))^2,$$

ce qui nous permet de réécrire

$$d(\mathbf{x}) = \sum_i p_i \epsilon^{(i)}(\mathbf{x}) - e(\mathbf{x}) = \overline{\epsilon}(\mathbf{x}) - e(\mathbf{x}).$$

Il en découle que $\overline{\epsilon}(\mathbf{x}) \geq d(\mathbf{x})$ qui peut être calculée sans connaitre h en chaque point puisque cette quantité ne dépend que des membres du comité.

La première étape de l'algorithme est de construire le comité. Pour ce faire, nous utilisons $\mathcal{Z}$ et nous créons m modèles différents $\widetilde{h}^{(1)}, \ldots, \widetilde{h}^{(m)}$. Nous pouvons maintenant utiliser la propriété sus-mentionnée selon laquelle l'algorithme d'optimisation de LS converge vers des minima locaux pour construire m modèles différents à partir d'un choix aléatoire des paramètres initiaux dans l'optimisation des réseaux. Nos modèles peuvent ensuite être utilisés pour estimer les quantités d'intérêt en tout point du domaine expérimental, qui sera noté $\mathcal{X}$.

Maintenant que le comité a été construit, nous allons comparer les réponses des différents modèles en tout point du domaine : pour cela, nous allons associer m sorties de modèles $\widetilde{h}^{(1)}(\mathbf{x}), \ldots, \widetilde{h}^{(m)}(\mathbf{x})$ à chaque point $\mathbf{x} \in \mathcal{X}$. L'idée sous-jacente du QBC est alors que si les réponses des modèles en chaque point sont globalement les mêmes, alors il n'y a aucune raison de penser que de rajouter des observations sera informatif. Dans le cas contraire, si les résultats des modèles sont très différents, cela signifie que la modélisation est difficile et que donc le rajout d'une observation peut être judicieux.

Concernant la fonction de désaccord entre les experts, nous pouvons la quantifier par la variance entre les réponses qui s'écrit

$$d(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \left(\widetilde{y}^{(i)} - \overline{y} \right)^2$$

où $\widetilde{y}^{(i)} = \widetilde{h}^{(i)}(\mathbf{x})$ et $\overline{y} = \frac{1}{m} \sum_{i=1}^{m} \widetilde{y}^{(i)}(\mathbf{x})$.

Afin d'illustrer l'applicabilité du QBC, reprenons notre exemple de la Figure 1.3. Une fois que nous avons calculé notre fonction de désaccord en chaque point, nous sélectionnons les k maxima locaux de la fonction les plus élevés (k étant le nombre de nouvelles observations qui est autorisé) et nous ajoutons les observations correspondantes à l'échantillon $\mathcal{Z}$. Il s'agit là de la première étape de notre algorithme. Nous reprenons ensuite l'échantillon $\mathcal{Z}$ ainsi mis à jour et nous réitérons l'algorithme jusqu'à ce que nous ne puissions plus rajouter d'observations ou que l'ajustement soit considéré comme satisfaisant. La Figure 1.4 représente une itération de l'algorithme associé à l'exemple de la Figure 1.3. La taille du comité est de 5 et nous avons choisi d'ajouter 5 observations à chaque itération. Le graphe de la variance illustre qu'il y a une réelle différence, différence comblée après 5 itérations. Nous pouvons en effet observer sur la Figure 1.5, obtenue après ces 5 itérations, que les experts semblent maintenant en accord et que tous ajustent bien la courbe initiale.

Cette approche très prometteuse du QBC va être mise en oeuvre dans le Chapitre 3 mais adaptée au contexte qui est le nôtre, à savoir celui des extrêmes. Compte tenu du fait qu'une loi GPD fait intervenir deux paramètres, un paramètre de forme et un paramètre d'échelle, nous avons alors deux fonctions de coût. Il s'agit là d'un véritable dilemme puisque les stations qui apportent de l'information sur le paramètre de forme n'en apportent pas nécessairement sur celui d'échelle. Dans le but d'éviter ce choix qui ne pourrait être qu'arbitraire, nous proposerons d'utiliser une autre fonction coût basée sur le quantile qui combine ces deux paramètres. Une comparaison sur la base de simulations et sur l'analyse d'un jeu de données réelles sera proposée dans le Chapitre 3 afin d'apprécier l'impact d'utiliser l'une ou l'autre de ces fonctions coûts sur la décision d'ajouter ou de

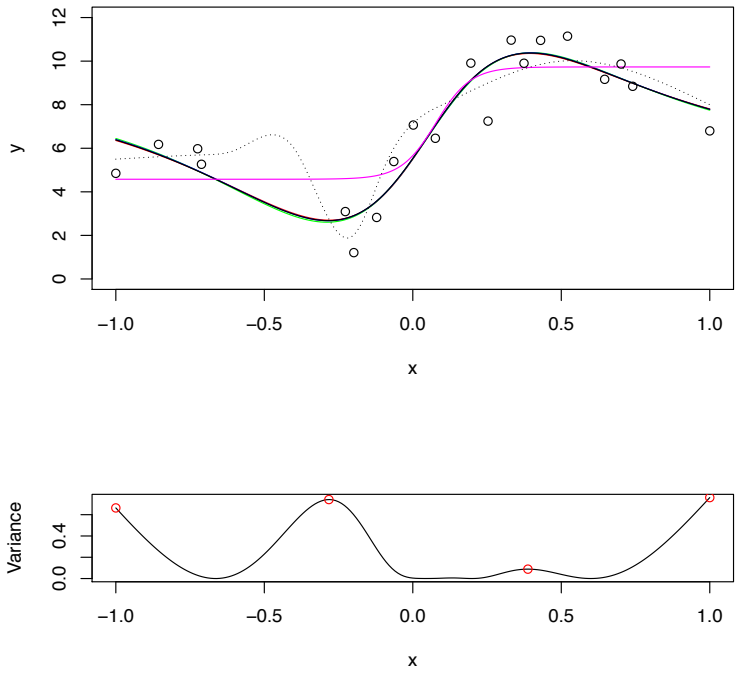

FIGURE 1.4 – Première itération de l'algorithme appliqué à l'exemple de la Figure 1.3. Le graphe du haut représente le comité des différents experts, la fonction d'intérêt est en pointillé, alors que les courbes en couleur représentent les différents experts. Le graphe du bas représente la fonction coût en chaque point, les points rouges représentant les maxima de la fonction d, i.e. les observations à rajouter.

supprimer une station.

1.4 Estimation robuste en présence de covariables

En pratique, on est souvent confronté à des points aberrants ("outliers") qui perturbent les techniques d'estimation. Typiquement, dans ce genre de situations, les estimateurs du modèle obtenus par maximum de vraisemblance sont très instables, ce qui nous incite à chercher à utiliser des méthodes robustes. Un traitement particulier de ces points

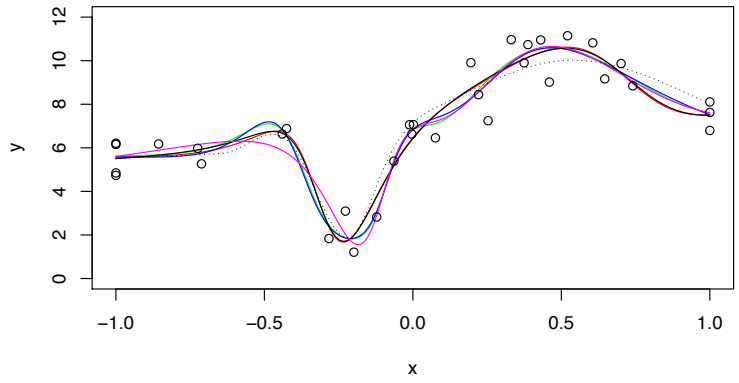

FIGURE 1.5 – Représentation du comité après 5 itérations de l'algorithme.

aberrants est donc nécessaire, par exemple en pondérant convenablement leur influence dans l'estimation.

Pour résoudre ce problème, des mesures de divergence entre deux densités f et g ont été introduites :

$$d_\alpha(f;g) = \int \left\{ g^{1+\alpha}(z) - \left(1+\frac{1}{\alpha}\right) f(z) g^\alpha(z) + \frac{1}{\alpha} f^{1+\alpha}(z) \right\} dz \quad \alpha > 0.$$

Quand $\alpha = 0$ l'intégrande n'est pas définie, mais doit être interprétée comme la version limite quand α tend vers 0, à savoir

$$d_0(f;g) = \int f(z) \ln \frac{f(z)}{g(z)} dz$$

qui n'est rien d'autre que la divergence de Kullback-Leibler.

Comme on peut le voir, ces mesures sont indexées par un unique paramètre, noté α, qui sert de compromis entre la robustesse et l'efficacité des estimateurs qui minimisent cette famille de divergences. Quand $\alpha = 0$, on retrouve l'estimateur du maximum de vraisemblance ; quand $\alpha = 1$ on obtient un estimateur robuste mais inefficace en minimisant la divergence qui, dans ce cas, n'est autre que l'erreur en moyenne quadratique. Quel que soit α, la procédure d'estimation présente l'avantage de ne pas nécessiter de lissage.

Différents exemples ont été étudiés dans la littérature pour investiguer le lien entre robustesse et efficacité. Il a été établi que les estimateurs avec un α petit possédaient des propriétés de grande robustesse avec une perte d'efficacité pas trop importante par rapport au maximum de vraisemblance. On peut noter également que ces estimateurs ne sont rien d'autres que des $M-$estimateurs.

Cette approche sera utilisée dans le Chapitre 4 afin d'estimer des paramètres de queue dans un contexte de théorie des valeurs extrêmes. Un tel sujet a été récemment étudié dans la littérature. On peut citer par exemple Brazauskas et Serfling (2000) et Vandewalle *et al.* (2007) pour des distributions Pareto strictes ou de type Pareto, Dupuis et Field (1998), Peng et Welsh (2001) et Juárez et Schucany (2004) pour des distributions GEV ou des GPD. En ce qui nous concerne, on se focalisera sur la classe de Gumbel. Bien que les comportements des queues soient différents, toutes ces distributions ont en commun un indice zéro et donc les différencier sur la base de ce paramètre uniquement est impossible. Pour résoudre ce problème, on se restreindra aux distributions de type Weibull pour lesquelles la queue est de la forme

$$\overline{F}(y) := 1 - F(y) = e^{-y^{1/\theta}\ell_F(y)}, y > 0$$

où $\theta > 0$ et ℓ_F est une fonction à variations lentes en l'infini. Le paramètre θ s'appelle le coefficient de Weibull. Différentes valeurs de ce paramètre permettent de couvrir une large partie de la classe de Gumbel et ainsi cette classe constitue un sous groupe très flexible. L'estimation de ce paramètre a été longuement étudiée dans la littérature (cf. e.g., Broniatowski, 1993 ; Beirlant *et al.*, 1995 ; Gardes et Girard, 2005, 2008b ; Diebolt *et al.*, 2008 ; Dierckx *et al.*, 2009 ; Goegebeur *et al.*, 2010 ou Goegebeur et Guillou, 2011) mais rarement en présence de covariables.

On se placera donc dans ce contexte de régression avec covariables aléatoires. Autrement

dit, on supposera disposer de couples (X_i, Y_i), $i = 1, ..., n$ indépendants de même loi que $(X, Y) \in \mathbb{R}^p \times \mathbb{R}_+$, et dont la loi conditionnelle de Y sachant $X = x$ est de type Weibull, i.e. avec une queue de la forme

$$\overline{F}(y; x) = e^{-y^{1/\theta(x)} \ell_F(y;x)}, y > 0. \tag{1.5}$$

C'est un sujet peu abordé dans la littérature. On peut néanmoins citer comme références : Wang et Tsai (2009) avec une approche paramétrique par le maximum de vraisemblance sous l'hypothèse du modèle de Hall (Hall, 1982), Daouia *et al.* (2011) dans le contexte de distribution de type Pareto et Daouia *et al.* (2013) dans le cas général d'un domaine d'attraction, mais sous des hypothèses restrictives sur F. Ici on considèrera le cas de distributions de type Weibull et l'approche utilisée sera basée sur une estimation locale dans un voisinage d'un point de l'espace des covariables. Cet ajustement local sera effectué par un critère de minimisation de la divergence. Ce critère a déjà été utilisé pour l'estimation robuste dans le cas de distributions de type Pareto, e.g. par Kim et Lee (2008), Dierckx *et al.* (2013a, b) mais pas dans le contexte de type Weibull. On proposera dans le Chapitre 4 un estimateur robuste du paramètre $\theta(x)$ introduit dans (1.5) et on établira ses propriétés asymptotiques sous des conditions de régularité convenables. Des simulations seront ensuite proposées pour illustrer son comportement à distance finie.

Cette thèse offre de nombreuses perspectives, à la fois théoriques et pratiques, dont une liste non exhaustive sera présentée en conclusion.

Chapitre 2

A non-parametric entropy-based approach to detect changes in climate extremes

Abstract

This chapter focuses primarily on temperature extremes measured at 24 european stations with at least 90 years of data. Here, the term extremes refers to rare excesses of daily maxima and minima. As mean temperatures in this region have been warming over the last century, it is mechanical that this positive shift can be detected also in extremes. After removing this warming trend, we focus on the question of determining if other changes are still detectable in such extreme events. As we do not want to hypothesize any parametric form of such possible changes, we propose a new non-parametric estimator based on the Kullback-Leibler divergence tailored for extreme events. The properties of our estimator are studied theoretically and tested with a simulation study. Our approach is also applied to seasonal extremes of daily maxima and minima for our 24 selected stations.

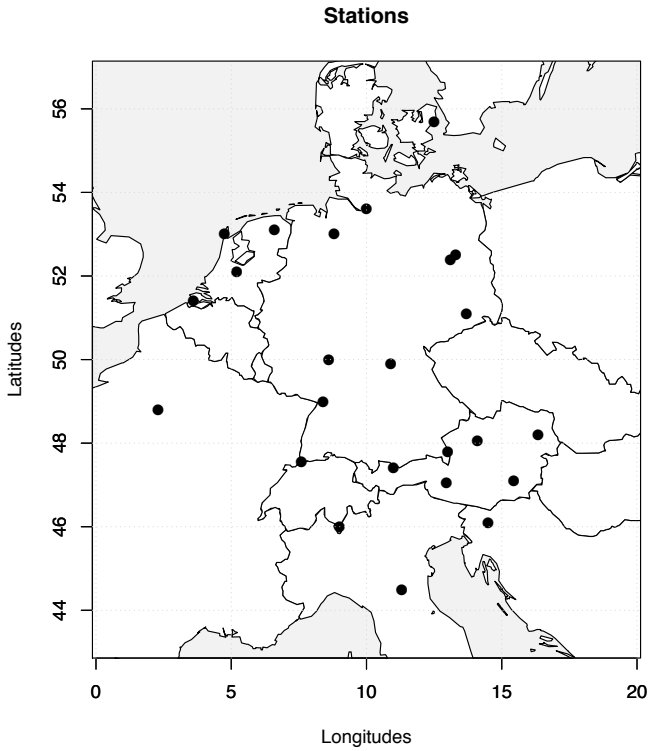

FIGURE 2.1 – Weather station locations described in Table 2.1 (source : ECA&D database).

2.1 Introduction

In a global warming context, climatologists, flood planners, insurers and risk modellers have been increasingly interested in determining whether the upper tail distribution of some meteorological quantity has changed over time at some specific places (Zwiers *et al.*, 2011 ; Kharin *et al.*, 2007). As a motivating example, we focus on 24 weather stations that have at least 90 years of daily maxima and minima temperature measurements, see Table 2.1 and black dots in Figure 2.1.

A typical inquiry in impact studies is to wonder if high temperatures over the current climatology (i.e. the last thirty years) significantly differ from the ones measured during previous time periods. As a positive shift in the mean behaviour of temperatures has

TABLE 2.1 – Characteristics of 24 weather stations from the European Climate Assessment & Dataset project `http://eca.knmi.nl/dailydata/predefinedseries.php`. The heights are expressed in meters.

			Austria			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Kremsmunster	+48 :03 :00	+014 :07 :59	383	1876	2011	-
Graz	+47 :04 :59	+015 :27 :00	366	1894	2011	2
Salzburg	+47 :48 :00	+013 :00 :00	437	1874	2011	5
Sonnblick	+47 :03 :00	+012 :57 :00	3106	1887	2011	-
Wien	+48 :13 :59	+016 :21 :00	199	1856	2011	2
			Denmark			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Koebenhavn	+55 :40 :59	+012 :31 :59	9	1874	2011	-
			France			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Montsouris	+48 :49 :00	+002 :19 :59	77	1900	2010	-
			Germany			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Bamberg	+49 :52 :31	+010 :55 :18	240	1879	2011	-
Berlin	+52 :27 :50	+013 :18 :06	51	1876	2011	1
Bremen	+53 :02 :47	+008 :47 :57	4	1890	2011	1
Dresden	+51 :07 :00	+013 :40 :59	246	1917	2011	-
Frankfurt	+50 :02 :47	+008 :35 :54	112	1870	2011	1
Hamburg	+53 :38 :06	+009 :59 :24	11	1891	2011	-
Karlsruhe	+49 :02 :21	+008 :21 :54	112	1876	2011	2
Potsdam	+52 :22 :59	+013 :04 :00	100	1893	2011	-
Zugspitze	+47 :25 :00	+010 :58 :59	2960	1901	2011	1
			Italy			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Bologna	+44 :30 :00	+011 :20 :45	53	1814	2010	-
			Netherlands			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
De Bilt	+52 :05 :56	+005 :10 :46	2	1906	2011	-
Den Helder	+52 :58 :00	+004 :45 :00	4	1901	2011	-
Eelde	+53 :07 :24	+006 :35 :04	5	1907	2011	-
Vlissingen	+51 :26 :29	+003 :35 :44	8	1906	2011	2
			Slovenia			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Ljubjana	+46 :03 :56	+014 :31 :01	299	1900	2011	5
			Switzerland			
Station name	Latitude	Longitude	Height	First year	Last year	Missing years
Basel	+47 :33 :00	+007 :34 :59	316	1901	2011	-
Lugano	+46 :00 :00	+008 :58 :00	300	1901	2011	-

been observed (e.g., see Figure 1 in Abarca-Del-Rio and Mestre, 2006), such a warming automatically translates into higher absolute temperatures (e.g., Shaby and Reich, 2012 ; Jarušková and Rencová, 2008 ; Dupuis, 2012). This mean behaviour being removed, is it still possible to detect changes in high extreme temperatures over the last century ? This important climatological question can be related to the investigation of Hoang *et al.* (2009) who wondered if the trends in extremes are only due to trends in mean and variance of the whole dataset. Our scope is different here. We neither aim at identifying smooth trends in extremes nor at linking changes between variances and upper quantiles. Our objective differs in the sense that we only want to determine if there is a change in extremes distributions. To explore such a question, we would like to assume very few distributional hypotheses, i.e. not imposing a specific parametric density, and to propose a fast statistical non-parametric approach that could be implemented to large datasets. Although no large outputs from Global climate models will be treated here, we keep in mind computational issues when proposing the statistical tools developed therein.

One popular approach used in climatology consists in building a series of so-called extreme weather indicators and in studying their temporal variabilities in terms of frequency and intensity (e.g., Alexander *et al.*, 2006 ; Frich *et al.*, 2002). A limit of working with such indices is that they often focus on the "moderate extremes" (90% quantile or below), but not on upper extremes (above the 95% quantile). In addition, their statistical properties have rarely been derived. A more statistically oriented approach to analyze large extremes is to take advantage of Extreme Value Theory (EVT). According to Fisher and Tippett's theorem (1928), if $(X_1, \ldots, X_n)$ is an independent and identically distributed sample of random variables and if there exists two sequences $a_n > 0$ and $b_n \in \mathbb{R}$ and a non degenerate distribution $H_{\mu,\sigma,\xi}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i=1,\ldots,n} \frac{X_i - b_n}{a_n} \leq x\right) = H_{\mu,\sigma,\xi}(x),$$

then $H_{\mu,\sigma,\xi}$ belongs to the class of distributions

$$H_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\} & \text{if } \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]_+\right\} & \text{if } \xi = 0 \end{cases}$$

which is called the Generalized Extreme Value distributions (GEV). The shape parameter, often called ξ in environmental sciences, is of primary importance. It characterizes the GEV tail behaviour (see Coles, 2001 ; Beirlant *et al.*, 2004, for more details). If one is ready to assume that temperature maxima for a given block size (day, month, season, etc) are approximately GEV distributed, then it is possible to study changes in the GEV parameters themselves (e.g., Fowler and Kilsby, 2003 ; Kharin *et al.*, 2007).

This GEV based method is attractive because it takes advantage of EVT and hypothesis testing can be clearly implemented. For example, Jarušková and Rencová (2008) studied test statistics for detecting changes in annual maxima and minima temperature series measured at five meteorological sites - Bruxelles, Cadiz, Milan, St. Petersburg, and Stockholm. Three limitations of such GEV-based approaches can be identified. It is tailored for maxima and this discards all but one observation per block, e.g. one value out of 365 days for annual maxima. It imposes a fixed GEV form that may be too restrictive for small blocks (the GEV being a limiting distribution). Not one but three parameters have to be studied to detect changes in a time series. Regarding the first limitation, a classical solution in EVT (e.g., Coles, 2001) is to work with excesses above a high threshold instead of block maxima. The tail (survival) distribution of such excesses is usually modelled by a Generalized Pareto (GP) tail (Pickands, 1975)

$$\overline{G}_{\sigma,\xi}(y) = \begin{cases} \left(1+\xi\frac{y}{\sigma}\right)^{-1/\xi}, & \text{if } \xi \neq 0, \\ \exp\left(-\frac{y}{\sigma}\right), & \text{if } \xi = 0, \end{cases}$$

where the scale parameter σ is positive and $y \geq 0$ if the shape parameter ξ is positive, and $y \in [0, -\sigma/\xi[$ when $\xi < 0$. Still, the two other limitations remain for the GP model.

In this chapter, we move away from imposing a fixed parametric density form. Having

no parametric density at our disposal obviously implies that it is impossible to monitor changes in parameters (e.g., Grigga and Tawn, 2012). Another strategy has to be followed to compare the distributional differences between extremes. In information theory (e.g., Burnham and Anderson, 1998), it is a common endeavour to compare the probability densities of two time periods by computing the entropy (Kullback Leibler directed divergence)

$$I(f;g) = \mathbb{E}_f \left\{ \ln \left(\frac{f(\mathbf{X})}{g(\mathbf{X})} \right) \right\},$$

where $\mathbf{X}$ represents the random vector with density f and g another density. Although not a true distance, this expectation provides relevant information on how close g is from f. Kullback (1968) coined the term "directed divergence" to distinguish it from the divergence defined by

$$D(f;g) = I(f;g) + I(g;f),$$

which is symmetric relative to f and g. We will follow this terminology. Working with the entropy presents a lot of advantages. It is a notion shared by many different communities : physics, climatology, statistics, computer sciences and so on. It is a concise one-dimensional summary. It is clearly related to model selection and hypothesis testing (e.g., Burnham and Anderson, 1998). For example, the popular Akaike criterion (Akaike, 1974) can be viewed throughout the Kullback Leibler divergence lenses. For some distributions, explicit expressions of the divergence can be derived. This is the case if g and f correspond to two Gaussian multivariate densities (e.g., Penny and Roberts, 2000). In terms of extremes, if g and f represent two univariate GP densities with identical scale parameters and two different positive shape parameters, ξ_f and ξ_g, then we can write

$$I(f;g) = -1 - \xi_f + sign(\xi_f - \xi_g) \left(1 + \frac{1}{\xi_g} \right) \left| \frac{\xi_f}{\xi_g} - 1 \right|^{-1/\xi_f} \int_{\frac{\xi_g}{\xi_f}}^{1} t^{-1/\xi_f} \left| 1 - t \right|^{1/\xi_f - 1} dt. \quad (2.1)$$

Although we will not assume that excesses follow explicitly a GP in our method, Equation (2.1) will be used in our simulations as a test case.

In this chapter, our goal is to provide and study a non-parametric estimator of the diver-

gence for large excesses. Fundamental features in an extreme value analysis are captured by the tail behaviour. In its original form, the divergence is not expressed in terms of tails but in function of probability densities. One important aspect of this work is to propose an approximation of the divergence in terms of the tail distributions, see Section 2.2. This leads to a new non-parametric divergence estimator tailored for excesses. Its properties will be studied in Section 2.3. The last section is dedicated to the analysis of a few simulations and of temperature extremes recorded at the stations plotted in Figure 2.1. All proofs are deferred to the Appendix.

2.2 Entropy for excesses

Our main interest resides in the upper tail behaviour and a first task is to make the divergence definition relevant to this extreme values context. This is done by replacing the densities f and g by densities of excesses above a certain high threshold u. We need a few notations to describe precisely this adapted divergence.

Definition 2.1 *Let X and Y be two absolutely continuous random variables with density f, resp. g, and tail $\overline{F}(x) = \mathbb{P}(X > x)$, resp. $\overline{G}(y) = \mathbb{P}(Y > y)$. Denote the random variable above the threshold u as $X_u = [X|X > u]$ with density $f_u(x) = f(x)/\overline{F}(u)$ and tail $\overline{F_u}(x) = \overline{F}(x)/\overline{F}(u)$ for $x \in (u, x_F)$ where x_F is the upper endpoint of F. The same type of notations can be used for $Y_u = [Y|Y > u]$. A suitable version of the directed divergence for extreme values is then*

$$I(f_u; g_u) = \mathbb{E}_{f_u}\left\{\ln\left(\frac{f_u(X_u)}{g_u(X_u)}\right)\right\} = \frac{1}{\overline{F}(u)}\int_u^{x_F} \ln\left(\frac{f_u(x)}{g_u(x)}\right) f(x)dx.$$

Assumption 2.1 *We always assume in this chapter that the densities f and g are chosen such that the two directed divergences $I(f_u; g_u)$ and $I(g_u; f_u)$ are finite and in particular, both upper endpoints are equal to $\tau = x_G = x_F$ in order to compute the ratios $\frac{f_u(x)}{g_u(x)}$ and $\frac{g_u(x)}{f_u(x)}$ for large $x > u$.*

If those assumptions are not satisfied in practice, this does not necessarily stop us from answering our climatological question : deciding if current temperature extremes over central Europe differ from past ones. If the difference $|x_F - x_G|$ is large, then the divergence is infinite and there is no need to develop complex statistical procedures to detect differences between current and past extremes. If the difference $|x_F - x_G|$ becomes smaller and smaller, it is more and more difficult to determine if the divergence is infinite from a given finite sample. For the limiting case, $x_F = x_G$, the divergence is finite and our estimate almost surely converges, see Theorem 2.1. This case corresponds to our main assumption and it is particularly relevant about temperature extremes over central Europe because it is physically possible that an upper temperature bound exists for this region, (e.g., Shaby and Reich, 2012 ; Jarušková and Rencová, 2008).

As previously mentioned, we would like to express the divergence in terms of survival functions which are more adapted to extremes than densities. The next proposition reaches this goal by providing an approximation of the divergence in function of $\overline{F}$ and $\overline{G}$.

Proposition 2.1 *If*

$$\lim_{u \to \tau} \int_u^\tau \left(\ln \frac{f(x)}{\overline{F}(x)} - \ln \frac{g(x)}{\overline{G}(x)} \right) (f_u(x) - g_u(x))\, dx = 0 \tag{2.2}$$

then the divergence $D(f_u; g_u) = I(f_u; g_u) + I(g_u; f_u)$ *is equivalent, as* $u \uparrow \tau$*, to the quantity*

$$K(f_u; g_u) = -L(f_u; g_u) - L(g_u; f_u)$$

where

$$L(f_u; g_u) = \mathbb{E}_f \left\{ \ln \frac{\overline{G}(X)}{\overline{G}(u)} \middle| X > u \right\} + 1. \tag{2.3}$$

For the special case of GP densities, we can explicitly compute from (2.1) the true divergence $D(f_u; g_u)$ and its approximation $K(f_u; g_u)$. In Figure 2.2, the relative error $|K(f_u; g_u) - D(f_u; g_u)| / D(f_u; g_u)$ when f and g correspond to two GP densities with a

unit scale parameter and $\xi_f = 0.1$ is displayed for different threshold values (x-axis). The solid, dotted and dash-dotted lines correspond to shape parameters $\xi_g = 0.2, 0.15$ and 0.12, respectively.

As the threshold increases, the relative error between $K(f_u; g_u)$ and $D(f_u; g_u)$ rapidly becomes small. The difference between $\xi_f = 0.1$ and ξ_g does not play an important role. The idea behind condition (2.2) is the following one. If $\ln\left(\frac{f(x)\cdot\overline{G}(x)}{\overline{F}(x)\cdot g(x)}\right)$ tends to a constant rapidly enough, then the integral $\int_u^\tau \text{cst} \times (f_u(x) - g_u(x))\, dx$ equals zero because $\int_u^\tau f_u(x)dx = \int_u^\tau g_u(x)dx = 1$. Is condition (2.2) satisfied for a large class of densities? The coming two subsections answer positively to this inquiry.

2.2.1 Checking condition (2.2)

In EVT, three types of tail behaviour (heavy, light and bounded) are possible and correspond to the GP sign of ξ, positive, null and negative, respectively. Those three cases have been extensively studied and they have been called the three domains of attraction, Fréchet, Gumbel and Weibull (e.g., see Chapter 2 of Embrechts *et al.*, 1997). The next two propositions focus on the validity of condition (2.2) for tails belonging to the Fréchet and Weibull domains of attraction, respectively. The Gumbel case that contains a lot of classical densities like the Gamma and Gaussian ones is more complex to deal with and we opt for a different approach based on stochastic ordering to check condition (2.2) for those types of densities.

Proposition 2.2 *Suppose that the random variables X and Y belong to the Fréchet max-domain of attraction, i.e. $\overline{F}$ and $\overline{G}$ are regularly varying,*

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\alpha} \text{ and } \lim_{t\to\infty} \frac{\overline{G}(tx)}{\overline{G}(t)} = x^{-\beta},$$

for all $x > 0$ and some $\alpha > 0$ and $\beta > 0$. We also impose the following classical second

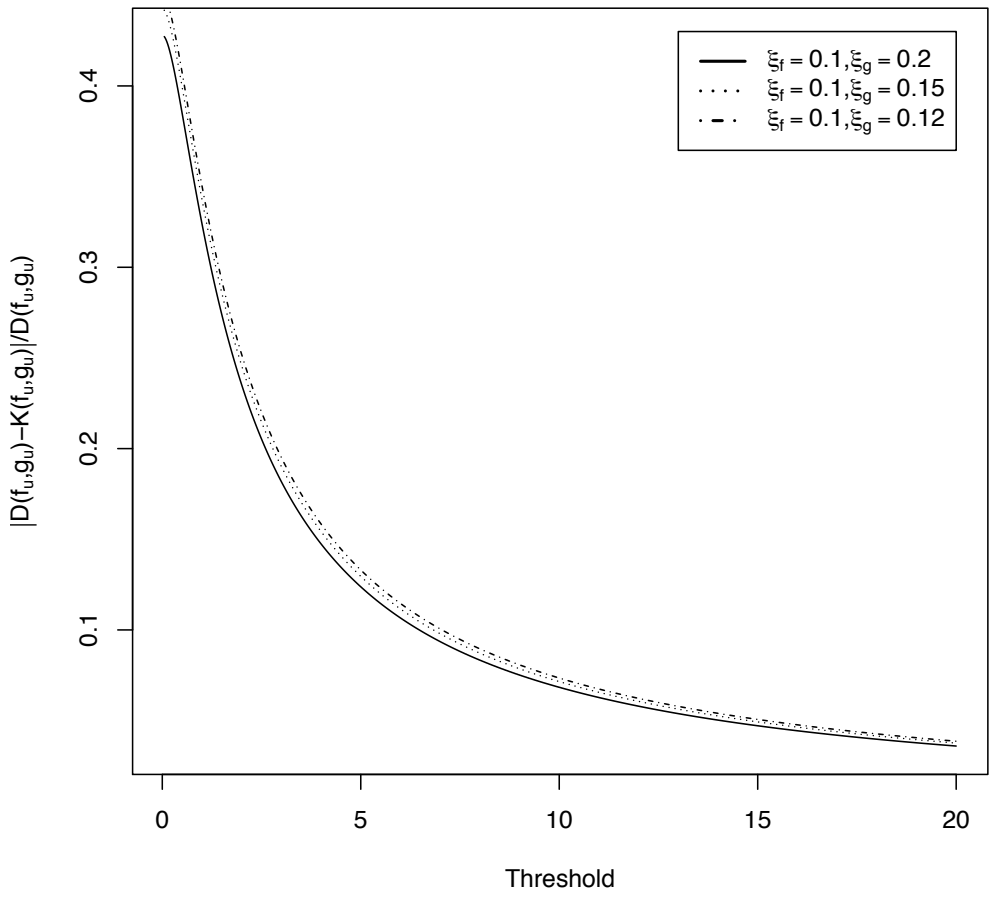

FIGURE 2.2 – Relative error between $K(f_u; g_u)$ and $D(f_u; g_u)$ in function of different thresholds, see Proposition 2.1 when f and g correspond to two GP densities with a unit scale parameter and different shape parameters $\xi_f = 0.1$ and $\xi_g = 0.2, 0.15$ and 0.12.

order condition (see e.g., de Haan and Stadtmüller, 1996)

$$\lim_{t\to\infty}\frac{\frac{\overline{F}(tx)}{\overline{F}(t)}-x^{-\alpha}}{q_F(t)}=x^{-\alpha}\frac{x^{\rho}-1}{\rho}\ \text{and}\ \lim_{t\to\infty}\frac{\frac{\overline{G}(tx)}{\overline{G}(t)}-x^{-\beta}}{q_G(t)}=x^{-\beta}\frac{x^{\eta}-1}{\eta},$$

for some $\rho<0$ *and* $\eta<0$ *and some functions* $q_F\neq 0$ *and* $q_G\neq 0$. *If the functions*

$$B(x)=\frac{xf(x)}{\overline{F}(x)}-\alpha\ \text{and}\ C(x)=\frac{xg(x)}{\overline{G}(x)}-\beta,$$

are eventually monotone, then condition (2.2) is satisfied.

Proposition 2.3 *Suppose that the random variables* X *and* Y *belong to the Weibull max-domain of attraction and have the same finite upper endpoints. Condition (2.2) is satisfied if the assumptions of Proposition 2.2 hold for the the tail functions* $\overline{F_*}(x):=\overline{F}\left(\tau-x^{-1}\right)$ *and* $\overline{G_*}\left(x\right):=\overline{G}\left(\tau-x^{-1}\right)$.

To treat the most classical densities belonging to the Gumbel domain of attraction, we need to recall the definition of the asymptotic stochastic ordering, (e.g., see Shaked and Shanthikumar, 1994). Having $X_u\geq_{st}Y_u$ for large u means $\mathbb{P}\left(X_u>t\right)\geq\mathbb{P}\left(Y_u>t\right)$, for all $t>u$ and for large u.

Proposition 2.4 *Suppose that* $X_u\geq_{st}Y_u$ *for large* u *and define*

$$\alpha(x)=\ln\left(\frac{f(x)}{\overline{F}(x)}\right)-\ln\left(\frac{g(x)}{\overline{G}(x)}\right).$$

If $\mathbb{E}(X_u)$, $\mathbb{E}(\alpha(X_u))$ *and* $\mathbb{E}(\alpha(Y_u))$ *are finite and the derivative* $\alpha'(\cdot)$ *is monotone and goes to zero as* $x\uparrow\tau$, *then (2.2) is satisfied.*

The proof of this proposition relies on a probabilistic version of the mean value theorem, (e.g., di Crescenzo, 1999). Applying this proposition can be straightforward in some important cases. For example, suppose that X and Y follow a standard exponential and standard normal distributions, respectively. We have $\mathbb{E}\left(X_u\right)=1+u$, $\mathbb{E}\left(Y_u\right)=\frac{\phi(u)}{\overline{\Phi}(u)}$ and $\alpha'\left(x\right)=x-\frac{\phi(x)}{\overline{\Phi}(x)}$ which is monotone and goes to zero as $x\uparrow\infty$. For large u, $X_u\geq_{st}Y_u$. Hence, condition (2.2) is satisfied.

Overall, condition (2.2) appears to be satisfied for most practical cases. In a nutshell, this condition tells us that our approximation can be used whenever the two densities of interest are comparable in their upper tails. For the few cases for which this condition is not satisfied, the discrepancy between the two tails is likely to be large and they can be easily handled for our climatological applications by just watching the two plotted time series under investigation and deduce that they are different.

2.3 Estimation of the divergence

In terms of inference, the key element given by Proposition 2.1 is captured by Equation (2.3). This expectation only depends on $\overline{G}(X)$ and consequently, one can easily plug in an empirical estimator of $\overline{G}$ to infer (2.3). More precisely, suppose that we have at our disposal two independent samples of size n and m, $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$. In our climate example, this could correspond to temperatures before 1980 and after 1980. To estimate $\overline{G}(t)$, we denote $\overline{G}_m(t) = \sum_{j=1}^m \mathbb{1}_{\{Y_j > t\}}/m$ the classical empirical tail. To avoid taking the logarithm of zero in $\ln \overline{G}(X)$, we slightly modify it by introducing

$$\widetilde{\overline{G}}_m(t) \;:=\; 1 - \frac{1}{m+1}\sum_{j=1}^{m} \mathbb{1}_{\{Y_j \leq t\}} = \frac{m}{m+1}\overline{G}_m(t) + \frac{1}{m+1}.$$

Our estimator of (2.3) is then simply defined by

$$\widehat{L}(f_u; g_u) = 1 + \frac{1}{N_n}\sum_{i=1}^{n} \ln\left(\frac{\widetilde{\overline{G}}_m(X_i \vee u)}{\widetilde{\overline{G}}_m(u)}\right), \text{ and } \widehat{K}(f_u; g_u) = -\widehat{L}(f_u; g_u) - \widehat{L}(g_u; f_u), \quad (2.4)$$

where N_n represents the number of data points above the threshold u in the sample $\mathbf{X}$. To avoid dividing by zero when calculating $\widehat{L}(f_u; g_u)$, we use the convention $0/0 = 0$ whenever N_n is equal to zero in (2.4). The estimator $\widehat{L}(f_u; g_u)$ is non-parametric and it has the advantage of being extremely fast to compute. Its asymptotic properties need to be derived. One non trivial element for this theoretical task comes from the mixing of the two samples in (2.4) that makes the random variables $\overline{G}_m(X_i \vee u)$ dependent.

Theorem 2.1 *Assume that F and G are continuous. Let $u < \tau$ fixed and suppose that the means*

$$\mathbb{E}_f\left(\ln\left(\frac{\overline{G}(X \vee u)}{\overline{G}(u)}\right)^2\right) \text{ and } \mathbb{E}_g\left(\ln\left(\frac{\overline{F}(Y \vee u)}{\overline{F}(u)}\right)^2\right)$$

are finite, $\frac{n}{m} \to c \in (0, \infty)$ and that there exists two non increasing sequences of positive numbers, k_n/n and ℓ_m/m, satisfying

$$k_n \geq \max\left(\ln n, 8n\overline{F}\left(\overline{G}^{\leftarrow}\left(\frac{\ell_m}{m}\right)\right)\right), \quad \frac{k_n}{n}\ln n \to 0 \quad \text{and} \quad \frac{\ell_m}{\ln\ln m} \to \infty.$$

Then we have

$$\widehat{L}(f_u; g_u) - L(f_u; g_u) = o(1) \text{ a.s.}$$

and

$$\widehat{K}(f_u; g_u) - K(f_u; g_u) = o(1) \quad a.s.$$

This theorem requires the existence of four sequences $k_n^{(1)}, \ell_m^{(1)}$ and $k_m^{(2)}, \ell_n^{(2)}$. In the specific case of strict Pareto tails $\overline{F}(x) = x^{-\alpha}$ and $\overline{G}(x) = x^{-\beta}$ with $\alpha, \beta > 0$, a possible choice for these sequences is

$$k_n^{(1)} = \frac{n}{\ln n \ln\ln n} \text{ and } \ell_m^{(1)} = \frac{m}{(8\ln m\ \ln\ln m)^{\beta/\alpha}}$$
$$k_m^{(2)} = \frac{m}{\ln m \ln\ln m} \text{ and } \ell_n^{(2)} = \frac{n}{(8\ln n\ \ln\ln n)^{\alpha/\beta}}.$$

2.4 Applications

2.4.1 Simulations

To compute $\widehat{K}(f_u; g_u)$ in (2.4) in our simulation study, we need to generate excesses from two different densities. A first choice is to choose two unit scale parameter GP densities with different shape parameters because we have the explicit expressions of $L(f; g)$, and $K(f; g)$ for such distributions (see Equation (2.1)). To explore the Fréchet case, we arbitrarily set $\xi_f = 0.15$. This corresponds to a typical value for daily precipitation

extremes (e.g., see Table 1 in Katz *et al.*, 2002). Concerning the shape parameter for g, it varies from $\xi_g = 0.05$ to 0.3. For each value of $\xi_f = 0.15$ and ξ_g, two samples with $m = n$ are simulated and our estimator $\widehat{K}(f_u; g_u)$ defined from (2.4) can be calculated. We repeat this experiment 500 times for three different sample sizes $n \in \{500, 1000, 5000\}$. The classical Mean Square Error (MSE) can be inferred from the 500 estimates of $\widehat{K}(f_u; g_u)$. The resulting MSE is plotted in the left panel of Figure 2.3. As expected, the MSE decreases as the sample size increases. The estimation of $K(f; g)$ improves whenever the two shape parameters are close to each other.

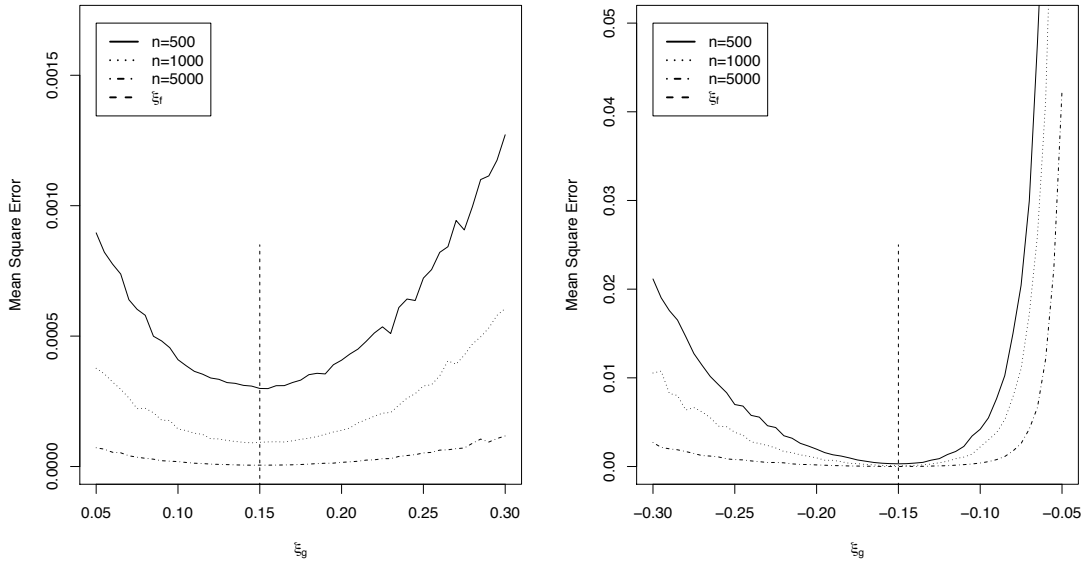

FIGURE 2.3 – Mean Square Error of $\widehat{K}(f_u; g_u)$ based on (2.4) and computed from 500 simulations of two GP distributed samples of sizes $n \in \{500, 1000, 5000\}$ and $n = m$. The left and right panels correspond to the Fréchet and Weibull cases, $\xi_f = 0.15$ and $\xi_f = -0.15$, respectively. The x-axis corresponds to different shape parameter values of ξ_g.

The same type of conclusions can be drawn for the Weibull domain. For this case, we set $\xi_f = -0.15$. Shape parameter values for temperature extremes usually belong to the

interval $[-0.3, -0.1]$ (e.g., see tables 6 and 7 in Jarušková and Rencová, 2008). In our simulations, ξ_g varies from -0.3 to -0.05 in the right panel of Figure 2.3. Overall, those MSE are small, but the two compared densities are GP distributed. To move away from this ideal situation, we keep the same GP density for $g(\cdot)$ with $\xi_g = 0.1$ (or $\xi_g = -0.1$ for the Weibull case) but f corresponds to a Burr with survival function $\overline{F}(x) = \left(\frac{1}{1+x^\tau}\right)^\lambda$ for $x > 0$ or to a reverse Burr defined by $\overline{F}(x) = \left(\frac{1}{1+(1-x)^{-\tau}}\right)^\lambda$ for $x < 1$. We fix $\lambda = 0.5$ and $\tau = 20$ in order to have $\xi = \frac{1}{\lambda\tau} = 0.1$ for the Burr and $\xi = -\frac{1}{\lambda\tau} = -0.1$ for the Reverse Burr. This design allows to assess the impact of the threshold choice that is represented in terms of quantiles on the x-axis of Figure 2.4.

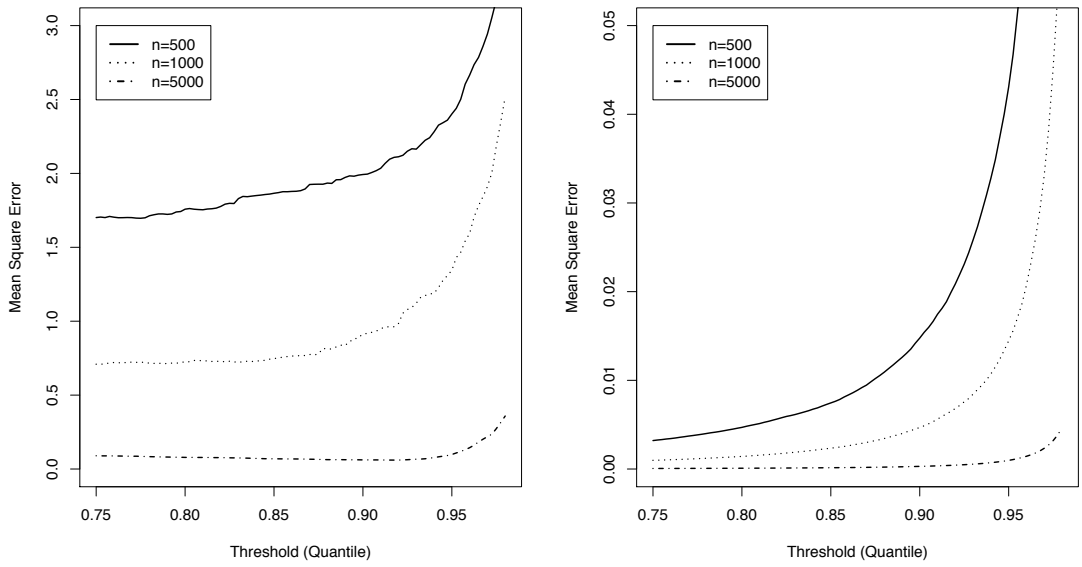

FIGURE 2.4 – Mean Square Error of $\widehat{K}(f_u; g_u)$ based on (2.4) and computed from 500 simulations with sizes $n \in \{500, 1000, 5000\}$ and $n = m$. The left panel compares a GP distribution with $\xi_f = 0.1$ and a Burr distribution. The right panel compares a GP distribution with $\xi_f = -0.1$ and a reverse Burr distribution. The x-axis corresponds to different thresholds (expressed as the mean of the 95% quantiles of the $\mathbf{X}$'s and $\mathbf{Y}$'s samples).

If the sample size is large, the MSE remains constant over a wide range of thresholds. For small sample sizes, special caution has to be applied for the heavy tail case. The right panel of Figure 2.4 dealing with the Weibull case clearly shows a rapid increase of the MSE for $n = 500$.

Concerning our main question on how to decide if two time series have different extremes, Table 2.2 displays the number of false positive (grey columns) and negative at a confidence level of 95% for four different GP distributions situations : the so-called Weibull-Weibull case ($\xi_f = -0.1$ and $\xi_g < 0$), the Fréchet-Fréchet case ($\xi_f = 0.1$ and $\xi_g > 0$), the Gumbel-Weibull case ($\xi_f = 0$ and $\xi_g < 0$) and the Gumbel-Fréchet case ($\xi_f = 0$ and $\xi_g > 0$). The scale parameter σ is taken as $-\xi$ in the Weibull-Weibull case (to make sure that the endpoints of the two distributions are the same) and set to one in all other cases. To see the influence of the sample sizes, $m = n$ can take five values (see the first column of Table 2.2). For each row of n and each column of ξ_g, the divergence $\widehat{K}(f_u; g_u)$ between the two samples is computed. To derive significance levels, we use a random permutation procedure (e.g., Davis *et al.*, 2012). By randomly permuting observations between two samples, the type I error under the null of no distributional difference can be controlled. Repeating 200 times this type of resampling leads to 200 values of $\widehat{K}_{H_0}(f_u; g_u)$ for which the 95% quantile level can be inferred. The original $\widehat{K}(f_u; g_u)$ can be compared to this quantile and a decision can then be made. The bold values in Table 2.2 correspond to the number of wrong decisions made by the test based on the divergence. As we have implemented our procedure on 1000 replicas, we expect on average to count 50 false positives at the 95% level, i.e. the grey column should contain a number close to 50. Outside of the grey columns, a value near zero indicates a good performance.

To benchmark our approach, we have also computed the classical non-parametric Kolmogorov-Smirnov, Wilcoxon rank sum and Wilcoxon signed rank tests (non bold values in Table 2.2) and they can be compared to our divergence based approach. Table 2.2 can teach us a few lessons. The Kolmogorov-Smirnov and Wilcoxon rank sum tests are

overall worse than the others, especially the Wilcoxon rank sum. The Wilcoxon signed rank test and our approach provide similar results for the Weibull-Weibull case. With a sample size of 200, one can distinguish $\xi_f = -0.1$ from $\xi_g \in \{-0.2, -0.05\}$, but not from $\xi_g \in \{-0.15, -0.08\}$, those values being too close to -0.1. For larger sample sizes, both approaches work well. The story is very different for the heavy tail case (Fréchet-Fréchet). One cannot expect differentiating $\xi_f = 0.1$ from any of the values of ξ_g of the table for small and moderate sample sizes. For $n = 10,000$, our divergence estimate is able to identify a difference when $\xi_g \in \{0.05, 0.15, 0.2\}$. This is not the case for the Wilcoxon signed rank test which is only able to detect a difference when $\xi_g = 0.2$. Finally, if we are in the Gumbel case ($\xi_f = 0$), the statistic $\widehat{K}(f_u; g_u)$ works adequately for $n = 200$ if $\xi_g \leq -0.3$. It is also the case for $n = 500$ and $\xi_g \geq 0.2$. In comparison, the Wilcoxon test has a much smaller validity range.

In summary, besides telling us that classical tests do not perform well but for Weibull-Weibull case, Table 2.2 emphasizes the difficulty of identifying small changes in non-negative shape parameters. For such a task, very large sample sizes are needed. Concerning our temperatures application, previous studies (e.g., Jarušková and Rencová, 2008) showed that the shape parameter of daily maxima is either below or equal to zero, i.e. we are in a Weibull-Weibull or a Gumbel-Weibull case. In the coming application section, we will deal with $3 \times 30 \times 90 = 8,100$ daily measurements per season (a season has three months, a month around thirty days and we have about 90 years of data for most stations). We will compare periods of 30 years and work with a 95% quantile threshold which will provide approximately 130 extremes per season for each period. According to Table 2.2, this will enable us to explore the Weibull-Weibull case and the Gumbel-Weibull case since the results given by the divergence-based test are acceptable in these two cases for this kind of sample size.

TABLE 2.2 – Number of false positive and negative out of 1000 replicas of two samples of sizes $n = m$ for a 95% level. The grey columns count the number of false positive (wrongly rejecting that f and g are equal). f and g correspond to a $GP(0, -\xi_f, \xi_f)$ and $GP(0, -\xi_g, \xi_g)$ density, respectively in the first row of the table and to a $GP(0, 1, \xi_f)$ and $GP(0, 1, \xi_g)$ density, respectively in the last three rows. The bold values correspond to the number of wrong decisions obtained with the divergence approach, and the three other columns correspond, from the left to the right, to the classical Kolmogorov-Smirnov, Wilcoxon rank sum, and Wilcoxon signed rank tests, respectively.

Weibull-Weibull case	$\xi_f = -0.1$																			
n \ ξ_g	-0.2				-0.15				-0.1				-0.08				-0.05			
50	262	168	138	**107**	691	605	563	**550**	26	41	48	**54**	889	845	841	**816**	241	153	123	**93**
100	42	16	11	**3**	445	318	266	**233**	50	70	69	**58**	824	739	719	**680**	24	15	11	**5**
200	0	0	0	**0**	110	56	32	**35**	30	49	47	**45**	627	530	476	**469**	0	0	0	**0**
500	0	0	0	**0**	0	0	0	**0**	52	57	51	**62**	189	131	95	**94**	0	0	0	**0**
1000	0	0	0	**0**	0	0	0	**0**	42	48	57	**50**	17	10	4	**5**	0	0	0	**0**
Fréchet-Fréchet case	$\xi_f = 0.1$																			
n \ ξ_g	0.05				0.08				0.1				0.15				0.2			
10	986	955	953	**946**	991	952	960	**939**	15	41	44	**49**	986	953	951	**954**	990	957	963	**950**
100	973	944	945	**935**	960	944	950	**941**	43	59	59	**49**	953	942	943	**930**	947	925	925	**899**
1000	937	923	901	**740**	946	957	955	**926**	45	34	34	**49**	943	931	908	**773**	854	835	740	**285**
10000	648	701	470	**9**	911	903	859	**597**	41	54	54	**55**	688	717	503	**16**	12	182	14	**0**
Gumbel-Weibull case	$\xi_f = 0$																			
n \ ξ_g	-0.5				-0.4				-0.3				-0.2				0			
50	804	806	680	**155**	868	847	757	**377**	921	893	847	**628**	949	934	912	**799**	28	45	50	**48**
100	513	635	390	**3**	727	730	556	**37**	896	853	743	**231**	933	906	862	**599**	38	53	45	**50**
200	63	332	104	**0**	348	534	280	**0**	709	722	502	**12**	888	870	777	**252**	37	48	52	**60**
500	0	28	0	**0**	1	159	19	**0**	134	424	138	**0**	616	712	490	**4**	58	55	52	**53**
1000	0	1	0	**0**	0	12	0	**0**	0	139	7	**0**	229	452	178	**0**	36	55	44	**54**
Gumbel-Fréchet case	$\xi_f = 0$																			
n \ ξ_g	0				0.2				0.3				0.4				0.5			
50	33	57	58	**63**	943	931	921	**853**	945	924	898	**746**	919	861	794	**589**	885	854	755	**473**
100	36	54	62	**74**	923	900	860	**695**	907	866	794	**470**	844	813	690	**253**	733	708	525	**100**
200	30	53	52	**57**	902	886	794	**412**	807	768	626	**120**	606	634	389	**18**	392	498	236	**1**
500	43	40	48	**55**	706	737	541	**52**	335	505	222	**0**	101	306	69	**0**	13	152	14	**0**
1000	45	40	51	**59**	423	541	262	**1**	34	213	31	**0**	0	71	0	**0**	0	10	0	**0**

2.4.2 Extreme temperatures

In geosciences, the yardstick period called a climatology is made of 30 years. So, we would like to know how temperature maxima climatologies have varied over different 30 year periods. To reach this goal, for any $t \in \{1, ..., 80\}$, we compare the period $[1900 + t, 1929 + t]$ with the current climatology [1981, 2010]. All our daily maxima and minima come from the ECA&D database (European Climate Assessment & Dataset project `http://eca.knmi.nl/dailydata/predefinedseries.php`). This database contains thousands of stations over Europe, but most measurement records are very short or incomplete and consequently, not adapted to the question of detecting changes in extremes. In this context, we only study stations that have at least 90 years of data, i.e. the black dots in Figure 2.1. As previously mentioned in the Introduction section, a smooth seasonal trend was removed in order to discard warming trends due to mean temperature changes. This was done by applying a classical smoothing spline with years as covariate for each season and station (R-package *mgcv*). The resulting trends appear to be coherent with mean temperature behaviour observed at the national and north-hemispheric levels (e.g., see Figure 5 in Abarca-Del-Rio and Mestre, 2006) : an overall warming trend with local changes around 1940 and around 1970.

As a threshold needs to be chosen, we set it as the mean of the 95% quantiles of the two climatologies of interest. We first focus on one single location, the Montsouris station in Paris, where daily maxima of temperatures have been recorded for at least one century. Figure 2.5 displays the estimated $\widehat{K}(f_u; g_u)$ on the y-axis and years on the x-axis with $t \in \{1, ..., 80\}$. We are going to use this example to explain how the grey circles on figures 2.6 and 2.7 have been obtained.

Similarly to our simulation study, a random permutation procedure with 200 replicas is run to derive the 95% confidence level. One slight difference with our simulation study is that instead of resampling days we have randomly resampled years in order to take care

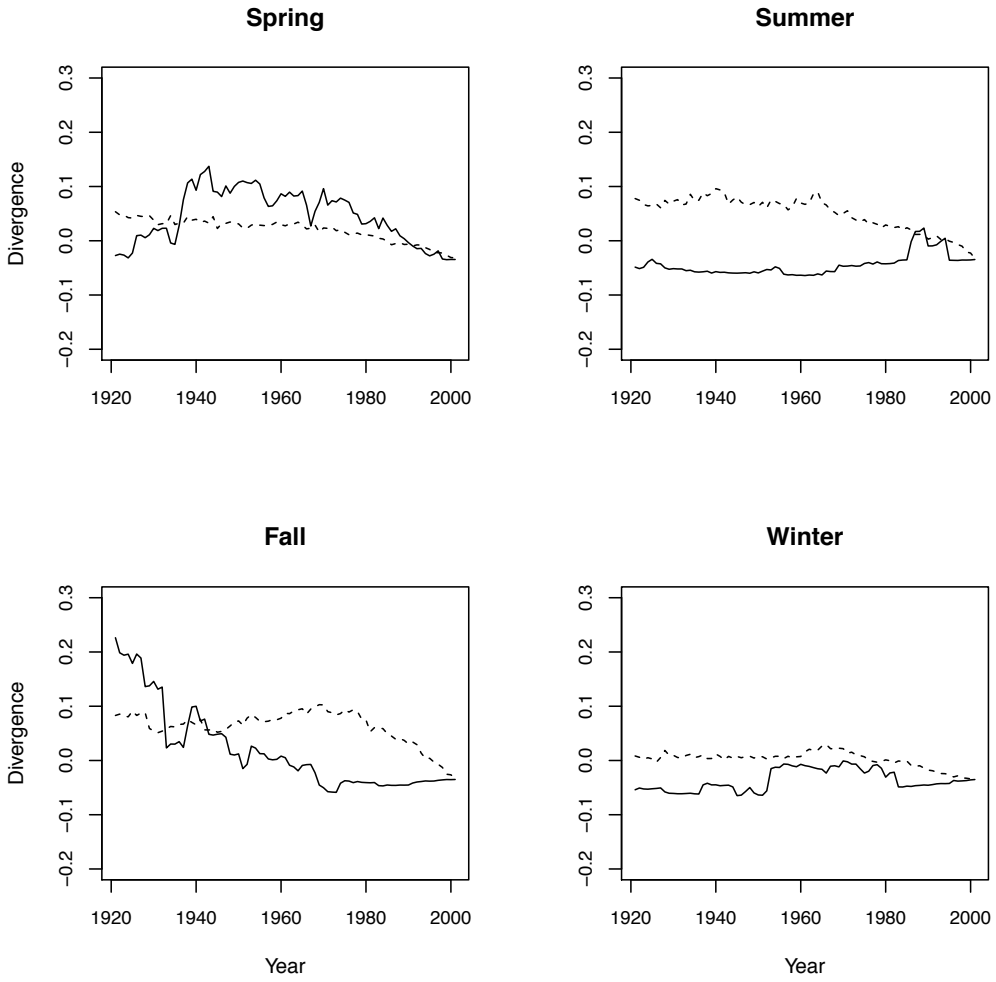

FIGURE 2.5 – Paris weather station : evolution of the divergence estimator (black curve), $\widehat{K}(f_u; g_u)$, in function of the years $[1900+t, 1929+t]$ with $t \in \{1, ..., 80\}$. The reference period is the current climatology, [1981, 2010]. The dotted line represents the 95% significant level obtained by a random permutation procedure.

of serial temporal correlations (it is unlikely that daily maxima are dependent from year to year). From Figure 2.5, the Fall season in Paris appears to be significantly different at the beginning of the 20th century than today. To quantify this information, it is easy to compute how long and how much the divergence is significantly positive. More precisely, we count the number of years for which $\widehat{K}(f_u; g_u)$ resides above the dotted line, and we sum up the divergence during those years (divided by the total number of years). Those two statistics can be derived for each station and for each season. In figures 2.6 and 2.7, the circle width and diameter correspond to the number of significant years and to the cumulative divergence over those years, respectively. For example, temperature maxima at the Montsouris station in Figure 2.5 often appear significantly different during Spring time (the border of the circle is thick in Figure 2.6) but the corresponding divergences are not very high on average. On the contrary, there are very few significant years during the Fall season, but the corresponding divergences are much higher (larger diameters with thinner border in Figure 2.6). This spatial representation tends to indicate that there are geographical and seasonal differences. For daily maxima, few locations witnessed significant changes in Summer. In contrast, the Winter season appears to have witnessed that extremes have changed during the last century. This is also true for the Spring season, but to a lesser degree. Daily minima divergences plotted in Figure 2.7 basically follows an opposite pattern, the Summer and Fall seasons appear to have undergone the most detectable changes.

2.5 Discussions

Recently, there have been a series of articles dealing with temperature extremes over Europe (e.g., Shaby and Reich, 2012 ; Jarušková and Rencová, 2008) and it is natural to wonder if our results differ from those past studies. The two main differences are the object of study and the variable of interest. Here, the latter corresponds to seasonal ex-

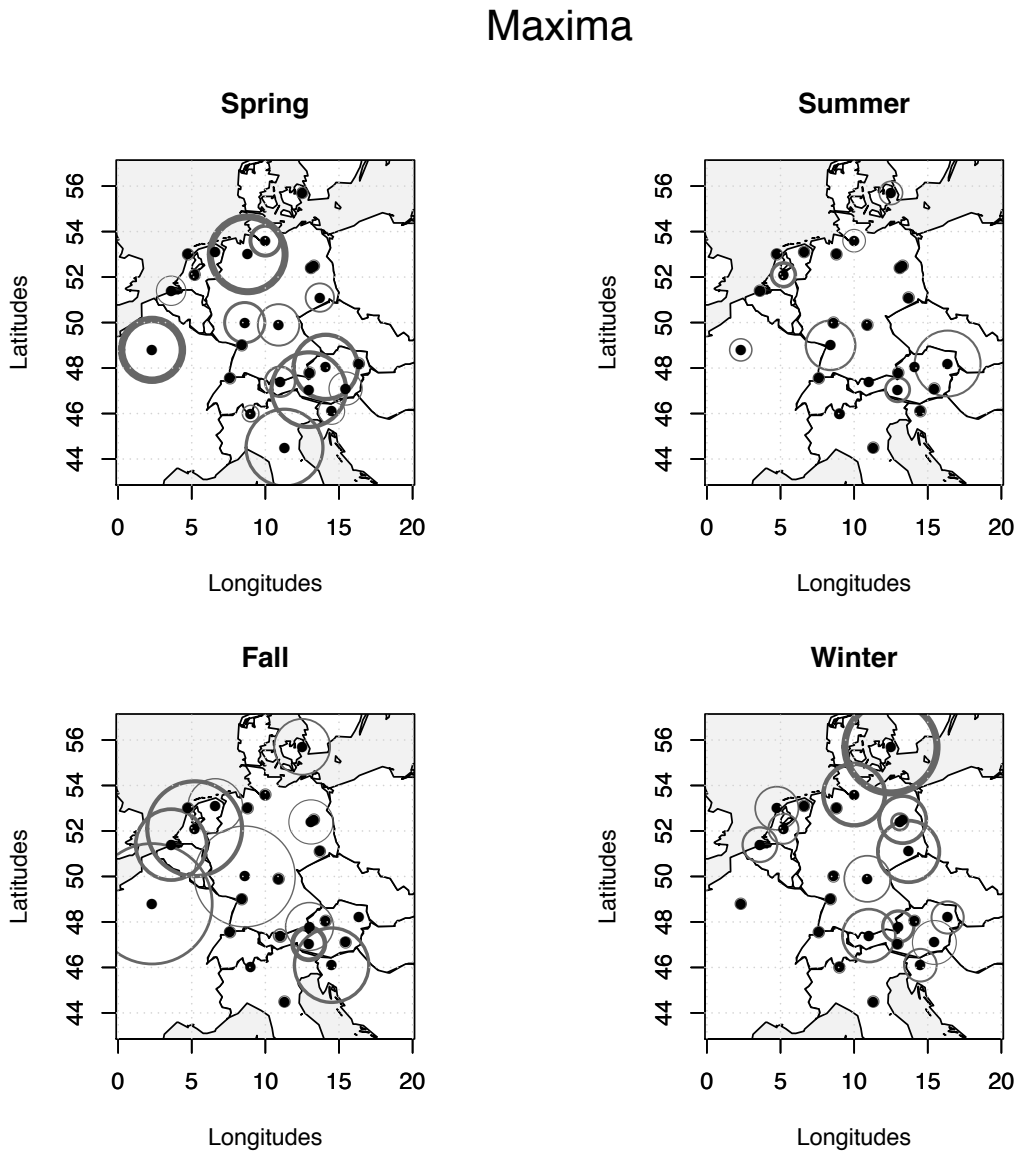

FIGURE 2.6 – The black dots represent the 24 locations described in Table 2.1 and come from the ECA&D database. The way the circles are built is explained in Section 4.2.

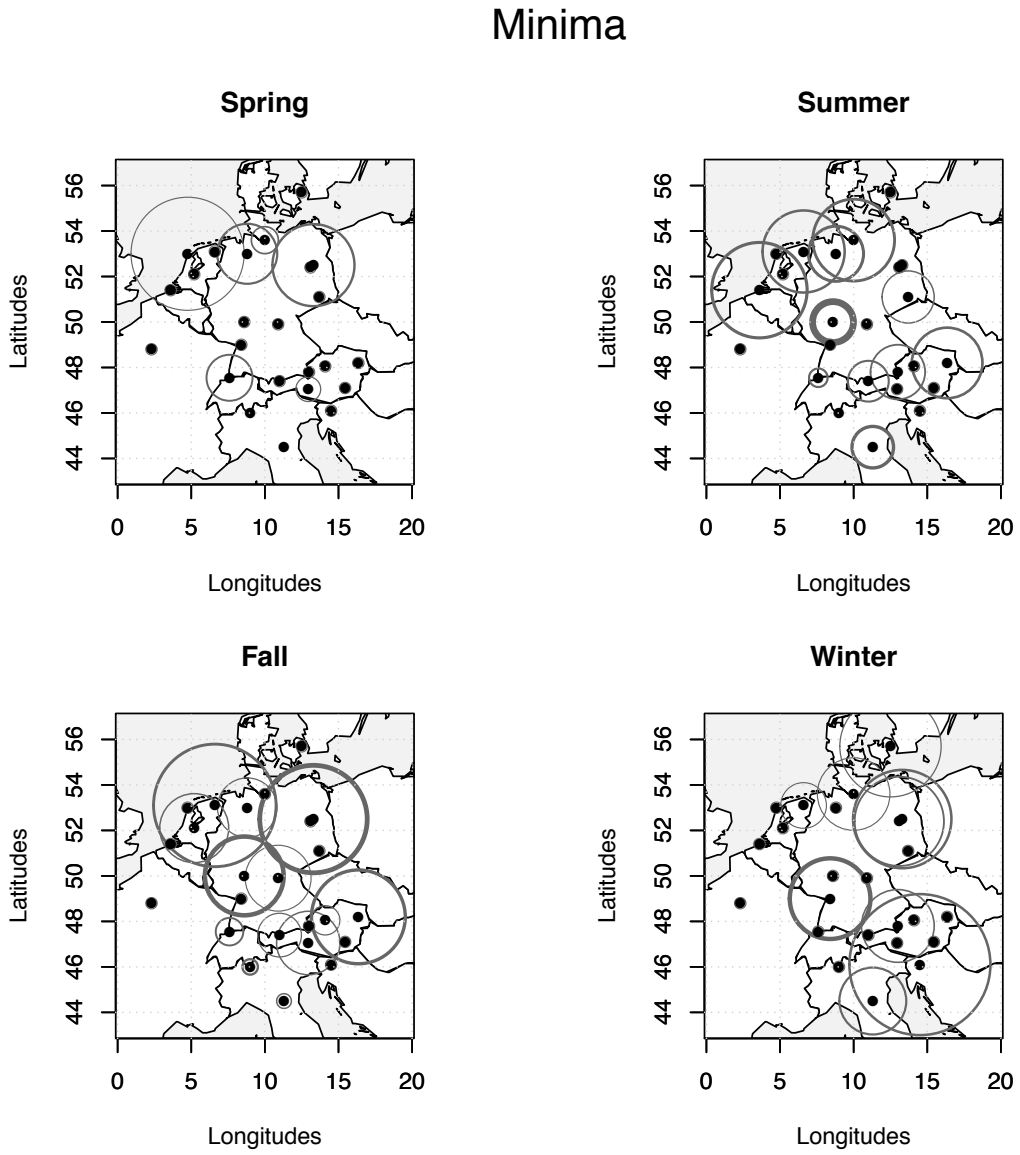

FIGURE 2.7 – Same as Figure 2.6, but for daily minima.

cesses obtained after removing a trend and the former focuses on determining if current excesses are different from the past ones (albeit the warming trend present in mean temperatures). Shaby and Reich (2012) aimed at a different objective. They solely focused on yearly maxima (not seasonal component) and took advantage of a flexible spatial max-stable model to pool information from around 1,000 stations (most of the sites appear after 1950 and this puts a stronger weight on the last 50 years). They found that "the distribution of extreme high temperatures seems to have shifted to the right (indicating warmer temperatures) in central Europe and Italy, while the distribution seems to have shifted to the left in Denmark, large swaths of Eastern Europe, and small pockets of Western Europe." It is not clear, if those shifts are due to changes in their GEV location parameters or to other alterations of the overall distribution shape. Hence, our study provides a complementary view by zooming on second order characteristics. Modifications of the distribution shape could have potentially dire straits consequences. We lack the spatial component for two reasons. Practically, stations with a very long record are very few and it is difficult to infer a reasonable spatial structure. Theoretically, statistical estimation techniques for excesses processes (e.g., Ferreira and de Haan, 2012) are still very rare, especially if we want to stay within a non-parametric framework. Future developments are needed to explore this theoretical and applied question.

Appendix

Proof of Proposition 2.1

For $x > u$, using the decomposition

$$\ln\left(\frac{f_u(x)}{g_u(x)}\right) = \ln\left(\frac{f(x)}{\overline{F}(x)}\right) + \ln\left(\frac{\overline{F}(x)}{\overline{F}(u)}\right) + \ln\left(\frac{\overline{G}(u)}{\overline{G}(x)}\right) + \ln\left(\frac{\overline{G}(x)}{g(x)}\right),$$

together with the fact that $\mathbb{E}_{f_u}\left\{\ln \overline{F}_u(X_u)\right\} = -1$, the Kullback Leibler distance $D(f_u; g_u)$ can be rewritten as

$$\begin{aligned} D(f_u; g_u) &= \int_u^\tau \left(\ln \frac{f(x)}{\overline{F}(x)} - \ln \frac{g(x)}{\overline{G}(x)}\right)(f_u(x) - g_u(x))\, dx - 2 \\ &\quad -\mathbb{E}_f\left\{\ln \frac{\overline{G}(X)}{\overline{G}(u)}\middle| X > u\right\} - \mathbb{E}_g\left\{\ln \frac{\overline{F}(Y)}{\overline{F}(u)}\middle| Y > u\right\}. \end{aligned}$$

Using (2.2), Proposition 2.1 follows. □

Proof of Proposition 2.2

Let $(u_n)_{n\in\mathbb{N}}$ be a sequence tending to infinity. We want to prove that

$$\int h_{u_n}(x)dx := \int \left(\ln \frac{xf(x)}{\alpha\overline{F}(x)} - \ln \frac{xg(x)}{\beta\overline{G}(x)}\right)\frac{f(x)}{\overline{F}(u_n)}\mathbb{1}_{\{x>u_n\}}dx \longrightarrow 0 \quad \text{as } n \to \infty.$$

Combining the remark following the proof of Theorem 1 in de Haan (1996) with the second order condition stated in Proposition 2.2 and the fact that $B(\cdot)$ and $C(\cdot)$ are eventually monotone, we deduce that these functions are of constant sign for large values of x, go to 0, and that their absolute value is regularly varying with index ρ (respectively η).

Thus

$$\begin{cases} \frac{xf(x)}{\alpha\overline{F}(x)} = 1 + x^\rho L_\rho(x) \\ \frac{xg(x)}{\beta\overline{G}(x)} = 1 + x^\eta L_\eta(x) \end{cases}$$

where $L_\rho(\cdot)$ and $L_\eta(\cdot)$ are two slowly varying functions. It is then clear that for $x > u_n$

$$h_{u_n}(x) \longrightarrow 0.$$

Now, remark that for a sufficiently large sequence u_n we have the bound

$$|h_{u_n}(x)| \leq x^\zeta L_\zeta(x)\frac{f(x)}{\overline{F}(x)}\mathbb{1}_{\{x>1\}} \leq Cx^{\zeta-1}L_\zeta(x)\mathbb{1}_{\{x>1\}}$$

where $\zeta < 0$ and C a suitable constant. Thus this bound is integrable.

Condition (2.2) follows by the dominated convergence theorem. □

Proof of Proposition 2.3

Similar to the proof of Proposition 2.2 with $\overline{F}$ and $\overline{G}$ replaced by $\overline{F_*}$ and $\overline{G_*}$ respectively.

□

Proof of Proposition 2.4

Note that

$$\Delta(u) := \int_u^\tau \alpha(x)\left(f_u(x) - g_u(x)\right)dx = \mathbb{E}\left[\alpha(X_u)\right] - \mathbb{E}\left[\alpha(Y_u)\right].$$

Moreover, the stochastic ordering implies that if $\mathbb{E}(X_u)$ and $\mathbb{E}(Y_u)$ exist, then we have the inequality

$$\mathbb{E}(X_u) \geq \mathbb{E}(Y_u).$$

Thus an application of a probabilistic version of the mean value theorem leads to

$$\Delta(u) = \mathbb{E}\left[\alpha'(Z_u)\right]\left\{\mathbb{E}(X_u) - \mathbb{E}(Y_u)\right\},$$

where Z_u corresponds to a non-negative random variable with density

$$f_{Z_u}(z) = \frac{\mathbb{P}\left[X > z \mid X > u\right] - \mathbb{P}\left[Y > z \mid Y > u\right]}{\mathbb{E}(X_u) - \mathbb{E}(Y_u)}, \quad \forall z > u$$

(see Theorem 4.1 in di Crescenzo, 1999). Note that we implicitly assume that $\mathbb{E}(X_u) > \mathbb{E}(Y_u)$. Otherwise Proposition 2.4 is trivial since $\mathbb{E}(X_u) = \mathbb{E}(Y_u)$ combined with $X_u \geq_{st} Y_u$ implies that $X_u = Y_u$ in distribution.

To conclude the proof, we only need to show that $\mathbb{E}[\alpha'(Z_u)] \to 0$ as $u \to \tau$. Since $\alpha'(\cdot)$ is monotone and tends to 0 at τ, $|\alpha'(\cdot)|$ is decreasing. Thus

$$\left|\mathbb{E}\left[\alpha'(Z_u)\right]\right| = \left|\int_u^\tau \alpha'(x) f_{Z_u}(x) dx\right| \leq \int_u^\tau \left|\alpha'(x) f_{Z_u}(x)\right| dx \leq \left|\alpha'(u)\right| \to 0.$$

The proof of Proposition 2.4 is then achieved. □

Proof of Theorem 2.1

We will use the notation

$$\widetilde{\overline{G}}(t) \;:=\; 1-\frac{n}{n+1}G(t)=\frac{n}{n+1}\overline{G}(t)+\frac{1}{n+1}.$$

We start by decomposing the difference $\widehat{L}(f_u;g_u)-L(f_u;g_u)$ into 6 terms :

$$\begin{aligned}
\widehat{L}(f_u;g_u)-L(f_u;g_u) &= \frac{1}{N_n}\sum_{i=1}^{n}\ln\left(\frac{\widetilde{\overline{G}}_m(X_i\vee u)}{\widetilde{\overline{G}}_m(u)}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X)}{\overline{G}(u)}\middle|X>u\right)\\
&= \frac{1}{N_n}\sum_{i=1}^{n}\ln\left(\frac{\widetilde{\overline{G}}(X_i\vee u)}{\widetilde{\overline{G}}(u)}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X)}{\overline{G}(u)}\middle|X>u\right)\\
&\quad+\frac{1}{N_n}\sum_{i=1}^{n}\ln\left(\frac{\widetilde{\overline{G}}_m(X_i\vee u)}{\widetilde{\overline{G}}(X_i\vee u)}\right)-\frac{n}{N_n}\ln\left(\frac{\widetilde{\overline{G}}_m(u)}{\widetilde{\overline{G}}(u)}\right)\\
&= \frac{1}{\overline{F}_n(u)}\left[\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{\widetilde{\overline{G}}(X_i\vee u)}{\widetilde{\overline{G}}(u)}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X\vee u)}{\overline{G}(u)}\right)\right]\\
&\quad+\left(\frac{1}{\overline{F}_n(u)}-\frac{1}{\overline{F}(u)}\right)\mathbb{E}_f\left(\ln\frac{\overline{G}(X\vee u)}{\overline{G}(u)}\right)\\
&\quad+\frac{1}{N_n}\sum_{i=1}^{n-k_n}\frac{\widetilde{\overline{G}}_m(X_{i,n}\vee u)-\widetilde{\overline{G}}(X_{i,n}\vee u)}{\widetilde{\overline{G}}(X_{i,n}\vee u)}\\
&\quad+\frac{1}{N_n}\sum_{i=1}^{n-k_n}\left[\ln\left(1+\frac{\widetilde{\overline{G}}_m(X_{i,n}\vee u)-\widetilde{\overline{G}}(X_{i,n}\vee u)}{\widetilde{\overline{G}}(X_{i,n}\vee u)}\right)-\frac{\widetilde{\overline{G}}_m(X_{i,n}\vee u)-\widetilde{\overline{G}}(X_{i,n}\vee u)}{\widetilde{\overline{G}}(X_{i,n}\vee u)}\right]\\
&\quad+\frac{1}{N_n}\sum_{i=n-k_n+1}^{n}\ln\left(\frac{\widetilde{\overline{G}}_m(X_{i,n}\vee u)}{\widetilde{\overline{G}}(X_{i,n}\vee u)}\right)\\
&\quad-\frac{n}{N_n}\ln\left(\frac{\widetilde{\overline{G}}_m(u)}{\widetilde{\overline{G}}(u)}\right)\\
&=: \sum_{\ell=1}^{6}Q_{\ell,n,m}.
\end{aligned}$$

We study each term separately.

Term $Q_{1,n,m}$.

$$\begin{aligned}
Q_{1,n,m} &= \frac{1}{\overline{F}_n(u)}\left[\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{\frac{n}{n+1}\overline{G}(X_i\vee u)+\frac{1}{n+1}}{\frac{n}{n+1}\overline{G}(u)+\frac{1}{n+1}}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X\vee u)}{\overline{G}(u)}\right)\right]\\
&= \frac{1}{\overline{F}_n(u)}\left[\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{\overline{G}(X_i\vee u)+\frac{1}{n}}{\overline{G}(u)+\frac{1}{n}}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X\vee u)}{\overline{G}(u)}\right)\right]\\
&= \frac{1}{\overline{F}_n(u)}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\ln\left(\frac{\overline{G}(X_i\vee u)+\frac{1}{n}}{\overline{G}(u)+\frac{1}{n}}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X\vee u)+\frac{1}{n}}{\overline{G}(u)+\frac{1}{n}}\right)\right\}\right]\\
&\quad+\frac{1}{\overline{F}_n(u)}\mathbb{E}_f\left(\ln\left(\frac{\overline{G}(X\vee u)+\frac{1}{n}}{\overline{G}(u)+\frac{1}{n}}\frac{\overline{G}(u)}{\overline{G}(X\vee u)}\right)\right)\\
&=: Q^{(1)}_{1,n,m}+Q^{(2)}_{1,n,m}.
\end{aligned}$$

Denote by

$$Z_i^{(n)} := \ln\left(\frac{\overline{G}(X_i\vee u)+\frac{1}{n}}{\overline{G}(u)+\frac{1}{n}}\right)-\mathbb{E}_f\left(\ln\frac{\overline{G}(X\vee u)+\frac{1}{n}}{\overline{G}(u)+\frac{1}{n}}\right).$$

Clearly $Z_i^{(n)}$ is an array of centred random variables that are identically distributed and rowwise independent. Thus, according to the strong law for arrays (see e.g., Chow and Teicher, 1978, p. 393), we have

$$\frac{1}{n}\sum_{i=1}^{n}Z_i^{(n)}=o(1)\text{ a.s.}$$

as soon as $\mathbb{E}_f(Z_i^{(1)})^2<\infty$. Now remark that $\ln\left(\frac{\overline{G}(u)+\frac{1}{n}}{\overline{G}(X\vee u)+\frac{1}{n}}\right)$ is a positive increasing function of n, thus

$$0\le\ln\left(\frac{\overline{G}(u)+1}{\overline{G}(X\vee u)+1}\right)\le\ln\left(\frac{\overline{G}(u)}{\overline{G}(X\vee u)}\right)\Longrightarrow\mathbb{E}_f(Z_i^{(1)})^2\le\mathbb{E}_f\left(\ln^2\left(\frac{\overline{G}(u)}{\overline{G}(X\vee u)}\right)\right)<\infty$$

by assumption. Consequently

$$Q^{(1)}_{1,n,m}=o(1)\text{ a.s.}$$

Now

$$Q^{(2)}_{1,n,m}=-\frac{1}{\overline{F}_n(u)}\left\{\mathbb{E}_f\left(\ln\left(\frac{\overline{G}(u)+\frac{1}{n}}{\overline{G}(X\vee u)+\frac{1}{n}}\right)\right)+\mathbb{E}_f\left(\ln\left(\frac{\overline{G}(X\vee u)}{\overline{G}(u)}\right)\right)\right\}.$$

Using again the fact that $\ln\left(\frac{\overline{G}(u)+\frac{1}{n}}{\overline{G}(X\vee u)+\frac{1}{n}}\right)$ is a positive increasing function of n, by the

dominated convergence theorem, we deduce that

$$\lim_n \mathbb{E}_f\left(\ln\left(\frac{\overline{G}(u)+\frac{1}{n}}{\overline{G}(X\vee u)+\frac{1}{n}}\right)\right) = \mathbb{E}_f\left(\lim_n\left(\ln\left(\frac{\overline{G}(u)+\frac{1}{n}}{\overline{G}(X\vee u)+\frac{1}{n}}\right)\right)\right) = \mathbb{E}_f\left(\ln\frac{\overline{G}(u)}{\overline{G}(X\vee u)}\right).$$

This implies that

$$Q^{(2)}_{1,n,m} = o(1) \text{ a.s.},$$

and thus

$$Q_{1,n,m} = o(1) \text{ a.s.}$$

Term $Q_{2,n,m}$.

By the strong law of large numbers, we have

$$Q_{2,n,m} = o(1) \text{ a.s.}$$

Term $Q_{3,n,m}$.

We need to use the sequences k_n and ℓ_m to treat this term. Since $k_n \geq \ln n$, $F^{\leftarrow}(1 - k_n/(8n))$ is eventually a.s. larger than $F_n^{\leftarrow}(1-k_n/n)$, hence than $X_{n-k_n,n}$.

Here is a quick way to see this : If we set $T_n = F^{\leftarrow}(1-p\varepsilon_n)$ for some $0<p<1$ and let 'Bin(r,q)' stand for a *binomial* (r,q) *random variable*, then the properties of quantile functions imply that $\mathbb{P}\{F_n^{\leftarrow}(1-\varepsilon_n) > F^{\leftarrow}(1-p\varepsilon_n)\} \leq \mathbb{P}\{\text{Bin}(n,p\varepsilon_n) > n\varepsilon_n\}$, which is dominated by $\left(\frac{enp\varepsilon_n}{n\varepsilon_n}\right)^{n\varepsilon_n} = (ep)^{n\varepsilon_n} = n^{n\varepsilon_n \ln(ep)/\ln(n)}$ (Giné and Zinn, 1984, Remark 4.7); if $p = 1/8$ and $n\varepsilon_n \geq \ln n$ then the series $\sum (ep)^{n\varepsilon_n}$ converges.

Thus, if we rewrite

$$\left|\frac{\widetilde{\overline{G}}_m(t)-\widetilde{\overline{G}}(t)}{\widetilde{\overline{G}}(t)}\right| = \frac{m}{n}\frac{n+1}{m+1}\frac{\overline{G}(t)}{\overline{G}(t)+\frac{1}{n}}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)} - \frac{1-\frac{n}{m}}{n+1}\frac{G(t)}{\overline{G}(t)}\right| \tag{2.5}$$

for n and m sufficiently large, we have

$$|Q_{3,n,m}| \;\leq\; \frac{2}{N_n}\sum_{i=1}^{n-k_n}\left|\frac{\overline{G}_m(X_{i,n}\vee u)-\overline{G}(X_{i,n}\vee u)}{\overline{G}(X_{i,n}\vee u)}\right| + \frac{2}{n+1}\left|1-\frac{n}{m}\right|\frac{1}{N_n}\sum_{i=1}^{n-k_n}\frac{1}{\overline{G}(X_{i,n}\vee u)}.$$

Remark now that

$$\begin{aligned}\frac{1}{N_n}\sum_{i=1}^{n-k_n}\frac{1}{\overline{G}(X_{i,n}\vee u)} &\leq \frac{1}{\overline{F}_n(u)}\frac{n-k_n}{n}\frac{1}{\overline{G}(X_{n-k_n,n}\vee u)}\\ &\leq \frac{1}{\overline{F}_n(u)}\frac{n-k_n}{n}\left\{\frac{1}{\overline{G}(X_{n-k_n,n})}+\frac{1}{\overline{G}(u)}\right\}\\ &\leq \frac{1}{\overline{F}_n(u)}\frac{n-k_n}{n}\left\{\frac{m}{\ell_m}+\frac{1}{\overline{G}(u)}\right\}\ a.s.\end{aligned}$$

Consequently, since $\frac{n}{m}\to c\in(0,\infty)$:

$$\begin{aligned}|Q_{3,n,m}| &\leq \frac{2}{N_n}\sum_{i=1}^{n-k_n}\left|\frac{\overline{G}_m(X_{i,n}\vee u)-\overline{G}(X_{i,n}\vee u)}{\overline{G}(X_{i,n}\vee u)}\right|+o(1)\ a.s.\\ &\leq \frac{n-k_n}{n}\frac{2}{\overline{F}_n(u)}\sup_{t\leq X_{n-k_n,n}\vee u}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right|+o(1)\ a.s.\\ &\leq \frac{n-k_n}{n}\frac{2}{\overline{F}_n(u)}\max\left(\sup_{t\leq X_{n-k_n,n}}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right|,\sup_{t\leq u}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right|\right)+o(1)\ a.s.\end{aligned}$$

Thus, by our choice of sequences k_n and ℓ_m, we have

$$\begin{aligned}\sup_{t\leq X_{n-k_n,n}}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right| &\leq \sup_{t\leq F^{\leftarrow}(1-k_n/(8n))}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right|\leq \sup_{t\leq\overline{G}^{\leftarrow}(\ell_m/m)}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right|\\ &= \sup_{t\geq\frac{\ell_m}{m}}\left|\frac{\mathcal{U}_m(t)-t}{t}\right|=o(1)\text{ a.s.}\end{aligned}$$

where $\mathcal{U}_m$ denotes the empirical distribution function of m uniform $(0,1)$ random variables (see Corollary 1 in Wellner, 1978).

Also $\forall\, T<\tau$, we have :

$$\sup_{t\leq T}\left|\frac{\overline{G}_m(t)-\overline{G}(t)}{\overline{G}(t)}\right|=o(1)\text{ a.s.}$$

which leads to

$$|Q_{3,n,m}|=o(1)\text{ a.s.}$$

Term $Q_{4,n,m}$.

Now, following the lines of proof of the term $Q_{3,n,m}$, for $i=1,...,k_n$, we have

$$\left|\frac{\widetilde{\overline{G}}_m(X_{i,n}\vee u)-\widetilde{\overline{G}}(X_{i,n}\vee u)}{\widetilde{\overline{G}}(X_{i,n}\vee u)}\right|=o(1)\text{ a.s.}$$

and thus using the inequality $\forall x \geq -\frac{1}{2}, |\ln(1+x) - x| \leq x^2$, we deduce that

$$|Q_{4,n,m}| \leq \frac{1}{N_n} \sum_{i=1}^{n-k_n} \left| \frac{\widetilde{\overline{G}}_m(X_{i,n} \vee u) - \widetilde{\overline{G}}(X_{i,n} \vee u)}{\widetilde{\overline{G}}(X_{i,n} \vee u)} \right|^2 = o(1) \text{ a.s.}$$

Term $Q_{5,n,m}$.

This term can be rewritten as :

$$Q_{5,n,m} = \frac{1}{N_n} \sum_{i=n-k_n+1}^{n} \ln \left(\frac{\frac{1}{m+1} + \frac{m}{m+1}\overline{G}_m(X_{i,n} \vee u)}{\frac{1}{n+1} + \frac{n}{n+1}\overline{G}(X_{i,n} \vee u)} \right).$$

Remark that

$$\frac{1}{m+1} \leq \frac{\frac{1}{m+1} + \frac{m}{m+1}\overline{G}_m(X_{i,n} \vee u)}{\frac{1}{n+1} + \frac{n}{n+1}\overline{G}(X_{i,n} \vee u)} \leq n+1$$

which implies that

$$|Q_{5,n,m}| \leq \frac{k_n}{n} \frac{1}{\overline{F}_n(u)} \max\left(\ln(n+1), \ln(m+1)\right) = o(1) \text{ a.s.}$$

Term $Q_{6,n,m}$.

Finally, remark that

$$Q_{6,n,m} = -\frac{1}{\overline{F}_n(u)} \ln \left(\frac{\frac{1}{m+1} + \frac{m}{m+1}\overline{G}_m(u)}{\frac{1}{n+1} + \frac{n}{n+1}\overline{G}(u)} \right) = -\frac{1}{\overline{F}_n(u)} \ln \left(\frac{m}{n} \frac{n+1}{m+1} \frac{\frac{1}{m} + \overline{G}_m(u)}{\frac{1}{n} + \overline{G}(u)} \right) = o(1) \text{ a.s.}$$

Combining all these results, Theorem 2.1 follows since

$$\widehat{K}(f_u; g_u) - K(f_u; g_u) = -\left(\widehat{L}(f_u; g_u) - L(f_u; g_u)\right) - \left(\widehat{L}(g_u; f_u) - L(g_u; f_u)\right).$$

□

Chapitre 3

Network design for heavy rainfall analysis

Abstract

The analysis of heavy rainfall distributional properties is a complex object of study in hydrology and climatology and it is essential for impact studies. In this chapter, we investigate the question of how to optimize the spatial design of a network of existing weather stations. Our main criterion for such an inquiry is the capability of the network to capture the statistical properties of heavy rainfall described by the Extreme Value Theory. We combine this theory with a machine learning algorithm based on neural networks and a Query By Committee approach. Our resulting algorithm is tested on simulated data and applied to high quality extreme daily precipitation measurements recorded in France at 331 weather stations during the time period 1980-2010.

3.1 Introduction

Weather, climate and hydrological extremes have always been of importance in human history. With our changing climate, there has been a growing research effort to understand, model and even predict extreme events at different time and spatial scales in atmospheric, hydrological and statistical sciences (e.g., Zwiers *et al.*, 2013). One driver for such a research endeavor resides in the increasing need of characterizing the frequency and intensity of extremes (see Alexander *et al.*, 2006 ; Groisman *et al.*, 2004). Such probabilistic knowledge is paramount for impact studies, assessment methods and adaptation strategies. In this context, high priority should be given to measuring uncertainties. This is directly linked to the issue, often overlooked in the statistical analysis of extreme events, of weather stations network design.

On the one hand, statisticians and climatologists would like to work with dense networks that have been measuring long atmospherical time series in order to accurately capture the spatial and temporal variability of extreme values. On the other hand, economical, regulatory and technical constraints demand a limited number of weather stations. Balancing those two opposing sides boils down to the question of how to design an optimal spatial network for extreme values.

Depending on financial resources and regulatory requirements, the goal of designing a network can be either defined as *augmenting* a network size to improve the spatial coverage or *reducing* it while maintaining the best spatial coverage. In this chapter, we discuss both cases but our application will only deal with the reduction size.

To highlight some of the statistical difficulties about thinning spatial networks for extremes, we study daily rainfall recorded by the national French weather service, Météo-France. When analyzing rainfall measurements, there is always a tradeoff between spatial coverage and time series length. If we had opted to analyze daily precipitation with at least fifty years of instrumental data, then only a few dozen stations with high quality

recordings would have been available. In this case, the question of reducing the network size would have been only theoretical with no practical value. Météo-France has the mandate to keep and maintain those high quality referenced stations. For this reason, we have decided to work on a shorter time period, 1980-2010, but with a large number of stations. With regards to extremes, this implies that fewer extremes in time will be analyzed and our design network has to take into account this higher estimation variability due to the reduced sample sizes. Figure 3.1 displays the sites of our 331 weather stations. Three different types of sites can be identified on this map. The set of dots represents 147 weather stations with top quality measurements. The second group of 110 triangles has also high quality recordings and rapid data quality checking. It will be kept for test and validation. The last group of 74 crosses is of lesser priority in the sense that data checking can take more than 48 hours. If Météo-France wants to know the loss of information induced by removing a few weather stations, it would certainly tap into the group $\mathcal{X}_+$. In contrast, stations in $\mathcal{X}_\circ$ and in $\mathcal{X}_\triangle$ correspond to reference points (airports, etc) and should not be candidates for removal. In Figure 3.1, we can observe that the sampling is not uniform in space. For example, the southern part along the Mediterranean sea has already a higher density of stations than the northern part along the Belgium border. This can be explained by atmospherical and geographical reasons. The southern part can witness complex weather systems linked to the pronounced orography (local effects) and coupled with a high population density (societal impacts). Extreme rainfall events there can be caused by southern winds, forcing warm and moist air to interact with mountainous areas, resulting in severe thunderstorms. Still, in terms of spatial design for extremes, is it better to remove stations in an already highly dense gauged region with very heavy rainfall or to discard an isolated weather station associated with less heavy precipitation ? This inquiry puts into light the two statistical tasks needed in our study : modeling the distributional properties of extreme precipitation and developing an algorithm to decide which stations could be removed without much information loss in terms of those characteristics of ex-

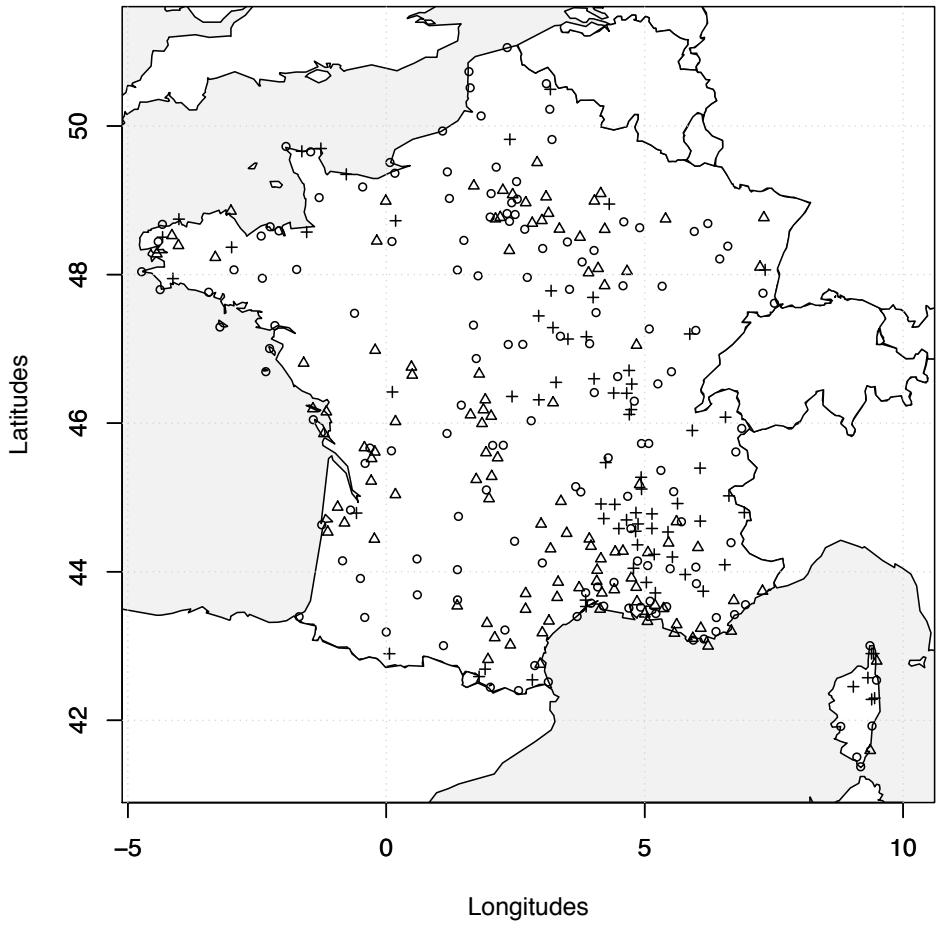

FIGURE 3.1 – Locations of 331 weather stations in France (source : Météo-France). The set $\mathcal{X}_\circ$ made of dots represents 147 weather stations with top quality measurements. The second group $\mathcal{X}_\triangle$ of triangles has also 110 stations of high quality recordings that we keep for test and validation. The last group $\mathcal{X}_+$ of 74 crosses is of lesser availability (data checking can take more than 48 hours) and contains the sites candidates for removal.

tremes. The first task is directly related to Extreme Value Theory (EVT) and the second one to the field of spatial network design.

For almost a century, extremal distribution features have been extensively studied by statisticians, climatologists and hydrologists through two main statistical quantities. The distribution of maxima over a block size (e.g., a year or a season) has been classically approximated by the so-called Generalized Extreme Value distribution (Fisher and Tippett, 1928) and excesses over a fixed high threshold have often been fitted by a Generalized Pareto (GP) distribution (see Pickands, 1975). In this work, we only focus on excesses above a threshold and we assume that they follow a GP distribution that is characterized by a scale and shape parameters, see Equation (3.1) for details.

In hydrology, the regional frequency analysis (RFA) described in Hosking and Wallis (1997) offers an interesting starting point to bring together GP distributions, heavy rainfall and spatial coherency. This RFA approach mainly consists in fitting a GP density at each location belonging to a climatically coherent region while the GP parameters are constrained over this region. In other words, a spatial structure is imposed on the GP parameters. This leads to the question on how to interpolate spatially GP parameters according to the network shown in Figure 3.1. Many avenues exist and we opt here for a well-known non-parametric regression method that will link the GP parameters with the geographical coordinates of the weather stations (the explanatory variables). The class of neural networks, adapted to spatial interpolation (Bishop, 2006 ; Ceresetti *et al.*, 2012) is a very flexible approach to model any continuous regression function. One essential reason for such a choice is that neural networks can be easily adapted to our problem of spatial network design.

The theory of spatial network design has been extensively studied in the literature (see Chapter 6 in Smith, 2001). At least two approaches can be distinguished. The first one is based on the theory of optimal design of experiments and has been initiated in the 50s (Kiefer and Wolfowitz, 1959), and applied to spatial statistics in the 80s (Fedorov and

Müller, 1989). It consists in fitting a model to the data and deciding to add stations at points where the variance of estimation is the highest (or to remove stations at which the variance of the model is the lowest). The second optimization method, based on the work of Caselton and Zidek (1984) is a Bayesian and information theoretic-based approach. In their work, they assume that observations come from an underlying random field with a multivariate normal distribution. The idea is to split the stations into two parts of fixed size. One of them corresponds to gauged sites and the other one to ungauged sites. The optimal design is the one for which the gauged sites bring the most information about the ungauged ones from an information-theoretical point of view. A Bayesian version of network design has been carefully investigated in environmental sciences, especially for dynamical monitoring networks (e.g., Berliner *et al.*, 1999 ; Nychka *et al.*, 1997 ; Wikle and Royle, 1999).

Those past approaches cannot be directly applied to our problem for two reasons. Firstly, we focus on extreme rainfall and their distributions may be skewed and heavy-tailed. Secondly, the spatial patterns of precipitation are complex and it is difficult to impose an explicit dynamical model to describe them. This implies that linear models or linear approximations of design networks (Cohn, 1996) may not be appropriate in our context of heavy rainfall.

To bypass the aforementioned limits, we will exploit the idea of Query by Committee (QBC) which is an algorithm that comes from machine learning. This algorithm was first introduced by Seung *et al.* (1992) for clustering problems and extended to the regression case by Krogh and Vedelsby (1995). It was originally introduced to add weather stations. The QBC is an iterative process that gradually adds locations in order to improve the quality of the dataset. It is particularly well suited in cases where new data are expensive to obtain and hence new observations have to be chosen carefully. We will adapt this method to handle the removal of weather stations.

This chapter is organized as follows. In Section 3.2 we quickly describe the extremal be-

havior of heavy rainfall in France. Section 3.3 presents the QBC in our spatial network context and its adaptation to extreme rainfall. The simulation study in Section 3.4 highlights the advantages, drawbacks and limitations of the QBC. In the last section, we apply our methodology to our French daily precipitation measurements.

3.2 Modeling of heavy rainfall

3.2.1 Extreme Value Theory

As already mentioned in the introduction, we are interested by rainfall excesses above a high threshold. As explained in the book of Coles (2001), EVT states that, if the threshold is high enough, the survival function, also called the tail distribution, can be well approximated by a GP tail defined as

$$\mathbb{P}(Z > z) = \begin{cases} \left(1 + \frac{\xi z}{\sigma}\right)^{1/\xi}, & \text{if } \xi \neq 0 \\ \exp\left(-\frac{z}{\sigma}\right), & \text{if } \xi = 0, \end{cases} \tag{3.1}$$

where the random variable Z represents thresholded rainfall excesses, $\sigma > 0$ is a scale parameter and ξ is a shape parameter. The distribution is defined for $z \in \mathbb{R}_+$ if $\xi \geq 0$ and for $z \in [0, -\sigma/\xi]$ if $\xi < 0$. The shape parameter governs the heaviness of the tail and is usually found to be positive when dealing with precipitation. Among the different estimators for ξ that have been proposed in the literature, two are widespread in hydrology and climatology, namely the Maximum Likelihood (ML) method (Smith, 1985) and the Probability Weighted Moments (PWM) method (see Hosking and Wallis, 1987). In this chapter, the estimates are derived by the Generalized Probability Weighted Moments (GPWM) method introduced by Diebolt *et al.* (2007) which is a refinement of the PWM one. Its use is motivated by a wider range of application ($\xi \in (-1, 3/2)$) compared to the PWM approach ($\xi \in (-1, 1/2)$). Unlike the ML estimator, it does not require any optimization and consequently, it is extremely fast and it does not provide divergent

values (this can happen when maximizing the likelihood). The scale and shape parameters estimates are simply defined as

$$\widehat{\sigma} = \frac{2.5\widehat{\mu}_1\widehat{\mu}_{1.5}}{2\widehat{\mu}_1 - 2.5\widehat{\mu}_{1.5}} \text{ and } \widehat{\xi} = \frac{4\widehat{\mu}_1 - (2.5)^2\widehat{\mu}_{1.5}}{2\widehat{\mu}_1 - 2.5\widehat{\mu}_{1.5}}, \tag{3.2}$$

where $\widehat{\mu}_s$ represents the empirical estimator of the GPWM $\mu_s = \mathbb{E}\left[Z\overline{G}^s_{\sigma,\xi}(Z)\right]$ with $\overline{G}_{\sigma,\xi}(z) = \mathbb{P}(Z > z)$.

To illustrate the spatial variability of those estimates with regards to our French rainfall data, the upper and lower left panels of Figure 3.2 display the estimated values of $\widehat{\sigma}$ and $\widehat{\xi}$ for the measurements taken from the network $\mathcal{X}_\circ$ shown in Figure 3.1. At each station, the threshold equals the 95% quantile after removing dry days. As expected for this country, the shape parameter roughly varies from zero to 0.5, this latter value occurring in the southern part of France. Note that the estimated scale parameters are strongly positive and consequently, the constraint $\sigma > 0$ will be satisfied for this example. Besides the analogies and the few differences between the two interpolated maps, it is clear that the spatial pattern of heavy rainfall in France is not uniform and local and complex features are present. This rapid and exploratory analysis leads to at least three questions : what is the uncertainty estimation for the two GP parameters ? are they correlated ? and how to model the spatial structures captured by each GP parameter ?

According to Diebolt *et al.* (2007), $\widehat{\xi}$ and $\widehat{\sigma}$ given by (3.2) are normally distributed with covariance matrix

$$\Sigma = \frac{1}{8(2\xi - 3)(4\xi - 7)} \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2} & \Sigma_{2,2} \end{pmatrix}, \tag{3.3}$$

with

$$\begin{aligned} \Sigma_{1,1} &= -(\xi - 2)(8\xi^2 - 10\xi + 13)\left(2\xi - 5\right)^2, \\ \Sigma_{1,2} &= -(2\xi - 5)(16\xi^3 - 92\xi^2 + 156\xi - 97)\sigma, \\ \Sigma_{2,2} &= \frac{(-32\xi^4 + 328\xi^3 - 1220\xi^2 + 1918\xi - 1093)\sigma^2}{(\xi - 2)}. \end{aligned}$$

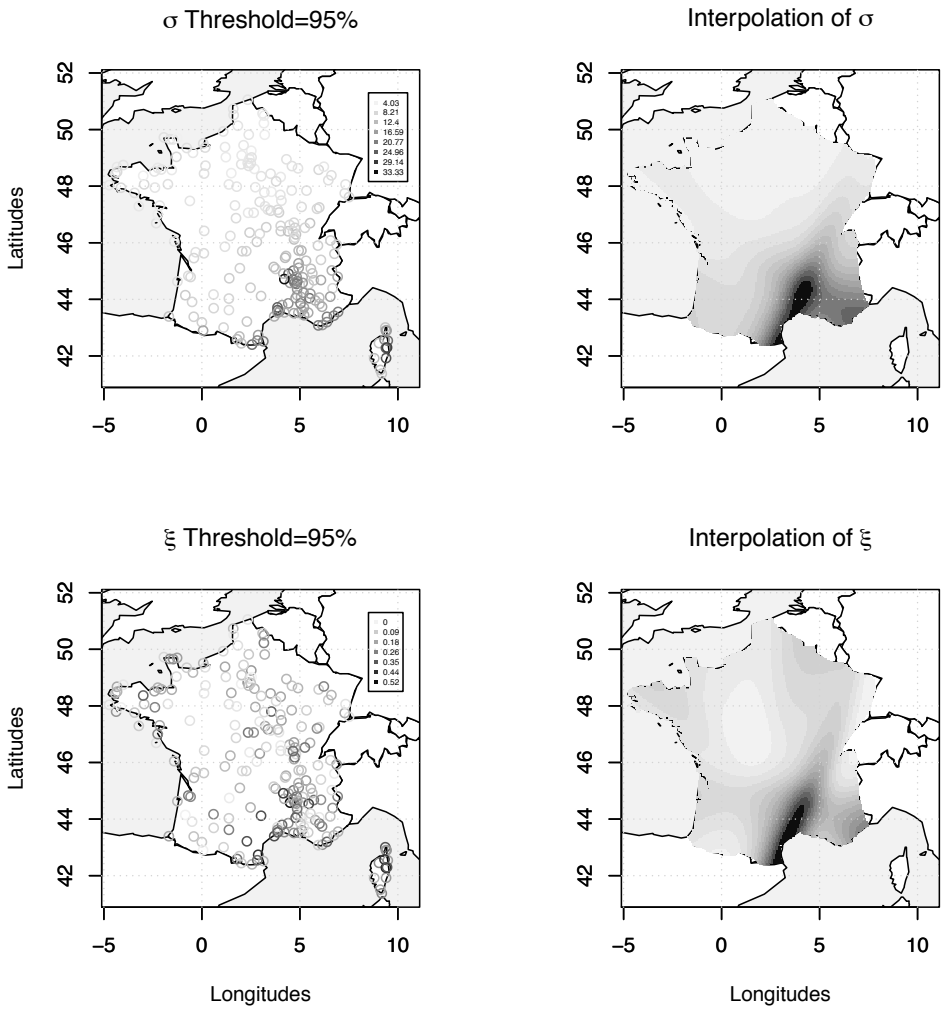

FIGURE 3.2 – Exploratory analysis of heavy rainfall in France. The left and right panels correspond to the individual and interpolated GP parameters obtained with (3.2), respectively. Each station belongs to the network $\mathcal{X}_\circ$ shown in Figure 3.1.

The matrix Σ is not diagonal and this means that the two estimators $\widehat{\sigma}$ and $\widehat{\xi}$ are correlated. To visualize this correlation effect, we implement a small simulation study by randomly generating 500 GP distributed samples of size 200, which is roughly the temporal length of our excesses dataset. The top left panel of Figure 3.3 shows the corresponding cloud of points for $\sigma = 25$ and $\xi = 0.5$, possible values for our rainfall application. As shown by this panel, this negative correlation effect is far from being negligible and we will have to take it into account in our design strategy. We also draw the boxplots corresponding to the relative estimation error for $\hat{\xi}$ and $\hat{\sigma}$. As expected in EVT, the shape parameter estimate has a larger variability than the inferred scale parameter. To assess the influence of the shape parameter on the correlation strength, the bottom panel displays the theoretical correlation between $\hat{\xi}$ and $\hat{\sigma}$ with $\sigma = 1$ and ξ varying from -0.5 to 1.5. Even the smallest absolute correlation value, at ξ around 1.0, is still large, around 0.5. In terms of spatial design, this implies that $\hat{\xi}$ and $\hat{\sigma}$ should be treated as a bivariate quantity.

3.2.2 Non-parametric regressions

As illustrated by the maps in Figure 3.2, it is not trivial to find some explanatory variables or parametric forms that could explain the spatial pattern of heavy rainfall in France. Hence, it makes sense to capture such complex spatial structures via a non-parametric approach. More precisely, we assume in the remaining of this chapter that the GP scale and shape parameters vary in space and should be viewed as follows

$$\begin{cases} \mathbb{R}^2 \to & \mathbb{R}_+^* \times [-1, 1.5] \\ \mathbf{x} \mapsto f(\mathbf{x}) = & (\sigma(\mathbf{x}), \xi(\mathbf{x})), \end{cases} \tag{3.4}$$

where $\mathbf{x} = (\text{lat}, \text{long})$ represents the latitude and longitude coordinates of the weather station $\mathbf{x}$. Basically, Diebolt *et al.* (2007) tells us that we can start our investigations

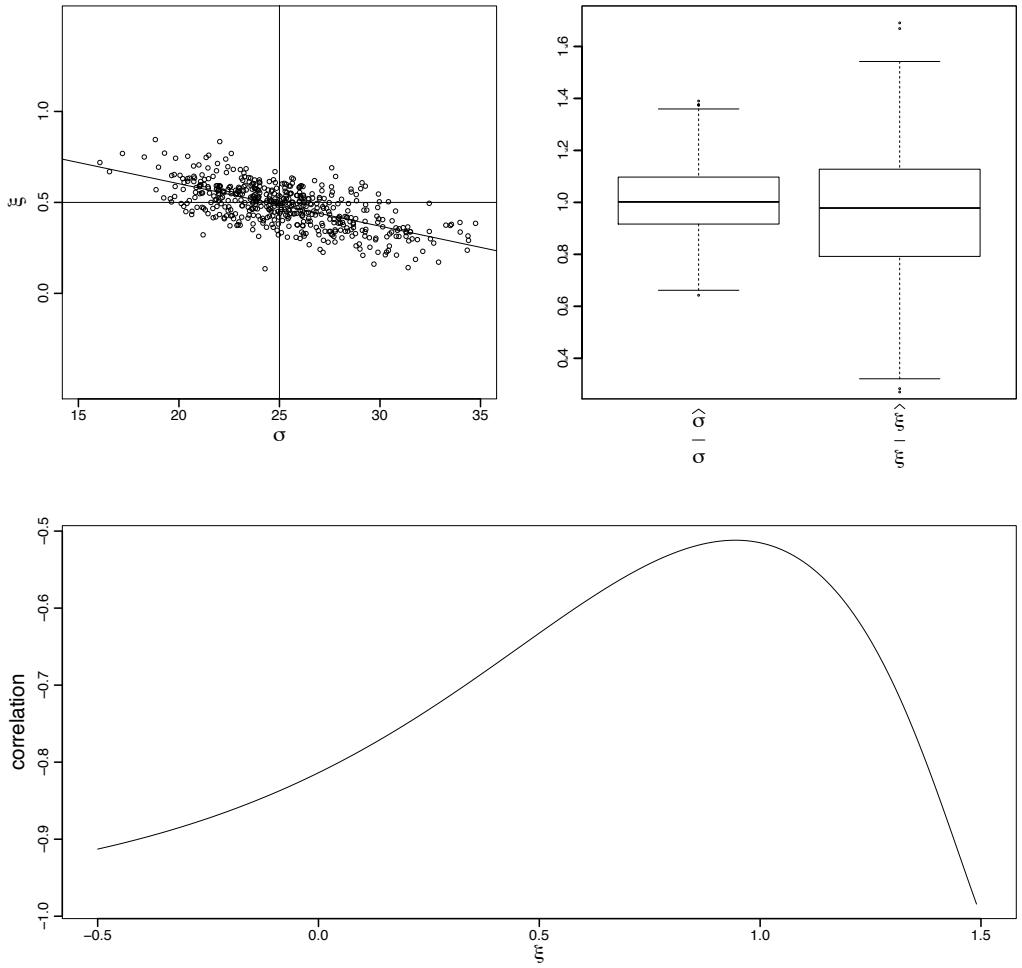

FIGURE 3.3 – Basic properties of the GP estimates defined by (3.2). The top left panel shows the strong negative correlation between 500 estimates of $\sigma = 25$ and $\xi = 0.5$ and sample size 200. The boxplots of the top right panel indicate the relative estimation error for $\hat{\xi}$ and $\hat{\sigma}$. The bottom panel displays the theoretical correlation between $\hat{\xi}$ and $\hat{\sigma}$ with $\sigma = 1$ and ξ varying from -0.5 to 1.5.

about spatial design with the simple non-parametric regression model

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x}),$$

where $\epsilon(\mathbf{x})$ corresponds to independent bivariate zero-mean Gaussian vectors with covariances defined from Equation (3.3). In this context, the vector $y(\mathbf{x}) = (\hat{\xi}(\mathbf{x}), \hat{\sigma}(\mathbf{x}))$ represents the estimates obtained from (3.2) at some selected weather station location $\mathbf{x}$. At this stage, we would like to emphasize a few points.

In contrast to classical regression problems, our "observational" vector $y(\mathbf{x})$ is not observed anymore. It represents the estimated GP parameters, so the temporal dimension has been squeezed into a distributional form.

As our heavy rainfall are assumed to be GP distributed, it is tempting to assume that those inferred GP parameters follow a bivariate Gaussian. But, as $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$ are allowed to vary in space, the two-dimensional covariance defined from Equation (3.3) will also vary in space. This brings additional modeling complexity. Instead of imposing the asymptotic covariance blueprint described by Equation (3.3), we prefer to integrate the uncertainty brought by the GP parameters estimation with resampling techniques, see Section 3.3.2.

It is also possible to add another level of complexity in the noise structure $\epsilon(\mathbf{x})$. For example, one could replace our noise independence assumption between $\epsilon(\mathbf{x}_j)$ and $\epsilon(\mathbf{x}_k)$ for $j \neq k$ by imposing a spatial covariance structure between the pair of stations (j, k). This is classically done in experimental design for linear models (e.g., Berliner *et al.*, 1999 ; Nychka *et al.*, 1997 ; Wikle and Royle, 1999). As $f(\mathbf{x})$ is considered non-parametric here, identifiability issues will raise quickly and we prefer to keep the assumption of independence in $\epsilon(\mathbf{x})$. In other words, we allow for a complex mean behavior in $f(\mathbf{x})$ and a simple covariance structure in $\epsilon(\mathbf{x})$, instead of a simple spatial trend and a complex covariance.

To estimate the unknown function $f(\mathbf{x})$ from $\{y(\mathbf{x}_i)\}$ with $i = 1, \ldots, n$, there exists a large

variety of non-parametric regression methods in the statistical literature (e.g., Hastie *et al.*, 2001). Keeping in mind our end goal (spatial design), we use one-hidden-layer neural networks with an hyperbolic tangent as activation function and two linear output neurons (two neurons because we jointly model the two GP parameters).

In a nutshell, a neural network scheme expresses the observations as an additive combination of non-linear blocks (here hyperbolic tangent functions that depend on the weather station latitude and longitude coordinates as explanatory variables), i.e.

$$\beta_0^{(H+1)} + \sum_{j=1}^{H} \beta_j^{(H+1)} \tanh\left(\beta_0^{(j)} + \beta_1^{(j)}\text{lat} + \beta_2^{(j)}\text{long}\right),$$

where H is called the number of hidden neurons and β represents the vector of parameters in $\mathbb{R}^{4H+1}$ to be estimated. The networks are optimized by minimizing a classical sum of squares between observed and predicted values. Among the different optimization routines available, we choose the standard backpropagation algorithm (see LeCun *et al.*, 1998). Concerning the number of hidden neurons H that characterizes the network complexity, we use a standard k-fold cross validation criterion (see Section 4.1 in Ceresetti *et al.*, 2012). As an example, the right panels of Figure 3.2 show one possible set of two maps obtained by a neural network with two linear outputs and jointly fitted to the two GP parameter values displayed on the left panels (those parameters have been centered and renormalized when fitting the neural network).

One of the features of neural networks that differentiates them from other interpolation techniques such as splines is that the optimization of the parameters converges to local minima. This means that a different initialization of the parameters will lead to a different network, probably with a different complexity. This drawback can be viewed as an asset in a spatial design context. It allows to generate a wide variety of regression functions for $f(\mathbf{x})$. The Query by Committee algorithm explained below takes advantage of this property.

3.3 Query by Committee and spatial design

The Query by Committee (QBC) approach is a machine learning algorithm introduced by Seung *et al.* (1992). It is based on neural networks and allows to create design of experiments. Its usefulness is emphasized in situations where it is expensive and time consuming to obtain new data (see Krogh and Vedelsby, 1995). We start by a presentation of its principle and theoretical basis.

The key ingredient of the QBC resides in its so-called committee of experts. This term simply means that m different experts are capable to produce m different models of the unknown function $f(\mathbf{x})$. This is very similar to Bayesian Model Averaging techniques (see e.g., Hoeting *et al.*, 1999 ; Sabourin *et al.*, 2013). Then, the idea is to search where these models disagree the most with respect to a common opinion. Mathematically, this latter term can be viewed as a weighted ensemble average (see Hansen and Salomon, 1990 ; Perrone and Cooper, 1993) defined by

$$\overline{f}(\mathbf{x}) = \sum_{k=1}^{m} p_k \hat{f}^{(k)}(\mathbf{x}), \tag{3.5}$$

where $(\hat{f}^{(1)}, \ldots, \hat{f}^{(m)})$ represents our m models (experts) and the non-negative weights p_k correspond to a priori knowledge about the quality of each expert. Here, we set them all at $1/m$.

A natural and important question is how to produce the committee of experts. As already mentioned, the optimization of neural networks converges to local minima and thus offers a simple way to build several different models with the same initial dataset. Another avenue that we will also use resides in the possibility to draw samples from our GPWM estimates, see Section 3.3.2.

To measure the disagreement among experts, we simply compute

$$\Delta(\mathbf{x}) = \frac{1}{m} \sum_{k=1}^{m} d(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})), \tag{3.6}$$

where $d(.,.)$ is a distance between the ensemble average and each individual member at

the site $\mathbf{x}$, see Section 3.3.1 for details. The QBC algorithm then advices to choose new weather stations at the local maxima of the disagreement function and starts the routine again with the updated datasets.

If s represents the number of stations that we want to add at each iteration, then the algorithm can be summed up as follows :

1. Take the initial training network $\mathcal{X}_\circ = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and its corresponding observational vector $\mathcal{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ with $\mathbf{y}_i = (\hat{\xi}(\mathbf{x}_i), \hat{\sigma}(\mathbf{x}_i))$.
2. Build a committee of m experts $(\hat{f}^{(1)}, \ldots, \hat{f}^{(m)})$ by regressing $\mathcal{Y}$ on $\mathcal{X}_\circ$ with m neural networks obtained by changing the initial conditions.
3. Find the s largest local maxima of the disagreement function $\Delta(\mathbf{x})$ among all experts over the whole territory and denote them by $\mathcal{X}' := (\mathbf{x}'_1, \ldots, \mathbf{x}'_s)$.
4. Create a new observational vector $\mathcal{Y}' = (\mathbf{y}'_1, \ldots, \mathbf{y}'_s)$ at the weather stations in $\mathcal{X}'$.
5. Add the elements of $\mathcal{X}'$ (resp. $\mathcal{Y}'$) to the initial training network $\mathcal{X}$ (resp. $\mathcal{Y}$).
6. Restart from Step 1 with the updated datasets until a stopping criterion is reached.

After each algorithmic loop, the average $\overline{f}(\mathbf{x})$ needed in Step 3 is re-computed with the updated datasets.

As already stated in the introduction, our main goal is to propose a design algorithm to reduce the number of stations. In its current form, the QBC method only adds weather stations. This issue can be solved by noticing that rainfall data are already available at the stations that could be removed. More precisely, the candidates for removal belong to the set $\mathcal{X}_+$ shown in Figure 3.1. Hence if we want to remove r stations from $\mathcal{X}_+$, we can progressively add stations from $\mathcal{X}_+$ to $\mathcal{X}_\circ$ until only r stations remain in $\mathcal{X}_+$. These last r stations will be considered the least informative and will be removed from the network. The choice of r, classically given by the decision maker, corresponds to our stopping criterion mentioned in Step 6.

To assess the quality of the final fit, we have kept the set $\mathcal{X}_{test}$ shown in Figure 3.1 for validation. The generalization error can be derived as

$$\Delta^2_{test} = \sum_{\mathbf{x}_i \in \mathcal{X}_{\triangle}} d(\hat{f}(\mathbf{x}_i), \overline{f}(\mathbf{x}_i)), \tag{3.7}$$

where $\hat{f}(\mathbf{x}_i)$ represents the estimate inferred at the location $\mathbf{x}_i$.

3.3.1 Choice of the disagreement function

An essential feature of the QBC resides in the choice of the distance $d(.,.)$ to compute the disagreement function defined by (3.6). If one is only interested in the shape parameter ξ, a pertinent choice could be

$$d_\xi(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})) = \left(\hat{\xi}^{(k)}(\mathbf{x}) - \overline{\xi}(\mathbf{x})\right)^2.$$

Similarly, optimizing the network with respect to σ could be done by opting for

$$d_\sigma(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})) = \left(\hat{\sigma}^{(k)}(\mathbf{x}) - \overline{\sigma}(\mathbf{x})\right)^2.$$

In hydrology, estimating high return levels represents a classical output and, in this context, a valuable distance could be

$$d_p(\hat{f}^{(k)}(\mathbf{x}), \overline{f}(\mathbf{x})) = \left(\widehat{q}_p^{(k)}(\mathbf{x}) - \overline{q}_p(\mathbf{x})\right)^2,$$

where $p \in (0,1)$ and the GP quantile is obtained from

$$\widehat{q}_p^{(k)}(\mathbf{x}) = \left((1-p)^{-\widehat{\xi}^{(k)}(\mathbf{x})} - 1\right) \frac{\widehat{\sigma}^{(k)}(\mathbf{x})}{\widehat{\xi}^{(k)}(\mathbf{x})}.$$

In the remaining of the chapter, we focus on the three distances : $d_\sigma(.,.)$, $d_\xi(.,.)$ and $d_p(.,.)$. To conclude this section, we note that the function $d(.,.)$ is called at two different places, for computing $\Delta(\mathbf{x})$ and calculating Δ_{test}. In terms of interpolation quality, i.e. goodness of fit, the objects of interest are the scale and shape surfaces. Hence, $d_\sigma(.,.)$ and $d_\xi(.,.)$ should be used in Δ_{test}, but the distance $d_p(.,.)$ does not give direct information about the two surfaces. Comparing experts is a different task and $d_p(.,.)$ can be plugged

in $\Delta(\mathbf{x})$.

3.3.2 GP parameters variability

As noted in Section 3.2.2, our "observational" vector $y(\mathbf{x}) = (\hat{\xi}(\mathbf{x}), \hat{\sigma}(\mathbf{x}))$ does not correspond to rainfall intensity but to estimated GP parameters. As GPWM are inferred quantities, it is possible to draw many realizations of shape and scale parameters in order to capture some variability inherent to transforming rainfall excesses into GP parameters. For each station $\mathbf{x}_i$, this means that we compute the GPWM estimates with the corresponding sample of excesses of length, say n_i. This gives us a pair $(\widehat{\sigma}(\mathbf{x}_i), \widehat{\xi}(\mathbf{x}_i))$ of estimates. Then we simulate ℓ samples of length n_i from a $GP(\widehat{\sigma}(\mathbf{x}_i), \widehat{\xi}(\mathbf{x}_i))$ distribution. The GPWM method is applied on each sample to provide ℓ shape and scale estimates. Those ℓ shape and scale parameters available at each station will be used as the initial input of our QBC algorithm.

Another approach to resample GP parameters could be to use the asymptotic Gaussian approximation proposed by Diebolt *et al.* (2007), see Equation (3.3). This will be faster but not much, the GPWM equations defined by (3.2) being explicit. By avoiding the asymptotic Gaussian approximation, we don't have to assume that the GP parameters estimates are perfectly normally distributed and consequently, we explore more accurately the uncertainties linked to the estimation with the GPWM method (the scatterplot in the upper left panel of Figure 3.3 may not be exactly Gaussian).

3.3.3 QBC for heavy rainfall

Before implementing the main steps of our modified QBC algorithm, the practitioner needs to make a few decisions.

One disagreement function $d(.,.)$ has to be selected. The total number of stations, say r, that should be removed from the set $\mathcal{X}_+$ has to be fixed. If we don't know how many sites

have to be disregarded, then r has to be estimated by computing a stopping criterion defined by Δ_{test} (e.g., see Gilardi, 2004).

The thinning of the network is based on adding progressively sites and, at the end of the process, removing the stations "that have not been added". Depending on the computational capacities available, the number of stations added at each QBC iteration, say s (or equivalently the number of iterations) has to be chosen. Adding only one station at each iteration, i.e. $s = 1$, is ideal, but it is the most expensive in terms of computational costs since the committee of experts has to be reconstructed by fitting m neural networks at each iteration.

The selection of m, the number of experts, and of ℓ, the number of replicas for exploring the GPWM estimates variability, also corresponds to a tradeoff between computational constraints and sampling quality.

All parameters being specified, we can summarize our modified QBC algorithm :

1. Compute the GP parameters with (3.2) at each station.
2. Apply the resampling technique explained in Section 3.3.2 to obtain ℓ observational vectors $\mathcal{Y}_j = (\mathbf{y}_{1,j}, \ldots, \mathbf{y}_{n,j})$ where $\mathbf{y}_{i,j} = (\hat{\xi}_j(\mathbf{x}_i), \hat{\sigma}_j(\mathbf{x}_i))$ with $j = 1, \ldots, \ell$ and $\mathcal{X}_\circ = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
3. For $j = 1, \ldots, \ell$, implement the modified QBC for each $\mathcal{Y}_j$, i.e.
   i. Take the initial training network $\mathcal{X}_\circ = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and its corresponding observational vector $\mathcal{Y}_j = (\mathbf{y}_{1,j}, \ldots, \mathbf{y}_{n,j})$.
   ii. Build a committee of m experts $(\hat{f}_j^{(1)}, \ldots, \hat{f}_j^{(m)})$ by regressing $\mathcal{Y}_j$ on $\mathcal{X}_\circ$ with m neural networks obtained by changing the initial conditions.
   iii. Find the s largest local maxima of $\Delta(\mathbf{x})$ that belong to $\mathcal{X}_+$ and denote them by $\mathcal{X}' := (\mathbf{x}'_1, \ldots, \mathbf{x}'_s) \subset \mathcal{X}_+$.
   iv. Create a new observational vector $\mathcal{Y}' = (\mathbf{y}'_1, \ldots, \mathbf{y}'_s)$ at the weather stations in $\mathcal{X}'$.

v. Add the elements of $\mathcal{X}'$ (resp. $\mathcal{Y}'$) to the initial training network $\mathcal{X}$ (resp. $\mathcal{Y}$).

vi. Restart from step (i) until no more than r stations can be added.

vii. Store the sites in $\mathcal{X}_+$ that have not been added in Step (vi).

4. Plot the frequency of the weather stations in $\mathcal{X}_+$ that have been stored in Step (vii).

From a statistical perspective, the final product of this experiment can be understood as a Binomial trial. We have a set of $dim(\mathcal{X}_+)$ individuals in $\mathcal{X}_+$ that can belong to two different groups of size r (stations removed) and $dim(\mathcal{X}_+) - r$ (stations kept), respectively. If all stations contain the same information (the null hypothesis), then the number of times, say $N(\mathbf{x})$, that a site $\mathbf{x}$ should be removed in ℓ trials follows a Binomial distribution

$$\mathbb{P}(N(\mathbf{x}) = i) = \binom{\ell}{i} (p(\mathbf{x}))^i (1 - p(\mathbf{x}))^{\ell - i}, \tag{3.8}$$

with

$$p(\mathbf{x}) = \frac{r}{dim(\mathcal{X}_+)} \text{ and } i = 1, \ldots, \ell.$$

For each station, we can identify if we fail to reject the null hypothesis ($p(\mathbf{x})$ does not vary with $\mathbf{x}$) for a given significant level, say 0.95.

3.4 Simulation study

In this simulation section, we mainly investigate the impacts of the disagreement functions, among four possibilities : $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a completely random disagreement.

To simplify tables and figures, we work with one dimensional simulations, i.e. $\mathbf{x} \in [-1, 1]$ and focus on two cases. The first case deals with a constant scale parameter and a varying shape parameter

$$\sigma(x) = 30 \text{ and } \xi(x) = \frac{1}{2}\left(1.3 - \exp\left(-16x^2\right)\right) - 0.1, \tag{3.9}$$

see the solid line in Figure 3.4. The second case corresponds to a dual situation, a constant

shape parameter and a varying scale parameter

$$\sigma(x) = 20\left(1.3 - \exp\left(-16x^2\right)\right) \text{ and } \xi(x) = 0.5, \tag{3.10}$$

see the solid line in Figure 3.5. In both cases, the varying functions are mostly constant but with a sharp drop around 0.5. If weather stations are uniformly distributed along the segment $[-1, 1]$, a well adjusted optimal network should keep its stations around $x = 0.5$ and propose to remove stations in regions where the parameters are fairly constant. Other setups with different varying shape parameters variations have also been studied and are available upon request (basically, our main conclusions remain identical).

Concerning the other parameters, we consider a set of $n = 55$ stations randomly drawn on $[-1, 1]$ and divided into our three groups with $dim(\mathcal{X}_\circ) = dim(\mathcal{X}_+) = 20$ and $dim(\mathcal{X}_\triangle) = 15$. Our goal is to remove $r = 5$ stations, i.e. to add $20 - 5 = 15$ stations that belong to $\mathcal{X}_+$. We set $s = 3$ to add three stations at each iteration. The number of experts m and the resampling of GP estimates ℓ are set to $m = \ell = 100$.

In Figure 3.4, the dots along the x-axis correspond to the 20 sites in $\mathcal{X}_\circ$. For this particular random draw, we can see that there are a little bit more sites on the interval $[0, 1]$ than on $[-1, 0]$. Consequently, it would make sense to remove stations on already dense segments. The set of stations which are candidates for removal is represented by 20 vertical grey lines, i.e. points in $\mathcal{X}_+$. From top to bottom, the four horizontally distinct groups of black segments correspond to the x-coordinates of stations removed by the QBC with respect to our four disagreement functions, $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a completely random one, respectively.

As expected, the completely random disagreement function produces the bottom group of black segments without any crosses. All grey lines are chosen with the same probability and the shape parameter fluctuations are not taken into account. In comparison, the three other disagreement functions do a much better job. For instance, no significant grey lines marked with a cross are around $x = 0$ where there is a lot of change in $\xi(x)$ and few

stations in $\mathcal{X}_\circ$, information is precious here and no stations should be removed in this vicinity. In addition, the asymmetrical distribution of circles belonging to $\mathcal{X}_\circ$ has been taken into account. Many more stations in $\mathcal{X}_+$ are removed on the right side of $x = 0$ than on the left one. Still, slight differences exist between the QBC outputs based on $d_\sigma(.,.)$, $d_\xi(.,.)$, or $d_p(.,.)$. In particular, $d_\xi(.,.)$ appears to capture well the two facts that a very high density of sites in $\mathcal{X}_\circ$ are located at $x \in [0.5, 1]$ and the shape parameter is constant over this region. If five stations have to be removed, the network should be thinned around this area. This feature is not as clear with $d_p(.,.)$, and to a lesser degree with $d_\sigma(.,.)$.

The same type of interpretation can be undertaken if we replace model (3.9) by model (3.10), see Figure 3.5. One particular aspect of this graph is that the random draw of sites in $\mathcal{X}_\circ$ had put three sites in a tiny neighborhood around $x = 0.4$. But no station around this point was removed with the distance $d_\sigma(.,.)$ and $d_\xi(.,.)$. This phenomenon can be explained by the sharp gradient in $\sigma(x)$ and the low number of points in $\mathcal{X}_+$ in the interval $[0, 0.5]$. One can also compare with the neighborhood of also three sites around $x = -0.6$. This is reassuring because it emphasizes that the density network cannot be the sole criterion for removal. Our QBC algorithm attempts to balance three constraints : the smoothness of the GP parameters and two spatial densities, of the unmovable network $\mathcal{X}_\circ$ and of the group of potentially removable stations $\mathcal{X}_+$.

Concerning the goodness of fit among distances, the values of Δ_{test} defined by (3.7) and computed at the first and last QBC iterations are displayed for our two models in Table 3.1. The bold values represent the best value within a block. The distance to find the optimal design via the QBC algorithm, see Equation (3.6), can be different from the distance used for assessment in Δ_{test} defined by (3.7).

For model (3.9), the expected result occurs, the smallest Δ_{test} happens when the distances to optimize, see (3.6), and to judge, see (3.7) are equal. As the number of added stations increases at each QBC step it makes sense that Δ_{test} is smaller at the last iteration.

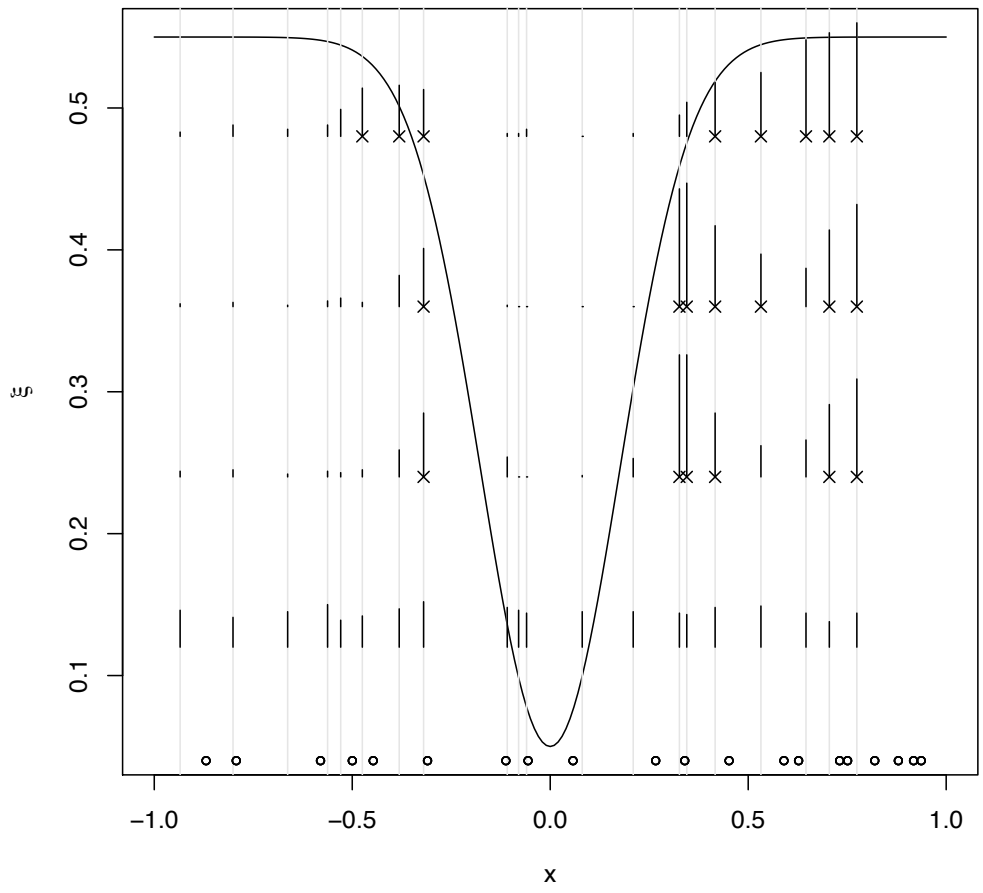

FIGURE 3.4 – QBC outputs for GP parameters following (3.9). The solid line represents the hidden true shape parameter $\xi(x)$. The dots along the x-axis correspond to the 20 sites in $\mathcal{X}_\circ$ and the locations candidate for removal are represented by the 20 vertical grey lines. The QBC is tuned to remove $r = 5$ stations out of 20. From top to bottom, the four horizontally distinct groups of black segments correspond to the x-coordinates of stations removed by the QBC with respect to our four disagreement functions, $d_\sigma(.,.)$ and $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a complete random one, respectively. Crosses indicate sites where the Binomial hypothesis with $p(\mathbf{x}) = 5/20$ and $\ell = 100$ trials, see Equation (3.8), is rejected at the 0.95 significant level. Each black segment height is proportional to the number of times that a site is chosen to be removed.

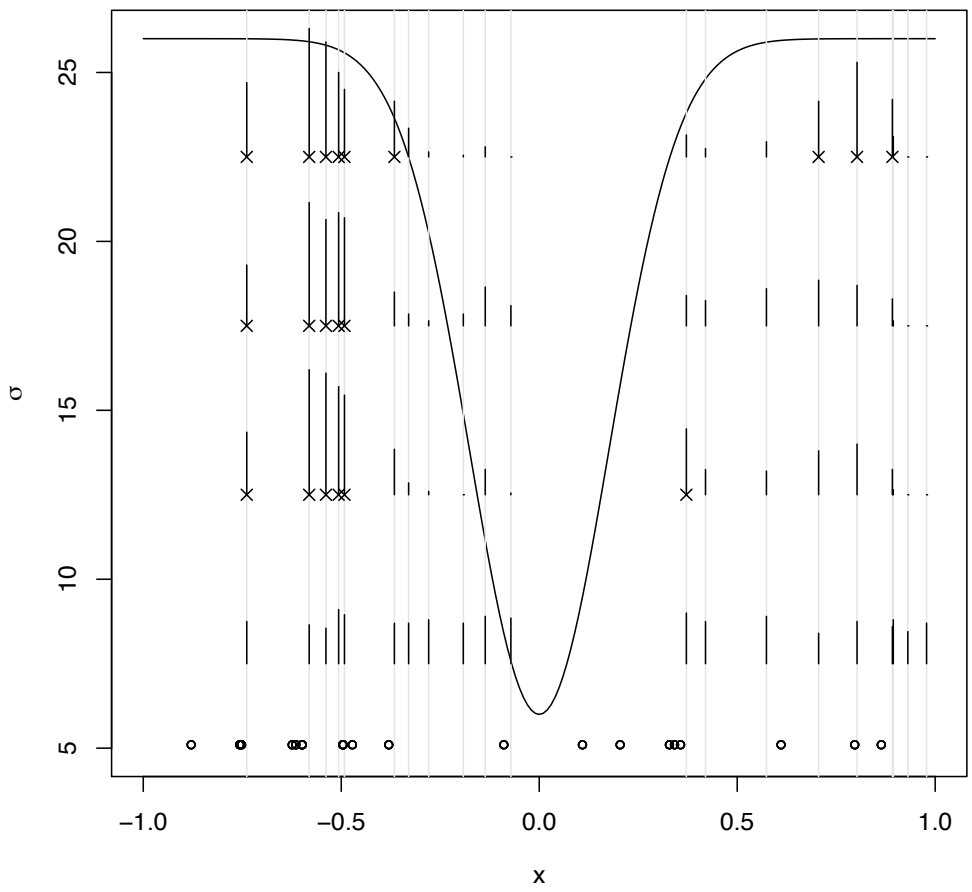

FIGURE 3.5 – Same as Figure 3.4, but for GP parameters following (3.10) and the solid line corresponds to the hidden true scale parameter $\sigma(x)$.

For model (3.10), there is no clear winner with respect to $d_\xi(.,.)$ in (3.7). This may be due to a few causes. The true shape parameter being small, the quantity of the type $(\hat{\xi} - \overline{\xi})^2$ can be minuscule and more decimals are needed (or a re-normalized version). A second reason is that the shape parameter is constant in model (3.10), so judging with $d_\xi(.,.)$ is not discriminatory enough. This is particularly relevant because one needs very large sample sizes to detect small changes around $\xi = 0.5$ (e.g., see Table 2 in Naveau *et al.*, 2013).

Distance chosen		Model (3.9)		Model (3.10)	
in (3.7)	in (3.6)	First step	Last step	First step	Last step
$d_\sigma(.,.)$	$d_\sigma(.,.)$	3.45	**3.30**	3.77	2.97
	$d_\xi(.,.)$	3.45	3.32	3.77	**2.95**
	$d_p(.,.)$	3.45	3.34	3.77	2.97
	Random	3.45	3.32	3.78	3.02
$d_\xi(.,.)$	$d_\sigma(.,.)$	0.159	0.147	0.142	**0.116**
	$d_\xi(.,.)$	0.160	**0.145**	0.142	**0.116**
	$d_p(.,.)$	0.160	0.146	0.142	**0.116**
	Random	0.160	0.150	0.143	**0.116**

TABLE 3.1 – Values of Δ_{test} defined by (3.7) at the first and last QBC iterations, for each distance function and for our two models. The bold values represent the best value within a block.

3.5 French heavy rainfall

Coming back to our French rainfall network in Figure 3.1, the number of stations in $\mathcal{X}_+$ is written as $74 = 15 \times 4 + 14$ to run our QBC algorithm in reasonable time (i.e. a few hours

on a desktop computer). With our notations, this means that we add at each iteration $s = 15$ sites and, the $r = 14$ remaining stations will be tagged as strong candidates for removal. This strategy will be repeated $\ell = 100$ times with the ℓ resampled GP estimates that will be used as inputs for the $m = 100$ experts.

From a computational point of view, it takes basically a little more than a couple of seconds to build a neural network with a desktop computer. Applying the algorithm 100 times, i.e. making $4 + 1$ steps each time and building committees of 100 experts at each step, implies that 50,000 neural networks are constructed during the whole process. This roughly corresponds to 100,000 seconds which means a little more than one day of computation.

Figure 3.6 provides four reduced networks, each one corresponding to a different disagreement function : $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a completely random disagreement. Clearly, the random option cannot take into account the spatial variability of the GP parameters observed in Figure 3.2 and each station in $\mathcal{X}_+$ is selected for removal with the same probability.

The other three distances produce similar maps and the statistically significant stations (with regards to the Binomial hypothesis with $p(\mathbf{x}) = 14/74$ and $\ell = 100$ trials) are basically spread along the northern coastline of France with a few points in the center of France, see the crosses. No stations are removed in the Pyrénées and Corsica. This makes sense because the orography is complex and the network in $\mathcal{X}_\circ$ is not very dense there. A few stations in the southern region near Avignon have also been selected for removal. This can be explained by the very high density of stations in $\mathcal{X}_\circ$.

The main message of Figure 3.6 appears to be that it is more reasonable to remove stations in the North, above the 46^o latitude, (even if the network density is already low there) than discard sites in the south (even if the network is already dense there). This can be explained by a statistical and atmospheric arguments. Intense heavy rainfall happen in the South and the GP parameter is more difficult to estimate for heavier tails.

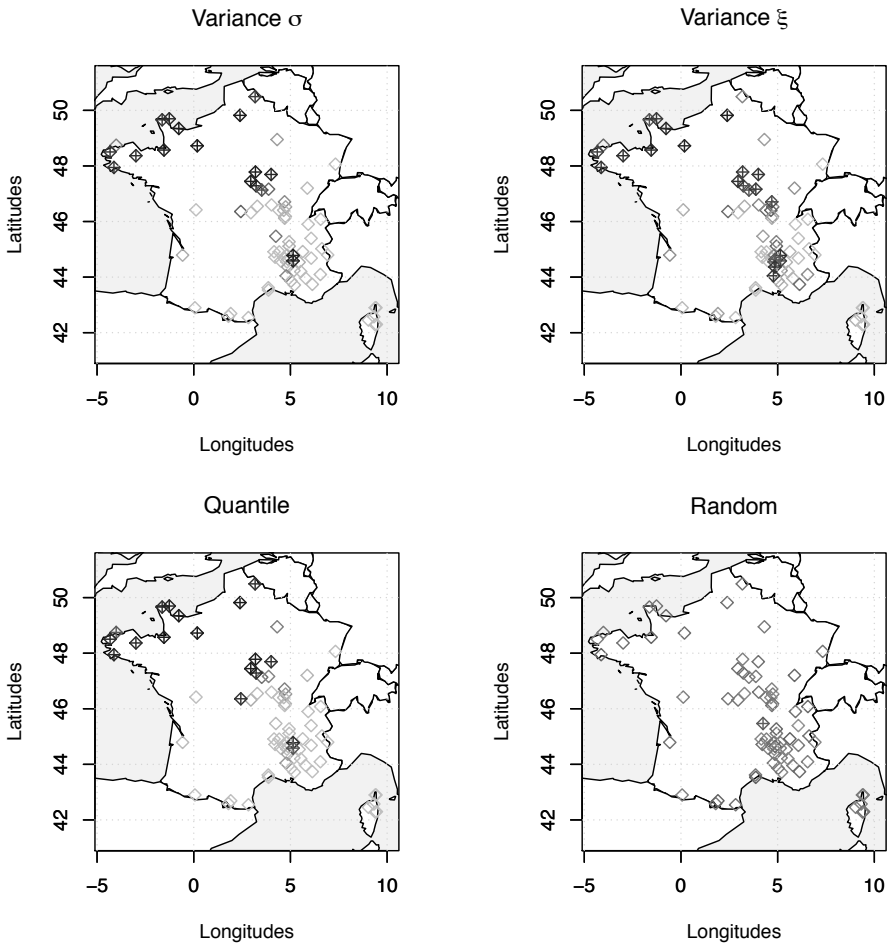

FIGURE 3.6 – Candidate sites for removal from the original set of 74 stations in $\mathcal{X}_+$, see Figure 3.1. Each panel from top left to bottom right corresponds to a different disagreement function : $d_\sigma(.,.)$, $d_\xi(.,.)$, $d_p(.,.)$ with $p = 0.95$ and a completely random disagreement. Crosses indicate sites where the Binomial hypothesis with $p(\mathbf{x}) = 14/74$ and $\ell = 100$ trials, see Equation (3.8), is rejected at the 0.95 significant level. Grey diamonds correspond to sites where we fail to reject the null hypothesis (i.e. less than 26 out of 100) and the color intensity is proportional to the number of times that a station was selected for removal, the lighter the smaller.

Concerning the goodness of fit, we only focus on $d_\sigma(.,.)$ in Table 3.2, the values for $d_\xi(.,.)$ being basically all equal to 0.124. As expected, the fit improves when adding stations through Step 3-v of our QBC, between the first and last step. The disagreement functions $d_\sigma(.,.)$ and $d_p(.,.)$ appear to provide the best value of Δ_{test}.

Distance chosen		QBC	
in (3.7)	in (3.6)	First step	Last step
$d_\sigma(.,.)$	$d_\sigma(.,.)$	4.44	**3.81**
	$d_\xi(.,.)$	4.44	3.83
	$d_p(.,.)$	4.44	**3.81**
	Random	4.45	3.89

TABLE 3.2 – Same as Table 3.1 but for our French heavy rainfall and for $d_\sigma(.,.)$.

To close our rainfall example, we repeat our QBC algorithm but with the "business as usual" hypothesis, i.e. we assume in (3.4) that heavy rainfall excesses follow a Gaussian distribution with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$, instead of our GP distribution assumption with parameters $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$. The normality assumption cannot be justified by EVT, but it is interesting for the decision maker to know if the classical Gaussian model leads to a different set of stations. Comparing Figure 3.6 and Figure 3.7 answers positively to this inquiry. The tip of Brittany, the furthest Western part of France, is treated differently, no stations removed in Finistère with the Gaussian hypothesis. In addition, the disagreement function $d_p(.,.)$ in Figure 3.7 produces a very different network thinning from the ones obtained with $d_\mu(.,.)$ and $d_\sigma(.,.)$. This was not the case with the GP assumption, the QBC generated similar reduced networks. This French example illustrates that network thinning for extremes should be treated carefully and classical statistical tools may not be appropriate to make the right decisions.

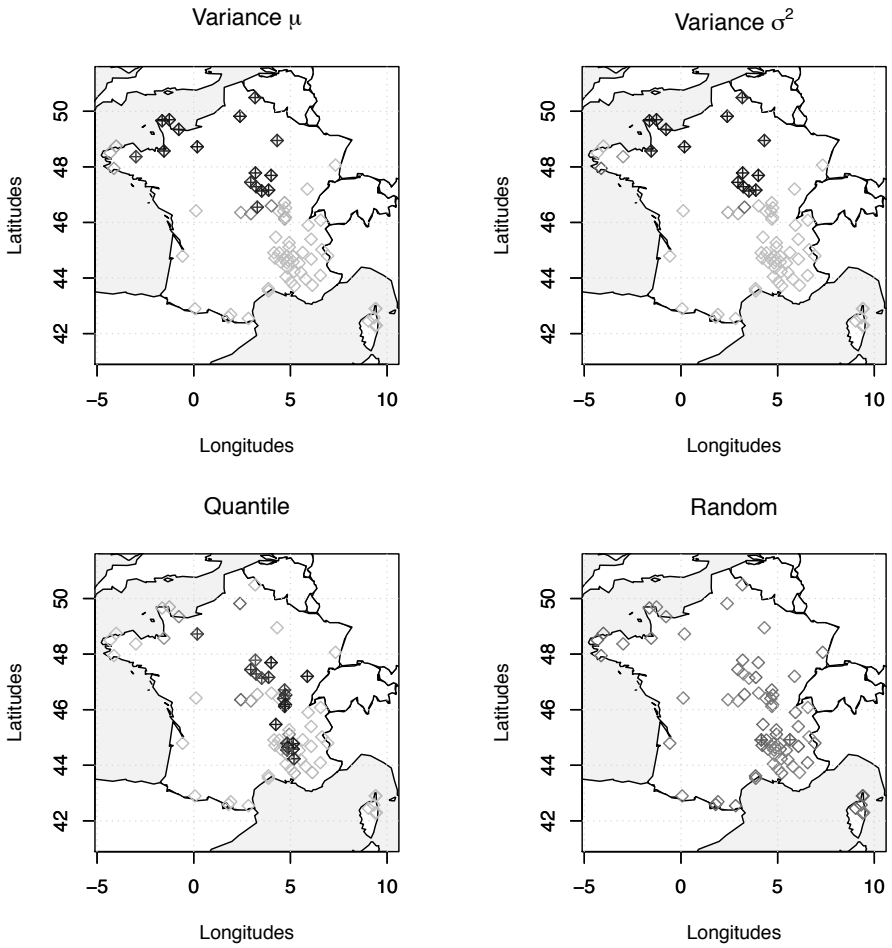

FIGURE 3.7 – Same as Figure 3.6, but with the hypothesis that heavy rainfall follow a Gaussian distribution with mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$.

3.6 Discussion

Our main goal here was to identify weather stations that can be potentially removed from an existing network. A variety of aspects were not taken into account in this study and could open other research avenues. We can list at least four items.

In our application, we only focus on heavy rainfall but a weather station measures other atmospheric variables. For example, extreme winds and heat waves clearly are of interest for impact studies. Hence, a multivariate statistical approach that would integrate many

atmospheric characteristics (and not only precipitation) should be implemented before taking any decisions concerning the removal of stations. This poses new mathematical challenges because multivariate EVT should be used to model the interactions between different types of extremes. Especially for heat waves which have very different spatial patterns (large scale phenomena) than heavy rainfall (local scale).

This example of heat waves also brings into light our assumption of spatial independence among our non-parametric regression residuals. For our case study, this is a reasonable hypothesis because we only want to remove a dozen of sites over France, i.e. the stations are apart by a few hundred kilometers. It is very unlikely, for heavy rainfall, that a strong spatial dependence is present in our residuals for this spatial range. But, it would have been different for heat waves and a covariance structure in $\epsilon(\mathbf{x})$ would be then necessary.

From a more mathematical perspective, it would also be interesting to take advantage of spatial GP processes of excesses. This probabilistic object can handle infinite dimensions and therefore, they are especially tailored to interpolate extreme values at locations without a station. This property is particularly useful for augmenting the network size. Although the probabilistic framework of such spatial processes is now well understood (e.g., Ferreira and de Haan, 2012), there is still a lot of open mathematical questions regarding the modeling and inference of such processes.

Finally combining EVT and machine learning appears to be a promising strategy to handle practical problems dealing with heavy rainfall modeling. One important idea of the QBC is the committee of experts. Here, we have used neural networks, but other non-parametric (such as Kernel regression methods) or even parametric approaches could be used. Augmenting the number of ways of building experts will allow to capture additional uncertainties. In other words, other experts (i.e. statistical models) could be easily added to the QBC blueprint. Still, theoretical work on the mathematical side and practical considerations are needed to insure that such a strategy is optimal at removing weather stations in an operational context.

Chapitre 4

Robust conditional Weibull-type estimation

Abstract

We study nonparametric robust tail coefficient estimation when the variable of interest, assumed to be of Weibull-type, is observed simultaneously with a random covariate. In particular, we introduce a robust estimator for the tail coefficient, using the idea of the density power divergence (see Basu *et al.*, 1998) based on the relative excesses above a high threshold. The main asymptotic properties of our estimator are established under very general assumptions. The finite sample performance of the proposed procedure is evaluated by a small simulation experiment.

4.1 Introduction

In practical data analysis, it is not unusual to encounter outliers which may have a disturbing effect on the estimation results. In such situations, the estimates of the model

according to the maximum likelihood approach are typically fairly unstable and this asks for robust methods. A special treatment of the outlying points is then required, for instance by an adequate downweighting of their influence on the estimation.

In this chapter, we focus on robust procedures in order to estimate some tail parameters in an extreme value context. Such a topic has been recently studied in the literature. We can mention among others Brazauskas and Serfling (2000) and Vandewalle *et al.* (2007) for strict Pareto and Pareto-type distributions, Dupuis and Field (1998), Peng and Welsh (2001), and Juárez and Schucany (2004) for generalized extreme value or generalized Pareto distributions. In the sequel, we consider the Gumbel class, which is a rich subclass of the max-domain of attraction. Although different types of tail behavior are possible, all these distributions have in common an extreme value index equal to zero and thus differentiating them on the basis of this parameter alone is impossible. To solve this issue, we restrict our study to Weibull-type distributions for which the distribution functions have the following form :

$$\overline{F}(y) := 1 - F(y) = e^{-y^{1/\theta}\ell_F(y)}, \qquad y > 0,$$

where $\theta > 0$ and ℓ_F is a slowly varying function at infinity, i.e. an ultimately positive function satisfying

$$\lim_{y\to\infty} \frac{\ell_F(\lambda y)}{\ell_F(y)} = 1, \qquad \text{for all } \lambda > 0.$$

Here θ denotes the Weibull-tail coefficient. Different values of it allow the Weibull-type distributions to cover a large part of the Gumbel class, and hence to constitute a flexible subgroup. The estimation of this coefficient has been extensively studied in the literature (see e.g., Broniatowski (1993), Beirlant *et al.* (1995), Gardes and Girard (2005, 2008b), Diebolt *et al.* (2008), Dierckx *et al.* (2009), Goegebeur *et al.* (2010) or Goegebeur and Guillou (2011) among others) but not much attention has been paid to the regression

context with covariates.

We will consider this framework of nonparametric regression estimation of conditional tails when the covariates are random. The case of random covariates is less explored in extreme value theory compared to the fixed covariates, and only few papers can indeed be mentioned : Wang and Tsai (2009) with a parametric maximum likelihood approach within the Hall subclass of Pareto-type models (Hall, 1982), Daouia *et al.* (2011) in the framework of Pareto-type distributions, and Daouia *et al.* (2013) in the general max-domain of attraction, but under rather restrictive assumptions on the underlying distribution function. Here, we consider the case of Weibull-type distributions and our approach will be based on local estimation within a narrow window around the point in the covariate space where the tail behavior of the response variable is of interest. This local fitting is performed by an adjustment of the robust minimum density power divergence (MDPD) estimation criterion, originally proposed by Basu *et al.* (1998), to the locally weighted regression setting. This criterion has already been used for robust estimation of Pareto-type distributions, see for instance Kim and Lee (2008), Dierckx *et al.* (2013a, b), but to the best of our knowledge it is new in the Weibull-type framework.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce our robust estimator of the conditional Weibull-tail coefficient and we state its main asymptotic properties. The finite sample performance of our procedure is illustrated on a small simulation study in Section 4.3. The proofs of all results can be found in the Appendix.

4.2 Construction and asymptotic properties

Let (X_i, Y_i), $i = 1, ..., n$, be independent copies of a random pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}_+$, where X has the density function f and the conditional survival function of Y given $X = x$ is

of Weibull-type with a tail coefficient $\theta(x) > 0$, that is

$$\overline{F}(y;x) = e^{-y^{1/\theta(x)}\ell_F(y;x)}, \qquad y > 0. \tag{4.1}$$

This model can also be defined in terms of the generalized inverse of F, denoted Q, i.e. $Q(q;x) := \inf\{y : F(y;x) \geq q\}$, $0 < q < 1$. Indeed, under (4.1), we have

$$Q(q;x) = (-\ln(1-q))^{\theta(x)}\,\ell(-\ln(1-q);x) \tag{4.2}$$

where ℓ is again a slowly varying function at infinity. The function $\theta(x)$ governs the tail behavior, with larger values indicating a slower tail decay. This function has to be adequately estimated from the data.

As is usual in an extreme value context, we base our estimation method on the relative excesses above a high threshold u_n, namely $Z := Y/u_n$, which admit, under model (4.1), the following conditional survival function :

$$\mathbb{P}\left(\frac{Y}{u_n} > t \middle| Y > u_n; x\right) = \frac{\overline{F}(tu_n;x)}{\overline{F}(u_n;x)} \simeq e^{-c_n(t^{1/\theta(x)}-1)} =: \overline{G}(t;c_n,\theta(x)) \quad \text{for } t > 1, \tag{4.3}$$

where $c_n := -\ln(\overline{F}(u_n;x))$. The approximation in (4.3) follows from the properties of slowly varying functions, and is valid for large values of u_n. We denote by g the density function associated to this distribution G.

The proposed estimation procedure works as follows. First we estimate c_n externally in a consistent way, cf infra. Then, we estimate $\theta(x)$ with the MDPD criterion combined with a kernel approach, and applied to the relative excesses above u_n. More precisely, we define the MDPD estimator as the value of θ minimizing the empirical density power divergence :

$$\widehat{\Delta}_\alpha(\theta;\widehat{c}_n) := \begin{cases} \frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\mathbb{1}_{\{Y_i>u_n\}}\left\{\int_1^\infty g^{1+\alpha}(z)dz - \left(1+\frac{1}{\alpha}\right)g^\alpha(Y_i/u_n)\right\} & \text{for } \alpha > 0 \\ -\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\mathbb{1}_{\{Y_i>u_n\}}\ln g(Y_i/u_n) & \text{for } \alpha = 0 \end{cases}$$

where $\widehat{c}_n$ is a consistent estimator for c_n, $K_h(x) := K(x/h)/h^p$, K is a joint density function on $\mathbb{R}^p$, $h = h_n$ is a positive, non-random sequence of bandwidths with $h \to 0$ if $n \to \infty$, $\mathbb{1}_{\{A\}}$ is the indicator function on the event A and u_n is a local non-random threshold sequence satisfying $u_n \to \infty$ if $n \to \infty$.

Note that a joint estimation of $\theta(x)$ and c_n with the MDPD method is practically feasible, but gives difficulties in the theoretical analysis concerning consistency and asymptotic normality, and in particular it requires the introduction of rather restrictive conditions. For this reason we opt to estimate c_n externally in a consistent way. Remark also that this density power divergence criterion is indexed by a single non-negative parameter, α, that controls the trade-off between robustness and efficiency. In particular it encompasses the maximum likelihood method, corresponding to $\alpha = 0$, which is efficient but not robust. Increasing the value of α increases the robustness and decreases the efficiency of the estimation.

The MDPD equation for θ is thus :

$$
\begin{aligned}
\widehat{\Delta}'_\alpha(\theta;\widehat{c}_n) &= \frac{1+\alpha}{n}\sum_{i=1}^{n} K_h(x-X_i)\mathbb{1}_{\{Y_i>u_n\}}\int_1^\infty g^\alpha(z)\frac{\partial g(z)}{\partial\theta}dz \\
&\quad -\frac{1+\alpha}{n}\sum_{i=1}^{n} K_h(x-X_i)\mathbb{1}_{\{Y_i>u_n\}} g^{\alpha-1}\left(\frac{Y_i}{u_n}\right)\frac{\partial g(Y_i/u_n)}{\partial\theta} \\
&= \frac{1+\alpha}{n}\sum_{i=1}^{n} K_h(x-X_i)\mathbb{1}_{\{Y_i>u_n\}}\left\{-\frac{\alpha\, e^{\widehat{c}_n(1+\alpha)}\widehat{c}_n^{\alpha\theta}}{\theta^{\alpha+1}(1+\alpha)^{2+\alpha(1-\theta)}}\Big[\theta\Psi(\alpha(1-\theta)+1,\widehat{c}_n(1+\alpha))\right. \\
&\quad \left.+(1-\theta\ln((\alpha+1)\widehat{c}_n)\Gamma(\alpha(1-\theta)+1,\widehat{c}_n(1+\alpha))\Big]\right\} \\
&\quad +\frac{\widehat{c}_n^{\alpha}}{\theta^{\alpha+1}}\frac{1+\alpha}{n}\sum_{i=1}^{n} K_h(x-X_i)e^{-\widehat{c}_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta}-1\right]}\left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta-1)}\mathbb{1}_{\{Y_i>u_n\}} \\
&\quad +\frac{\widehat{c}_n^{\alpha}}{\theta^{\alpha+2}}\frac{1+\alpha}{n}\sum_{i=1}^{n} K_h(x-X_i)e^{-\widehat{c}_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta}-1\right]}\left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta-1)}\ln\frac{Y_i}{u_n}\mathbb{1}_{\{Y_i>u_n\}} \\
&\quad -\frac{\widehat{c}_n^{\alpha+1}}{\theta^{\alpha+2}}\frac{1+\alpha}{n}\sum_{i=1}^{n} K_h(x-X_i)e^{-\widehat{c}_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta}-1\right]}\left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta-1)+1/\theta}\ln\frac{Y_i}{u_n}\mathbb{1}_{\{Y_i>u_n\}} \qquad (4.4)
\end{aligned}
$$

where $\Gamma(a,b)$ denotes the incomplete Gamma function

$$\Gamma(a,b) := \int_b^\infty z^{a-1}\, e^{-z} dz$$

and $\Psi(a,b)$ its derivative with respect to the first argument

$$\Psi(a,b) := \int_b^\infty \ln z\, z^{a-1}\, e^{-z} dz.$$

In the functions Γ and Ψ, b is positive whereas $a \in \mathbb{R}$.

The MDPD estimator for $\theta(x)$ satisfies the estimating equation

$$\widehat{\Delta}'_\alpha(\theta; \widehat{c}_n) = 0, \tag{4.5}$$

where $\widehat{c}_n := -\ln(\widehat{\overline{F}}(u_n;x))$,

$$\widehat{\overline{F}}(u_n;x) := \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\mathbb{1}_{\{Y_i>u_n\}}}{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)},$$

is a kernel estimator for $\overline{F}(u_n;x)$, as considered also in Daouia *et al.* (2013) and de Wet *et al.* (2013).

In view of (4.4) we start by considering the following locally weighted sums of power-transformed excesses over a high threshold :

$$T_n(K,\alpha,\beta,r;x) := \frac{1}{n}\sum_{i=1}^n K_h(x-X_i) e^{-c_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta(x)}-1\right]} \left(\frac{Y_i}{u_n}\right)^\beta \left(\ln\frac{Y_i}{u_n}\right)_+^r \mathbb{1}_{\{Y_i>u_n\}} \tag{4.6}$$

where $\alpha \geq 0$, $\beta \in \mathbb{R}$, $r \geq 0$, and $(x)_+ := \max\{0,x\}$.

To obtain the limiting behavior of (4.6) one has to impose some more structure on the tail of the distribution. Typically one invokes a so-called second order condition, specifying the rate of convergence of $\ell(\lambda y;x)/\ell(y;x)$, where ℓ is the slowly varying function appearing in (4.2), to its limit, being one, as $y \to \infty$.

Assumption $(\mathcal{R})$ *There exists a constant $\rho(x) < 0$ and a rate function $b(.;x)$ satisfying*

$b(y;x) \to 0$ *as* $y \to \infty$*, such that for all* $\lambda \geq 1$*, we have*

$$\ln\left(\frac{\ell(\lambda y;x)}{\ell(y;x)}\right) = b(y;x)D_{\rho(x)}(\lambda)(1+o(1))$$

with $D_{\rho(x)}(\lambda) := \int_1^\lambda t^{\rho(x)-1}dt$*, and where* $o(1)$ *is uniform in* $\lambda \geq 1$, as $y \to \infty$.

As shown in Geluk and de Haan (1987), $(\mathcal{R})$ implies that $|b(y;x)|$ is regularly varying with index $\rho(x)$, i.e. $|b(\lambda y;x)|/|b(y;x)| \to \lambda^{\rho(x)}$ as $y \to \infty$ for all $\lambda > 0$, so $\rho(x)$ governs the rate of the first order convergence of $\ell(\lambda y;x)/\ell(y;x)$ to one. If $|\rho(x)|$ is small then the convergence is slow and the estimation of tail quantities is generally difficult. Condition $(\mathcal{R})$ is well accepted in the extreme value literature, see e.g. Gardes and Girard (2008a).

As a first step in the theoretical study of estimators for $\theta(x)$, we consider the local behavior of the following conditional expectation :

$$m(u_n,\alpha,\beta,r;x) = \mathbb{E}\left(e^{-c_n\alpha\left[\left(\frac{Y}{u_n}\right)^{1/\theta(x)}-1\right]}\left(\frac{Y}{u_n}\right)^{\beta}\left(\ln\frac{Y}{u_n}\right)_+^r \mathbb{1}_{\{Y>u_n\}}\middle| X=x\right).$$

Lemma 4.1 *Case (i),* $\alpha=\beta=r=0$ *:*

$$m(u_n,0,0,0;x) = \overline{F}(u_n;x).$$

Case (ii), $(\alpha,\beta,r) \in \mathbb{R}^+\times\mathbb{R}\times\mathbb{R}^+\setminus(0,0,0)$ *: assume (4.2) and Assumption* $(\mathcal{R})$*. We have for* $u_n \to \infty$ *that*

$$\begin{aligned} m(u_n,\alpha,\beta,r;x) &= \overline{F}(u_n;x)\frac{\Gamma(1+r)}{(1+\alpha)^{1+r}}\theta^r(x)\left\{c_n^{-r}+\frac{\theta(x)\beta}{1+\alpha}c_n^{-1}\mathbb{1}_{\{r=0\}}+\frac{r-\alpha}{1+\alpha}\frac{b(c_n;x)}{\theta(x)}c_n^{-r}\right. \\ &\quad \left.+o(b(c_n;x)c_n^{-r})+O\left(\frac{1}{c_n^{(1+\alpha)\wedge 2}}\mathbb{1}_{\{r=0\}}\right)+O\left(\frac{1}{c_n^{(1+r+\alpha-r\varepsilon)\wedge(1+r)}}\mathbb{1}_{\{r>0\}}\right)\right\}, \end{aligned}$$

for ε *sufficiently small.*

We now turn to the derivation of the asymptotic expansion for the unconditional moment.

Let

$$m_n(K,\alpha,\beta,r;x) := \mathbb{E}\left[K_h(x-X)e^{-c_n\alpha\left[\left(\frac{Y}{u_n}\right)^{1/\theta(x)}-1\right]}\left(\frac{Y}{u_n}\right)^{\beta}\left(\ln\frac{Y}{u_n}\right)_+^{r}\mathbb{1}_{\{Y>u_n\}}\right].$$

Note that since $T_n(K,\alpha,\beta,r;x)$ is an average of independent and identically distributed (i.i.d.) terms we also have that $m_n(K,\alpha,\beta,r;x) = \mathbb{E}(T_n(K,\alpha,\beta,r;x))$.

We introduce the following further conditions. Let $\|.\|$ be a norm on $\mathbb{R}^p$.

Assumption $(\mathcal{F})$ There exists $M_f > 0$ and $\eta_f > 0$ such that $|f(x)-f(x')| \leq M_f\|x-x'\|^{\eta_f}$ for all $x, x' \in \mathbb{R}^p$.

Assumption $(\mathcal{K})$ K is a bounded density function on $\mathbb{R}^p$, with support Ω included in the unit hypersphere in $\mathbb{R}^p$.

Finally we introduce a condition on the oscillation of the response distribution in a neighborhood of the point x where the estimation will take place. This condition is formulated in terms of the conditional excess function :

Assumption $(\mathcal{M})$ The conditional excess function $m(u_n,\alpha,\beta,r;x)$ satisfies for $u_n \to \infty$, $h \to 0$, and some $\bar{\alpha} > 0$, $R > 0$, $\xi > 0$ that

$$\Phi_n(x) := \sup_{\alpha\in[0,\bar{\alpha}]}\sup_{\beta\in[\alpha/\theta(x)-\xi,\alpha/\theta(x)+\xi]}\sup_{r\in[0,R]}\sup_{z\in\Omega}\left|\frac{m(u_n,\alpha,\beta,r;x-hz)}{m(u_n,\alpha,\beta,r;x)}-1\right| \to 0 \text{ as } n\to\infty.$$

The following lemma gives then the asymptotic expansion of $m_n(K,\alpha,\beta,r;x)$.

Lemma 4.2 *Assume (4.2),* $(\mathcal{R})$, $(\mathcal{F})$, $(\mathcal{K})$ *and* $(\mathcal{M})$. *For all* $x \in \mathbb{R}^p$ *where* $f(x) > 0$, *we have that if* $u_n \to \infty$ *and* $h \to 0$ *as* $n \to \infty$ *then*

$$m_n(K,\alpha,\beta,r;x) = m(u_n,\alpha,\beta,r;x)f(x)(1+O(h^{\eta_f})+O(\Phi_n(x))).$$

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ *be a sample of independent copies of the random vector* (X,Y)

where $Y|X = x$ satisfies (4.2) and $X \sim f$. If in addition to the previous assumptions one also has that $nh^p\overline{F}(u_n;x) \to \infty$ as $n \to \infty$ then

$$\widetilde{T}_n(K,\alpha,\beta,r;x) := \frac{c_n^r T_n(K,\alpha,\beta,r;x)}{\overline{F}(u_n;x)f(x)} \xrightarrow{\mathbb{P}} \frac{\theta^r(x)\Gamma(1+r)}{(1+\alpha)^{1+r}} \text{ as } n \to \infty.$$

The consistency of $\widehat{\overline{F}}(u_n;x)$ follows now easily from Lemma 4.2.

Corollary 4.1 *Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be i.i.d. random vectors and assume $(\mathcal{F})$, $(\mathcal{K})$ and $(\mathcal{M})$. For all $x \in \mathbb{R}^p$ where $f(x) > 0$ we have that if $h \to 0$, $u_n \to \infty$ with $nh^p\overline{F}(u_n;x) \to \infty$ as $n \to \infty$, then*

$$\frac{\widehat{\overline{F}}(u_n;x)}{\overline{F}(u_n;x)} \xrightarrow{\mathbb{P}} 1.$$

Note that this result holds for a general conditional survival function $\overline{F}(y;x)$, where $X \sim f$, i.e. the assumption of conditional Weibull-type behavior, and hence $(\mathcal{R})$, is not needed. In our context we have then $\widehat{c}_n - c_n = -\ln \widehat{\overline{F}}(u_n;x)/\overline{F}(u_n;x) \xrightarrow{\mathbb{P}} 0$, by a straightforward application of the continuous mapping theorem.

The following theorem states the existence and consistency of sequences of solutions to the estimating equation (4.5). From now on we denote the true value of $\theta(x)$ by $\theta_0(x)$.

Theorem 4.1 *Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample of independent copies of the random vector (X,Y) where $Y|X = x$ satisfies (4.2), $X \sim f$, and assume $(\mathcal{R})$, $(\mathcal{F})$, $(\mathcal{K})$ and $(\mathcal{M})$ hold. For all $x \in \mathbb{R}^p$ where $f(x) > 0$, we have that if $h_n \to 0$, $u_n \to \infty$ with $nh_n^p\overline{F}(u_n;x) \to \infty$, then with probability tending to 1 there exists sequences of solutions $(\widehat{\theta}_n(x))_{n\in\mathbb{N}}$ of the estimating equation (4.5), such that $\widehat{\theta}_n(x) \xrightarrow{\mathbb{P}} \theta_0(x)$, as $n \to \infty$.*

We now derive the limiting distribution of a vector of statistics of the form (4.6), when properly normalized. This result will form the basis for proving the asymptotic normality of the MDPD estimator. Let

$$\mathbb{T}'_n := (\widetilde{T}_n(K_1,\alpha_1,\beta_1,r_1;x),\ldots,\widetilde{T}_n(K_J,\alpha_J,\beta_J,r_J;x))$$

for some positive integer J and let Σ be a $(J \times J)$ covariance matrix with elements

$$\sigma_{j,k} := \frac{\theta_0^{r_j+r_k}(x)\|K_jK_k\|_1\Gamma(1+r_j+r_k)}{(1+\alpha_j+\alpha_k)^{1+r_j+r_k}}.$$

Theorem 4.2 *Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample of independent copies of the random vector (X,Y) where $Y|X=x$ satisfies (4.2) and $X\sim f$, and assume $(\mathcal{R})$, $(\mathcal{F})$, $(\mathcal{M})$ hold and kernel functions $K_1,\ldots,K_J$ satisfying $(\mathcal{K})$. For all $x\in\mathbb{R}^p$ where $f(x)>0$, we have that if $h\to 0$, $u_n\to\infty$ for $n\to\infty$, with $nh^p\overline{F}(u_n;x)\to\infty$, then*

$$\sqrt{nh^p\overline{F}(u_n;x)}\ [\mathbb{T}_n-\mathbb{E}(\mathbb{T}_n)]\rightsquigarrow N_J\left(0,\frac{1}{f(x)}\,\Sigma\right).$$

With the result of Theorem 4.2, we can now establish the asymptotic normality of $\widehat{\theta}_n(x)$, when properly normalized.

Theorem 4.3 *Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample of independent copies of the random vector (X,Y) where $Y|X=x$ satisfies (4.2), $X\sim f$, and assume $(\mathcal{R})$, $(\mathcal{F})$, $(\mathcal{K})$ and $(\mathcal{M})$ hold. Consider $(\widehat{\theta}_n(x))_{n\in\mathbb{N}}$, a consistent sequence of estimators for $\theta_0(x)$ satisfying (4.5). For all $x\in\mathbb{R}^p$ where $f(x)>0$, we have that if $h\to 0$ and $u_n\to\infty$ as $n\to\infty$ with $nh^p\overline{F}(u_n;x)\to\infty$, $\sqrt{nh^p\overline{F}(u_n;x)}b(c_n;x)\to\lambda\in\mathbb{R}$, $\sqrt{nh^p\overline{F}(u_n;x)}h_n^{\eta_f}\to 0$, $\sqrt{nh^p\overline{F}(u_n;x)}\Phi_n(x)\to 0$, and $\sqrt{nh^p\overline{F}(u_n;x)}/c_n^{(1+\alpha-\varepsilon)\wedge 1}\to 0$ (for some small $\varepsilon>0$) then*

$$\sqrt{nh^p\overline{F}(u_n;x)}\left(\widehat{\theta}_n(x)-\theta_0(x)\right)$$
$$\rightsquigarrow N\left(\lambda,\frac{\theta_0^2(x)(1+\alpha)^2\|K\|_2^2}{(1+\alpha^2)^2(1+2\alpha)^3}(1+4\alpha+9\alpha^2+14\alpha^3+13\alpha^4+8\alpha^5+4\alpha^6)\right).$$

Note that the mean of the limiting distribution in Theorem 4.3 depends only on λ and not on α nor on the weight function K or parameters related to the distribution of Y given $X=x$. This is in line with the usual asymptotic normality result in the univariate case, see e.g. Girard (2004), Gardes and Girard (2008b) and Goegebeur *et al.* (2010). The asymptotic variance in Theorem 4.3 is increasing in α, which reflects the decreasing efficiency of the MDPD estimation method when α increases. The maximum likelihood estimator, corresponding to $\alpha=0$, has an asymptotic variance equal to $\theta_0^2(x)\|K\|_2^2$, which, apart

from the factor $\|K\|_2^2$, coincides with the asymptotic variance of the Hill type estimator proposed by Girard (2004) in the univariate context. In Figure 4.1 we show the asymptotic standard deviation as a function of α when $\theta_0(x) = 1$ and $K(u) = 0.5\ \mathbb{1}_{u\in[-1,1]}$.

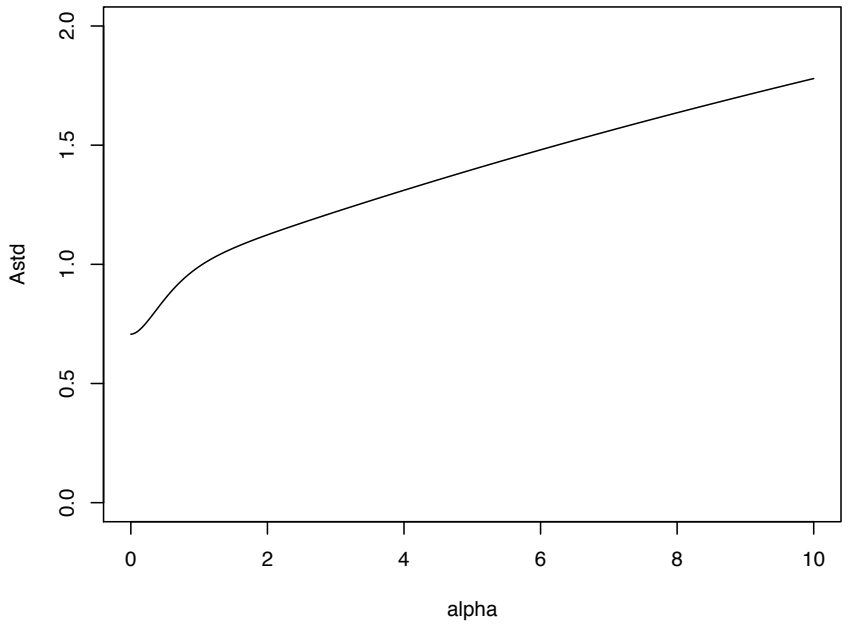

FIGURE 4.1 – Asymptotic standard deviation of the MDPD estimator for $\theta_0(x)$ as a function of α when $\theta_0(x) = 1$ and $K(u) = 0.5\ \mathbb{1}_{u\in[-1,1]}$.

4.3 Simulation results

The aim of this section is to illustrate the efficiency of our robust estimation method on a small simulation study. As is clear from the above discussion, the computation of the estimator requires a selection for the bandwidth parameter h and the threshold u_n. We select the threshold as usual in extreme value theory, that is, we take the $(k+1)$ largest response observation in the ball $B(x,h)$ for any fixed value of x. We propose a

data driven method to determine (h, k) and we compare it with a theoretical one, called Oracle strategy, which requires the knowledge of the function $\theta(x)$. These two methods are similar to those used in Goegebeur *et al.* (2013) and are briefly recalled as follows :

The Oracle strategy, proposed in Gardes *et al.* (2010) consists in selecting (h, k) satisfying

$$(h_0, k_0) := \underset{h,k}{\operatorname{argmin}} \, \Delta\left(\widehat{\theta}(\cdot), \theta(\cdot)\right), \tag{4.7}$$

where

$$\Delta^2\left(\widehat{\theta}(\cdot), \theta(\cdot)\right) := \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}\left(z_m\right) - \theta\left(z_m\right)\right)^2,$$

and $z_1, \ldots, z_M$ are points regularly spread in the covariate space. The retained value $\widehat{\theta}(x)$ is the one corresponding to this pair (h_0, k_0).

The data driven method does not require any prior knowledge about $\theta(x)$ and thus can be directly applied to a real dataset. First, the method involves the selection of an optimal bandwidth h. To this aim, we can use the following cross validation criterion introduced by Yao (1999), implemented by Gannoun *et al.* (2002), and studied by Daouia *et al.* (2011, 2013) in an extreme value context :

$$h_{cv} := \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{1}_{\{Y_i\leq Y_j\}} - \widetilde{F}_{n,-i}\left(Y_j|X_i\right)\right)^2, \tag{4.8}$$

where $\mathcal{H}$ is a grid of values for h and

$$\widetilde{F}_{n,-i}\left(y|x\right) := \frac{\sum_{j=1,j\neq i}^{n} K_h\left(x - X_j\right)\mathbb{1}_{\{Y_j\leq y\}}}{\sum_{j=1,j\neq i}^{n} K_h\left(x - X_j\right)}.$$

Once the bandwidth h has been chosen, we compute $\widehat{\theta}(x)$ for each $k = 5, \ldots, k_{max}$ where k_{max} is an appropriate maximum value of the number of exceedances. The retained value $\widehat{\theta}$ is the median of these estimates of θ.

In the simulations, we use the following smooth and symmetric function

$$\theta(x) = \frac{1}{2}\left(\frac{1}{10} + \sin(\pi x)\right)\left(\frac{11}{10} - \frac{1}{2}\exp\left\{-64\left(x - \frac{1}{2}\right)^2\right\}\right)$$

proposed by Daouia *et al.* (2011), though originally in the context of Pareto-type tails, and the estimators are based on the bi-quadratic kernel function

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \mathbf{1}\{x \in [-1, 1]\}.$$

The function $\theta(x)$ is differentiable and has several stationary points. As such, it is more challenging than a monotone function. We assume that X is uniformly distributed on $[0, 1]$ and we consider four different settings for the conditional distribution of Y given $X = x$:

• The Weibull distribution $\mathcal{W}(\xi(x), \lambda)$,

$$1 - F(y; x) = e^{-\lambda y^{\xi(x)}}, \quad y > 0; \xi(x), \lambda > 0,$$

for which $\theta(x) = 1/\xi(x)$ and $\rho(x) = -\infty$. We consider the case $\lambda = 1$.

• The extended Weibull distribution $\mathcal{EW}(\xi(x), \lambda)$ (Klüppelberg and Villaseñor, 1993),

$$1 - F(y; x) = r(y)e^{-y^{\xi(x)}},$$

where $\xi(x) > 0$ and $r(y)$ is a regularly varying function at infinity with index λ. Here $\theta(x) = 1/\xi(x)$ and $\rho(x) = -1$. We choose $r(y) = 1/y$.

• The perturbed Weibull distribution $\widetilde{\mathcal{W}}(\xi(x), \lambda)$ (Dierckx *et al.*, 2009),

$$1 - F(y; x) = e^{-y^{\xi(x)}\left(C + Dy^{\lambda}\right)}, \quad \xi(x) > 0, \lambda < 0, C > 0, D \in \mathbb{R},$$

having $\theta(x) = 1/\xi(x)$ and $\rho(x) = \lambda\theta(x)$. In this case we use $\lambda = -5$, $C = 1$ and $D = -1$.

• A contaminated distribution with $F_\varepsilon(y; x) = (1 - \varepsilon)F(y; x) + \varepsilon\breve{F}(y; x)$ where the

distribution function F is one of the three above mentioned distributions and the contaminating distribution is a shifted Weibull distribution, i.e. it has distribution function $\breve{F}(y;x) = 1 - e^{-(y^\beta - y_c^\beta)}, y > y_c$. We choose $\beta = 4/3$, two different values for ε, namely 0.005 and 0.01, and $y_c = 1.2$ times the 95% quantile of the uncontaminated distribution F.

Each time $N = 100$ samples of size $n = 1000$ are generated. Both methods are implemented on $M = 37$ values of x, equally spaced in $[0, 1]$, namely $\{0.05, 0.075, ..., 0.925, 0.95\}$. In all the settings, the minimization (4.7) is performed on a grid $\mathcal{H} = \{0.05, 0.075, \ldots, 0.15\}$ and $k_{max} = 25$.

As indicators of efficiency we compute the bias, together with the mean square error and the standard error

$$\text{Bias}(\widehat{\theta}(.)) := \frac{1}{MN}\sum_{m=1}^{M}\sum_{i=1}^{N}\left|\widehat{\theta}^{(i)}(z_m) - \theta(z_m)\right|, \quad \text{MSE}(\widehat{\theta}(.)) := \frac{1}{MN}\sum_{m=1}^{M}\sum_{i=1}^{N}\left[\widehat{\theta}^{(i)}(z_m) - \theta(z_m)\right]^2,$$

$$\text{Sd}(\widehat{\theta}(.)) := \sqrt{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{N-1}\sum_{i=1}^{N}\left(\widehat{\theta}^{(i)}(z_m) - \overline{\widehat{\theta}}(z_m)\right)^2},$$

where $\overline{\widehat{\theta}}(z_m) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}^{(i)}(z_m)$ and $\widehat{\theta}^{(i)}(z_m)$ is the estimate of $\theta(z_m)$ obtained with the i^{th} sample and our estimator $\widehat{\theta}(x)$ is evaluated at points $z_1, ..., z_M$ regularly spaced in $[0, 1]$.

The boxplots based on the 100 simulations in the uncontaminated case are given in Figure 4.2 for the Oracle approach and Figure 4.3 for the data driven method. Overall, both methods perform quite well, with, as expected, a better performance with the Oracle approach, which is also confirmed by the bias, the standard deviation and the MSE given in Table 4.1 In addition, we can observe that all these indicators slightly increase with α whatever the method used. The increase in standard deviation when α increases is in line with the decreasing efficiency of the MDPD method for increasing α, as mentioned before.

Now we consider setting 4 where the distribution is contaminated. Figures 4.4 and 4.5 summarize the results for $\varepsilon = 0.005$ with the Oracle and data driven approach, respectively. Again the Oracle strategy outperforms the data driven method, but the latter is still acceptable and fits quite well the curve of the function θ. Similar comments can be made in terms of bias, standard deviation and MSE, reported in Tables 4.2 and 4.3, for $\varepsilon = 0.005$ and $\varepsilon = 0.01$, respectively. Indeed, based on these tables we can draw the following conclusions :

• The Oracle strategy clearly outperforms the data driven method, and, as expected, the estimates obtained with both methods deteriorate in terms of bias, MSE and standard deviation when the proportion of contamination ε increases.

• Unlike the uncontaminated cases where the smallest value of α, here 0.1, gave the best results, we see now that a larger value of α is needed in order to deal with the contamination. For the Oracle method the results in Tables 4.2 and 4.3 suggest to take $\alpha = 0.25$ or $\alpha = 0.5$, while for the data driven method a larger value of α seems to be needed, say $\alpha = 1$ or $\alpha = 2$.

Summarized, we can state that the proposed MDPD estimator for the tail coefficient of Weibull-type distributions is a promising new estimator. Indeed, on uncontaminated data the estimation results do not indicate the typical bias-issues that occur in practice when estimating Weibull-type tails, as experienced in, among others, Gardes and Girard (2008b) and Goegebeur *et al.* (2010) in the univariate context. Also in case of contamination the method continues to work well, despite the fact that the contamination was quite severe in terms of shift and tail heaviness of the contaminating distribution.

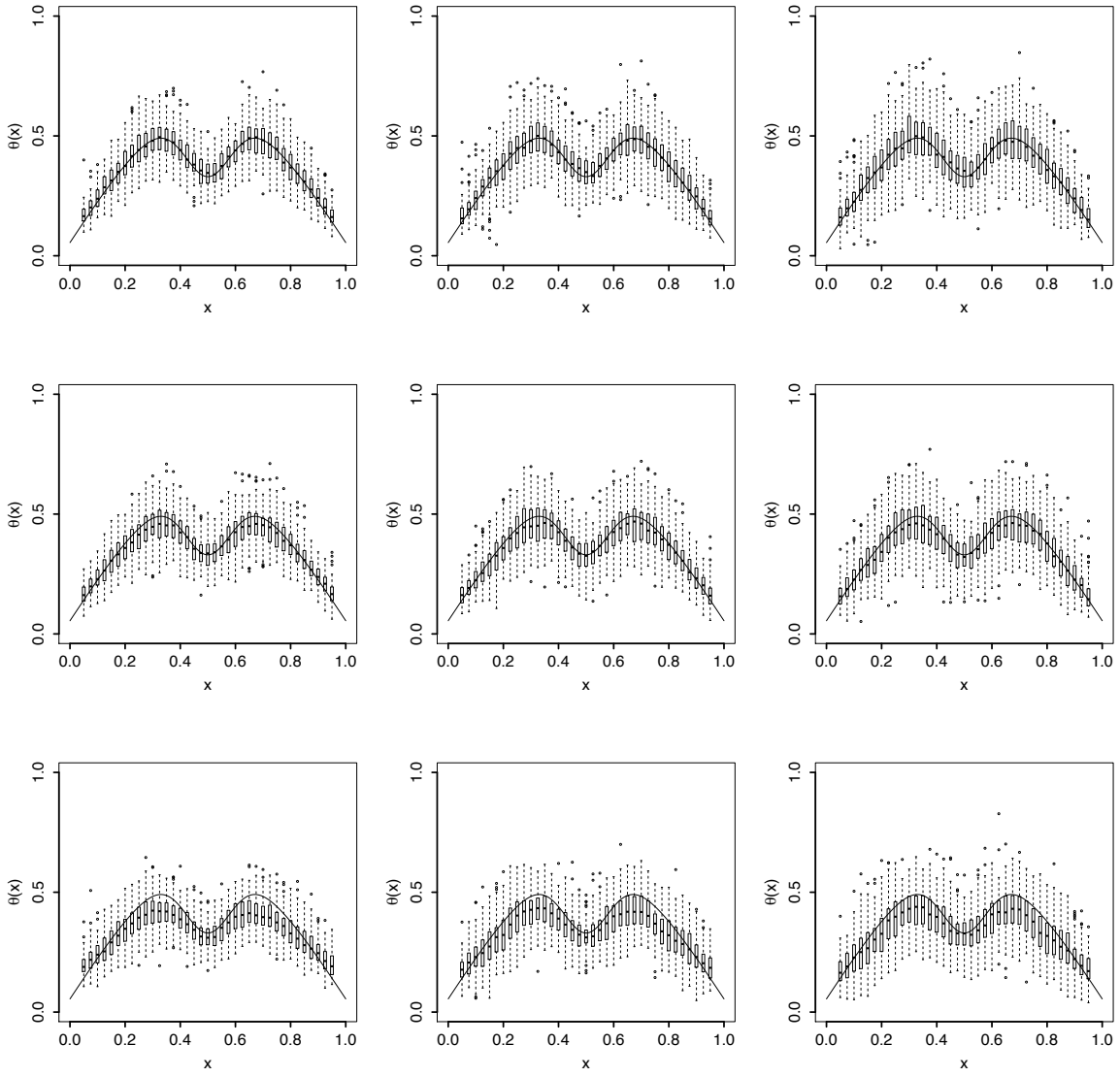

FIGURE 4.2 – Results for the Oracle method for the three distributions and different values of α. Row 1 : Strict Weibull, row 2 : Extended Weibull, row 3 : Perturbed Weibull ; Column 1 : $\alpha = 0.1$, column 2 : $\alpha = 0.5$, column 3 : $\alpha = 1$.

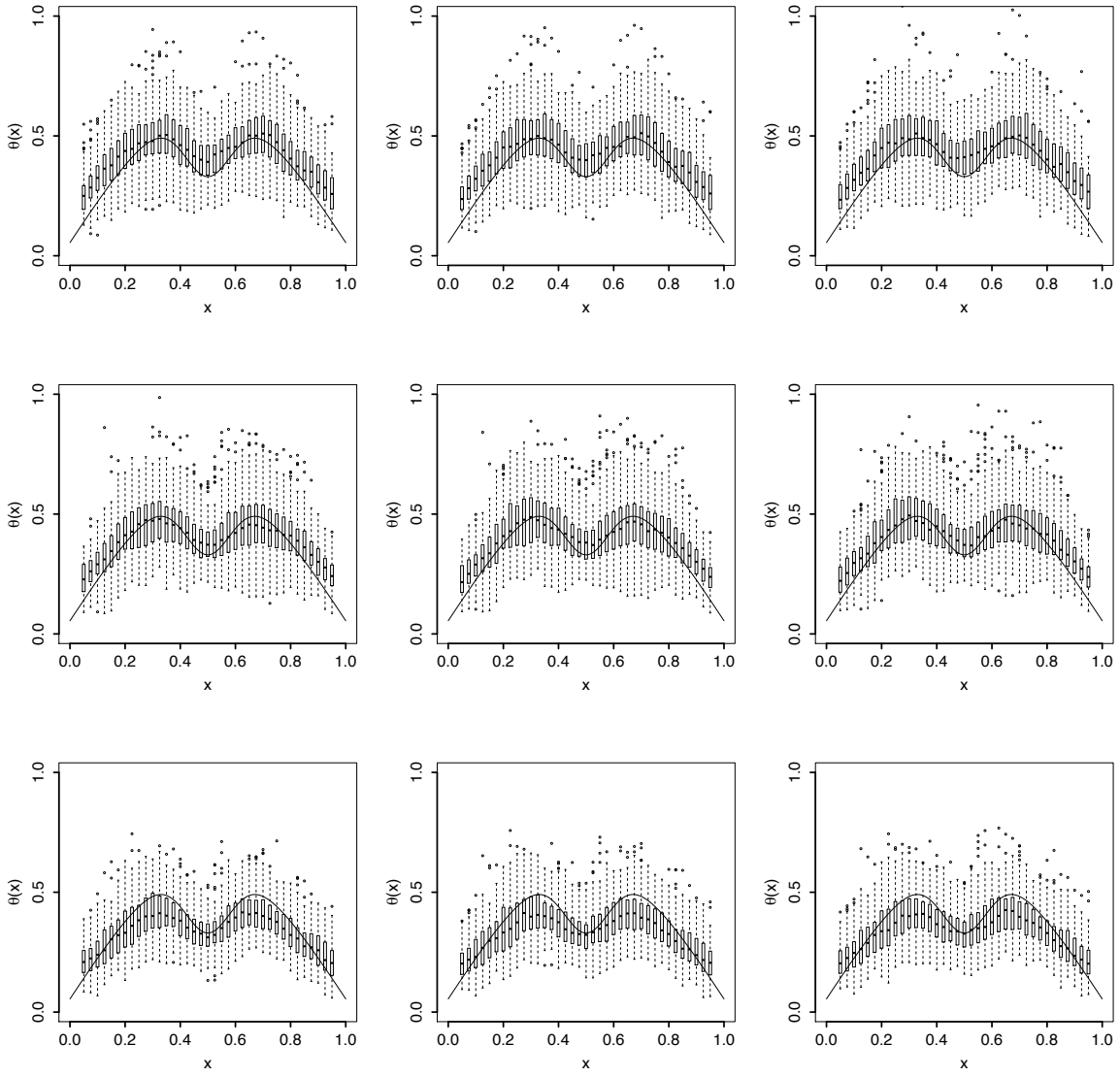

FIGURE 4.3 – Results for the data driven strategy for the three distributions and different values of α. Row 1 : Strict Weibull, row 2 : Extended Weibull, row 3 : Perturbed Weibull ; Column 1 : $\alpha = 0.1$, column 2 : $\alpha = 0.5$, column 3 : $\alpha = 1$.

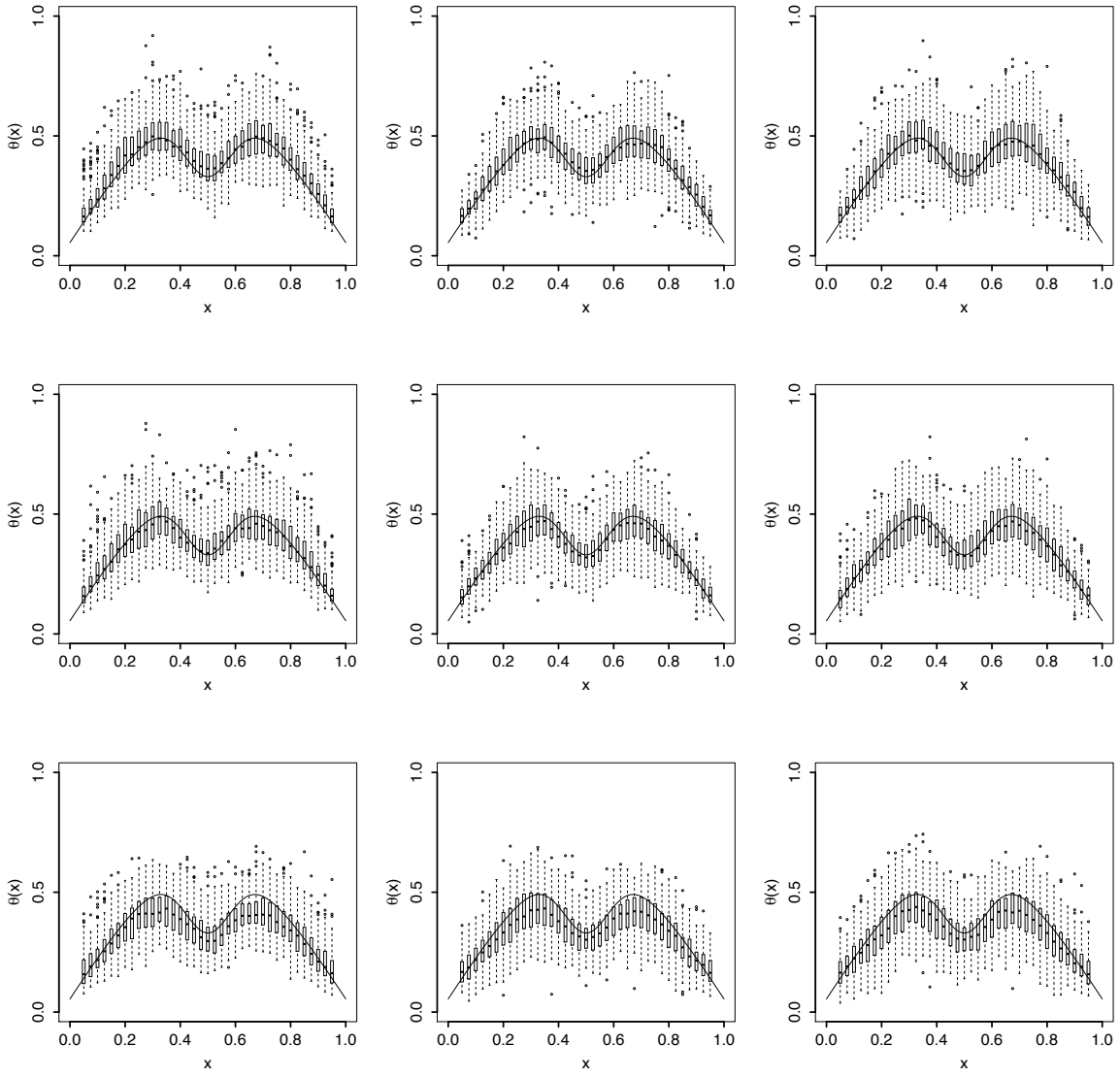

FIGURE 4.4 – Results for the Oracle method for the three contaminated distributions with $\varepsilon = 0.005$, $\beta = 4/3$ and different values of α. Row 1 : contaminated Strict Weibull, row 2 : contaminated Extended Weibull, row 3 : contaminated Perturbed Weibull ; Column 1 : $\alpha = 0.1$, column 2 : $\alpha = 0.5$, column 3 : $\alpha = 1$.

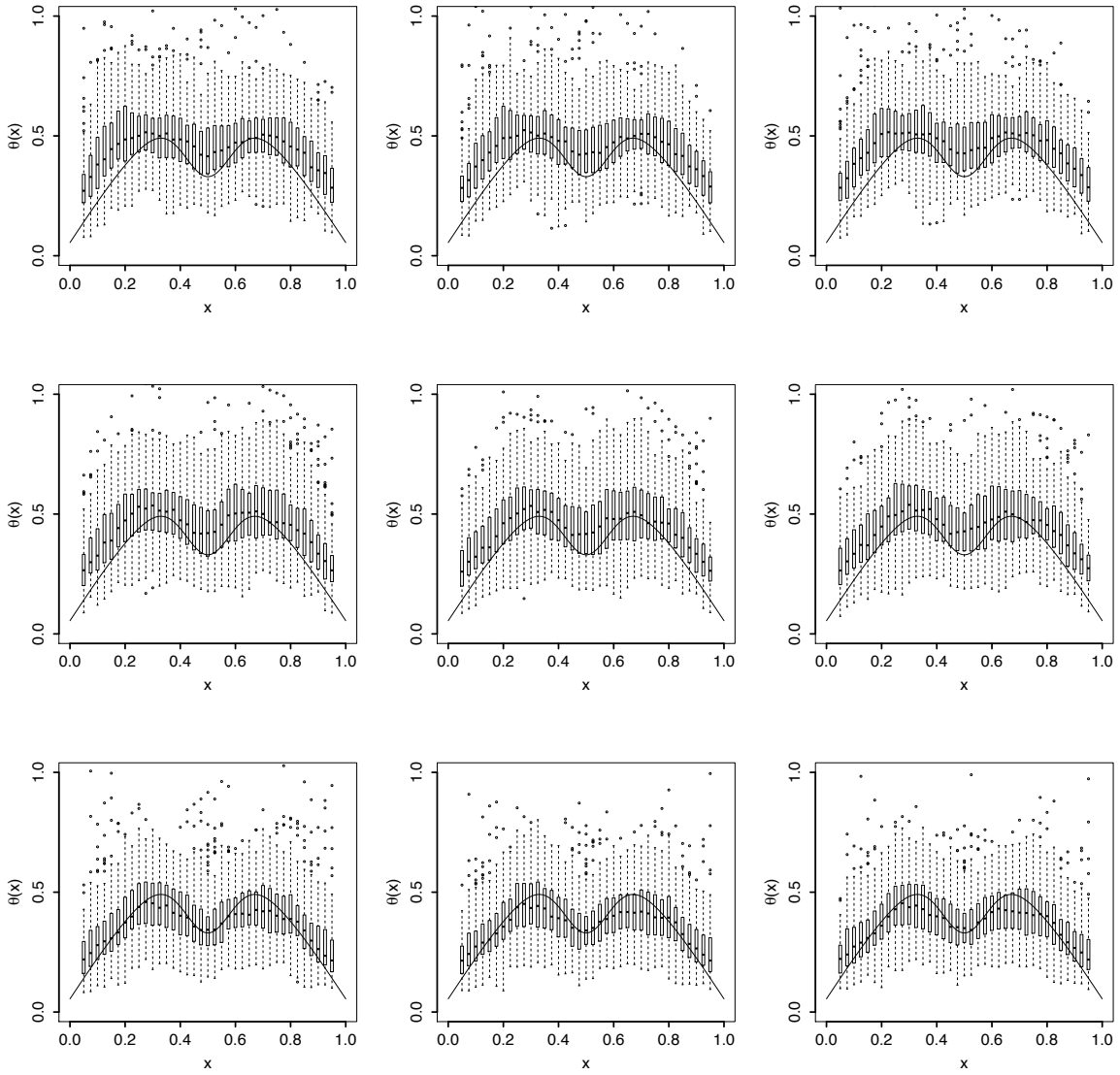

FIGURE 4.5 – Results for the data driven strategy for the three contaminated distributions with $\varepsilon = 0.005$, $\beta = 4/3$ and different values of α. Row 1 : contaminated Strict Weibull, row 2 : contaminated Extended Weibull, row 3 : contaminated Perturbed Weibull ; Column 1 : $\alpha = 0.1$, column 2 : $\alpha = 0.5$, column 3 : $\alpha = 1$.

Distribution	Value of α	Oracle			Data Driven		
		Bias	MSE	Sd	Bias	MSE	Sd
$\mathcal{W}(1/\theta(x), 1)$	0.1	0.051	0.004	0.065	0.095	0.014	0.097
	0.25	0.053	0.005	0.067	0.095	0.014	0.099
	0.5	0.059	0.006	0.076	0.096	0.015	0.104
	1	0.070	0.008	0.088	0.098	0.016	0.109
	2	0.079	0.010	0.100	0.101	0.017	0.111
$\mathcal{EW}(1/\theta(x), -1)$	0.1	0.053	0.005	0.065	0.089	0.013	0.100
	0.25	0.055	0.005	0.068	0.089	0.013	0.102
	0.5	0.060	0.006	0.075	0.091	0.014	0.107
	1	0.068	0.007	0.085	0.093	0.014	0.110
	2	0.068	0.007	0.085	0.093	0.014	0.110
$\widetilde{\mathcal{W}}(1/\theta(x), -5)$	0.1	0.062	0.006	0.062	0.077	0.009	0.084
	0.25	0.063	0.006	0.066	0.077	0.009	0.084
	0.5	0.069	0.007	0.074	0.079	0.010	0.086
	1	0.075	0.009	0.086	0.083	0.010	0.090
	2	0.082	0.010	0.092	0.084	0.011	0.091

TABLE 4.1 – Bias, mean square error and standard deviation of the MDPD estimator for the three distributions, the two approaches and different values of α in the case where there is no contamination.

Distribution F	Value of α	Oracle			Data Driven		
		Bias	MSE	Sd	Bias	MSE	Sd
$\mathcal{W}(1/\theta(x), 1)$	0.1	0.068	0.008	0.086	0.173	0.057	0.170
	0.25	0.062	0.007	0.079	0.148	0.041	0.151
	0.5	0.064	0.007	0.081	0.130	0.031	0.138
	1	0.072	0.009	0.091	0.126	0.028	0.133
	2	0.080	0.011	0.102	0.130	0.029	0.132
$\mathcal{EW}(1/\theta(x), -1)$	0.1	0.069	0.008	0.089	0.179	0.058	0.175
	0.25	0.062	0.006	0.079	0.150	0.040	0.154
	0.5	0.064	0.007	0.081	0.130	0.029	0.141
	1	0.071	0.008	0.089	0.124	0.026	0.135
	2	0.076	0.009	0.095	0.125	0.026	0.134
$\widetilde{\mathcal{W}}(1/\theta(x), -5)$	0.1	0.074	0.008	0.082	0.142	0.044	0.175
	0.25	0.069	0.008	0.077	0.117	0.029	0.151
	0.5	0.074	0.009	0.082	0.102	0.020	0.132
	1	0.079	0.010	0.090	0.098	0.017	0.121
	2	0.083	0.011	0.095	0.097	0.016	0.119

TABLE 4.2 – Bias, mean square error and standard deviation of the MDPD estimator for the three distributions, the two approaches and different values of α in the contaminated case with $\varepsilon = 0.005$ and $\beta = 4/3$.

Distribution F	Value of α	Oracle			Data Driven		
		Bias	MSE	Sd	Bias	MSE	Sd
$\mathcal{W}(1/\theta(x), 1)$	0.1	0.082	0.012	0.099	0.257	0.117	0.208
	0.25	0.069	0.008	0.087	0.228	0.094	0.200
	0.5	0.068	0.008	0.087	0.193	0.069	0.182
	1	0.076	0.010	0.097	0.177	0.057	0.167
	2	0.084	0.012	0.107	0.178	0.056	0.162
$\mathcal{EW}(1/\theta(x), -1)$	0.1	0.078	0.011	0.100	0.251	0.108	0.214
	0.25	0.067	0.008	0.087	0.209	0.078	0.197
	0.5	0.067	0.008	0.086	0.171	0.054	0.177
	1	0.074	0.009	0.093	0.157	0.044	0.160
	2	0.080	0.010	0.100	0.157	0.043	0.154
$\widetilde{\mathcal{W}}(1/\theta(x), -5)$	0.1	0.083	0.011	0.097	0.198	0.078	0.204
	0.25	0.074	0.009	0.086	0.160	0.055	0.190
	0.5	0.075	0.009	0.086	0.130	0.036	0.168
	1	0.080	0.010	0.093	0.118	0.027	0.149
	2	0.084	0.011	0.099	0.114	0.025	0.141

TABLE 4.3 – Bias, mean square error and standard deviation of the MDPD estimator for the three distributions, the two approaches and different values of α in the contaminated case with $\varepsilon = 0.01$ and $\beta = 4/3$.

Appendix

Proof of Lemma 4.1

The case $\alpha = \beta = r = 0$ is trivial, so we only consider case (ii). Let $p_n := F(u_n; x)$. Remark that

$$\begin{aligned} m(u_n, \alpha, \beta, r; x) &= \mathbb{E}\left(e^{-c_n\alpha\left[\left(\frac{Q(U;x)}{Q(p_n;x)}\right)^{1/\theta(x)}-1\right]} \left(\frac{Q(U;x)}{Q(p_n;x)}\right)^{\beta} \left(\ln\frac{Q(U;x)}{Q(p_n;x)}\right)_+^{r} \mathbb{1}_{\{Q(U;x)>Q(p_n;x)\}} \right) \\ &= \int_{p_n}^{\widetilde{p}_n} e^{-c_n\alpha\left[\left(\frac{Q(u;x)}{Q(p_n;x)}\right)^{1/\theta(x)}-1\right]} \left(\frac{Q(u;x)}{Q(p_n;x)}\right)^{\beta} \left(\ln\frac{Q(u;x)}{Q(p_n;x)}\right)^{r} du \\ &\quad + \int_{\widetilde{p}_n}^{1} e^{-c_n\alpha\left[\left(\frac{Q(u;x)}{Q(p_n;x)}\right)^{1/\theta(x)}-1\right]} \left(\frac{Q(u;x)}{Q(p_n;x)}\right)^{\beta} \left(\ln\frac{Q(u;x)}{Q(p_n;x)}\right)^{r} du \\ &=: m^{(1)}(u_n, \alpha, \beta, r; x) + m^{(2)}(u_n, \alpha, \beta, r; x), \end{aligned}$$

where U is a uniform $[0,1]$ random variable and $\widetilde{p}_n := 1 - \frac{1-p_n}{\ln\frac{e}{1-p_n}}$.

We will study the two terms separately. First remark that

$$\frac{Q(u;x)}{Q(p_n;x)} = \left(1 + \frac{-\ln\frac{1-u}{1-p_n}}{-\ln(1-p_n)}\right)^{\theta(x)} \frac{\ell\left(\left(1 + \frac{-\ln\frac{1-u}{1-p_n}}{-\ln(1-p_n)}\right)(-\ln(1-p_n)); x\right)}{\ell(-\ln(1-p_n); x)}. \tag{4.9}$$

Thus by the change of variable $z = \frac{1-u}{1-p_n}$, Assumption $(\mathcal{R})$ and the bound $\frac{\rho(x)-1}{2}z^2 \leq D_{\rho(x)}(1+z) - z \leq 0$, for $z \geq 0$, we deduce that

$$\begin{aligned} m^{(1)}(u_n, \alpha, \beta, r; x) &= (1-p_n) \int_{\frac{1-\widetilde{p}_n}{1-p_n}}^{1} e^{-c_n\alpha\left[\left(1+\frac{-\ln z}{c_n}\right)\left(1+b(c_n;x)D_{\rho(x)}\left(1+\frac{-\ln z}{c_n}\right)(1+o(1))\right)^{1/\theta(x)}-1\right]} \\ &\quad \times \left(1 + \frac{-\ln z}{c_n}\right)^{\theta(x)\beta} \left(1 + b(c_n;x)D_{\rho(x)}\left(1 + \frac{-\ln z}{c_n}\right)(1+o(1))\right)^{\beta} \\ &\quad \times \left(\ln\left[\left(1 + \frac{-\ln z}{c_n}\right)^{\theta(x)} \left(1 + b(c_n;x)D_{\rho(x)}\left(1 + \frac{-\ln z}{c_n}\right)(1+o(1))\right)\right]\right)^{r} dz \\ &= (1-p_n) \int_{\frac{1-\widetilde{p}_n}{1-p_n}}^{1} z^{\alpha} \left[\theta^r(x)\left(\frac{-\ln z}{c_n}\right)^{r} + \theta^r(x)\left(\theta(x)\beta - \frac{r}{2}\right)\left(\frac{-\ln z}{c_n}\right)^{r+1}\right. \\ &\quad + r\theta^{r-1}(x)\left(\frac{-\ln z}{c_n}\right)^{r} b(c_n;x)(1+o(1)) - \alpha\theta^{r-1}(x)\frac{(-\ln z)^{r+1}}{c_n^r} b(c_n;x)(1+o(1)) \\ &\quad \left. + O\left(\left(\frac{-\ln z}{c_n}\right)^{r+2}\right)\right] dz. \end{aligned}$$

Now remark that

$$\int_{\frac{1-\widetilde{p}_n}{1-p_n}}^{1} z^{\alpha}(-\ln z)^r dz = \frac{1}{(1+\alpha)^{r+1}}\left\{\Gamma(r+1) - \Gamma(r+1,(1+\alpha)\ln(1+c_n))\right\}.$$

Thus

$$\begin{aligned} m^{(1)}(u_n,\alpha,\beta,r;x) &= (1-p_n)\frac{\Gamma(1+r)}{(1+\alpha)^{1+r}}\theta^r(x)\left\{c_n^{-r} + \frac{\theta(x)\beta}{1+\alpha}\,c_n^{-1}\mathbb{1}_{\{r=0\}} + \frac{r-\alpha}{1+\alpha}\frac{b(c_n;x)}{\theta(x)}c_n^{-r}\right. \\ &\quad -c_n^{-1-\alpha}\mathbb{1}_{\{r=0\}} + o(b(c_n;x)c_n^{-r}) + O\left(\frac{\ln c_n}{c_n^{2+\alpha}}\mathbb{1}_{\{r=0\}}\right) + O\left(\frac{1}{c_n^2}\mathbb{1}_{\{r=0\}}\right) \\ &\quad \left. +O\left(\frac{(\ln c_n)^r}{c_n^{1+r+\alpha}}\mathbb{1}_{\{r>0\}}\right) + O\left(\frac{1}{c_n^{1+r}}\mathbb{1}_{\{r>0\}}\right)\right\}. \end{aligned}$$

Now, concerning the $m^{(2)}(u_n,\alpha,\beta,r;x)$ term, using the monotonicity of Q and of the exponential function leads to the inequality

$$\begin{aligned} m^{(2)}(u_n,\alpha,\beta,r;x) &\le e^{-c_n\alpha\left[\left(\frac{Q(\widetilde{p}_n;x)}{Q(p_n;x)}\right)^{1/\theta(x)}-1\right]}\int_{\widetilde{p}_n}^{1}\left(\frac{Q(u;x)}{Q(p_n;x)}\right)^{\beta}\left(\ln\frac{Q(u;x)}{Q(p_n;x)}\right)^{r}du \\ &=: T_1\times T_2. \end{aligned}$$

Clearly, using (4.9), Assumption $(\mathcal{R})$ and the bound for $D_{\rho(x)}(1+.)$, we have

$$\left(\frac{Q(\widetilde{p}_n;x)}{Q(p_n;x)}\right)^{1/\theta(x)} = 1 + \frac{b\,(c_n;x)}{\theta(x)}\frac{\ln(1+c_n)}{c_n}(1+o(1)) + \frac{\ln(1+c_n)}{c_n}.$$

This implies that

$$T_1 = e^{-\alpha\ln(1+c_n)}e^{-\frac{\alpha}{\theta(x)}b(c_n;x)\ln(1+c_n)(1+o(1))} = c_n^{-\alpha}(1+o(1))$$

since $\rho(x)<0$.

Now, concerning the term T_2, using the tail quantile function $U(y;x) := Q\left(1-\frac{1}{y};x\right)$, $y>1$, combined with the change of variables $z = \frac{1-p_n}{1-u}$, we deduce that

$$\begin{aligned} T_2 &= (1-p_n)\left(\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)}\right)^{r}\int_{1+c_n}^{\infty}\left[1+\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)}\frac{U\left(\frac{z}{1-p_n};x\right)-U\left(\frac{1}{1-p_n};x\right)}{a\left(\frac{1}{1-p_n};x\right)}\right]^{\beta}\frac{1}{z^2} \\ &\quad\times\left(\frac{\ln U\left(\frac{z}{1-p_n};x\right)-\ln U\left(\frac{1}{1-p_n};x\right)}{a\left(\frac{1}{1-p_n};x\right)/U\left(\frac{1}{1-p_n};x\right)}\right)^{r}dz, \end{aligned}$$

where a is the positive function that appears in the max-domain of attraction condition

$$\frac{U(tx)-U(t)}{a(t)} \to \ln x, \text{ as } t \to \infty, \text{for all } x > 0.$$

We have to study two cases depending on the sign of β.

First case : $\beta \leq 0$**.** Using the fact that $U(.)$ is an increasing function combined with Corollary B.2.10 in de Haan and Ferreira (2006, p. 376), we deduce that for p_n sufficiently large and ε sufficiently small that

$$T_2 \leq (1-p_n)\left(\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)}\right)^r O\left(c_n^{r\varepsilon-1}\right) = O\left(\frac{1-p_n}{c_n^{1+r-r\varepsilon}}\right),$$

where we have also used that

$$\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)} = O\left(\frac{1}{c_n}\right), \quad \text{as } p_n \uparrow 1,$$

see e.g. the proof of Lemma 1 in de Wet *et al.* (2013).

Second case : $\beta > 0$**.** Using again Corollary B.2.10 in de Haan and Ferreira (2006, p. 376) we have for p_n sufficiently large, δ and $\widetilde{\delta}$ positive constants, and ε and $\widetilde{\varepsilon}$ sufficiently small that

$$\begin{aligned}
T_2 &\leq (1-p_n)\delta^r \left(\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)}\right)^{r+\beta} \widetilde{\delta}^{\beta}\left[1+\frac{U\left(\frac{1}{1-p_n};x\right)}{a\left(\frac{1}{1-p_n};x\right)}\frac{1}{\widetilde{\delta}(1+c_n)^{\widetilde{\varepsilon}}}\right]^{\beta}\int_{1+c_n}^{\infty} z^{\beta\widetilde{\varepsilon}+r\varepsilon-2}dz \\
&= (1-p_n)\delta^r \frac{1}{(1+c_n)^{\widetilde{\varepsilon}\beta}}\left(\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)}\right)^{r}\left[1+\frac{a\left(\frac{1}{1-p_n};x\right)}{U\left(\frac{1}{1-p_n};x\right)}\widetilde{\delta}(1+c_n)^{\widetilde{\varepsilon}}\right]^{\beta}\int_{1+c_n}^{\infty} z^{\beta\widetilde{\varepsilon}+r\varepsilon-2}dz \\
&= O\left(\frac{1-p_n}{c_n^{1+r-r\varepsilon}}\right).
\end{aligned}$$

Finally

$$m^{(2)}(u_n,\alpha,\beta,r;x) = O\left(\frac{1-p_n}{c_n^{1+r+\alpha-r\varepsilon}}\right).$$

Combining all these results together leads to Lemma 4.1. □

Proof of Lemma 4.2

From the rule of repeated expectations we have that

$$m_n(K,\alpha,\beta,r;x) = \mathbb{E}(K_h(x-X)m(u_n,\alpha,\beta,r;X)).$$

Straightforward operations give

$$\begin{aligned}
m_n(K,\alpha,\beta,r;x) &= \int_\Omega K(z)m(u_n,\alpha,\beta,r;x-hz)f(x-hz)dz \\
&= m(u_n,\alpha,\beta,r;x)f(x) + m(u_n,\alpha,\beta,r;x)\int_\Omega K(z)(f(x-hz)-f(x))dz \\
&\quad + f(x)\int_\Omega K(z)(m(u_n,\alpha,\beta,r;x-hz)-m(u_n,\alpha,\beta,r;x))dz \\
&\quad + \int_\Omega K(z)(m(u_n,\alpha,\beta,r;x-hz)-m(u_n,\alpha,\beta,r;x))(f(x-hz)-f(x))dz \\
&=: m(u_n,\alpha,\beta,r;x)f(x) + T_3 + T_4 + T_5.
\end{aligned}$$

We now analyze each of the terms separately. By $(\mathcal{F})$ and $(\mathcal{K})$ we have that

$$\begin{aligned}
|T_3| &\leq m(u_n,\alpha,\beta,r;x)M_f\int_\Omega K(z)\|hz\|^{\eta_f}dz \\
&= O(m(u_n,\alpha,\beta,r;x)h^{\eta_f}),
\end{aligned}$$

and, by $(\mathcal{M})$ and $(\mathcal{K})$,

$$\begin{aligned}
|T_4| &\leq f(x)m(u_n,\alpha,\beta,r;x)\int_\Omega K(z)\left|\frac{m(u_n,\alpha\beta,r;x-hz)}{m(u_n,\alpha,\beta,r;x)}-1\right|dz \\
&= O(m(u_n,\alpha,\beta,r;x)\Phi_n(x)).
\end{aligned}$$

Using similar arguments one obtains $T_5 = O(m(u_n,\alpha,\beta,r;x)h^{\eta_f}\Phi_n(x))$. This proves the statement about the unconditional expectation.

For what concerns the convergence in probability, we have already from the first part of the proof that

$$\mathbb{E}\left(\widetilde{T}_n(K,\alpha,\beta,r;x)\right) = \frac{\theta^r(x)\Gamma(1+r)}{(1+\alpha)^{1+r}}\,(1+o(1)).$$

Also, again by using the result from the first part of the proof

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\widetilde{T}_n(K,\alpha,\beta,r;x)\right) &= \frac{c_n^{2r}\,\mathbb{V}\mathrm{ar}\left(K_h(x-X)e^{-c_n\alpha\left[\left(\frac{Y}{u_n}\right)^{1/\theta(x)}-1\right]}\left(\frac{Y}{u_n}\right)^{\beta}\left(\ln\frac{Y}{u_n}\right)_+^{r}\mathbb{1}_{\{Y>u_n\}}\right)}{n(\overline{F}(u_n;x)f(x))^2}\\ &= \frac{\theta^{2r}(x)\|K\|_2^2\Gamma(1+2r)}{(1+2\alpha)^{1+2r}nh^p\overline{F}(u_n;x)f(x)}\,(1+o(1)).\end{aligned}$$

Thus

$$\mathbb{V}\mathrm{ar}\left(\widetilde{T}_n(K,\alpha,\beta,r;x)\right)\to 0$$

under the assumptions of the lemma, and the convergence in probability follows. □

Proof of Corollary 4.1

First note that

$$\widehat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i),$$

is a classical kernel density estimator for f. As shown in Parzen (1962), if $nh^p\to\infty$, then for all $x\in\mathbb{R}^p$ where $f(x)>0$ one has that $\widehat{f}_n(x)\xrightarrow{\mathbb{P}} f(x)$. The result follows then by noting that

$$\frac{\widehat{\overline{F}}(u_n;x)}{\overline{F}(u_n;x)} = \frac{f(x)}{\widehat{f}_n(x)}\widetilde{T}_n(K,0,0,0;x).$$

□

Proof of Theorem 4.1

To prove the theorem we will adjust the arguments used to prove existence and consistency of solutions of the likelihood estimating equation, see e.g. Theorem 3.7 and Theorem 5.1 in Chapter 6 of Lehmann and Casella (1998), to the MDPD framework. Rescale the objective function $\widehat{\Delta}_\alpha(\theta;\widehat{c}_n)$ as

$$\widetilde{\Delta}_\alpha(\theta;\widehat{c}_n) := \frac{\widehat{\Delta}_\alpha(\theta;\widehat{c}_n)}{\overline{F}(u_n;x)f(x)c_n^{\alpha}}.$$

First, we will show that

$$\mathbb{P}_{\theta_0(x)}(\widetilde{\Delta}_\alpha(\theta_0(x);\widehat{c}_n) < \widetilde{\Delta}_\alpha(\theta;\widehat{c}_n)) \to 1 \tag{4.10}$$

as $n \to \infty$, for any θ sufficiently close to $\theta_0(x)$.

By Taylor's theorem

$$\begin{aligned}\widetilde{\Delta}_\alpha(\theta;\widehat{c}_n) - \widetilde{\Delta}_\alpha(\theta_0(x);\widehat{c}_n) &= \widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)(\theta-\theta_0(x)) + \frac{1}{2}\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)(\theta-\theta_0(x))^2 \\ &\quad + \frac{1}{6}\widetilde{\Delta}'''_\alpha(\widetilde{\theta};\widehat{c}_n)(\theta-\theta_0(x))^3,\end{aligned}$$

where $\widetilde{\theta}$ is a value between θ and $\theta_0(x)$. The term $\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)$ can be obtained from (4.4). Write $\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n) =: R_1 + R_2 + R_3 - R_4$. For analyzing the term R_1, we use the recursive relationships

$$\begin{aligned}\Gamma(a,b) &= e^{-b}b^{a-1} + (a-1)\Gamma(a-1,b), \\ \Psi(a,b) &= e^{-b}b^{a-1}\ln b + (a-1)\Psi(a-1,b) + \Gamma(a-1,b),\end{aligned}$$

Lemma 4.2, and the consistency of $\widehat{\overline{F}}(u_n;x)$, giving

$$R_1 \xrightarrow{\mathbb{P}} -\frac{\alpha}{\theta_0^{\alpha+1}(x)(1+\alpha)}.$$

For R_2 we rearrange the terms to obtain

$$\begin{aligned}R_2 &= \frac{1+\alpha}{\theta_0^{\alpha+1}(x)}(1+o_{\mathbb{P}}(1))\left\{\frac{T_n(K,\alpha,\alpha(1/\theta_0(x)-1),0;x)}{\overline{F}(u_n;x)f(x)}\right. \\ &\quad \left. + \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)\left[e^{-\widehat{c}_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)}-1\right]} - e^{-c_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)}-1\right]}\right]\left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta_0(x)-1)}\mathbb{1}_{\{Y_i>u_n\}}}{\overline{F}(u_n;x)f(x)}\right\} \\ &=: \frac{1+\alpha}{\theta_0^{\alpha+1}(x)}(R_{2,1}+R_{2,2})(1+o_{\mathbb{P}}(1)).\end{aligned}$$

By Lemma 4.2 we have that $R_{2,1} \xrightarrow{\mathbb{P}} (1+\alpha)^{-1}$. For the term $R_{2,2}$, use the mean value

theorem to obtain, with $\widetilde{c}_n$ being a random value between c_n and $\widehat{c}_n$,

$$\begin{aligned} R_{2,2} &= \alpha \ln \frac{\widehat{\overline{F}}(u_n;x)}{\overline{F}(u_n;x)} \left[\frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i) e^{-\widetilde{c}_n \alpha \left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)} - 1 \right]} \left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta_0(x)-1)+1/\theta_0(x)} \mathbb{1}_{\{Y_i>u_n\}}}{\overline{F}(u_n;x) f(x)} \right. \\ &\quad \left. - \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i) e^{-\widetilde{c}_n \alpha \left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)} - 1 \right]} \left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta_0(x)-1)} \mathbb{1}_{\{Y_i>u_n\}}}{\overline{F}(u_n;x) f(x)} \right] \\ &=: \alpha \ln \frac{\widehat{\overline{F}}(u_n;x)}{\overline{F}(u_n;x)} (R_{2,2,1} - R_{2,2,2}), \end{aligned}$$

which can be easily bounded as follows

$$\begin{aligned} R_{2,2,1} &\leq \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i) \left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta_0(x)-1)+1/\theta_0(x)} \mathbb{1}_{\{Y_i>u_n\}}}{\overline{F}(u_n;x) f(x)} = O_{\mathbb{P}}(1), \\ R_{2,2,2} &\leq \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-X_i) \left(\frac{Y_i}{u_n}\right)^{\alpha(1/\theta_0(x)-1)} \mathbb{1}_{\{Y_i>u_n\}}}{\overline{F}(u_n;x) f(x)} = O_{\mathbb{P}}(1), \end{aligned}$$

and therefore, by the consistency of $\widehat{\overline{F}}(u_n;x)$, the convergence $R_{2,2} \xrightarrow{\mathbb{P}} 0$ follows. Combining all results gives

$$R_2 \xrightarrow{\mathbb{P}} \frac{1}{\theta_0^{\alpha+1}(x)}.$$

The terms R_3 and R_4 can be analyzed in an analogous way and yield

$$R_3 \xrightarrow{\mathbb{P}} 0 \qquad \text{and} \qquad R_4 \xrightarrow{\mathbb{P}} \frac{1}{\theta_0^{\alpha+1}(x)(1+\alpha)}.$$

Thus $\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n) \xrightarrow{\mathbb{P}} 0$. Let $|\theta - \theta_0(x)| = r$, $r > 0$. With probability tending to 1 we have that

$$\left| \widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)(\theta - \theta_0(x)) \right| < r^3.$$

We now turn to the analysis of $\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)$. Let

$$\phi(a,b) := \int_b^\infty \ln^2 z \; z^{a-1} e^{-z} dz,$$

and

$$\widehat{T}_n(K,\alpha,\beta,r;x) := \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i) e^{-\widehat{c}_n\alpha\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)}-1\right]} \left(\frac{Y_i}{u_n}\right)^{\beta} \left(\ln\frac{Y_i}{u_n}\right)_+^{r} \mathbb{1}_{\{Y_i>u_n\}}.$$

Note that the function $\phi(a,b)$ satisfies the recursive relationship

$$\phi(a,b) = e^{-b}b^{a-1}\ln^2 b + (a-1)\phi(a-1,b) + 2\Psi(a-1,b). \tag{4.11}$$

After tedious calculations one obtains the following expression for $\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)$:

$$\begin{aligned}
&\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n) = \\
&\quad \frac{T_n(K,0,0,0;x)}{\overline{F}(u_n;x)f(x)} \frac{e^{\widehat{c}_n(1+\alpha)}\widehat{c}_n^{\alpha\theta_0(x)}}{\theta_0^{\alpha+2}(x)(1+\alpha)^{1+\alpha(1-\theta_0(x))}c_n^\alpha} \\
&\quad \times \big\{\alpha(1+\alpha)\Gamma(\alpha(1-\theta_0(x))+1,\widehat{c}_n(1+\alpha)) + 2\alpha^2\theta_0(x)\Psi(\alpha(1-\theta_0(x))+1,\widehat{c}_n(1+\alpha)) \\
&\quad +\alpha^2\theta_0^2(x)\phi(\alpha(1-\theta_0(x))+1,\widehat{c}_n(1+\alpha)) \\
&\quad -2\alpha^2\theta_0(x)\ln(\widehat{c}_n(1+\alpha))[\Gamma(\alpha(1-\theta_0(x))+1,\widehat{c}_n(1+\alpha)) + \theta_0(x)\Psi(\alpha(1-\theta_0(x))+1,\widehat{c}_n(1+\alpha))] \\
&\quad +\alpha^2\theta_0^2(x)\ln^2(\widehat{c}_n(1+\alpha))\Gamma(\alpha(1-\theta_0(x))+1,\widehat{c}_n(1+\alpha))\big\} \\
&\quad -\frac{(\alpha+1)^2\widehat{c}_n^\alpha}{\theta_0^{\alpha+2}(x)c_n^\alpha}\frac{\widehat{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1),0;x)}{\overline{F}(u_n;x)f(x)} - \frac{2(\alpha+1)^2\widehat{c}_n^\alpha}{\theta_0^{\alpha+3}(x)c_n^\alpha}\frac{\widehat{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1),1;x)}{\overline{F}(u_n;x)f(x)} \\
&\quad +\frac{2(\alpha+1)^2\widehat{c}_n^{\alpha+1}}{\theta_0^{\alpha+3}(x)c_n^\alpha}\frac{\widehat{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1)+1/\theta_0(x),1;x)}{\overline{F}(u_n;x)f(x)} \\
&\quad -\frac{\alpha(1+\alpha)\widehat{c}_n^\alpha}{\theta_0^{\alpha+4}(x)c_n^\alpha}\frac{\widehat{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1),2;x)}{\overline{F}(u_n;x)f(x)} \\
&\quad +\frac{(1+2\alpha)(1+\alpha)\widehat{c}_n^{\alpha+1}}{\theta_0^{\alpha+4}(x)c_n^\alpha}\frac{\widehat{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1)+1/\theta_0(x),2;x)}{\overline{F}(u_n;x)f(x)} \\
&\quad -\frac{\alpha(1+\alpha)\widehat{c}_n^{\alpha+2}}{\theta_0^{\alpha+4}(x)c_n^\alpha}\frac{\widehat{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1)+2/\theta_0(x),2;x)}{\overline{F}(u_n;x)f(x)}.
\end{aligned}$$

By a line of argumentation similar to that used for $\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)$ and also using (4.11) one obtains that under the conditions of the theorem

$$\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n) \xrightarrow{\mathbb{P}} \frac{1+\alpha^2}{\theta_0^{\alpha+2}(x)(1+\alpha)^2}. \tag{4.12}$$

Write

$$\begin{aligned}\frac{1}{2}\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)(\theta-\theta_0(x))^2 &= \frac{1+\alpha^2}{2\theta_0^{\alpha+2}(x)(1+\alpha)^2}(\theta-\theta_0(x))^2 \\ &\quad +\frac{1}{2}\left(\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)-\frac{1+\alpha^2}{\theta_0^{\alpha+2}(x)(1+\alpha)^2}\right)(\theta-\theta_0(x))^2.\end{aligned}$$

The random part in the right-hand side of the above display is in absolute value less than r^3 with probability tending to 1. There exist thus a $\delta_1 > 0$ and an $r_0 > 0$ such that for $r < r_0$

$$\frac{1}{2}\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)(\theta-\theta_0(x))^2 > \delta_1 r^2$$

with probability tending to 1.

For the third order derivative, one can show that $|\widetilde{\Delta}'''_\alpha(\theta;\widehat{c}_n)| \leq M(\boldsymbol{V})$, where $\boldsymbol{V} := [(X_1,Y_1),\ldots,(X_n,Y_n)]$, for $\theta \in (\theta_0(x)-r,\theta_0(x)+r)$, with $M(\boldsymbol{V}) \xrightarrow{\mathbb{P}} M$, which is bounded. The derivation is straightforward but lengthy and is therefore omitted. We can thus conclude that with probability tending to 1

$$\frac{1}{6}|\widetilde{\Delta}'''_\alpha(\widetilde{\theta};\widehat{c}_n)(\theta-\theta_0(x))^3| < \frac{1}{3}\,Mr^3.$$

Overall, we have that with probability tending to 1

$$\widetilde{\Delta}_\alpha(\theta;\widehat{c}_n)-\widetilde{\Delta}_\alpha(\theta_0(x);\widehat{c}_n) > \delta_1 r^2-(1+M/3)r^3,$$

which is positive if $r < \delta_1/(1+M/3)$, and thus (4.10) follows.

To complete the proof we adjust the line of argumentation of Theorem 3.7 in Chapter 6 of Lehmann and Casella (1998). Let $\delta > 0$ be such that $\theta_0(x)-\delta > 0$, and define

$$S_n(\delta) = \left\{\boldsymbol{v} : \widetilde{\Delta}_\alpha(\theta_0(x);\widehat{c}_n) < \widetilde{\Delta}_\alpha(\theta_0(x)-\delta;\widehat{c}_n) \text{ and } \widetilde{\Delta}_\alpha(\theta_0(x);\widehat{c}_n) < \widetilde{\Delta}_\alpha(\theta_0(x)+\delta;\widehat{c}_n)\right\}.$$

For any $\boldsymbol{v} \in S_n(\delta)$, since $\widetilde{\Delta}_\alpha(\theta;\widehat{c}_n)$ is differentiable with respect to θ, we have that there exists a $\widehat{\theta}_{n,\delta}(x) \in (\theta_0(x)-\delta,\theta_0(x)+\delta)$ where $\widetilde{\Delta}_\alpha(\theta;\widehat{c}_n)$ achieves a local minimum, so

$\widetilde{\Delta}'_\alpha(\widehat{\theta}_{n,\delta}(x);\widehat{c}_n)=0$. By the first part of the proof of the theorem, $\mathbb{P}_{\theta_0(x)}(S_n(\delta))\to 1$ for any δ small enough, and hence there exists a sequence $\delta_n\downarrow 0$ such that $\mathbb{P}_{\theta_0(x)}(S_n(\delta_n))\to 1$ as $n\to\infty$. Now let $\widehat{\theta}^*_n(x)=\widehat{\theta}_{n,\delta_n}(x)$ if $\boldsymbol{v}\in S_n(\delta_n)$ and arbitrary otherwise. Since $\boldsymbol{v}\in S_n(\delta_n)$ implies $\widetilde{\Delta}'_\alpha(\widehat{\theta}^*_n(x);\widehat{c}_n)=0$ we have that

$$\mathbb{P}_{\theta_0(x)}(\widetilde{\Delta}'_\alpha(\widehat{\theta}^*_n(x);\widehat{c}_n)=0)\geq\mathbb{P}_{\theta_0(x)}(S_n(\delta_n))\to 1,$$

as $n\to\infty$, which establishes the existence part. For the consistency of the solution sequence, note that for any fixed $\delta>0$ and n large enough

$$\mathbb{P}_{\theta_0(x)}(|\widehat{\theta}^*_n(x)-\theta_0(x)|<\delta)\geq\mathbb{P}_{\theta_0(x)}(|\widehat{\theta}^*_n(x)-\theta_0(x)|<\delta_n)\geq\mathbb{P}_{\theta_0(x)}(S_n(\delta_n))\to 1,$$

as $n\to\infty$, whence the consistency of the estimator sequence. □

Proof of Theorem 4.2

Let $r_n:=\sqrt{nh^p\overline{F}(u_n;x)}$. To prove the theorem we will make use of the Cramér-Wold device (see e.g., Severini, 2005, p. 337) according to which it is sufficient to show that

$$\Lambda_n:=\xi' r_n[\mathbb{T}_n-\mathbb{E}(\mathbb{T}_n)]\rightsquigarrow N\left(0,\frac{1}{f(x)}\,\xi'\Sigma\xi\right),$$

for all $\xi\in\mathbb{R}^J$.

Take an arbitrary $\xi\in\mathbb{R}^J$. A straightforward rearrangement of terms leads to

$$\begin{aligned}\Lambda_n &= \sum_{i=1}^n\sqrt{\frac{h^p}{n\overline{F}(u_n;x)}}\frac{1}{f(x)}\left[\sum_{j=1}^J\xi_j c_n^{r_j}K_{j,h}(x-X_i)e^{-c_n\alpha_j\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_i}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_i}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_i>u_n\}}\right.\\ &\quad\left.-\mathbb{E}\left(\sum_{j=1}^J\xi_j c_n^{r_j}K_{j,h}(x-X_i)e^{-c_n\alpha_j\left[\left(\frac{Y_i}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_i}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_i}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_i>u_n\}}\right)\right]\\ &=: \sum_{i=1}^n W_i.\end{aligned}$$

By the model assumptions, $W_1,\ldots,W_n$ are i.i.d. random variables, and therefore

$\mathbb{V}\mathrm{ar}(\Lambda_n) = n\mathbb{V}\mathrm{ar}(W_1)$. We have

$$\mathbb{V}\mathrm{ar}(W_1) = \frac{h^p}{n\overline{F}(u_n;x)f^2(x)} \sum_{j=1}^{J}\sum_{k=1}^{J} \xi_j\xi_k c_n^{r_j+r_k}\mathbb{C}_{j,k},$$

with

$$\begin{aligned}
\mathbb{C}_{j,k} \;:=\; & \mathbb{E}\left[K_{j,h}(x-X_1)K_{k,h}(x-X_1)e^{-c_n(\alpha_j+\alpha_k)\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_j+\beta_k}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_j+r_k}\mathbb{1}_{\{Y_1>u_n\}}\right] \\
& -\mathbb{E}\left[K_{j,h}(x-X_1)e^{-c_n\alpha_j\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_1>u_n\}}\right]\times \\
& \mathbb{E}\left[K_{k,h}(x-X_1)e^{-c_n\alpha_k\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_k}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_k}\mathbb{1}_{\{Y_1>u_n\}}\right].
\end{aligned}$$

By using the results of Lemmas 4.1 and 4.2 we have then

$$\mathbb{C}_{j,k} = \frac{\overline{F}(u_n;x)f(x)}{h^p c_n^{r_j+r_k}} \frac{\|K_jK_k\|_1\Gamma(1+r_j+r_k)\theta_0^{r_j+r_k}(x)}{(1+\alpha_j+\alpha_k)^{1+r_j+r_k}}\,(1+o(1)),$$

which gives that $\mathbb{V}\mathrm{ar}(\Lambda_n) = 1/f(x)\xi'\Sigma\xi(1+o(1))$. To establish the convergence in distribution to a normal random variable we have to verify the Lyapounov condition for triangular arrays of random variables (Billingsley, 1995, p. 362). In the present context this simplifies to verifying that $n\mathbb{E}|W_1|^3 \to 0$. We have

$$\begin{aligned}
\mathbb{E}|W_1|^3 \;\le\; & \left(\frac{h^p}{n\overline{F}(u_n;x)}\right)^{3/2}\frac{1}{f^3(x)}\times \\
& \left\{\mathbb{E}\left[\left(\sum_{j=1}^{J}|\xi_j|c_n^{r_j}K_{j,h}(x-X_1)e^{-c_n\alpha_j\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_1>u_n\}}\right)^3\right]\right. \\
& +3\mathbb{E}\left[\left(\sum_{j=1}^{J}|\xi_j|c_n^{r_j}K_{j,h}(x-X_1)e^{-c_n\alpha_j\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_1>u_n\}}\right)^2\right] \\
& \times\mathbb{E}\left[\sum_{j=1}^{J}|\xi_j|c_n^{r_j}K_{j,h}(x-X_1)e^{-c_n\alpha_j\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_1>u_n\}}\right] \\
& \left.+4\left[\mathbb{E}\left(\sum_{j=1}^{J}|\xi_j|c_n^{r_j}K_{j,h}(x-X_1)e^{-c_n\alpha_j\left[\left(\frac{Y_1}{u_n}\right)^{1/\theta_0(x)}-1\right]}\left(\frac{Y_1}{u_n}\right)^{\beta_j}\left(\ln\frac{Y_1}{u_n}\right)_+^{r_j}\mathbb{1}_{\{Y_1>u_n\}}\right)\right]^3\right\}.
\end{aligned}$$

Again by using Lemmas 4.1 and 4.2 we obtain that

$$\mathbb{E}|W_1|^3 = O\left(\left(n\sqrt{nh^p\overline{F}(u_n;x)}\right)^{-1}\right),$$

and hence, $n\mathrm{E}|W_1|^3 \to 0$. □

Proof of Theorem 4.3

Apply a Taylor series expansion to the estimating equation $\widetilde{\Delta}'_\alpha(\widehat{\theta}_n(x);\widehat{c}_n) = 0$ around $\theta_0(x)$. This gives

$$0 = \widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n) + \widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n)(\widehat{\theta}_n(x) - \theta_0(x)) + \frac{1}{2}\widetilde{\Delta}'''_\alpha(\widetilde{\theta}_n(x);\widehat{c}_n)(\widehat{\theta}_n(x) - \theta_0(x))^2$$

where $\widetilde{\theta}_n(x)$ is a random value between $\widehat{\theta}_n(x)$ and $\theta_0(x)$. A straightforward rearrangement of the terms leads then to

$$\begin{aligned} r_n(\widehat{\theta}_n(x) - \theta_0(x)) &= -\frac{1}{\widetilde{\Delta}''_\alpha(\theta_0(x);\widehat{c}_n) + \frac{1}{2}\widetilde{\Delta}'''_\alpha(\widetilde{\theta}_n(x);\widehat{c}_n)(\widehat{\theta}_n(x) - \theta_0(x))}\, r_n\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n) \\ &= -\frac{\theta_0^{\alpha+2}(x)(1+\alpha)^2}{1+\alpha^2}\, r_n\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)(1+o_{\mathbb{P}}(1)) \end{aligned} \tag{4.13}$$

by (4.12), the consistency of $\widehat{\theta}_n(x)$ and the boundedness of the third derivative. Another application of Taylor's theorem gives

$$r_n\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n) = r_n\widetilde{\Delta}'_\alpha(\theta_0(x);c_n) - \frac{\partial}{\partial\widehat{c}_n}\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)\bigg|_{\widetilde{c}_n} r_n \ln\frac{\widehat{\overline{F}}(u_n;x)}{\overline{F}(u_n;x)}$$

with $\widetilde{c}_n$ being a random value between $\widehat{c}_n$ and c_n. Direct computations allow us to prove that, under our assumptions, and using the second part of Lemma 4.2 and arguments similar to those used in the proof of Theorem 4.1, we have

$$\frac{\partial}{\partial\widehat{c}_n}\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n)\bigg|_{\widetilde{c}_n} = o_{\mathbb{P}}(1).$$

In addition, by Theorem 2 in de Wet *et al.* (2013), we deduce that

$$\begin{aligned} r_n\widetilde{\Delta}'_\alpha(\theta_0(x);\widehat{c}_n) &= r_n\widetilde{\Delta}'_\alpha(\theta_0(x);c_n)+o_{\mathbb{P}}(1) \\ &= -\frac{\alpha}{\theta_0^{\alpha+1}(x)(1+\alpha)}\, r_n\left[\widetilde{T}_n(K,0,0,0;x)-1\right] \\ &\quad +\frac{1+\alpha}{\theta_0^{\alpha+1}(x)}\, r_n\left[\widetilde{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1),0;x)-\frac{1}{1+\alpha}\right] \\ &\quad -\frac{1+\alpha}{\theta_0^{\alpha+2}(x)}\, r_n\left[\widetilde{T}_n(K,\alpha,\alpha(1/\theta_0(x)-1)+1/\theta_0(x),1;x)-\frac{\theta_0(x)}{(1+\alpha)^2}\right]+o_{\mathbb{P}}(1). \end{aligned} \tag{4.14}$$

Finally, combining (4.13) and (4.14) with Theorem 4.2 and the delta-method, Theorem 4.3 follows. □

Conclusions et Perspectives

Différents problèmes relatifs à la théorie des valeurs extrêmes sont abordés dans cette thèse. Avant de rentrer dans le vif du sujet, les notions essentielles à la compréhension de nos travaux sont présentées dans le Chapitre 1, notions qui couvrent différents domaines allant des valeurs extrêmes à la théorie de l'information en passant par les réseaux de neurones.

Puis, dans le Chapitre 2, nous nous intéressons à un problème crucial en climatologie, à savoir identifier ou non un changement au cours du temps dans la distribution des valeurs extrêmes. Pour cela nous avons proposé une approche non-paramétrique basée sur la divergence de Kullback-Leibler que nous avons adaptée au contexte des extrêmes en l'exprimant à l'aide de fonctions de répartition au lieu de densités. Grâce à des outils relevant des processus empiriques, nous avons établi la consistance en probabilité de l'estimateur proposé et nous avons illustré son comportement sur la base de simulations et de données réelles climatiques.

Dans le Chapitre 3, nous nous sommes aussi attachés à résoudre un problème lié aux extrêmes dans un contexte environnemental. Les stations météo disséminées sur tout le territoire permettent de mesurer les précipitations en temps réel à une multitude d'endroits. De ce fait, elles peuvent être utilisées pour étendre notre compréhension du comportement des précipitations à l'échelle du territoire tout entier. En particulier, une question pratique importante est de savoir où ajouter (resp. retirer) des stations pour gagner (resp. perdre) le plus (resp. le moins) d'information sur le comportement des extrêmes ? Afin

de répondre à cette question, une approche, basée sur les réseaux de neurones, appelée Query By Committee a été mise en oeuvre. Cette méthodologie a été appliquée à un jeu de données réelles de 331 stations en France. Nous avons ainsi mis en exergue les avantages, inconvénients et limites de cette approche.

Finalement, dans le Chapitre 4, nous avons cherché à estimer de façon robuste le paramètre de queue d'une distribution de type Weibull en présence de covariables aléatoires. Pour cela, nous avons utilisé des mesures de divergence entre deux densités, mesures dépendant d'un paramètre unique qu'il faut choisir de façon adéquate pour préserver un bon compromis entre efficacité et robustesse. Une estimation locale dans le voisinage d'un point de l'espace des covariables a été utilisée combinée à un critère de minimisation de la divergence pour l'ajustement.

Plusieurs perspectives à échéance plus ou moins proche apparaissent à l'issue de cette thèse. On peut citer notamment les points suivants :

• L'estimateur de la divergence approchée est consistant en probabilité. La preuve de sa normalité asymptotique est encore à ce jour un problème ouvert. Si on utilise une approche basée sur les processus empiriques comme pour montrer sa consistance, nous avons alors besoin d'hypothèses (d'existence de moments) qui ne sont pas satisfaites dans notre contexte des valeurs extrêmes.

• D'un point de vue pratique dans le Chapitre 2, nous avons cherché à répondre à la question si oui ou non il y avait un changement au cours du temps dans la distribution des extrêmes, mais en aucun cas, nous ne nous sommes posés la question de savoir dans quel sens s'effectuait ce changement. Bien entendu la question serait très intéressante d'un point de vue pratique, mais on ne peut pas directement y répondre avec l'approche proposée compte tenu du fait que l'entropie fournit uniquement une mesure de l'écart entre deux distributions, mais n'est pas une mesure orientée (la divergence étant toujours positive). Moyennant une modification appropriée de la divergence, serait-il possible d'appréhender la question ?

• Des données réelles climatiques concernant 24 stations météorologiques en Europe ont servi à illustrer notre méthodologie dans le Chapitre 2. Cependant, si on pouvait disposer d'un plus grand nombre de stations on pourrait espérer identifier ou non la présence d'une structure spatiale dans l'évolution du comportement des extrêmes de température sur le continent.

• Le Query By Committee est une approche qui s'est avérée très prometteuse dans notre contexte des valeurs extrêmes, ceci dit son efficacité a été illustrée dans le Chapitre 3 uniquement sur la base de simulations et sur l'application à un jeu de données réelles. Il serait donc particulièrement judicieux de bâtir une théorie qui permette de consolider cette méthode d'un point de vue théorique. D'un point de vue pratique, seules les précipitations ont été considérées alors que, bien entendu, les stations météo enregistrent d'autres caractéristiques climatiques comme la température ou la vitesse du vent. Une question intéressante est donc de savoir si les stations à enlever et/ou à rajouter se situent aux mêmes endroits quelle que soit la caractéristique considérée.

• Dans le Chapitre 3, nous ne nous sommes intéressés qu'au problème de supprimer des stations et pas d'en ajouter. Cette question est bien plus délicate et soulève un certain nombre de problèmes. En effet, elle nécessite une optimisation de la fonction de désaccord d, ce qui requiert un temps de calcul plus élevé. Un moyen d'y remédier pourrait consister à ajouter des stations sur une grille au lieu du territoire tout entier, ce qui nécessiterait le calcul de d uniquement sur un nombre fini de points. De plus, nous ne disposons pas d'information sur le comportement des extrêmes aux endroits où il n'y a pas de stations et par conséquent nous ne pourrons pas faire plus d'un pas à notre algorithme. Pour pallier ce problème, il serait intéressant d'utiliser des sorties de modèles climatiques qui nous permettraient alors d'avoir cette information.

• Comme le montre le résultat de normalité asymptotique du Chapitre 4 (Théorème 4.3), notre estimateur robuste de l'indice est asymptotiquement biaisé. Il serait bon de pouvoir proposer un estimateur débiaisé. Pour cela, il faudrait utiliser une approximation plus

précise que celle fournie par $\overline{G}$ en exploitant la condition du second ordre et appliquer ensuite à nouveau la méthode MDPD, comme cela a été proposée par Dierckx *et al.* (2013b) dans le cas de distributions de type Pareto.

Bibliographie

[1] Abarca-Del-Rio, R. et Mestre, O. (2006). Decadal to secular time scales variability in temperature measurements over France. *Geophysical Research Letters* **33**, 1–4.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

[3] Alexander, L.V., Zhang, X., Peterson, T.C., Cesar, J., Gleason, B., Tank, A.M.G.K, Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Kumar, K.R., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D.B., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M. et Vazquez-Aguirre, J.L. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research* **111**.

[4] Balkema, A. et de Haan, L. (1974). Residual life time at a great age. *Annals of Probability* **2**, 792–801.

[5] Basu, A., Harris, I.R., Hjort, N.L. et Jones, M.C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **85**, 549–559.

[6] Beirlant, J., Broniatowski, M., Teugels, J.L. et Vynckier, P. (1995). The mean residual life function at great age : applications to tail estimation. *Journal of Statistical Planning and Inference* **45**, 21–48.

[7] Beirlant, J., Goegebeur, Y., Segers, J et Teugels, J. (2004). *Statistics of extremes : Theory and applications*. John Wiley & Sons : New York.

[8] Berliner, M., Lu, Z.-Q. et Snyder, C. (1999). Statistical design for adaptative weather observations. *Journal of Atmospheric Sciences* **56**, 2536-2552.

[9] Billingsley, P. (1995). *Probability and Measure*. Wiley series in Probability and Mathematical Statistics.

[10] Bishop, C.M. (2006). *Pattern recognition and machine learning*. Science and Statistics.

[11] Brazauskas, V. et Serfling, R. (2000). Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. *Extremes* **3**, 231–249.

[12] Broniatowski, M. (1993). On the estimation of the Weibull tail coefficient. *Journal of Statistical Planning and Inference* **35**, 349–366.

[13] Burnham, K.P. et Anderson, D.R. (1998). *Model Selection and Inference : A Practicle Information-Theoretical Approach*. Springer-Verlag.

[14] Caselton, W.F. et Zidek, J.V. (1984). Optimal monitoring network designs. *Statistics and Probability Letters* **2**, 223-227.

[15] Ceresetti, D., Ursu, E., Carreau, J., Anquetin, S., Creutin, J.D., Gardes, L., Girard, S. et Molinié, G. (2012). Evaluation of classical spatial-analysis schemes of extreme rainfall. *Natural hazards and earth system sciences* **12**, 3229-3240.

[16] Chow, Y.S. et Teicher, H. (1978). *Probability theory : independence, interchangeability, martingales*. Springer.

[17] Cohn, D.A. (1996). Neural network exploration using optimal experiment design. *Neural Networks* **9**, 1071-1083.

[18] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

[19] Daouia, A., Gardes, L. et Girard, S. (2013). On kernel smoothing for extremal quantile regression. *À paraître dans Bernoulli*.

[20] Daouia, A., Gardes, L., Girard, S. et Lekina, A. (2011). Kernel estimators of extreme level curves. *Test* **20**, 311–333.

[21] Davis, R.A., Mikosch, T. et Cribben, I. (2012). Towards Estimating Extremal Serial Dependence via the Bootstrapped Extremogram. *Journal of Econometrics* **170**, 142–152.

[22] de Haan, L. (1996). Von Mises-Type Conditions in Second Order Regular Variation. *Journal of Mathematical Analysis and Applications* **197**, 400-410.

[23] de Haan, L. et Ferreira, A. (2006). *Extreme Value Theory : An Introduction*. Springer.

[24] de Haan, L. et Stadtmüller, U. (1996). Generalized Regular Variation of Second Order. *Journal of the Australian Mathematical Society* **61**, 381-395.

[25] de Wet, T., Goegebeur, Y., Guillou, A. et Osmann, M. (2013). Kernel regression with Weibull-type tails. *Soumis*.

[26] di Crescenzo, A. (1999). A Probabilistic Analogue of the Mean Value Theorem and its Applications to Reliability Theory. *Journal of Applied Probability* **36**, 706-719.

[27] Diebolt, J., Gardes, L., Girard, S. et Guillou, A. (2008). Bias-reduced estimators of the Weibull tail-coefficient. *Test* **17**, 311–331.

[28] Diebolt, J., Guillou, A. et Rached, I. (2007). Approximation of the distribution of excesses through a generalized probability weighted moments method. *Journal of Statistical Planning and Inference* **137**, 841–857.

[29] Dierckx, G., Beirlant, J., De Waal, D. et Guillou, A. (2009). A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference* **139**, 1905–1920.

[30] Dierckx, G., Goegebeur, Y. et Guillou, A. (2013a). An asymptotically unbiased minimum density power divergence estimator for the Pareto-tail index. *À paraître dans Journal of Multivariate Analysis*.

[31] Dierckx, G., Goegebeur, Y. et Guillou, A. (2013b). Local robust and asymptotically unbiased estimation of conditional Pareto-type tails. *Soumis*.

[32] Dupuis, D.J. (2012). Modeling Waves of Extreme Temperature : The Changing Tails of Four Cities. *Journal of the American Statistical Association* **107**, 24-39.

[33] Dupuis, D.J. et Field, C. (1998). Robust estimation of extremes. *Canadian Journal of Statistics* **26**, 119–215.

[34] Embrechts, P., Klüppelberg, C. et Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag.

[35] Fedorov, V.V. et Müller, W. (1989). Comparison of two approaches in the optimal design of an observation network. *Statistics* **20**, 339-351.

[36] Ferreira, A. et de Haan, L. (2012). The Generalized Pareto process ; with application. *arXiv*.

[37] Fisher, R.A. et Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or the smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* **24**, 180–190.

[38] Fowler, H.J. et Kilsby, C.G. (2003). A regional frequency analysis of United Kingdom extreme rainfall from 1961 to 2000. *International Journal of Climatology* **23**, 1313-1334.

[39] Frich, P., Alexander, L.V., Della-Marta, P., Gleason, B., Haylock, M., Tank, A.M.G.K. et Peterson, T.C. (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research* **19**, 193-212.

[40] Gannoun, A., Girard, S., Guinot, C. et Saracco, J. (2002). Reference ranges based on nonparametric quantile regression. *Statistics in Medicine* **21**, 3119–3135.

[41] Gardes, L. et Girard, S. (2005). Estimating extreme quantiles of Weibull tail distributions. *Communications in Statistics : Theory & Methods* **34**, 1065–1080.

[42] Gardes, L. et Girard, S. (2008a). A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis* **99**, 2368–2388.

[43] Gardes, L. et Girard, S. (2008b). Estimation of the Weibull-tail coefficient with linear combination of upper order statistics. *Journal of Statistical Planning and Inference* **138**, 1416–1427.

[44] Gardes, L., Girard, S. et Lekina, A. (2010). Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis* **101**, 419–433.

[45] Geluk, J.L. et de Haan, L. (1987). *Regular Variation, Extensions and Tauberian Theorems*. CWI Tract 40, Center for Mathematics and Computer Science, Amsterdam.

[46] Gilardi, N. (2004). Design of experiments by committee of neural networks. *Technical Report, Institut Français du Pétrole*.

[47] Giné, E. et Zinn, J. (1984). Some Limit Theorems For Empirical Processes. *The Annals of Probability* **12**, 929–989.

[48] Girard, S. (2004). A Hill type estimator of the Weibull tail coefficient. *Communications in Statistics : Theory & Methods* **33**, 205–234.

[49] Goegebeur, Y., Beirlant, J. et de Wet, T. (2010). Generalized kernel estimators for the Weibull tail coefficient. *Communications in Statistics : Theory & Methods* **39**, 3695–3716.

[50] Goegebeur, Y. et Guillou, A. (2011). A weighted mean excess function approach to the estimation of Weibull-type tails. *Test* **20**, 138–162.

[51] Goegebeur, Y., Guillou, A. et Schorgen, A. (2013). Nonparametric regression estimation of conditional tails – the random covariate case. *À paraître dans Statistics*.

[52] Greenwood, J.A., Landwehr, J.M., Matalas, N.C. et Wallis, J.R. (1979). Probability weighted moments : definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, **15**, 1049–1054.

[53] Grigga, O. et Tawn, J. (2012). Threshold models for river flow extremes. *Environmetrics* **23**, 295-305.

[54] Groisman, P.Y., Knight, R.W., Easterling, D.R., Karl, T.R., Hegerl, G.C. et Razuvaev, V.N. (2004). Trends in intense precipitation in the climate record. *Journal of Climate* **18**, 1326-1350.

[55] Hall, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **44**, 37–42.

[56] Hansen, L.K. et Salomon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 993–1001.

[57] Hastie, T., Tibshirani, R. et Friedman, J. (2001). *The elements of statistical learning*. Springer.

[58] Hoang, T.H., Parey, S. et Dacunha-Castelle, D. (2009). Multidimensional trends : The example of temperature. *European Physical Journal (Special Topics)* **174**, 113-124.

[59] Hoeting, J.A., Madigan, D., Raftery, A.E. et Volinsky, C.T. (1999). Bayesian model averaging : A tutorial. *Statistical Science* **14**, 382-417.

[60] Hosking, J.R.M. et Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, **29**, 339–349.

[61] Hosking, J.R.M. et Wallis, J.R. (1997). *Regional frequency analysis : an approach based on L-Moments*. Cambridge University Press.

[62] Jarušková, D. et Rencová, M. (2008). Analysis of annual maximal and minimal temperatures for some European cities by change point methods. *Environmetrics* **19**, 221-233.

[63] Juárez, S. et Schucany, W. (2004). Robust and efficient estimation for the generalized Pareto distribution. *Extremes* **7**, 237–251.

[64] Katz, R., Parlange, M. et Naveau, P. (2002). Extremes in hydrology. *Advances in Water Resources* **25**, 1287-1304.

[65] Kharin, V.V., Zwiers, F.W., Zhang, X. et Hegerl, G.C. (2007). Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *Journal of Climate* **20**, 1419-1444.

[66] Kiefer, J. et Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics* **30**, 271-294.

[67] Kim, M. et Lee, S. (2008). Estimation of a tail index based on minimum density power divergence. *Journal of Multivariate Analysis* **99**, 2453–2471.

[68] Klüppelberg, C. et Villaseñor, J.A. (1993). Estimation of distribution tails - a semiparametric approach. *Blätter der Deutschen Gesellschaft für Versicherungsmathematik* **21**, 213–235.

[69] Krogh, A. et Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. *Advances in Neural Information Processing Systems* **7**, 231–238.

[70] Kullback, S. (1968). *Information theory and statistics.* Dover books on mathematics. Dover Publications.

[71] LeCun, Y., Bottou, L., Orr, G.B. et Müller, K.-R. (1998). *Efficient Backprop*, Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science LNCS 1524, Springer Verlag.

[72] Lehmann, E.L. et Casella, G. (1998). *Theory of Point Estimation.* Springer.

[73] Naveau, P., Guillou, A. et Rietsch, T. (2013). A non-parametric entropy-based approach to detect changes in climate extremes. *À paraître dans Journal of the Royal Statistical Society : Series B, Statistical Methodology.*

[74] Nychka, D., Yang, Q. et Royle, J.A. (1997). Constructing spatial designs using regression subset selection. *Statistics for the environment-3 : Pollution assessment and control*, 131-154.

[75] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065–1076.

[76] Peng, L. et Welsh, A. (2001). Robust estimation of the generalized Pareto distribution. *Extremes* **4**, 53–65.

[77] Penny, W.D. et Roberts, S.J. (2000). Variational Bayes for 1-Dimensional Mixture Models. *Technical report, Department of Engineering Science, Oxford University.*

[78] Perrone, M. P. et Cooper, L.N. (1993). When networks disagree : Ensemble methods for hybrid neural networks. *Neural Networks for Speech and Image Processing*, 126–142.

[79] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics* **3**, 119-131.

[80] Sabourin, A., Naveau, P. et Fougères, A.-L. (2013). Bayesian model averaging for multivariate extremes. *Extremes*, 1-26.

[81] Seung, H.S., Opper, M. et Sompolinsky, H. (1992). Query by committee. *Proc. of the Fifth Workshop on Computational Learning Theory*, 287–294.

[82] Severini, T. (2005). *Elements of Distribution Theory*. Cambridge Series in Statistical and Probabilistic Mathematics.

[83] Shaby, B. et Reich, B.J. (2012). Bayesian spatial extreme value analysis to assess the changing risk of concurrent high temperatures across large portions of European cropland. *Environmetrics*, **23**, 638-648.

[84] Shaked, M. et Shanthikumar, J.G. (1994). *Stochastic orders and their applications*, Academic Press (Boston).

[85] Smith, R.L. (1984). Threshold methods for sample extremes. *Statistical Extremes and Applications*, ed. J. Tiago de Oliveira, Dordrecht : D. Reidel, 621–638.

[86] Smith, R.L. (1985). Maximum likelihood estimation in a class of non regular cases. *Biometrika* **72**, 69-90.

[87] Smith, R.L. (2001). *Environmental statistics*. University of North California.

[88] Vandewalle, B., Beirlant, J., Christmann, A. et Hubert, M. (2007). A robust estimator

for the tail index of Pareto-type distributions. *Computational Statistics and Data Analysis* **51**, 6252–6268.

[89] Wang, H. et Tsai, C.L. (2009). Tail index regression. *Journal of the American Statistical Association* **104**, 1233–1240.

[90] Wellner, J.A. (1978). Limit Theorems for the Ratio of the Empirical Distribution Function to the True Distribution Function. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* **45**, 73-88.

[91] Wikle, C.K. et Royle, J.A. (1999). Space-time dynamic desing of environmental monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics* **4**, 489–507.

[92] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks* **5**, 241–259.

[93] Yao, Q. (1999). Conditional predictive regions for stochastic processes. *Technical report, University of Kent at Canterbury.*

[94] Zwiers, F.W., Alexander, L.V., Hegerl, G.C., Kossin, J.P., Knuston, T.R., Naveau, P., Nicholls, N., Schär, C., Seneviratne, S.I. et Zhang, X. (2013). Climate extremes : challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. *Climate Science for Serving Society*, 339-389.

[95] Zwiers, F.W., Zhang, X. et Feng, Y. (2011). Anthropogenic influence on long return period daily temperature extremes at regional scales. *Journal of Climate* **24**, 881-892.

Dans cette thèse nous apportons plusieurs contributions, à la fois théoriques et appliquées, à la théorie des valeurs extrêmes.

Les deux premiers chapitres de cette thèse s'attachent à répondre à des questions cruciales en climatologie. La première question est de savoir si un changement dans le comportement des extrêmes de température peut être détecté entre le début du siècle et aujourd'hui. Pour cela nous proposons d'utiliser la divergence de Kullback Leibler, que nous adaptons au contexte des extrêmes. Des résultats théoriques et des simulations permettent de valider l'approche proposée, dont les performances sont ensuite illustrées sur un jeu de données réelles. La deuxième question quant à elle combine les réseaux de neurones à la théorie des valeurs extrêmes afin de savoir où ajouter (resp. retirer) des stations dans un réseau pour gagner (resp. perdre) le plus (resp. le moins) d'information sur le comportement des extrêmes. Un algorithme, le Query By Committee, issu de la théorie du machine learning est développé puis appliqué à un jeu de données réelles. Les avantages, inconvénients et limites de cette approche sont ainsi mis en exergue.

La dernier chapitre de la thèse traite d'un sujet plus théorique, à savoir l'estimation robuste du paramètre de queue d'une distribution de type Weibull en présence de covariables aléatoires. Nous proposons un estimateur robuste en utilisant un critère de minimisation de la divergence entre deux densités et étudions ses propriétés asymptotiques. Des simulations illustrent les performances de l'estimateur à distance finie.

Cette thèse offre de nombreuses perspectives dont une liste non exhaustive est dressée en conclusion.

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
UMR 7501
Université de Strasbourg et CNRS
7 Rue René Descartes
67 084 STRASBOURG CEDEX

Tél. 03 68 85 01 29
Fax 03 68 85 03 28

www-irma.u-strasbg.fr
irma@math.unistra.fr

IRMA
http ://www-irma.u-strasbg.fr/

ISSN 0755-3390