HAL open science

Contributions à la théorie des valeurs extrêmes et à la réduction de dimension pour la régression

Laurent Gardes

To cite this version:

Laurent Gardes. Contributions à la théorie des valeurs extrêmes et à la réduction de dimension pour la régression. Mathématiques [math]. Université Joseph-Fourier - Grenoble I, 2010. tel-00540747

HAL Id: tel-00540747

https://theses.hal.science/tel-00540747v1

Submitted on 29 Nov 2010

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

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

UNIVERSITÉ JOSEPH FOURIER – GRENOBLE I
LABORATOIRE JEAN KUNTZMANN

Mémoire d'habilitation présenté par

Laurent GARDES

en vue de l'obtention du diplôme d'

HABILITATION À DIRIGER DES RECHERCHES

de l'UNIVERSITÉ JOSEPH FOURIER

(Spécialité : Informatique et Mathématiques Appliquées)

Titre

Contributions à la théorie des valeurs extrêmes et à la réduction de dimension pour la régression

Soutenue le 17 novembre 2010 devant le jury composé de :

John Einmahl	Professeur, Université de Tilburg
Stéphane Girard	Chargé de Recherches, INRIA Rhône-Alpes
Armelle Guillou	Professeur, Université de Strasbourg
Clémentine Prieur	Professeur, Université de Grenoble 1
Holger Rootzén	Professeur, Université de Göteborg
Jérôme Saracco	Professeur, Université de Bordeaux 4

Habilitation préparée au sein de l'équipe MISTIS
(INRIA Rhône-Alpes, Laboratoire Jean Kuntzmann)

à Mickaël et Leslie,

à Paul.

Remerciements

Je tiens tout d'abord à remercier John Einmahl, Holger Rootzen et Jérôme Saracco pour avoir écrit un rapport sur mon habilitation à diriger des recherches. Ils ont pris sur leur temps precieux pour lire très attentivement mon manuscript et participer à ma soutenance. Merci !

Merci aussi à Clémentine Prieur pour avoir présidé mon jury et pour sa lecture conciencieuse de mon document.

Un grand merci à Stéphane Girard et Armelle Guillou pour avoir pris part à la soutenance et surtout pour les nombreuses collaborations scientifiques que nous avons eues. Stéphane et Armelle, je vous suis très reconnaissant pour toute l'aide que vous m'avez apportée ainsi que pour l'amitié que vous m'avez témoignée depuis ma thèse.

Merci aussi à tous mes co-auteurs ainsi qu'à toutes les personnes avec qui j'ai travaillé. Un remerciement tout particulier à l'équipe Mistis qui m'a accueillie et dans laquelle j'ai pu travailler dans une ambiance très chaleureuse.

Enfin, un grand merci à toute ma famille et plus particulièrement à ma femme Muriel et mes enfants Mickaël et Leslie qui m'ont toujours encouragé.

Table des matières

Introduction **i**

1 Estimation de quantiles extrêmes pour des lois à queue de type Weibull **1**
1.1 Introduction . . . 1
1.2 Introduction à la théorie des valeurs extrêmes . . . 2
1.2.1 Convergence en loi du maximum d'un échantillon . . . 2
1.2.2 Caractérisation des domaines d'attraction . . . 3
1.2.2.1 Domaine d'attraction de Fréchet . . . 3
1.2.2.2 Domaine d'attraction de Weibull . . . 4
1.2.2.3 Domaine d'attraction de Gumbel . . . 5
1.3 Inférence sur les lois à queue de type Weibull . . . 5
1.3.1 Estimation de l'indice de queue de Weibull . . . 6
1.3.1.1 Utilisation de poids . . . 7
1.3.1.2 Utilisation d'autres suites de normalisation . . . 8
1.3.1.3 Un estimateur de θ débiaisé . . . 12
1.3.2 Estimation de quantiles extrêmes . . . 14
1.3.2.1 Etude de la famille d'estimateurs $\mathcal{Q}_{\alpha_n}$ . . . 15
1.3.2.2 Un estimateur de $q(\alpha_n)$ débiaisé . . . 16
1.4 Autres domaines d'attraction . . . 17
1.4.1 Domaine d'attraction de Weibull . . . 17
1.4.2 Ensemble des domaines d'attraction . . . 18
1.5 Conclusion et perspectives . . . 20

2 Estimation de quantiles extrêmes conditionnels **25**
2.1 Introduction . . . 25
2.2 Cas d'une loi conditionnelle à queue lourde avec une covariable déterministe . . . 26
2.2.1 Estimation de l'indice des valeurs extrêmes conditionnel . . . 27
2.2.1.1 Définition de la famille d'estimateurs . . . 27
2.2.1.2 Résultat de normalité asymptotique . . . 28
2.2.1.3 Quelques choix de fonctions poids . . . 29
2.2.2 Estimation de quantiles extrêmes conditionnels . . . 33
2.2.3 Application à l'estimation de niveaux de retour . . . 36
2.3 Cas d'une loi conditionnelle à queue lourde avec une covariable aléatoire . . . 42
2.3.1 Estimation de petites probabilités . . . 43
2.3.2 Estimation de quantiles extrêmes conditionnels . . . 44

2.3.3 Application à l'estimation de l'indice des valeurs extrêmes conditionnel . 45
2.4 Estimation de support 46
2.5 Conclusion et perspectives 48

3 Réduction de dimension pour la régression 53
3.1 Introduction et motivations 53
3.2 Régularisation de la méthode SIR 57
3.2.1 Modèle de régression inverse 58
3.2.2 Estimation des paramètres par maximum de vraisemblance 59
3.2.3 Régularisation par introduction d'un a priori Gaussien 60
3.2.4 Discussion sur une autre méthode de régularisation 64
3.3 Illustration sur simulation 64
3.4 Application à l'étude d'images hyperspectrales du sol martien 68
3.5 Conclusion et perspectives 73

Conclusion générale 77

Quelques notations

$\xrightarrow{d}$, $\xrightarrow{\mathrm{P}}$	Convergence en distribution et en probabilité.
$\mathbb{I}$	Fonction indicatrice.
$x \propto y$, $x \approx y$	x proportionnel à y, x proche de y.
A^t	Transposée de la matrice A.
$\|.\|_2$	Norme euclidienne.
$\arg\max$, $\arg\min$	Argument du maximum, argument du minimum.

Introduction

Dans ce mémoire d'Habilitation à Diriger des Recherches (HDR), je fais la synthèse de mes travaux de recherche effectués essentiellement après ma thèse de doctorat obtenue en 2003. Mes activités de recherche peuvent être divisées en trois thèmes :

1) Inférence sur les lois à queue de type Weibull,
2) estimation de quantiles extrêmes pour des lois conditionnelles à queue lourde,
3) réduction de dimension pour la régression.

Chacun de ces thèmes fait l'objet d'un chapitre. Les chapitres 1 et 2 s'inscrivent dans le domaine de la théorie des valeurs extrêmes. Cette théorie a pour objectif l'étude des queues de distribution et est utilisée notamment pour estimer des quantiles dits *extrêmes* c'est à dire dont l'ordre tend vers zéro avec la taille de l'échantillon. Le chapitre 3 porte sur la réduction de dimension pour la régression. La question à laquelle nous essayons de répondre dans ce chapitre est la suivante : comment trouver le meilleur sous-espace où projeter des variables explicatives vivant dans des espaces de grande dimension tout en conservant le maximum d'information sur les variables à prédire ?
Je donne ci-dessous un bref résumé du contenu de chaque chapitre.

Dans le Chapitre 1 nous nous intéressons à une famille particulière de lois : les lois à queue de type Weibull. Ces lois possèdent une fonction de survie qui décroit à une vitesse exponentielle (on parle aussi de *queue légère*). Des exemples de telles lois sont les lois exponentielle, normale, gamma, etc ... La vitesse de convergence de la queue de distribution est controlée par un paramètre de forme appelé *indice de queue de Weibull*. Nous introduisons et étudions le comportement asymptotique d'estimateurs de cet indice et des quantiles extrêmes. Nous présentons aussi une méthode de réduction du biais basée sur un modèle de régression exponentiel.

Dans de nombreuses applications, la quantité d'intérêt est mesurée simultanément avec une covariable. Par exemple, en hydrologie, on mesure la quantité horaire de précipitation Y tombée en fonction de la position géographique x et on souhaite par exemple estimer une carte de période de retour. En d'autres termes, il s'agit d'estimer les quantiles extrêmes de la loi de Y conditionnellement au fait que la mesure a été effectuée à la position géographique x. Ce problème est traité dans le Chapitre 2 où l'on fait l'hypothèse que la loi conditionnelle de Y est une loi à queue lourde c'est à dire dont la fonction de survie décroit comme une fonction puissance. Nous proposons des estimateurs des quantiles extrêmes de la loi conditionnelle qui sont à présent fonction de la covariable. Nous les utilisons également pour étudier le comportement des pluies extrêmes dans la région Cévennes-Vivarais. Cette étude fait partie du projet MEDUP (Forecast

and projection in climate scenario of mediterranean intense events : Uncertainties and propagation on environment) financé par le programme ANR, Vulnérabilité, Milieux et Climats (VMC).

Enfin, dans le Chapitre 3, nous proposons de régulariser la méthode *Sliced Inverse Regression* (SIR) en introduisant un a priori Gaussien. La méthode SIR est une méthode de réduction de dimension ayant pour objectif de trouver le meilleur sous-espace sur lequel projeter les variables explicatives $X_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ afin d'expliquer au mieux les variables à prédire $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$. Lorsque la dimension d est grande, la matrice de variance-covariance $\hat{\Sigma}$ des variables explicatives X_i est généralement mal conditionnée rendant alors l'application de la méthode SIR délicate (car elle est basée sur l'inversion de la matrice $\hat{\Sigma}$). Dans ce cadre, nous proposons une méthode de régularisation de SIR en introduisant un a priori Gaussien. Cette régularisation a pour effet l'amélioration du conditionnement de la matrice de variance-covariance des variables explicatives. La méthode SIR régularisée est utilisée pour estimer les propriétés physiques du sol martien à partir d'images hyperspectrales. Cette application fait partie du projet VAHINE (Visualisation et analyse d'images hyperspectrales multi-dimensionnelles en astrophysique) financé par le programme ANR, Masse de Données et COnnaissances (MDCO).

Plusieurs domaines de la Statistique sont abordés dans ce mémoire : la statistique des valeurs extrêmes pour les Chapitres 1 et 2, la statistique fonctionnelle dans le Chapitre 2 et enfin la réduction de dimension et la régression non paramétrique dans le Chapitre 3. Les problèmes traités dans le Chapitre 1 sont des problèmes de statistique théorique. Les Chapitres 2 et 3 relèvent à la fois de la statistique théorique (estimation, résultats de normalité asymptotique) et de la statistique appliquée (en hydrologie et en planétologie).

Pour faciliter la lecture de ce document, j'ai choisi de ne pas donner les démonstrations des différents théorèmes énoncés. Le lecteur pourra, s'il le souhaite, les consulter dans les articles cités en référence. Pour les mêmes raisons, les simulations permettant de vérifier la validité des estimateurs proposés ne sont pas toutes présentées dans ce mémoire. Je donne ci-dessous les publications associées à chaque chapitre.

Publications associées au Chapitre 1

- J. Diebolt, L. Gardes, S. Girard & A. Guillou. Bias-reduced estimators of the Weibull tail-coefficient, *Test*, **17**, 311-331, (2008).
- J. Diebolt, L. Gardes, S. Girard & A. Guillou. Bias-reduced extreme quantiles estimators of Weibull tail-distributions, *Journal of Statistical Planning and Inference*, **138**, 1389-1401, (2008).
- L. Gardes & S. Girard. Estimation of the Weibull tail-coefficient with linear combination of upper order statistics, *Journal of Statistical Planning and Inference*, **138**, 1416-1427, (2008).
- L. Gardes & S. Girard. Comparison of Weibull tail-coefficient estimators, *REVSTAT - Statistical Journal*, **4(2)**, 163-188, (2006).
- L. Gardes & S. Girard. *Asymptotic properties of a Pickands type estimator of the extreme value index*, In Louis R. Velle, editor, Focus on probability theory, Nova Science, New-York, 133-149, (2006).

- L. Gardes & S. Girard. Asymptotic distribution of a Pickands-type estimator of the extreme value index, *Comptes Rendus de l'Académie des Sciences*, **t. 341**, Série I, 53-58, (2005).
- L. Gardes & S. Girard. Estimating extreme quantiles of Weibull tail-distributions, *Communication in Statistics - Theory and Methods*, **34**, 1065-1080, (2005).

Publications associées au Chapitre 2

- A. Daouia, L. Gardes, S. Girard & A. Lekina. Kernel estimators of extreme level curves, *Test*, à paraître.
- L. Gardes & S. Girard. Conditional extremes from heavy-tailed distributions : an application to the estimation of extreme rainfall return levels, *Extremes*, **13(2)**, 177-204, (2010).
- L. Gardes, S. Girard & A. Lekina. Functional nonparametric estimation of conditional extreme quantiles, *Journal of Multivariate Analysis*, **101**, 419-433, (2010).
- L. Gardes & S. Girard. A moving window approach for nonparametric estimation of the conditional tail index, *Journal of Multivariate Analysis*, **99**, 2368-2388, (2008).
- L. Gardes. Estimating the Support of a Poisson process via the Faber-Schauder basis and extreme values, *Publication de l'Institut de Statistique de l'Université de Paris*, **XXXXVI**, 43-72, (2002).

Publications associées au Chapitre 3

- C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes & S. Girard. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression, *Journal of Geophysical Research - Planets*, **114**, E06005, (2009).
- C. Bernard-Michel, L. Gardes & S. Girard. Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, **19**, 85-98, (2009).
- C. Bernard-Michel, L. Gardes & S. Girard. A Note on Sliced Inverse Regression with Regularizations, *Biometrics*, **64**, 982-984, (2008).

Chapitre 1

Estimation de quantiles extrêmes pour des lois à queue de type Weibull

1.1 Introduction

Dans ce chapitre, nous étudions le comportement des valeurs extrêmes d'un échantillon de variables aléatoires unidimensionnelles. Nous nous concentrons essentiellement sur une famille particulière de lois : les lois à queue de type Weibull. Ces lois ont une fonction de survie qui décroit vers zéro à la vitesse exponentielle. Nous donnerons une définition plus précise de cette famille dans le paragraphe 1.3. Notre principal objectif est de proposer des estimateurs de quantiles extrêmes. Plus précisément, disposant d'un échantillon $X_1, \ldots, X_n$ de n variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition commune $F(.)$, nous souhaitons estimer le réel $q(\alpha_n)$ défini par

$$q(\alpha_n) = \bar{F}^{\leftarrow}(\alpha_n), \text{ avec } \alpha_n \to 0 \text{ lorsque } n \to \infty,$$

où (α_n) est une suite connue et $\bar{F}^{\leftarrow}(u) = \inf\{x,\ \bar{F}(x) \leq u\}$ est l'inverse généralisée de la fonction de survie $\bar{F}(.) = 1 - F(.)$. Un problème similaire à l'estimation de $q(\alpha_n)$ est l'estimation de "petites probabilités" p_n. Autrement dit, pour une suite de réels (x_n) fixée, nous souhaitons estimer la probabilité p_n définie par

$$p_n = \bar{F}(x_n),\ x_n \to \infty \text{ lorsque } n \to \infty.$$

Ce sont les hydrologues qui ont été parmi les premiers à s'intéresser à ces deux problèmes. Disposant d'un échantillon de hauteurs d'eau annuelles d'un cours d'eau, ils se sont posés les deux questions suivantes :

1) quelle est la hauteur d'eau qui est atteinte pour une faible probabilité donnée ?
2) pour une "grande" hauteur d'eau fixée, qu'elle est la probabilité d'observer une hauteur d'eau qui lui sera supérieure ?

Les questions 1) et 2) se rapportent donc respectivement à l'estimation d'un quantile extrême et d'une "petite probabilité". La difficulté principale réside dans le fait que l'on considère un ordre de quantile $\alpha_n \to 0$ (ou de manière équivalente un seuil $x_n \to \infty$). En effet, si par exemple $n\alpha_n \to 0$ lorsque $n \to \infty$, il est facile de montrer que $\mathbb{P}(X_{n,n} < q(\alpha_n)) \to 1$ où $X_{n,n} = \max(X_1, \ldots, X_n)$. La quantité $q(\alpha_n)$ n'appartient donc pas à l'intervalle de variation

de nos observations. En conséquence, l'estimateur de $q(\alpha_n)$ ne peut être obtenu en inversant simplement la fonction de répartition empirique

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\},$$

car $\hat{F}_n(x) = 1$ pour $x \geq X_{n,n}$. L'estimation de quantiles extrêmes et/ou de "petites probabilités" est requise dans de nombreux domaines d'application parmi lesquels citons la fiabilité [22], la finance [23], les assurances [7, 11] et la climatologie [49]. Pour répondre à ces deux questions, nous devons donc étudier de près le comportement de la queue de distribution de $F(.)$ en utilisant la théorie des valeurs extrêmes. Nous présentons les éléments essentiels de cette théorie dans la section 1.2. Dans la section 1.3, nous proposons des estimateurs de quantiles extrêmes pour la famille des lois à queue de type Weibull. D'autres types de queue de distribution sont aussi considérés dans la section 1.4.

1.2 Introduction à la théorie des valeurs extrêmes

Lorsque l'on s'intéresse à la partie centrale d'un échantillon, le résultat clé est le Théorème de la Limite Centrale (abrégé en TLC) donnant la loi asymptotique de la somme des observations. Par contre, si l'on souhaite étudier les valeurs extrêmes de cet échantillon, le TLC ne présente que peu d'intérêt. On utilise plutôt un résultat établissant la loi asymptotique du maximum de l'échantillon. Ce résultat est énoncé dans le paragraphe 1.2.1. Il permet de classer la plupart des lois en trois domaines d'attraction. La caractérisation des fonctions de répartition dans chacun de ces domaines est donnée dans le paragraphe 1.2.2.

1.2.1 Convergence en loi du maximum d'un échantillon

Le résultat ci-dessous établit la loi asymptotique du maximum $X_{n,n} = \max(X_1, \ldots, X_n)$ de l'échantillon. Il a été démontré notamment par Gnedenko [38].

Théorème 1.1 *S'il existe deux suites $(a_n > 0)$, (b_n) et un réel γ tels que*

$$\mathbb{P}\left\{\frac{X_{n,n} - b_n}{a_n} \leq x\right\} \to H_\gamma(x),$$

lorsque $n \to \infty$ alors

$$H_\gamma(x) = \begin{cases} \exp[-(1+\gamma x)_+^{-1/\gamma}] & \text{si } \gamma \neq 0, \\ \exp(-\mathrm{e}^{-x}) \text{ si } \gamma = 0, \end{cases}$$

où $y_+ = \max(0, y)$.

La fonction de répartition $H_\gamma(.)$ est la fonction de répartition de la loi des valeurs extrêmes. Cette loi dépend du seul paramètre γ appelé l'indice des valeurs extrêmes. Selon le signe de γ, on définit trois domaines d'attraction :

- si $\gamma > 0$, on dit que $F(.)$ appartient au domaine d'attraction de Fréchet. Il contient les lois dont la fonction de survie décroit comme une fonction puissance. On parle aussi de lois à queue lourde. Dans ce domaine d'attraction, on trouve les lois de Pareto, de Student, de Cauchy, etc ...
- si $\gamma = 0$, $F(.)$ est dans le domaine d'attraction de Gumbel qui regroupe les lois ayant une fonction de survie à décroissance exponentielle. C'est le cas des lois normale, gamma, exponentielle, etc ...
- si $\gamma < 0$, $F(.)$ appartient au domaine d'attraction de Weibull. Ce domaine contient les lois dont le point terminal $x_F = \inf\{x,\ F(x) \geq 1\}$ est fini. C'est le cas par exemple des lois uniformes, lois beta, etc ...

Un classement de nombreuses lois par domaine d'attraction est disponible dans [23, Tableaux 3.4.2-3.4.4]. Nous allons à présent rappeler les théorèmes de caractérisation des trois domaines d'attraction ci-dessus.

1.2.2 Caractérisation des domaines d'attraction

La caractérisation des domaines d'attraction fait largement appel à la notion de fonctions à variations régulières. Rappelons qu'une fonction $U(.)$ est à variations régulières d'indice $\delta \in \mathbb{R}$ à l'infini (on notera par la suite $U(.) \in \mathcal{RV}_\delta$) si pour tout $\lambda > 0$,

$$\lim_{x\to\infty} \frac{U(\lambda x)}{U(x)} = \lambda^\delta.$$

Si $\delta = 0$, on dit que la fonction $U(.)$ est à variations lentes ($U(.) \in \mathcal{RV}_0$). On montre facilement que toute fonction à variations régulières d'indice $\delta \in \mathbb{R}$ s'écrit,

$$U(x) = x^\delta L(x),\ L \in \mathcal{RV}_0.$$

Rappelons enfin que toutes les fonctions à variations lentes $L(.)$ s'écrivent sous la forme :

$$L(x) = c(x) \exp\left\{\int_1^x \frac{\Delta(u)}{u} du\right\},$$

où $c(x) \to c > 0$ et $\Delta(x) \to 0$ lorsque $x \to \infty$. Cette représentation des fonctions à variations lentes est connue sous le nom de *représentation de Karamata* (voir [10, Théorème 1.3.1]). De plus, si la fonction $c(.)$ est constante, la fonction $L(.)$ est dite normalisée. De nombreux résultats sur les fonctions à variations régulières sont donnés dans le livre de Bingham *et al.* [10].

1.2.2.1 Domaine d'attraction de Fréchet

Le résultat ci-dessous énoncé par Gnedenko [38] et dont on trouvera une démonstration simple dans le livre de Resnick [48, Proposition 1.11] assure que toute fonction appartenant au domaine d'attraction de Fréchet est une fonction à variations régulières.

Théorème 1.2 *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Fréchet (avec un indice des valeurs extrêmes $\gamma > 0$) si et seulement si la fonction de survie $\bar{F}(.) \in \mathcal{RV}_{-1/\gamma}$.*

Autrement dit, une fonction de répartition $F(.)$ appartenant au domaine d'attraction de Fréchet s'écrit sous la forme :

$$F(x) = 1 - x^{-1/\gamma} L(x),\ L(.) \in \mathcal{RV}_0. \tag{1.1}$$

Les suites de normalisation (a_n) et (b_n) sont données dans ce cas par $a_n = \bar{F}^{\leftarrow}(1/n)$ et $b_n = 0$ (voir [48, Proposition 1.11]). Il faut aussi noter que toutes les fonctions de répartition du domaine d'attraction de Fréchet ont un point terminal infini. On peut montrer (voir [10, Théorème 1.5.12]) que l'équation (1.1) est équivalente à :

$$q(\alpha) = \alpha^{-\gamma} \ell(\alpha^{-1}),\ \ell(.) \in \mathcal{RV}_0, \tag{1.2}$$

où $\alpha \in [0, 1]$. De nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes γ et des quantiles extrêmes $q(\alpha_n)$ pour des lois à queue lourde. L'estimateur le plus connu de $\gamma > 0$ est l'estimateur proposé par Hill [84] et défini par

$$\hat{\gamma}_n^H = \frac{1}{k_n} \sum_{i=1}^{k_n} \log(X_{n-i+1,n}) - \log(X_{n-k_n,n}), \tag{1.3}$$

où $X_{1,n} \leq \ldots \leq X_{n,n}$ est l'échantillon ordonné associé aux variables aléatoires $X_1, \ldots, X_n$ et (k_n) est une suite d'entiers telle que $1 < k_n < n$. D'autres estimateurs de cet indice ont été proposés notamment par Beirlant *et al.* [52, 4] qui utilisent un modèle de régression exponentiel pour débiaiser l'estimateur de Hill et par Feuerverger *et al.* [73] qui introduisent un estimateur des moindres carrés. L'utilisation d'un noyau dans l'estimateur de Hill a été étudiée par Csörgő *et al.* [13]. Un estimateur efficace de l'indice des valeurs extrêmes a été proposé par Falk *et al.* [25]. Une liste plus détaillée des différents travaux sur l'estimation de l'indice des valeurs extrêmes est effectuée par Csörgő *et al.* [14]. Concernant l'étude du quantile extrême d'ordre α_n, Weissman [98] propose l'estimateur

$$\hat{q}_n^W(\alpha_n) = X_{n-k_n+1,n} \left(\frac{k_n}{n\alpha_n} \right)^{\hat{\gamma}_n^H}. \tag{1.4}$$

1.2.2.2 Domaine d'attraction de Weibull

Le résultat suivant (voir Gnedenko [38], Resnick [48, Proposition 1.13]) montre que l'on passe du domaine d'attraction de Fréchet à celui de Weibull par un simple changement de variable dans la fonction de répartition.

Théorème 1.3 *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Weibull (avec un indice des valeurs extrêmes $\gamma < 0$) si et seulement si son point terminal x_F est fini et si la fonction de répartition $F_*(.)$ définie par*

$$F_*(x) = \begin{cases} 0 & \text{si } x < 0 \\ F(x_F - 1/x) & \text{si } x \geq 0, \end{cases}$$

appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $-\gamma > 0$.

Ainsi, une fonction de répartition $F(.)$ du domaine d'attraction de Weibull s'écrit pour $x \leq x_F$:

$$F(x) = 1 - (x_F - x)^{-1/\gamma} L((x_F - x)^{-1}),\ L(.) \in \mathcal{RV}_0. \tag{1.5}$$

De manière équivalente, le quantile d'ordre $\alpha \in [0,1]$ associé s'écrit :

$$q(\alpha) = x_F - \alpha^{-\gamma} \ell(1/x),\ \ell(.) \in \mathcal{RV}_0. \tag{1.6}$$

Les suites de normalisation (a_n) et (b_n) sont données par $a_n = x_F - \bar{F}^{\leftarrow}(1/n)$ et $b_n = x_F$. Ce domaine d'attraction a été considéré notamment par Falk [24] et Hall *et al.* [42] pour estimer le point terminal d'une distribution. Dans la section 1.4, nous présentons des estimateurs de l'indice γ et des quantiles extrêmes adaptés à ce type de loi.

1.2.2.3 Domaine d'attraction de Gumbel

La caractérisation des fonctions de répartition du domaine d'attraction de Gumbel est plus complexe. Le résultat ci-dessous est démontré notamment dans Resnick [48, Proposition 1.4].

Théorème 1.4 *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Gumbel si et seulement si il existe $z < x_F \leq \infty$ tel que*

$$\bar{F}(x) = c(x) \exp\left\{ - \int_z^x \frac{1}{a(t)} dt \right\},\ z < x < x_F, \tag{1.7}$$

où $c(x) \to c > 0$ lorsque $x \to x_F$ et $a(.)$ est une fonction positive et dérivable de dérivé $a'(.)$ telle que $a'(x) \to 0$ lorsque $x \to x_F$.

Le domaine d'attraction de Gumbel regroupe une grande diversité de lois comptant parmi elles la plupart des lois usuelles (loi normale, exponentielle, gamma, log-normale). Cette famille étant difficile à étudier dans toute sa généralité, de nombreux auteurs se sont concentrés sur une sous-famille : les lois à queue de type Weibull. Leur définition est donnée dans la section suivante.

1.3 Inférence sur les lois à queue de type Weibull

Les lois à queue de type Weibull correspondent au cas particulier où l'on suppose dans (1.7) que la dérivée de la fonction $a(.)$ est à variations régulières avec un indice strictement négatif. Plus précisément, si on suppose que $a'(.) \in \mathcal{RV}_{-1/\theta}$ où $\theta > 0$ est appelé l'indice de queue de Weibull, on montre facilement que l'équation (1.7) s'écrit :

$$\bar{F}(x) = \exp\left\{ -x^{1/\theta} L(x) \right\},\ L(.) \in \mathcal{RV}_0. \tag{1.8}$$

Une fonction de répartition s'écrivant selon le modèle (1.8) est dite à queue de type Weibull d'indice $\theta > 0$. L'équation (1.8) est équivalente à

$$q(\alpha) = (-\log \alpha)^{\theta} \ell(-\log \alpha),\ \ell(.) \in \mathcal{RV}_0, \tag{1.9}$$

où $\alpha \in [0,1]$. Cette famille de lois contient par exemple les lois normale, Gamma, exponentielle, etc ... Par contre, la loi log-normale qui appartient au domaine d'attraction de Gumbel n'est pas une loi à queue de type Weibull. Dans la suite, on considère un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et distribuées selon le modèle (1.8). On note $X_{1,n} \leq \ldots \leq X_{n,n}$ l'échantillon ordonné associé. Le paragraphe 1.3.1 est consacré à l'estimation de l'indice de queue de Weibull. L'estimation des quantiles extrêmes est discutée dans le paragraphe 1.3.2. Les résultats présentés dans ces deux paragraphes ont été publiés en collaboration avec J. Diebolt, S. Girard et A. Guillou : pour le paragraphe 1.3.1 dans *REVSTAT* [33], *Journal of Statistical Planning and Inference* [34] et *Test* [19] et pour le paragraphe 1.3.2 dans *Communication in Statistics - Theory and Methods* [31] et *Journal of Statistical Planning and Inference* [20].

1.3.1 Estimation de l'indice de queue de Weibull

Les lois à queue de type Weibull appartiennent évidemment au domaine d'attraction de Gumbel (*i.e.* avec un indice des valeurs extrêmes $\gamma = 0$). L'indice des valeurs extrêmes ne fournit donc aucune information sur la vitesse de décroissance de la fonction de survie à l'intérieur de cette famille de loi. C'est l'indice de queue de Weibull θ qui nous donne cette information : une valeur de θ proche de zéro (resp. l'infini) correspond à une décroissance rapide (resp. lente) de la queue de distribution. La connaissance de ce paramètre est donc essentielle si l'on souhaite par exemple estimer un quantile extrême. Il existe dans la littérature de nombreux estimateurs de l'indice θ. Berred [9] propose un estimateur basé sur des valeurs records. D'autres estimateurs utilisent les k_n plus grandes observations de l'échantillon parmi lesquels citons ceux proposés par Broniatowski [12] et Beirlant *et al.* [3]. Un estimateur intéressant de l'indice de queue de Weibull a été proposé par Beirlant *et al.* [8]. Il est défini par :

$$\hat{\theta}_n^B = \sum_{i=1}^{k_n-1} \log((X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Big/ \sum_{i=1}^{k_n-1} (\log\log(n/i) - \log\log(n/k_n)) \,, \tag{1.10}$$

où (k_n) est une suite d'entiers tels que $1 < k_n < n$. Son expression est proche de celle de l'estimateur proposé par Hill [84] (voir l'équation (1.3)). Les propriétés asymptotiques de cet estimateur ont été étudiées par Girard [37]. Nous présentons dans ce paragraphe deux familles d'estimateurs englobant $\hat{\theta}_n^B$ (voir les sous-paragraphes 1.3.1.1 et 1.3.1.2) et nous proposons un estimateur débiaisé de l'indice θ (voir le sous-paragraphe 1.3.1.3). Les résultats asymptotiques sont obtenus (entre autres) sous les hypothèses suivantes.

(H.1) La suite (k_n) vérifie $k_n \to \infty$ et $n/k_n \to \infty$ lorsque $n \to \infty$.

(H.2) Il existe un paramètre $\rho < 0$ et une fonction $b(.)$ vérifiant $b(x) \to 0$ lorsque $x \to \infty$ tels que pour tout $1 < A < \infty$

$$\lim_{x\to\infty} \sup_{\lambda\in[1,A]} \left| \frac{\log(\ell(\lambda x)/\ell(x))}{b(x)K_\rho(\lambda)} - 1 \right| = 0,$$

où $K_\rho(\lambda) = \int_1^\lambda t^{\rho-1}dt$ et $\ell(.)$ est la fonction à variations lentes introduite dans (1.9).

L'hypothèse **(H.1)** assure que le nombre de statistiques d'ordre conservées k_n est assez grand

$(k_n \to \infty)$ pour obtenir des estimateurs stables, mais pas trop $(n/k_n \to \infty)$ pour que les observations utilisées restent dans la queue de distribution. Le choix de la suite (k_n) est donc un compromis entre le biais et la variance de l'estimateur.
L'hypothèse **(H.2)** est très souvent utilisée pour étudier le comportement asymptotique d'estimateurs d'indice ou de quantiles extrêmes. Elle est notamment nécessaire pour démontrer la normalité asymptotique de l'estimateur de Hill défini en (1.3). On peut montrer (voir par exemple [36]) que la fonction $b(.)$ (appelée aussi fonction de biais) est à variations régulières d'indice $\rho < 0$. Le paramètre ρ (appelé paramètre du second ordre) contrôle donc la vitesse de convergence de $\ell(\lambda x)/\ell(x)$ vers 1. Une valeur de ρ proche de 0 implique une faible vitesse de convergence.

1.3.1.1 Utilisation de poids

Nous définissons une famille d'estimateurs de θ en incorporant des poids dans l'estimateur $\hat{\theta}_n^B$ défini par (1.10). Les estimateurs obtenus sont donc des combinaisons linéaires de statistiques d'ordre c'est à dire des L-estimateurs. Plus précisément, nous introduisons la famille d'estimateurs $\Theta_1 = \{\hat{\theta}_n(\zeta),\ \zeta = (\zeta_{1,n}, \ldots, \zeta_{k_n-1,n})\}$ avec

$$\hat{\theta}_n(\zeta) = \sum_{i=1}^{k_n-1} \zeta_{i,n} \log((X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Bigg/ \sum_{i=1}^{k_n-1} \zeta_{i,n}(\log\log(n/i) - \log\log(n/k_n)) \ , \tag{1.11}$$

où $\zeta_{i,n} = W(i/k_n) + \varepsilon_{i,n}$, $(\varepsilon_{i,n})$, $i = 1, \ldots, k_n - 1$ étant une suite non-aléatoire. La fonction déterministe $W(.)$ doit satisfaire les deux hypothèses de régularité ci dessous :

(H.3) la fonction $W(.)$ est définie et admet une dérivé continue sur l'intervalle $]0, 1[$.

(H.4) Il existe $M > 0$, $0 \leq q < 1/2$ et $p < 1$ tels que pour tout $x \in]0, 1[$, $|W(x)| \leq Mx^{-q}$ et $|W'(x)| \leq Mx^{-p-q}$.

Ces conditions sont essentielles pour établir la normalité asymptotique des L-estimateurs (voir par exemple [45]). Nous établissons à présent (voir [34, Théorème 1]) la normalité asymptotique des estimateurs appartenant à cette famille. On pose :

$$\|\varepsilon\|_{n,\infty} = \max_{1,\ldots,k_n-1} |\varepsilon_{i,n}|,\ \mu(W) = \int_0^1 W(x)\log(1/x)dx,$$

$$\sigma^2(W) = \int_0^1 \int_0^1 W(x)W(y)\frac{\min(x,y) - xy}{xy}dxdy.$$

Théorème 1.5 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *à* **(H.4)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to \Lambda \in \mathbb{R}$ *et* $k_n^{1/2} \max\{1/\log(n), \|\varepsilon\|_{n,\infty}\} \to 0$ *lorsque* $n \to \infty$ *alors,*

$$k_n^{1/2}(\hat{\theta}_n(\zeta) - \theta - b(\log(n/k_n))) \xrightarrow{d} \mathcal{N}(0, \theta^2\sigma^2(W)/\mu^2(W)).$$

Dans le cas où $\liminf \|\varepsilon\|_{n,\infty} \log(n) \leq 1$ avec $\Lambda \neq 0$, le Théorème 1.5 n'est valable que si $\rho > -1$ ce qui correspond à une vitesse de convergence lente dans la condition **(H.2)**. Nous donnons à

présent deux choix possibles pour les poids $\zeta_{i,n}$, $i = 1, \ldots, k_n - 1$.

- En prenant $\zeta_{i,n} = 1$ pour tout $i = 1, \ldots, k_n - 1$ (*i.e.* $W(x) = 1$ pour tout $x \in]0, 1[$ et $\varepsilon_{i,n} = 0$ pour tout $i = 1, \ldots, k_n - 1$), on retrouve l'estimateur $\hat{\theta}_n^B$ proposé par Beirlant *et al.* [8]. Le résultat de normalité asymptotique établi par Girard [37] est une conséquence directe du Théorème 1.5.

  Corollaire 1.1 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $k_n^{1/2}/\log(n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n^B - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

- Nous proposons un nouvel estimateur de θ en utilisant la remarque suivante : d'après (1.9),

$$\frac{\log q(\alpha)}{\log\log(1/\alpha)} = \theta + \frac{\log \ell(\log(1/\alpha))}{\log\log(1/\alpha)}.$$

  Ainsi, comme $\log(\ell(x))/\log(x) \to 0$ lorsque $x \to \infty$ (voir [10, Propriété 1.3.6]), on en déduit que pour $\alpha \to 0$,

$$\log q(\alpha) \sim \theta \log\log(1/\alpha). \tag{1.12}$$

  Ainsi, les points $(\log\log(n/i), \log(X_{n-i+1,n}))$, $i = 1, \ldots, k_n - 1$ sont approximativement répartis sur une droite de pente θ. Nous proposons d'estimer θ par l'estimateur des moindres carrés. Nous montrons (voir [34, Corollaire 2]) qu'il appartient à notre famille d'estimateurs avec les poids :

$$\zeta_{i,n} = \zeta_{i,n}^Z = \log\log(n/i) - \frac{1}{k_n - 1} \sum_{i=1}^{k_n - 1} \log\log(n/i) = W(i/k_n) + \varepsilon_{i,n},$$

  où $W(x) = -(\log(x)+1)$ et, uniformément en $i = 1, \ldots, k_n - 1$, $\varepsilon_{i,n} = \mathrm{O}(\log^2(k_n)/\log(n)) + \mathrm{O}(\log(k_n)/k_n)$. L'estimateur $\hat{\theta}_n(\zeta^Z)$ ainsi obtenu est similaire à l'estimateur de Zipf introduit par Kratz *et al.* [87] et Schultze *et al.* [94] dans le cas de lois à queue lourde. La normalité asymptotique de $\hat{\theta}_n(\zeta^Z)$ est une conséquence directe du Théorème 1.5.

  Corollaire 1.2 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $k_n^{1/2} \log^2(k_n)/\log(n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(\zeta^Z) - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2)$.

1.3.1.2 Utilisation d'autres suites de normalisation

Dans l'estimateur $\hat{\theta}_n^B$, la somme des écarts entre les logarithmes des statistiques d'ordre est normalisée par la suite :

$$T_n^{(1)} = \sum_{i=1}^{k_n - 1} \log\log(n/i) - \log\log(n/k_n)). \tag{1.13}$$

Nous proposons de remplacer cette suite $(T_n^{(1)})$ par une suite positive quelconque (T_n). Ceci nous conduit à définir la famille d'estimateurs $\Theta_2 = \{\hat{\theta}_n(T_n),\ T_n > 0\}$ avec

$$\hat{\theta}_n(T_n) = \frac{1}{T_n} \sum_{i=1}^{k_n-1} \log((X_{n-i+1,n}) - \log(X_{n-k_n+1,n})). \tag{1.14}$$

Nous montrons dans [33, Théorème 2.1] un résultat de normalité asymptotique pour cette famille d'estimateurs. Posons,

$$u_n = \frac{k_n \int_0^\infty \log(1 + x/t)\mathrm{e}^{-x} dx}{T_n} - 1.$$

Théorème 1.6 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $T_n k_n \log(n/k_n) \to 1$ *et* $k_n^{1/2} b(\log(n/k_n)) \to \Lambda \in \mathbb{R}$ *lorsque* $n \to \infty$ *alors*

$$k_n^{1/2}(\hat{\theta}_n(T_n) - \theta - b(\log(n/k_n)) - \theta u_n) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

La meilleure vitesse de convergence de $\hat{\theta}_n(T_n)$ est obtenue lorsque $\Lambda \neq 0$. Dans ce cas, on peut montrer (voir [33, Proposition 2.2]) que k_n est équivalente à $\Lambda^2(\log n)^{-2\rho}\ell^*(\log(n))$ où $\ell^*(.)$ est une fonction à variations lentes. Sous l'hypothèse supplémentaire $\liminf \|\varepsilon\|_{n,\infty} \log(n) \leq 1$, les estimateurs de la famille Θ_1 ont la même vitesse de convergence que ceux de la famille Θ_2. Le Théorème 1.6 nous permet de déterminer l'erreur moyenne quadratique asymptotique (AMSE) de $\hat{\theta}_n(T_n)$. Elle est donnée par

$$AMSE(\hat{\theta}_n(T_n)) = (\theta u_n + b(\log(n/k_n)))^2 + \frac{\theta^2}{k_n}. \tag{1.15}$$

Le terme de variance asymptotique est donc identique pour tous les estimateurs de la famille. Le biais dépend par contre du choix de la suite (T_n). Le choix idéal pour cette suite de normalisation serait donc de prendre T_n de telle sorte que $\theta u_n + b(\log(n/k_n)) = 0$. Malheureusement, θ et la fonction de biais $b(.)$ sont inconnus et il est donc impossible de définir une telle suite (T_n). Nous donnons à présent quelques choix possibles pour cette suite.

- Comme nous l'avons déjà mentionné, le choix $(T_n) = (T_n^{(1)})$ (voir l'équation (1.13)) conduit à l'estimateur $\hat{\theta}_n^B$. La normalité asymptotique de $\hat{\theta}_n(T_n^{(1)}) = \hat{\theta}_n^B$ est une conséquence directe du Théorème 1.6.

  Corollaire 1.3 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $\log(k_n)/\log(n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(T_n^{(1)}) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

  Dans le corollaire 1.1 du sous-paragraphe précédent, ce résultat est obtenu sous la condition supplémentaire $k_n^{1/2}/\log(n) \to 0$.

- Un choix naturel pour (T_n) est de prendre la suite annulant une partie du biais asymptotique : la partie dépendant de u_n. Pour ce faire, il suffit de prendre $(T_n) = (T_n^{(2)})$ avec

$$T_n^{(2)} = k_n \int_0^\infty \log\left(1 + \frac{x}{\log(n/k_n)}\right) \mathrm{e}^{-x} dx.$$

En remarquant que $T_n^{(2)} = n/k_n \mathrm{E}_1(\log(n/k_n))$, où $\mathrm{E}_1(z)$ dénote l'exponentielle intégrale calculée au point z (voir [1, Chapitre 5, p. 225-233]), le calcul de la suite $T_n^{(2)}$ est simple à effectuer. Il est aussi intéressant de noter que

$$T_n^{(1)} = \sum_{i=1}^{k_n} \log\left(1 - \frac{\log(i/k_n)}{\log(n/k_n)}\right)$$

est en fait l'approximation par des sommes de Riemann de $T_n^{(2)}$ car en intégrant par partie on montre que

$$T_n^{(2)} = \int_0^1 \log\left(1 - \frac{\log(x)}{\log(n/k_n)}\right) dx.$$

On déduit du Théorème 1.6 le résultat suivant :

Corollaire 1.4 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(T_n^{(2)}) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

- Une des hypothèses du Théorème 1.6 étant que $T_n k_n \log(n/k_n) \to 1$, un choix simple est de prendre $T_n = T_n^{(3)} = (k_n \log(n/k_n))^{-1}$. La normalité asymptotique de l'estimateur ainsi obtenu est donnée ci-dessous :

Corollaire 1.5 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $k_n^{1/2}/\log(n/k_n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(T_n^{(3)}) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

Nous allons à présent comparer ces trois estimateurs en fonction de leur erreur moyenne quadratique asymptotique définie par l'équation (1.15).

Proposition 1.1 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to \Lambda \in \mathbb{R}$, *plusieurs situations sont possibles.*

i) $b(.)$ *est asymptotiquement strictement positive. Posons* $\beta_1 = 2 \lim_{x\to\infty} x b(x)$.

Si $\beta_1 > \theta$ *alors, pour* n *assez grand,*

$$AMSE(\hat{\theta}_n(T_n^{(3)})) < AMSE(\hat{\theta}_n(T_n^{(2)})) < AMSE(\hat{\theta}_n(T_n^{(1)})).$$

Si $\beta_1 < \theta$ *alors, pour* n *assez grand,*

$$AMSE(\hat{\theta}_n(T_n^{(2)})) < \min(AMSE(\hat{\theta}_n(T_n^{(1)})), AMSE(\hat{\theta}_n(T_n^{(3)}))).$$

ii) $b(.)$ *est asymptotiquement strictement négative. Posons* $\beta_2 = -4 \lim_{n\to\infty} b(\log n)\dfrac{k_n}{\log k_n}$.

Si $\beta_2 > \theta$, *alors, pour* n *assez grand,*

$$AMSE(\hat{\theta}_n(T_n^{(1)})) < AMSE(\hat{\theta}_n(T_n^{(2)})) < AMSE(\hat{\theta}_n(T_n^{(3)})).$$

Si $\beta_2 < \theta$, *alors, pour* n *assez grand,*

$$AMSE(\hat{\theta}_n(T_n^{(2)})) < \min(AMSE(\hat{\theta}_n(T_n^{(1)})), AMSE(\hat{\theta}_n(T_n^{(3)}))).$$

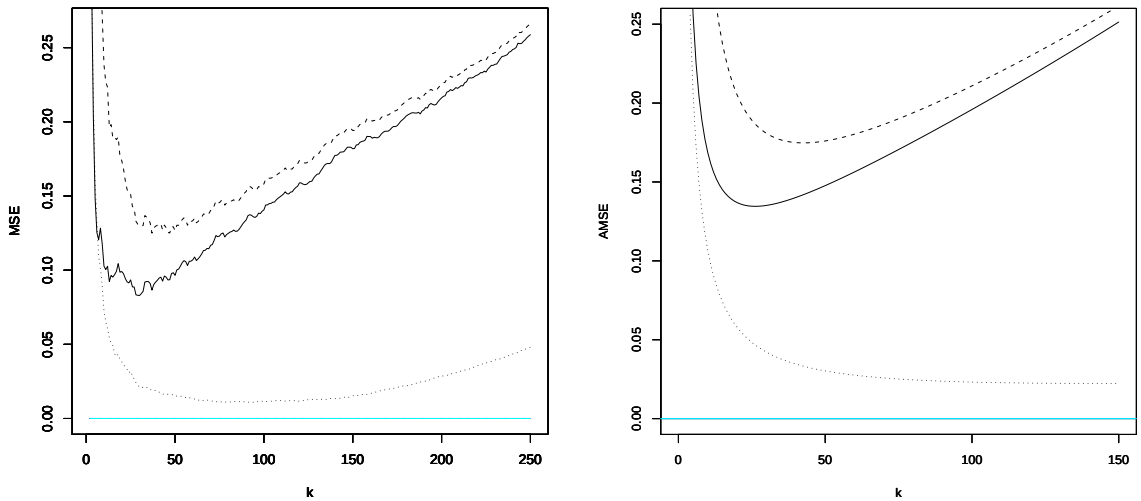

FIG. 1.1 – Comparaison des estimateurs $\hat{\theta}_n(T_n^{(1)})$ (trait plein), $\hat{\theta}_n(T_n^{(2)})$ (tirés) et $\hat{\theta}_n(T_n^{(3)})$ (pointillés). En abscisse, k_n et en ordonnée les erreurs moyenne quadratique empiriques (gauche) et asymptotiques (droite).

Comme l'on pouvait s'y attendre, il n'y a pas d'estimateur qui soit préférable dans toutes les situations. Dans le cas i), β_1 ne dépend pas de la suite (k_n). Le classement entre les trois estimateurs $\hat{\theta}_n(T_n^{(1)})$, $\hat{\theta}_n(T_n^{(2)})$ et $\hat{\theta}_n(T_n^{(3)})$ dépend donc uniquement de la loi des observations. Si la fonction biais $b(.)$ converge rapidement vers zéro alors $\beta_1 < \theta$ et ainsi l'utilisation de l'estimateur $\hat{\theta}_n(T_n^{(2)})$ est préférable. Au contraire, si $b(.)$ converge lentement vers zéro, $\beta_1 > \theta$ et l'estimateur $\hat{\theta}_n(T_n^{(3)})$ sera de meilleure qualité. Dans le cas ii), β_2 dépend de la suite (k_n). Si k_n est petite (par exemple $k_n \propto -1/b(\log(n))$) alors $\beta_2 = 0$ et l'estimateur $\hat{\theta}_n(T_n^{(2)})$ est préférable. Inversement, si k_n est grande (par exemple $k_n \propto (b(log(n)))^{-2}$) alors $\beta_2 = \infty$ et $\hat{\theta}_n(T_n^{(1)})$ sera asymptotiquement le meilleur estimateur. Les comparaisons ci-dessus sont valables uniquement asymptotiquement. Afin de voir si elles sont aussi valides pour une taille d'échantillon finie, nous simulons $N = 200$ échantillons de taille $n = 500$ d'une loi $\Gamma(0.5, 1)$. Pour chaque échantillon, nous calculons les estimateurs $\hat{\theta}_{n,i}(T_n^{(1)})$, $\hat{\theta}_{n,i}(T_n^{(2)})$ et $\hat{\theta}_{n,i}(T_n^{(3)})$ $(i = 1, \ldots, N)$ pour différentes valeurs de k_n. Nous en déduisons les erreurs en moyenne quadratique empiriques définies par :

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{n,i}(T_n^{(j)}) - \theta)^2, \ j = 1, \ 2, \ 3.$$

Ces erreurs empiriques sont comparées aux erreurs asymptotiques des trois estimateurs données par l'équation (1.15). Les résultats sont présentés dans la Figure 1.1 où l'on s'aperçoit que le comportement des erreurs empiriques et celui des erreurs asymptotiques sont très similaires. Les résultats de comparaison des trois estimateurs données dans le Théorème 1.1 semblent donc aussi utilisables pour une taille d'échantillon finie. D'autres simulations avec différentes lois ont été faites (voir [33]) conduisant à la même conclusion.

1.3.1.3 Un estimateur de θ débiaisé

Nous proposons à présent un estimateur débiaisé de l'indice de queue de Weibull θ. Il est basé sur un modèle de régression exponentiel inspiré de ceux proposés par Beirlant *et al.* [4, 52] et Feuerverger *et al.* [73] pour des lois du domaine d'attraction de Fréchet. Plus précisément, on définit les variables aléatoires

$$Z_j = j\log(n/j)(\log(X_{n-j+1,n}) - \log(X_{n-j,n})),\ j = 1, \dots, k_n.$$

Nous établissons dans [19, Corollaire 2.1] le modèle suivant :

$$Z_j = \left(\theta + \left(\frac{\log(n/k_n)}{\log(n/j)}\right) b(\log(n/k_n))\right) f_j + \mathrm{o_P}(b(\log(n/k_n))),\ j = 1, \dots, k_n \tag{1.16}$$

où f_j, $j = 1, \dots, k_n$ sont des variables aléatoires indépendantes de loi exponentielle de paramètre 1 et le terme $\mathrm{o_P}(b(\log(n/k_n)))$ ne dépend pas de j. On obtient à partir du modèle (1.16) l'approximation

$$Z_j \approx \theta + b(\log(n/k_n))x_j + \eta_j,\ j = 1, \dots, k_n \tag{1.17}$$

où η_j est un terme d'erreur aléatoire centré et $x_j = \log(n/k_n)/\log(n/j)$. En estimant les paramètres θ et $b(\log(n/k_n))$ du modèle de régression linéaire (1.17) par la méthode des moindres carrés ordinaires, nous définissons l'estimateur de θ débiaisé par :

$$\hat{\theta}_n^D = \frac{1}{k_n}\sum_{j=1}^{k_n} Z_j - \frac{\hat{b}(\log(n/k_n))}{k_n}\sum_{j=1}^{k_n} x_j, \tag{1.18}$$

où

$$\hat{b}(\log(n/k_n)) = \sum_{j=1}^{k_n}\left(x_j - \frac{1}{k_n}\sum_{j=1}^{k_n} x_j\right) Z_j \Bigg/ \sum_{j=1}^{k_n}\left(x_j - \frac{1}{k_n}\sum_{j=1}^{k_n} x_j\right)^2. \tag{1.19}$$

La normalité asymptotique de $\hat{\theta}_n^D$ est donnée par le résultat ci-dessous (voir [19, Théorème 3.1]).

Théorème 1.7 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si la fonction $b(.)$ est telle que $x|b(x)| \to \infty$ lorsque $x \to \infty$ et si*

$$\frac{k_n^{1/2}}{\log(n/k_n)} b(\log(n/k_n)) \to \tilde{\Lambda} \in \mathbb{R},$$

avec en plus, si $\tilde{\Lambda} = 0$, $\dfrac{\log^2(k_n)}{\log(n/k_n)} \to 0$ et $\dfrac{k_n^{1/2}}{\log(n/k_n)} \to \infty$, on a :

$$\frac{k_n^{1/2}}{\log(n/k_n)}(\hat{\theta}_n^D - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

L'hypothèse $x|b(x)| \to \infty$ implique que dans la condition **(H.2)** la vitesse de convergence est lente (et plus particulièrement que $\rho \geq -1$). Les estimateurs non débiaisés de θ auront donc tendance à avoir un biais important dans ce cas. On peut montrer (voir [19, Tableau 1]) que les lois normale, Gamma satisfont cette hypothèse mais pas les lois de Weibull. En prenant $\Lambda \neq 0$ dans les Théorèmes 1.5 et 1.6 et $\tilde{\Lambda} \neq 0$ dans le Théorème 1.7, on montre que l'estimateur débiaisé $\hat{\theta}_n^D$ admet la même vitesse de convergence que les estimateurs des familles Θ_1 et Θ_2 avec en plus un biais asymptotique nul.

Nous effectuons une simulation afin de comparer le comportement de l'estimateur non débiaisé $\hat{\theta}_n^B$ avec celui de l'estimateur débiaisé $\hat{\theta}_n^D$. Nous générons $N = 100$ échantillons de taille $n = 500$ selon une loi normale centrée et réduite (pour laquelle $\theta = 1/2$). En fonction du nombre de statistiques ordonnées k_n, nous comparons sur la Figure 1.2 les moyennes empiriques des deux estimateurs ainsi que leurs erreurs moyennes quadratiques empiriques. Nous notons un bon comportement de l'estimateur débiaisé en terme de biais ainsi qu'en terme de variance. La méthode de débiaisage proposée semble donc efficace à taille d'échantillon finie. Des simulations sur d'autres lois ont confirmé ce bon comportement (voir [19, Section 4])

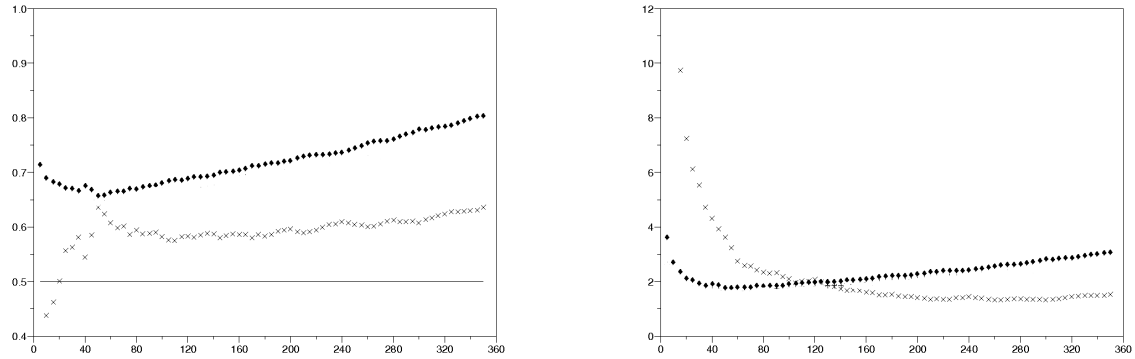

FIG. 1.2 – Comparaison des estimateurs $\hat{\theta}_n^D$ ($\times$) et $\hat{\theta}_n^B$ ($\diamond$). En abscisse, k_n et en ordonnée les erreurs en moyenne quadratique empiriques (gauche) et asymptotiques (droite).

En ne tenant pas compte du terme de biais dans le modèle de régression (1.16), nous obtenons un estimateur non débiaisé de θ défini par :

$$\frac{1}{k_n} \sum_{j=1}^{k_n} Z_j.$$

Nous montrons dans [19, Théorème 2.2] que l'erreur moyenne quadratique asymptotique (AMSE) de cet estimateur est donnée par :

$$AMSE(k_n) = \frac{\theta^2}{k_n} + \left(\frac{b(\log(n/k_n))}{k_n} \sum_{j=1}^{k_n} \frac{\log(n/k_n)}{\log(n/j)} \right)^2.$$

Un choix possible pour k_n est alors de prendre $k_n^{opt} = \arg\min_{k_n} AMSE(k_n)$. Nous pouvons estimer cette erreur par la quantité $\widehat{AMSE}(k_n)$ obtenue en remplaçant θ et $b(\log(n/k_n)$ par les estimateurs $\hat{\theta}_n^D$ et $\hat{b}(\log(n/k_n)$ définis précédemment. Le nombre k_n^{opt} est estimé par :

$$\hat{k}_n = \arg\min_{k_n} \widehat{AMSE}(k_n).$$

Comme l'ont fait remarquer récemment Asimit *et al.* [2], $AMSE(k_n) \sim \theta^2/k_n + b^2(\log(n))$. Ainsi, la sélection du nombre d'observations k_n n'est pas justifiée théoriquement puisque $k_n^{opt} \sim n$. Cependant, les simulations effectuées dans [19, Section 4] montrent que l'utilisation de la valeur $\hat{k}_n$ pour estimer l'indice Θ conduit à de bons résultats.

1.3.2 Estimation de quantiles extrêmes

Toujours pour des lois à queue de type Weibull, nous nous intéressons à présent au problème d'estimation d'un quantile extrême $q(\alpha_n)$ lorsque l'ordre α_n converge vers zéro. Le principal estimateur de $q(\alpha_n)$ disponible dans la littérature a été proposé par Beirlant *et al.* [8]. Il est basé sur l'approximation (1.12) qui assure que pour n assez grand, on a sous l'hypothèse **(H.1)** :

$$\log q(\alpha_n) \approx \theta \log\log(1/\alpha_n) \text{ et } \log q(k_n/n) \approx \theta \log\log(n/k_n).$$

En soustrayant membre à membre les deux approximations ci-dessus et en appliquant la fonction exponentielle, on montre facilement que :

$$q(\alpha_n) \approx q(k_n/n) \left(\frac{\log(1/\alpha_n)}{\log(n/k_n)}\right)^{\theta}.$$

En estimant $q(k_n/n)$ par $X_{n-k_n+1,n}$ qui est le quantile associé à la fonction de répartition empirique et θ par $\hat{\theta}_n^B$, Beirlant *et al.* [8] proposent l'estimateur suivant :

$$\hat{q}^B(\alpha_n) = X_{n-k_n+1,n} \left(\frac{\log(1/\alpha_n)}{\log(n/k_n)}\right)^{\hat{\theta}_n^B}.$$

La construction de cet estimateur est similaire à celle de l'estimateur proposé par Weissman [98] (voir équation 1.4) pour des lois du domaine d'attraction de Fréchet. Un autre estimateur de $q(\alpha_n)$ a été proposé par Beirlant *et al.* [3]. Il est défini par :

$$\hat{q}^{B*}(\alpha_n) = X_{n-k_n+1,n} \left(1 + \frac{\hat{\sigma}_n \log(k_n/(n\alpha_n))}{\hat{\theta}_n^{B*} X_{n-k_n+1,n}}\right)^{\hat{\theta}_n^{B*}},$$

avec

$$\hat{\sigma}_n = \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} (X_{n-i+1,n} - X_{n-k_n+1,n}) \text{ et } \hat{\theta}_n^{B*} = \frac{\log n/k_n}{X_{n-k_n+1,n}} \hat{\sigma}_n.$$

Etant donné que

$$1 + \frac{\hat{\sigma}_n \log(k_n/(n\alpha_n))}{\hat{\theta}_n^{B*} X_{n-k_n+1,n}} = \frac{\log(1/\alpha_n)}{\log(n/k_n)},$$

l'estimateur $\hat{q}^{B*}(\alpha_n)$ est en fait l'estimateur $\hat{q}^{B}(\alpha_n)$ pour lequel l'indice θ n'est pas estimé par $\hat{\theta}_n^B$ mais par $\hat{\theta}_n^{B*}$. Nous proposons dans [31] d'unifier l'étude du comportement asymptotique de ces deux estimateurs. Plus généralement, nous nous intéressons à la famille d'estimateurs du quantile extrême $q(\alpha_n)$ définie par $\mathcal{Q}_{\alpha_n} = \{\hat{q}(\alpha_n, \hat{\theta}_n),\ \hat{\theta}_n \text{ estimateur de } \theta\}$ avec

$$\hat{q}(\alpha_n, \hat{\theta}_n) = X_{n-k_n+1,n}\tau_n^{\hat{\theta}_n},\ \tau_n = \frac{\log(1/\alpha_n)}{\log(n/k_n)},$$

où $\hat{\theta}_n$ est un estimateur quelconque de θ.

1.3.2.1 Etude de la famille d'estimateurs $\mathcal{Q}_{\alpha_n}$

Nous nous proposons d'établir la loi asymptotique des estimateurs de la famille $\mathcal{Q}_{\alpha_n}$. Deux situations peuvent se présenter : soit l'estimateur $\hat{\theta}_n$ converge rapidement vers θ de telle sorte que la loi asymptotique de $\hat{q}(\alpha_n, \hat{\theta}_n)$ est donnée par celle de la statistique d'ordre $X_{n-k_n+1,n}$. Cette situation se présente lorsque la condition ci-dessous est satisfaite :

(H.5) Il existe une suite β_n telle que $\log(n/k_n)k_n^{1/2}(\hat{\theta}_n - \beta_n - \theta) \xrightarrow{\mathrm{P}} 0$.

Soit la vitesse de convergence de $\hat{\theta}_n$ vers θ est inférieure à $\log(n/k_n)k_n^{1/2}$ et dans ce cas la loi asymptotique est celle de l'estimateur de l'indice θ. Cette situation est décrite par la condition

(H.6) Il existe deux suites ϑ_n et β_n ainsi qu'une loi non dégénérée $\mathcal{D}$ telles que :
$\vartheta_n = o(\log(n/k_n)k_n^{1/2})$ et $\vartheta_n(\hat{\theta}_n - \beta_n - \theta) \xrightarrow{d} \mathcal{D}$.

Dans les deux situations, la suite (β_n) représente le biais asymptotique de l'estimateur de l'indice θ. Le théorème suivant établit la loi asymptotique des estimateurs de la famille $\{\hat{q}(\alpha_n, \hat{\theta}_n)\}$ dans les deux situations décrites ci-dessus.

Théorème 1.8 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si $\tau_n \to \tau \in]1, \infty[$ alors :*

- *sous la condition* **(H.5)** *et si $k_n^{1/2}\log(n/k_n)b(\log(n/k_n)) \to 0$,*

$$\log(n/k_n)k_n^{1/2}\tau^{-\beta_n}\left(\frac{\hat{q}(\alpha_n, \hat{\theta}_n)}{q(\alpha_n)} - \tau_n^{\beta_n}\right) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

- *sous la condition* **(H.6)** *et si $\vartheta_n b(\log(n/k_n)) \to 0$,*

$$\frac{\vartheta_n}{\log(\tau)}\tau^{-\beta_n}\left(\frac{\hat{q}(\alpha_n, \hat{\theta}_n)}{q(\alpha_n)} - \tau_n^{\beta_n}\right) \xrightarrow{d} \mathcal{D}.$$

Le meilleur estimateur du quantile extrême $q(\alpha_n)$ est obtenu en utilisant un estimateur de θ satisfaisant l'hypothèse **(H.5)** avec $\beta_n = 0$. Malheureusement, à notre connaissance, un tel estimateur de θ n'existe pas. A titre d'exemple, l'estimateur de θ proposé par Broniatowski [12] satisfait la condition **(H.5)** mais avec un biais asymptotique β_n non nul. Pour la grande majorité

des estimateurs de l'indice θ (notamment ceux introduits dans le paragraphe 1.3.1), c'est la condition **(H.6)** qui est satisfaite avec un biais asymptotique pouvant être annulé.
La condition $\tau_n \to \tau \in]1, \infty[$ implique que l'on peut choisir un ordre α_n proportionnel à $n^{-\tau}$. Ainsi, plus τ est grand plus le quantile estimable est extrême. En contre partie, une grande valeur de τ augmente la variance asymptotique de l'estimateur dans les deux situations.
On peut montrer que la vitesse de convergence de $\hat{q}(\alpha_n, \hat{\theta}_n)$ lorsque la condition **(H.6)** est satisfaite avec un biais asymptotique $\beta_n = 0$ est de l'ordre de $(\log(n))^{-\rho-\epsilon}$ où $\epsilon \in]0, -\rho[$ peut être choisi aussi petit que l'on veut.

1.3.2.2 Un estimateur de $q(\alpha_n)$ débiaisé

Nous utilisons les estimateurs de θ et du biais $b(\log(n/k_n))$ définis dans le sous-paragraphe 1.3.1.3, équations (1.18) et (1.19) pour proposer un estimateur débiaisé du quantile extrême $q(\alpha_n)$. Plus précisément, on se base sur le résultat suivant : sous la condition **(H.2)**, si $\tau_n \to \tau \in]1, \infty[$, on a lorsque $n \to \infty$

$$q(\alpha_n) \sim q(k_n/n)\tau_n^{\theta} \exp\{b(\log(n/k_n))K_\rho(\tau_n)\}.$$

En estimant $q(k_n/n)$ par la statistique d'ordre $X_{n-k_n+1,n}$, θ par $\hat{\theta}_n^D$ (voir équation (1.18)), $b(\log(n/k_n))$ par $\hat{b}(\log(n/k_n))$ (voir équation (1.19)) et ρ par un estimateur $\hat{\rho}_n$, nous proposons l'estimateur

$$X_{n-k_n+1,n}\tau_n^{\hat{\theta}_n^D} \exp\{\hat{b}(\log(n/k_n))K_{\hat{\rho}_n}(\tau_n)\}.$$

Si on néglige le terme de correction $\exp\{\hat{b}(\log(n/k_n))K_{\hat{\rho}_n}(\tau_n)\}$ dans l'expression ci-dessus, on retrouve l'estimateur non débiaisé $\hat{q}(\alpha_n, \hat{\theta}_n^D)$ appartenant à la famille $\mathcal{Q}_{\alpha_n}$. Concernant le paramètre ρ, plusieurs estimateurs ont été proposés pour des modèles différents (citons les travaux de Gomes [40], Gomes *et al.* [41], Feuerverger *et al.* [73] Peng *et al.* [46] et Beirlant *et al.* [4]). Dans le résultat suivant (voir [20, Théorème 1]), nous montrons que l'on peut remplacer $\hat{\rho}_n$ par une valeur arbitraire $\rho^\natural < 0$ et obtenir un estimateur

$$\hat{q}^D(\alpha_n) = X_{n-k_n+1,n}\tau_n^{\hat{\theta}_n^D} \exp\{\hat{b}(\log(n/k_n))K_{\rho^\natural}(\tau_n)\}$$

asymtotiquement normal.

Théorème 1.9 *On se place sous le modèle (1.8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si la fonction $b(.)$ est telle que $x|b(x)| \to \infty$ lorsque $x \to \infty$, si $\tau_n \to \tau \in]1, \infty[$ et si*

$$\frac{k_n^{1/2}}{\log(n/k_n)} b(\log(n/k_n)) \to \tilde{\Lambda} \in \mathbb{R},$$

avec en plus, si $\tilde{\Lambda} = 0$, $\dfrac{\log^2(k_n)}{\log(n/k_n)} \to 0$ et $\dfrac{k_n^{1/2}}{\log(n/k_n)} \to \infty$, on a :

$$\frac{k_n^{1/2}}{\log(n/k_n)} \left(\frac{\hat{q}^D(\alpha_n)}{q(\alpha_n)} - 1 \right) \xrightarrow{d} \mathcal{N}(\tilde{\Lambda}\mu(\tau), \theta^2\sigma^2(\tau)),$$

avec $\sigma^2(\tau) = (K_{\rho^\natural}(\tau) - \log(\tau))^2$ et $\mu(\tau) = (K_{\rho^\natural}(\tau) - K_\rho(\tau))^2$.

Si $\tilde{\Lambda} \neq 0$ et si $\rho^\natural = \rho$ alors l'estimateur $\hat{q}^D(\alpha_n)$ est sans biais avec une vitesse de convergence de l'ordre de $\log^{-\rho^\natural}(n)\ell^*(\log(n))$ où $\ell^*(.)$ est une fonction à variations lentes. Cette vitesse est meilleure que celle obtenue pour les estimateurs de la famille $\mathcal{Q}_{\alpha_n}$ lorsque $\hat{\theta}_n$ satisfait l'hypothèse **(H.6)** avec un biais asymptotique $\beta_n = 0$. Evidemment un mauvais choix de $\rho^\natural$ conduit à un estimateur du quantile extrême biaisé. Notons cependant que les lois à queue de type Weibull usuelles (loi normale, Gamma) ont un paramètre du second ordre $\rho = -1$. En pratique, nous prenons donc une valeur $\rho^\natural$ égale à -1.

1.4 Autres domaines d'attraction

Dans cette section, nous regroupons les travaux effectués sur l'estimation de l'indice γ et sur l'estimation de quantiles extrêmes pour des lois autres que les lois à queue de type Weibull. Dans le paragraphe 1.4.1 nous nous intéressons aux lois dans le domaine d'attraction de Weibull (lois à support borné). Les résultats présentés dans ce paragraphe sont issus de ma thèse [76] et ne sont donc pas détaillés ici. Un estimateur de l'indice des valeurs extrêmes γ valable quel que soit le domaine d'attraction est proposé dans le paragraphe 1.4.2. Ce travail a fait l'objet de deux publications en collaboration avec Stéphane Girard : en tant que chapitre dans un livre [32] et comme Note dans les Comptes-Rendus de l'Académie des Sciences [30].

1.4.1 Domaine d'attraction de Weibull

Dans tout ce paragraphe, on considère un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de même fonction de répartition $F(.)$ satisfaisant (1.5). On note par $X_{1,n} \leq \ldots \leq X_{n,n}$ les statistiques d'ordre associées. Très souvent, un estimateur de quantiles extrêmes pour une loi à support borné $F(.)$ est utilisé comme estimateur du point terminal x_F. En effet, l'estimateur ainsi obtenu présente l'avantage d'être plus robuste aux valeurs aberrantes de l'échantillon que l'estimateur naïf consistant à prendre le maximum des observations. Cette technique est notamment utilisée pour faire de l'estimation de support (voir Chapitre 2). Dans le cadre de ma thèse dirigée par P. Jacob et S. Girard, nous avons proposé notamment deux estimateurs de l'indice des valeurs extrêmes $\gamma < 0$ ainsi qu'un estimateur de quantiles extrêmes adaptés au domaine d'attraction de Weibull. Le premier estimateur de γ est défini par

$$\hat{\gamma}_n^{T_1} = -\frac{\log((1-u_n)/(1-v_n)}{\tau_{u_n}/\tau_{v_n}},$$

avec

$$\tau_{u_n} = \mathbb{I}\{X_{n,n} > 0\} \sum_{i=1}^{n} \mathbb{I}\{X_i \geq u_n X_{n,n}\} \text{ et } \tau_{v_n} = \mathbb{I}\{X_{n,n} > 0\} \sum_{i=1}^{n} \mathbb{I}\{X_i \geq v_n X_{n,n}\}.$$

Son existence est assurée par [76, Lemme 3.3]. Le second estimateur de γ est défini comme étant la solution de l'équation en ξ :

$$\frac{1-u_n}{1-v_n}\left(\frac{\tau_{u_n}^{-\xi}-1}{\tau_{v_n}^{-\xi}-1}\right) = 1.$$

L'existence et l'unicité de la solution de cette équation sont assurées par [76, Lemme 3.4]. La convergence faible des estimateurs $\hat{\gamma}_n^{T_1}$ et $\hat{\gamma}_n^{T_2}$ est montrée dans [76, Théorèmes 3.5 et 3.7]. La normalité asymptotique de $\hat{\gamma}_n^{T_1}$ est quant à elle établie dans [76, Théorème 3.8] sous l'hypothèse restrictive $\gamma < -1/2$.
Concernant l'estimation de quantiles extrêmes, nous définissons un estimateur de $q(\alpha_n)$ par

$$\hat{q}_n^T(\alpha_n, \hat{\gamma}_n) = X_{n,n} \frac{v_n(\tau_{u_n}^{-\hat{\gamma}_n} - (n\alpha_n)^{-\hat{\gamma}_n}) - u_n(\tau_{v_n}^{-\hat{\gamma}_n} - (n\alpha_n)^{-\hat{\gamma}_n})}{\tau_{u_n}^{-\hat{\gamma}_n} - \tau_{v_n}^{-\hat{\gamma}_n}},$$

où $\hat{\gamma}_n$ est un estimateur faiblement consistant de γ. La consistance faible de l'estimateur est donnée dans [76, Théorème 4.1]). Nous montrons dans [76, Théorème 4.3] que si $\alpha_n < 1/n$, $\mathbb{P}\{\hat{q}_n^T(\alpha_n, \hat{\gamma}_n) > X_{n,n}\} \to 1$ lorsque $n \to \infty$. Ceci est une condition minimale pour que l'estimateur naïf $X_{n,n}$ ne soit pas toujours préférable à $\hat{q}_n^T(\alpha_n, \hat{\gamma}_n)$.

1.4.2 Ensemble des domaines d'attraction

Dans ce paragraphe, nous proposons un estimateur de l'indice des valeurs extrêmes $\gamma \in \mathbb{R}$ valable quel que soit le domaine d'attraction auquel appartient la loi étudiée. Pour ce faire, on dispose d'un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de même fonction de répartition $F(.)$ (l'échantillon ordonné associé est noté $X_{1,n} \leq \ldots \leq X_{n,n}$). Dans ce cadre, l'estimateur le plus connu a été introduit par Dekkers *et al.* [18]. C'est une adaptation de l'estimateur de Hill au cas $\gamma \in \mathbb{R}$. Un estimateur débiaisé a été proposé par Beirlant *et al.* [5]. Nous pouvons aussi citer les estimateurs introduits par Feueverger *et al.* [73], Gomes [40] et de Haan *et al.* [16]. Un autre estimateur intéressant de $\gamma \in \mathbb{R}$ a été proposé par Pickands [90]. Il est défini par :

$$\hat{\gamma}_n^P = \frac{1}{\log(2)} \log\left(\frac{X_{n-k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-2k_n+1,n} - X_{n-4k_n+1,n}}\right), \tag{1.20}$$

où k_n est un entier strictement positif et inférieur à $n/4$. Les propriétés asymptotiques de cet estimateur ont été étudiées par Dekkers *et al.* [67]. L'expression de l'estimateur de Pickands découle d'un résultat sur le quantile $q(\alpha)$ associé à la fonction de répartition $F(.)$ (voir de Haan [65]). Si la fonction de répartition $F(.)$ appartient à l'un des trois domaines d'attraction alors, uniformément localement en $x, y > 0$, $y \neq 1$,

$$\lim_{t\to\infty} \frac{q(1/(tx)) - q(1/t)}{q(1/(ty)) - q(1/t)} = \frac{K_\gamma(x)}{K_\gamma(y)}, \tag{1.21}$$

où $K_t(x) = \int_1^x u^{t-1} du$. En remplaçant dans (1.21) le quantile $q(\alpha)$ par son estimateur empirique $\inf\{x,\ \hat{F}_n(x) \geq 1-\alpha\}$ où $\hat{F}_n(.)$ est la fonction de répartition empirique, t par $n/(2k_n)$, x par $1/2$ et y par 2, on obtient (pour n assez grand) l'approximation :

$$\frac{X_{n-4k_n+1,n} - X_{n-2k_n+1,n}}{X_{n-k_n+1,n} - X_{n-2k_n+1,n}} \approx -2^{-\gamma}. \tag{1.22}$$

L'estimateur de Pickands est la solution de l'équation (1.22) en γ. Un des inconvénients de $\hat{\gamma}_n^P$ est qu'il n'utilise pas l'information apportée par la plus grande observation. Nous proposons dans [30, 32] un estimateur de γ proche de l'estimateur de Pickands mais utilisant l'observation

$X_{n,n}$. Pour ce faire, on remplace dans (1.21) le quantile $q(.)$ par son estimateur empirique, x par $1/k_n$, y par c/k_n où $1 < k_n < n$, $c > 1$ et t par n. Pour n assez grand, on obtient alors l'équation en γ :

$$\frac{X_{n-k_n+1,n} - X_{n,n}}{X_{n-\lfloor k_n c\rfloor+1,n} - X_{n,n}} \approx \frac{K_\gamma(1/k_n)}{K_\gamma(c/k_n)}, \tag{1.23}$$

où $\lfloor x\rfloor$ dénote la partie entière de x. On définit l'estimateur de γ noté $\hat{\gamma}_n^N$ comme étant la solution de l'équation (1.23) en γ. Nous montrons dans [32, Lemme 1] que cette solution existe et est unique. L'estimateur $\hat{\gamma}_n^N$ présente la particularité de ne pas être toujours asymptotiquement normal (voir [32, Théorème 2]). Pour démontrer ce résultat, nous introduisons les conditions suivantes :

(H.7) la fonction quantile $q(.)$ admet une dérivée négative et il exite une fonction à variations lentes $\ell^*(.)$ telle que pour $\alpha \in]0,1[$, $q'(\alpha) = -\alpha^{-(1+\gamma)}\ell^*(1/\alpha)$.

On pose $\delta = \min(-\gamma, 1/2)$ et on introduit les variables aléatoires $Z_{k_n} = \bar{F}(X_{n-k_n+1,n})/\bar{F}(X_{n,n})$ et $N_n = 1/\bar{F}(X_{n,n})$.

(H.8) La fonction à variations lentes $\ell^*(.)$ est telle que :

$$K_\delta(k_n/c) \sup_{t\in[1,Z_{k_n/c}]} \left| \frac{\ell^*(tN_n/Z_{k_n/c})}{\ell^*(N_n/Z_{k_n/c})} - 1 \right| \xrightarrow{\mathrm{P}} 0.$$

Les conditions **(H.7)** et **(H.8)** sont des hypothèses du second ordre sur la fonction quantile $q(.)$. Des conditions similaires sont utilisées par Dekkers *et al.* [67] pour établir la normalité asymptotique de l'estimateur de Pickands. La condition **(H.11)** contrôle la vitesse de convergence du rapport $\ell^*(tx)/\ell^*(x)$ vers 1. La loi asymptotique de $\hat{\gamma}_n^N$ est donnée dans le résultat ci-dessous.

Théorème 1.10 *Posons $V_{k_n}(\gamma) = K_\delta(k_n)\{(\log(k_n)-1)\mathbb{I}\{\gamma \geq 0\}+1\}$. Si la suite (k_n) satisfait l'hypothèse* **(H.1)** *et si les conditions* **(H.7)** *et* **(H.8)** *sont satisfaites alors, pour tout $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\{V_{k_n}(\gamma)(\hat{\gamma}_n^N - \gamma) \leq t\} = \begin{cases} \exp(-\mathrm{e}^{-t}) & si\ \gamma > 0, \\ \exp(-\mathrm{e}^{-t/2}) & si\ \gamma = 0, \\ \exp\{-[1 + t\log(c)/K_\gamma(1/c)]^{-1/\gamma}\} & si\ -1/2 < \gamma < 0, \\ \Phi\{-tc^{-\gamma}\log(c)/(2\gamma\sigma)\} & si\ \gamma < -1/2, \end{cases}$$

où $\sigma = c^{-\gamma}(c-1)^{1/2}$ et $\Phi(.)$ est la fonction de répartition associée à la loi normale centrée et réduite.

Le cas $\gamma = -1/2$ n'est pas traité dans le Théorème 1.10. Nous avons cependant montré que dans ce cas $V_{k_n}(\gamma)(\hat{\gamma}_n^N - \gamma)$ converge vers une loi non dégénérée n'admettant pas d'expression explicite pour sa fonction de répartition. Nous montrons aussi dans [32, Corollaire 1] que sous certaines conditions, si $\gamma < 0$, alors la vitesse de convergence $V_{k_n}(\gamma)$ de l'estimateur $\hat{\gamma}_n^N$ est de l'ordre d'une puissance de n alors que si $\gamma \geq 0$, la vitesse de convergence est dans le meilleur des cas de l'ordre de $\log^2(n)$.

1.5 Conclusion et perspectives

Dans ce chapitre, nous nous sommes intéressés à la famille des lois à queue de type Weibull. Nous avons proposé plusieurs estimateurs de l'indice de queue et de quantiles extrêmes. Ce type de lois intervenant dans de nombreuses applications (hydrologie notammment), des publications récentes leur sont consacrées. Citons par exemple Dierckx *et al.* [21] qui utilisent la moyenne des excès au dessus d'un seuil pour estimer l'indice de queue de Weibull et Goegebeur *et al.* [39] qui proposent des tests d'adéquation pour ces types de lois.

Les directions de recherche associées à cette thématique sont nombreuses. La première consiste à relier les résultats d'estimation obtenus dans ce chapitre avec ceux obtenus pour des lois à queue lourde. Pour ce faire, en collaboration avec S. Girard et A. Guillou, nous avons introduit dans un travail soumis [35] une famille de lois englobant notamment les lois du domaine d'attraction de Fréchet et les lois à queue de type Weibull. La fonction de survie de ces lois est donnée par

$$\bar{F}(x) = \exp\big(-K_\tau^{-1}(\log(x^{1/\theta}\ell(x))\big), \tag{1.24}$$

où $\theta > 0$, $\ell(.)$ est une fonction à variations lentes et

$$K_\tau(x) = \int_1^x u^{\tau-1}du,\ \tau \in [0,1].$$

Ainsi, si $\tau = 0$, $\bar{F}(.)$ est la fonction de survie d'une loi à queue de type Weibull d'indice θ. Si $\tau = 1$, la fonction de survie est celle d'une loi du domaine d'attraction de Fréchet avec pour indice des valeurs extrêmes θ. Si τ est connu, nous proposons d'estimer le paramètre θ du modèle (1.24) par

$$\hat{\theta}_n = \frac{1}{\mu_{1,\tau}(\log(n/k_n))}\frac{1}{k_n-1}\sum_{i=1}^{k_n-1}(\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})),$$

avec pour $t > 0$,

$$\mu_{1,\tau}(\log(n/k_n)) = \int_0^\infty (K_\tau(x+t) - K_\tau(t))\mathrm{e}^{-x}dx.$$

Nous montrons la normalité asymptotique de cet estimateur. Nous retrouvons donc en particulier les résultats de normalité de l'estimateur de l'indice des valeurs extrêmes proposé par Hill [84] et l'estimateur de l'indice de queue de Weibull proposé par Beirlant *et al.* [8].
Un autre axe de recherche possible est d'utiliser ce résultat pour proposer des tests d'hypothèses sur les queues de distributions. Pour un jeu de données réelles, on pourrait notamment décider s'il est issu d'une loi à queue lourde ou d'une loi à queue de type Weibull.
A plus long terme, nous envisageons d'étudier les propriètés asymptotiques des estimateurs de l'indice des valeurs extrêmes et des quantiles extrêmes lorsque les observations ne sont pas indépendantes. La dépendance des observations extrêmes a été étudiée notamment par Falk *et al.* [26, 27]. Le choix du nombre de statistiques ordonnés ainsi que l'estimation du paramètre du second ordre ρ sont des problèmes toujours ouverts aussi bien dans le cas de lois à queue lourde que pour des lois à queue de type Weibull.

Bibliographie du Chapitre 1

[1] M. Abramowitz and J. Stegan. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Dover, New York (1972).

[2] V. Asimit, D. Li, and L. Peng. Pitfalls in using Weibull tailed distributions. *Journal of Statistical Planning and Inference* (2010). To appear.

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

[4] J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, **2**, 177–200 (1999).

[5] J. Beirlant, G. Dierckx, and A. Guillou. Estimation of the extreme value index and regression on generalized quantile plots. *Bernoulli*, **11(6)**, 949–970 (2005).

[6] J. Beirlant, G. Dierckx, A. Guillou, and C. Stărică. On exponential representations of log-spacings of extreme order statistics. *Extremes*, **5**, 157–180 (2002).

[7] J. Beirlant and J. Teugels. Modelling large claims in non-life insurance. *Insurance : Mathematics and Economics*, **11**, 17–29 (1992).

[8] J. Beirlant, J. Teugels, and P. Vynckier. *Practical analysis of extreme values.* Leuven University Press, Leuven, Belgium (1996).

[9] M. Berred. Record values and the estimation of the Weibull tail-coefficient. *Comptes-Rendus de l'Académie des Sciences*, **T. 312, Série I**, 943–946 (1991).

[10] N. Bingham, C. Goldie, and J. Teugels. *Regular Variation.* Cambridge University Press (1987).

[11] E. Brodin and H. Rootzén. Univariate and bivariate gpd methods for predicting extreme wind storm losses. *Insurance : Mathematics and Economics*, **44**, 345–356 (2009).

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

[13] S. Csörgő, P. Deheuvels, and D. Mason. Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, **13**, 1050–1077 (1985).

[14] S. Csörgő and L. Viharos. Estimating the tail index. In B. Szyszkowicz, editor, *Asymptotic Methods in Probability and Statistics*, pages 833–881. North-Holland, Amsterdam (1998).

[15] L. de Haan. *Slow variation and characterization of domains of attraction.* Statistical Extremes and Application. Reidel, Dorchrecht, j. tiago de oliveira edition (1984).

[16] L. de Haan and S. Resnick. A simple asymptotic estimate for the index of a stable distribution. *Journal of the Royal Statistical Society, Series B*, **42**, 83–87 (1980).

[17] A. Dekkers and L. de Haan. On the estimation of the extreme value index and large quantile estimation. *The Annals of Statistics*, **17**, 1795–1832 (1989).

[18] A. Dekkers, J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, **17**, 1833–1855 (1989).

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

[20] J. Diebolt, L. Gardes, S. Girard, and A. Guillou. Bias-reduced extreme quantiles estimators of Weibull tail-distributions. *Journal of Statistical Planning and Inference*, **138**, 1389–1401 (2008).

[21] G. Dierckx, J. Beirlant, D. D. Waal, and A. Guillou. A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference*, **139(6)**, 1905–1920 (2009).

[22] O. Ditlevsen. Distribution arbitrariness in structural reliability. In Balkema, editor, *Structural Safety and Reliability*, pages 1241–1247. Rotterdam (1998).

[23] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*. Springer (1997).

[24] M. Falk. Some best parameter estimates for distributions with finite endpoint. *Statistics*, **27**, 115–125 (1995).

[25] M. Falk and F. Marohn. Efficient estimation of the shape parameter in Pareto models with partially known scale. *Statistics & Decisions*, **15**, 229–239 (1997).

[26] M. Falk and R. Michel. Testing for tail independence in extreme value models. *Annals of the Institute of Statistical Mathematics*, **58**, 261–290 (2006).

[27] M. Falk and R. Reiss. Efficient estimation of the canonical dependence function. *Extremes*, **6**, 61–82 (2003).

[28] A. Feuerverger and P. Hall. Estimating a tail exponent by modelling departure from a Pareto distribution. *The Annals of Statistics*, **27**, 760–781 (1999).

[29] L. Gardes. *Estimation d'une fonction quantile extrême*. Ph.D. thesis, Université Montpellier II (2003).

[30] L. Gardes and S. Girard. Asymptotic distribution of a pickands-type estimator of the extreme value index. *Comptes-Rendus de l'Académie des Sciences*, **t. 341, Série I**, 53–58 (2005).

[31] L. Gardes and S. Girard. Estimating extreme quantiles of Weibull tail-distributions. *Communication in Statistics - Theory and Methods*, **34**, 1065–1080 (2005).

[32] L. Gardes and S. Girard. Asymptotic properties of a Pickands type estimator of the extreme value index. In L. R. Velle, editor, *Focus on probability theory*, pages 133–149. Nova Science, New York (2006).

[33] L. Gardes and S. Girard. Comparison of Weibull tail-coefficient estimators. *REVSTAT - Statistical Journal*, **4(2)**, 163–188 (2006).

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

[35] L. Gardes, S. Girard, and A. Guillou. Weibull tail-distributions revisited : a new look at some tail estimators (2009). Http ://hal.archives-ouvertes.fr/hal-00340661/fr/.

[36] J. Geluk and L. D. Haan. *Regular variation, extensions and Tauberian theorems.* Center for Mathematics and Computer Science, Amsterdam, Netherlands (1987).

[37] S. Girard. A hill type estimate of the Weibull tail-coefficient. *Communication in Statistics - Theory and Methods*, **33(2)**, 205–234 (2004).

[38] B. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics*, **44**, 423–453 (1943).

[39] Y. Goegebeur and A. Guillou. Goodness-of-fit testing for Weibull-type behavior. *Journal of Statistical Planning and Inference*, **140(6)**, 1417–1436 (2010).

[40] M. Gomes. Asymptotic unbiased estimators of the tail index based on external estimation of the second order parameter. *Extremes*, **5(1)**, 5–31 (2002).

[41] M. Gomes, M. Martins, and M. Neves. Improving second order reduced bias extreme value index estimation. *Revstat*, **5(2)**, 177–207 (2007).

[42] P. Hall and B. Park. New methods for bias correction at endpoints and boundaries. *The Annals of Statistics*, **30(5)**, 1460–1479 (2002).

[43] B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**, 1163–1174 (1975).

[44] M. Kratz and S. Resnick. The qq-estimator and heavy tails. *Stochastic Models*, **12**, 699–724 (1996).

[45] D. Mason. Asymptotic normality of linear combinations of order statistics with a smooth score function. *The Annals of Statistics*, **9(4)**, 899–908 (1981).

[46] L. Peng and Q. Yongcheng. Estimating the first and second order parameters of a heavy tailed distribution. *Australian and New Zealand Journal of Statistics*, **46(2)**, 305–312 (2004).

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

[48] S. Resnick. *Extreme values, regular variation and point processes.* Springer Series in Operations Research and Financial Engineering (1987).

[49] H. Rootzén and T. Tajvidi. Can losses caused by wind storms be predicted from meteorological observations ? *Scandinavian Actuarial Journal*, **5**, 162–175 (2001).

[50] J. Schultze and J. Steinebach. On least squares estimates of an exponential tail coefficient. *Statistics and Decisions*, **14**, 353–372 (1996).

[51] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, **73**, 812–815 (1978).

Chapitre 2

Estimation de quantiles extrêmes conditionnels

2.1 Introduction

Dans ce chapitre, nous nous intéressons au cas où la variable aléatoire d'intérêt Y est mesurée conjointement avec une covariable x (aléatoire ou non). Cette situation se présente dans de nombreux domaines d'application. Citons par exemple l'hydrologie où la variable Y représente le niveau horaire de pluie (en mm) tombée en un point géographique caractérisé par sa position $x = (\text{latitude}, \text{longitude}, \text{altitude})$. Dans d'autres applications la covariable est une courbe. Par exemple en astrophysique, Y est la quantité d'un certain paramètre physique et x est une courbe hyperspectrale (voir le Chapitre 3 pour plus de détails). Notre objectif est de proposer dans ce cadre un estimateur de quantiles extrêmes qui sont alors des fonctions de la covariable. Dans la suite, ces quantiles seront appelés *quantiles extrêmes conditionnels*. Dans le cas classique (*i.e.* sans covariable), de nombreux estimateurs ont été proposés (nous en donnons quelques exemples dans le chapitre 1). La littérature sur l'estimation de quantiles extrêmes conditionnels est plus récente. Davison *et al.* [64] proposent de modéliser les excès par une loi de Pareto généralisée dont les paramètres sont des fonctions de la covariable x. Une forme paramétrique est supposée sur ces fonctions et l'estimation est effectuée par maximum de vraisemblance ou moindres carrés. Smith [95] utilise une méthode similaire en modélisant les maxima par une loi des valeurs extrêmes. Le cas de séries temporelles (*i.e.* lorsque la covariable est le temps) est considéré par Hall *et al.* [83] et Davison *et al.* [63]. Les paramètres de la loi conditionnelle de Y sachant l'instant de mesure sont estimés en maximisant une vraisemblance pondérée par une fonction noyau. Nous considérons ici principalement le cas où la loi conditionnelle de Y sachant la covariable est une loi à queue lourde. Plus précisément, la probabilité pour que $Y \geq y$ sachant que la covariable associée est égale à x est donnée par :

$$\bar{F}(y, x) = y^{-1/\gamma(x)} L(y, x),$$

où $\gamma(.)$ est une fonction à valeurs strictement positives appelée *indice des valeurs extrêmes conditionnel*. De plus, pour tout x fixé, la fonction $L(., x)$ est une fonction à variations lentes. De manière équivalente, le quantile extrême conditionnel d'ordre α est défini par :

$$q(\alpha, x) = \bar{F}^{\leftarrow}(\alpha, x) = \alpha^{-1/\gamma(x)} \ell(1/\alpha, x), \tag{2.1}$$

où pour x fixé, $\ell(.,x)$ est une autre fonction à variations lentes. Le cas de lois conditionnelles à queue lourde a été considéré notamment par Beirlant *et al.* [53, 54]. Ils imposent une forme paramétrique à la fonction $\gamma(.)$ et en estiment les paramètres par maximum de vraisemblance. Dans les sections 2.2 et 2.3 nous proposons des estimateurs de l'indice des valeurs extrêmes conditionnel et de quantiles extrêmes conditionnels pour le modèle (2.1) en ne faisant aucune hypothèse paramétrique. Le cas d'une covariable déterministe est considéré dans la section 2.2. Dans la section 2.3, nous considérons le cas d'une covariable aléatoire. Enfin, dans la section 2.4, nous considérons le cas où la loi conditionnelle de Y sachant la covariable est dans le domaine d'attraction de Weibull (*i.e.* à support borné). Estimer le quantile extrême conditionnel d'ordre zéro revient donc à estimer le support de la loi conditionnelle.

2.2 Cas d'une loi conditionnelle à queue lourde avec une covariable déterministe

Nous nous intéressons ici au cas où la covariable x mesurée simultanément avec la variable d'intérêt Y est déterministe et appartient à un espace métrique E muni d'une distance d. Nous supposons que le quantile extrême conditionnel d'ordre $\alpha \in]0,1[$ de Y sachant x est donné par (2.1). La variable Y sera supposée positive. Pour tout point $t \in E$, notre objectif est d'estimer la quantité $q(\alpha_n, t)$ lorsque $\alpha_n \to 0$. Pour ce faire, une estimation préalable de $\gamma(t)$ est nécessaire. Nous disposons d'observations indépendantes $(Y_1, x_1), \ldots, (Y_n, x_n)$ obtenues selon le modèle (2.1). Dans un premier temps, nous sélectionnons les observations dont la covariable est suffisamment proche du point t auquel on souhaite effectuer l'estimation. Plus précisément, on définit la boule de centre t et de rayon r par :

$$B(t,r) = \{x \in E,\ d(x,t) \leq r\}.$$

Pour estimer $\gamma(t)$ et $q(\alpha_n, t)$, nous utilisons les variables Y_i pour lesquelles la covariable associée x_i appartient à la boule $B(t, h_{n,t})$ où $(h_{n,t})$ est une suite positive convergeant vers zéro avec la taille de l'échantillon. Le nombre de variables ainsi sélectionnées est donné par :

$$m_{n,t} = \sum_{i=1}^{n} \mathbb{I}\{x_i \in B(t, h_{n,t})\}.$$

Remarquons que la proportion $m_{n,t}/n$ de points sélectionnés peut être rapprochée de la notion de probabilité de petite boule définie notamment dans Ferraty *et al.* [72]. Cette proportion mesure la concentration des covariables autour du point t. Les variables ainsi sélectionnées sont notées $\{Z_1, \ldots, Z_{m_{n,t}}\}$ et $Z_{1,m_{n,t}} \leq \ldots \leq Z_{m_{n,t},m_{n,t}}$ est l'échantillon ordonné associé. Cette méthode de sélection est appelée dans la suite *méthode des fenêtres mobiles*. Dans [78], nous proposons de sélectionner les variables par la méthode des plus proches voisins. Plus précisément, on se fixe un nombre $m_{n,t}$, et on sélectionne les variables $\{Z_1, \ldots, Z_{m_{n,t}}\}$ associées aux $m_{n,t}$ covariables les plus proches (au sens de la distance d) de t. Les deux approches (fenêtres mobiles ou plus proches voisins) sont équivalentes. Pour simplifier la rédaction, les résultats asymptotiques seront présentés uniquement pour la sélection par fenêtres mobiles. L'application (voir paragraphe 2.2.3) est faite en utilisant la méthode des plus proches voisins. Dans le paragraphe 2.2.1, nous proposons une famille d'estimateurs de l'indice des valeurs extrêmes

conditionnel. Les résultats obtenus sont publiés dans *Extremes* [78] et *Journal of Multivariate Analysis* [77] en collaboration avec S. Girard. L'estimation de $q(\alpha_n, t)$ est traitée dans le paragraphe 2.2.2 et a fait l'objet d'une publication dans *Journal of Multivariate Analysis* [79] en collaboration avec S. Girard et A. Lekina. L'étude du comportement asymptotique de ces estimateurs requiert entre autres les hypothèses suivantes. Soit $t \in E$ la valeur de la covariable pour laquelle on souhaite estimer l'indice des valeurs extrêmes conditionnel et le quantile extrême conditionnel. Nous supposerons dans toute la suite que $\gamma(t) > 0$.

(A.1) La fonction à variations lentes $\ell(., t)$ introduite dans (2.1) est normalisée.

Cette hypothèse est équivalente à supposer que la fonction à variations lentes $\ell(., t)$ s'écrit sous la forme :

$$\ell(1/\alpha, t) = c(t) \exp\left\{ \int_1^{\alpha^{-1}} \frac{\Delta(u,t)}{u} du \right\},$$

avec $\Delta(u, t) \to 0$ lorsque $u \to \infty$. Les deux hypothèses suivantes contrôlent le comportement au voisinage de l'infini de la fonction $\Delta(., t)$.

(A.2) La fonction $|\Delta(., t)|$ est à variations régulières d'indice $\rho(t) < 0$.

Autrement dit, pour tout $v > 0$, $|\Delta(vy, t)/\Delta(y, t)| \to v^{\rho(t)}$ lorsque $y \to \infty$. Les hypothèses **(A.1)** et **(A.2)** impliquent que

$$\log\left(\frac{\ell(v/\alpha, t)}{\ell(1/\alpha, t)}\right) = \Delta(1/\alpha, t)\frac{v^{\rho(t)} - 1}{\rho(t)}(1 + o(1)),$$

lorsque $\alpha \to 0$. La fonction $\Delta(., t)$ détermine donc la vitesse de convergence du rapport $\ell(v/\alpha, t)/\ell(1/\alpha, t)$ vers 1. Les conditions **(A.1)** et **(A.2)** (dites conditions du second-ordre) sont essentielles pour démontrer les résultats de convergence en loi des estimateurs. Une valeur du paramètre du second-ordre $\rho(t)$ proche de zéro conduit généralement à un biais important pour les estimateurs des quantiles extrêmes conditionnels.

(A.3) La fonction $|\Delta(., t)|$ est asymptotiquement décroissante.

2.2.1 Estimation de l'indice des valeurs extrêmes conditionnel

2.2.1.1 Définition de la famille d'estimateurs

Nous introduisons la famille suivante d'estimateurs de $\gamma(t)$:

$$\hat{\gamma}_n(t, W) = \sum_{i=1}^{k_{n,t}} i \log\left(\frac{Z_{m_{n,t}-i+1, m_{n,t}}}{Z_{m_{n,t}-i, m_{n,t}}}\right) W(i/k_{n,t}, t) \Bigg/ \sum_{i=1}^{k_{n,t}} W(i/k_{n,t}, t)\ , \tag{2.2}$$

où $(k_{n,t})$ est une suite d'entiers telle que $1 < k_{n,t} < m_{n,t}$ et $W(., t)$ est une fonction définie sur $]0, 1[$ et telle que $\int_0^1 W(s, t)ds \neq 0$. Des exemples de telles fonctions seront donnés dans le sous-paragraphe 2.2.1.3. Cette famille d'estimateurs est une extension de celle proposée par

Beirlant *et al.* [52] dans le cas non conditionnel (sans covariable). Pour faciliter l'écriture des estimateurs, nous posons dans la suite :

$$\{C_{i,n}(t),\ i=1,\dots,k_{n,t}\} = \left\{ i \log\left(\frac{Z_{m_{n,t}-i+1,m_{n,t}}}{Z_{m_{n,t}-i,m_{n,t}}}\right),\ i=1,\dots,k_{n,t}\right\}.$$

La normalité asymptotique est établie dans le sous-paragraphe 2.2.1.2

2.2.1.2 Résultat de normalité asymptotique

Les hypothèses requises sur la fonction poids $W(.,t)$ sont données ci-dessous. Elles sont également utilisées dans Beirlant *et al.* [52] pour démontrer la normalité asymptotique de leurs estimateurs.

(B.1) Il existe une fonction $u(.,t)$ définie sur $]0,1[$ telle que pour tout $s \in]0,1[$,

$$sW(s,t) = \int_0^s u(\xi,t)d\xi,$$

avec pour tout $j=1,\dots,k_{n,t}$,

$$\left| k_{n,t} \int_{(j-1)/k_{n,t}}^{j/k_{n,t}} u(\xi,t)d\xi \right| < g\left(\frac{j}{k_{n,t}+1},t\right),$$

où $g(.,t)$ est une fonction positive, continue, définie sur $]0,1[$ telle que

$$\int_0^1 \max(1,\log(1/s))g(s,t)ds < \infty.$$

(B.2) Il existe une constante $\delta_1 > 0$ telle que

$$\int_0^1 |W(s,t)|^{2+\delta_1} ds < \infty.$$

Le comportement asymptotique des estimateurs $\hat{\gamma}_n(t,W)$ dépend de celui de la fonction log-quantile au voisinage du point t. Nous introduisons une mesure de l'oscillation de la fonction log-quantile : soit $a \in]0,1/2[$, posons

$$\omega_n(a) = \sup\left\{ \left|\log\frac{q(\alpha,x)}{q(\alpha,x')}\right|,\ \alpha \in]a,1-a[,\ (x,x') \in B(t,h_{n,t})^2 \right\}.$$

Nous donnons à présent le résultat de normalité asymptotique que nous avons obtenu (voir [77, Théorème 2]).

Théorème 2.1 *On se place sous le modèle (2.1). Sous les hypothèses* **(A.1)**, **(A.2)**, **(A.3)**, **(B.1)** *et* **(B.2)**, *si* $k_{n,t} \to \infty$, $m_{n,t}/k_{n,t} \to \infty$ *et s'il existe* $\delta_2 > 0$ *tel que*

$$k_{n,t}^2 \omega_n(m_{n,t}^{-(1+\delta_2)}) \to 0, \text{ et } k_{n,t}^{1/2}\Delta(m_{n,t}/k_{n,t},t) \to \xi(t) \in \mathbb{R},$$

alors

$$k_{n,t}^{1/2}(\hat{\gamma}_n(t,W) - \gamma(t) - \Delta(m_{n,t}/k_{n,t},t)\mathcal{AB}(t,W)) \xrightarrow{d} \mathcal{N}(0,\gamma^2(t)\mathcal{AV}(t,W)),$$

avec

$$\mathcal{AB}(t,W) = \int_0^1 W(s,t)s^{-\rho(t)}ds \text{ et } \mathcal{AV}(t,W) = \int_0^1 W^2(s,t)ds.$$

Les hypothèses $k_{n,t} \to \infty$ et $m_{n,t}/k_{n,t} \to \infty$ sont classiques en théorie des valeurs extrêmes. Ce sont les adaptations au cas conditionnel de l'hypothèse **(H.1)** utilisée dans le chapitre 1. Remarquons aussi qu'elles impliquent que $m_{n,t} \to \infty$ lorsque $n \to \infty$. Le biais asymptotique de $\hat{\gamma}_n(t,W)$ est donné par $\Delta(m_{n,t}/k_{n,t},t)\mathcal{AB}(t,W)$. Le facteur $\Delta(m_{n,t}/k_{n,t},t)$ dépend uniquement de la loi des observations. Le facteur $\mathcal{AB}(t,W)$ peut être controlé par un choix judicieux de la fonction poids $W(.,t)$ (voir sous-paragraphe 2.2.1.3). La variance asymptotique est quant à elle donnée par $\gamma^2(t)\mathcal{AV}(t,W)/k_{n,t}$. Une grande valeur de $\gamma(t)$ (*i.e.* une queue très lourde) conduit donc à une importante variance asymptotique. Ici encore, le facteur $\mathcal{AV}(t,W)$ peut être minimisé par un bon choix de la fonction $W(.,t)$ (voir sous-paragraphe 2.2.1.3). La condition $k_{n,t}^{1/2}\Delta(m_{n,t}/k_{n,t},t) \to \xi(t) \in \mathbb{R}$ impose au biais asymptotique d'être du même ordre que la variance asymptotique. Enfin, la condition $k_{n,t}^2\omega_n(m_{n,t}^{-(1+\delta_2)}) \to 0$ est nécessaire pour contrôler les variations de la loi des estimateurs dans la boule $B(t,h_{n,t})$.

2.2.1.3 Quelques choix de fonctions poids

Deux choix classiques Le choix le plus simple consiste à prendre dans (2.2) la fonction de poids constante $W^H(s,t) = 1$ pour tout $s \in [0,1]$. Ceci conduit à l'estimateur

$$\hat{\gamma}_n^H(t) = \hat{\gamma}_n(t,W^H) = \frac{1}{k_{n,t}}\sum_{i=1}^{k_{n,t}} C_{i,n}(t),$$

qui est une adaptation directe de l'estimateur de Hill [84]. La fonction $W^H(.,t)$ satisfait les hypothèses **(B.1)** et **(B.2)** et le Théorème 2.1 s'applique avec $\mathcal{AB}(t,W^H) = (1-\rho(t))^{-1}$ et $\mathcal{AV}(t,W^H) = 1$. Nous avons aussi montré (voir [77, Proposition 2]) que cet estimateur possède la plus petite variance asymptotique parmi les estimateurs de la famille (2.2). Nous avons aussi proposé une méthode d'estimation de $\gamma(t)$ inspirée de celle utilisée dans un cadre non conditionnel par Kratz *et al.* [87] et Schultze *et al.* [94] pour construire un estimateur de l'indice des valeur extrême appelé *estimateur de Zipf*. Nous nous sommes basés sur la remarque suivante : pour une valeur de $k_{n,t}$ pas trop grande et pour $h_{n,t}$ proche de zéro, les points

$$\left(\tau_{i,n}(t) = \sum_{j=i}^{m_{n,t}} \frac{1}{j}, \log(Z_{m_{n,t}-i+1,m_{n,t}})\right),\ i = 1,\dots,k_{n,t}$$

sont approximativement situés sur une droite de pente $\gamma(t)$. L'estimateur des moindres carrés associé est donné par :

$$\sum_{i=1}^{k_{n,t}}(\tau_{i,n}(t) - \bar{\tau}_n(t))\log(Z_{m_{n,t}-i+1,m_{n,t}}) \Bigg/ \sum_{i=1}^{k_{n,t}}(\tau_{i,n}(t) - \bar{\tau}_n(t))\tau_{i,n}(t) \tag{2.3}$$

où

$$\bar{\tau}_n(t) = \frac{1}{k_{n,t}} \sum_{i=1}^{k_{n,t}} \tau_{i,n}(t).$$

En remarquant que (2.3) peut se réécrire sous la forme

$$\sum_{i=1}^{k_{n,t}} \left(\frac{1}{i} \sum_{j=1}^{i} (\tau_{j,n}(t) - \bar{\tau}_n(t)) \right) C_{i,n}(t) \Bigg/ \sum_{i=1}^{k_{n,t}} \left(\frac{1}{i} \sum_{j=1}^{i} (\tau_{j,n}(t) - \bar{\tau}_n(t)) \right),$$

et en montrant que (voir [77, Démonstration du Corollaire 4])

$$\frac{1}{i} \sum_{j=1}^{i} (\tau_{j,n}(t) - \bar{\tau}_n(t)) \sim -\log(i/k_{n,t}),$$

nous avons proposé d'utiliser la fonction poids définie par $W^Z(s,t) = -\log(s)$ conduisant à l'estimateur de $\gamma(t)$ défini par :

$$\hat{\gamma}_n^Z(t) = \hat{\gamma}_n(t, W^Z) = \frac{1}{k_{n,t}} \sum_{i=1}^{k_{n,t}} \log(k_{n,t}/i) C_{i,n}(t).$$

La fonction $W^Z(.,t)$ satisfaisant les hypothèses **(B.1)** et **(B.2)**, le Théorème 2.1 établit la normalité asymptotique de $\hat{\gamma}_n^Z(t)$ avec $\mathcal{AB}(t, W^Z) = (1-\rho(t))^{-2}$ et $\mathcal{AV}(t, W^Z) = 2$. Cet estimateur possède évidemment une variance asymptotique plus grande que celle de $\hat{\gamma}_n^H(t)$. Par contre, l'estimateur $\hat{\gamma}_n^Z(t)$ donnant des poids plus importants aux grandes observations, il possède un biais asymptotique plus faible que l'estimateur $\hat{\gamma}_n^H(t)$.

Estimateur asymptotiquement sans biais de variance minimale Nous avons montré dans [77, Proposition 3] que la fonction poids $W(.,t)$ solution du problème d'optimisation :

$$\text{minimiser} \int_0^1 W^2(s,t)ds \text{ sous les contraintes } \int_0^1 W(s,t)s^{-\rho(t)}ds = 0 \text{ et } \int_0^1 W(s,t)ds = 1,$$

est définie par :

$$W^{\text{opt}}(s,t) = \frac{\rho(t)-1}{\rho^2(t)} \big(\rho(t) - 1 + (1-2\rho(t))s^{-\rho(t)}\big).$$

La fonction $W^{\text{opt}}(.,t)$ vérifiant les hypothèses **(B.1)** et **(B.2)**, son utilisation dans (2.2) donne lieu à un estimateur asymptotiquement sans biais ($\mathcal{AB}(t, W^{\text{opt}}) = 0$) et ayant la plus petite variance asymptotique possible (ici $\mathcal{AV}(t, W^{\text{opt}}) = (1-1/\rho(t))^2$). Cependant cet estimateur n'est pas utilisable en pratique car la valeur du paramètre $\rho(t)$ est inconnue. Nous nous intéressons à présent à l'estimateur obtenu en remplaçant dans la fonction $W^{\text{opt}}(.,t)$ le paramètre $\rho(t)$ par une valeur arbitraire ρ^*.

Estimateur obtenu en remplaçant $\rho(t)$ par ρ^* Nous considérons la famille de fonctions poids définie par :

$$\left\{W_{\rho^*}^{\text{opt}}(s,t) = \frac{\rho^*-1}{(\rho^*)^2}\big(\rho^*-1+(1-2\rho^*)s^{-\rho^*}\big),\ \rho^*<0\right\}.$$

Les fonctions de cette famille satisfaisant les hypothèses **(B.1)** et **(B.2)**, le Théorème 2.1 assure la normalité des estimateurs associés avec un biais asymptotique proportionnel à

$$\mathcal{AB}(t,W_{\rho^*}^{\text{opt}}) = \frac{(1-\rho^*)(\rho^*-\rho(t))}{\rho^*(1-\rho(t))(1-\rho^*-\rho(t))},$$

et une variance asymptotique proportionnelle à

$$\mathcal{AV}(t,W_{\rho^*}^{\text{opt}}) = (1-1/\rho^*)^2.$$

Evidemment si $\rho^*=\rho(t)$ le biais asymptotique est nul. De plus si on choisit $\rho^*<\rho(t)$ alors l'estimateur obtenu avec le poids $W_{\rho^*}^{\text{opt}}(.,t)$ (noté $\hat{\gamma}_n^{\text{opt}}(t,\rho^*)$) possède une variance asymptotique plus faible que l'estimateur asymptotiquement sans biais de variance minimale. Le choix de ρ^* reste cependant difficile. Comme le suggérent Feuerverger *et al.* [73], un choix possible est de poser $\rho^*=-1$. Pour ce choix, la variance asymptotique de $\hat{\gamma}_n^{\text{opt}}(t,-1)$ sera 4 fois plus importante que celle de l'estimateur $\hat{\gamma}_n^H(t)$. Concernant le biais asymptotique, on a :

$$\mathcal{AB}(t,W^Z) \le \mathcal{AB}(t,W^H) \le \mathcal{AB}(t,W_{-1}^{\text{opt}}) \text{ si } \rho(t)\le -4,$$

$$\mathcal{AB}(t,W^Z) \le \mathcal{AB}(t,W_{-1}^{\text{opt}}) \le \mathcal{AB}(t,W^H) \text{ si } \rho(t)\in[-4,-(1+\sqrt{33})/4\approx -1.686],$$

$$\mathcal{AB}(t,W_{-1}^{\text{opt}}) \le \mathcal{AB}(t,W^Z) \le \mathcal{AB}(t,W^H) \text{ si } \rho(t)\in[-(1+\sqrt{33})/4,0].$$

Ainsi pour des valeurs de $\rho(t)$ proche de zéro (plus précisément supérieures à -1.686), l'estimateur $\hat{\gamma}_n^{\text{opt}}(t,-1)$ possède un biais asymptotique plus faible que ceux de $\hat{\gamma}_n^H(t)$ et $\hat{\gamma}_n^Z(t)$.

Famille de poids log-gamma Nous nous sommes aussi intéressés dans [78] aux fonctions poids de la forme :

$$W(s,t) = p(s,a,\lambda) = \frac{\lambda^{-a}}{\Gamma(a)}s^{-1/\lambda-1}(-\log s)^{a-1},$$

où $a\ge 1$ et $\lambda\in]0,1]$. La fonction $p(.,a,\lambda)$ est une densité sur l'intervalle $]0,1[$ introduite notamment par Consul *et al.* [59] sous le nom de densité *log-gamma*. En prenant $a=\lambda=1$ on retrouve la fonction $W^H(.,t)$ et avec $a=2$ et $\lambda=1$, on obtient la fonction poids $W^Z(.,t)$. La fonction $p(.,a,\lambda)$ satisfait les hypothèses **(B.1)** et **(B.2)** avec :

$$\mathcal{AB}(t,p(.,a,\lambda)) = \mathcal{AB}(a,\lambda,\rho(t)) = (1-\lambda\rho(t))^{-a} \text{ et}$$

$$\mathcal{AV}(t,p(.,a,\lambda)) = \mathcal{AV}(a,\lambda) = \frac{\Gamma(2a-1)}{\lambda\Gamma^2(a)}(2-\lambda)^{1-2a}.$$

Une méthode pour choisir au mieux les paramètres a et λ est de minimiser en ces paramètres la moyenne asymptotique du carré des erreurs donnée d'après le Théorème 2.1 par :

$$\Delta^2\left(\frac{m_{n,t}}{k_{n,t}},t\right)\mathcal{AB}^2(a,\lambda,\rho(t)) + \gamma^2(t)\frac{\mathcal{AV}(a,\lambda)}{k_{n,t}}.$$

Cette erreur n'est pas calculable en pratique car le paramètre $\rho(t)$ et la fonction $\Delta(.,t)$ sont inconnus. Nous proposons de remplacer le carré du biais asymptotique $\mathcal{AB}^2(a,\lambda,\rho(t))$ par sa moyenne sur toutes les valeurs possibles de $\rho(t)$:

$$\mathcal{MSB}(a,\lambda) = \int_{-\infty}^{0} \mathcal{AB}^2(a,\lambda,\rho)d\rho = \frac{1}{\lambda(2a-1)},$$

et de majorer l'erreur obtenue de la façon suivante : en posant $\pi(a,\lambda) = \mathcal{MSB}(a,\lambda)\mathcal{AV}(a,\lambda)$, on a pour tout $\lambda \in]0,1]$ et $a \in [1, a_{\max}]$,

$$\begin{aligned}\Delta^2\left(\frac{m_{n,t}}{k_{n,t}}, t\right)\mathcal{MSB}(a,\lambda) + \gamma^2(t)\frac{\mathcal{AV}(a,\lambda)}{k_{n,t}} &= \frac{\pi(a,\lambda)}{k_{n,t}}\left\{\frac{k_{n,t}\Delta^2(m_{n,t}/k_{n,t},t)}{\mathcal{AV}(a,\lambda)} + \frac{\gamma^2(t)}{\mathcal{MSB}(a,\lambda)}\right\} \\ &\leq \frac{\pi(a,\lambda)}{k_{n,t}}\left\{\xi^2(t) + o(1) + \gamma^2(t)(2a_{\max}-1)\right\}.\end{aligned}$$

Nous proposons alors de trouver les paramètres a et λ minimisant le produit $\pi(a,\lambda)$. Les résultats de ce problème d'optimisation sont donnés par

$$\lambda_\pi = \frac{4}{1+2a_\pi} \text{ et } a_\pi \approx 2,19.$$

En terme de biais asymptotique, l'estimateur $\hat{\gamma}_n^\pi(t)$ obtenu en utilisant la fonction poids $p(.,a_\pi,\lambda_\pi)$ est meilleur que l'estimateur $\hat{\gamma}_n^H(t)$ mais mois bon que $\hat{\gamma}_n^Z(t)$. Concernant la variance asymptotique, le classement est inversé. Pour la famille des fonctions de poids log-gamma, nous avons montré que si l'on fixe $\mathcal{MSB}(a,\lambda)$ à la valeur b, la variance asymptotique est alors proportionnelle à :

$$b\frac{\Gamma(2a)}{\Gamma^2(a)}\left\{2 - \frac{1}{b(2a-1)}\right\}^{1-2a}, \ a \geq \max\{1,(1+b)/(2b)\}.$$

Nous pouvons donc calculer pour une valeur fixée b de $\mathcal{MSB}(a,\lambda)$ la variance asymptotique optimale obtenue en minimisant la quantité ci-dessus :

$$\mathcal{OAV}(b) = \min_{a\geq\max\{1,(1+b)/(2b)\}} b\frac{\Gamma(2a)}{\Gamma^2(a)}\left\{2 - \frac{1}{b(2a-1)}\right\}^{1-2a}.$$

La Figure 2.1 ci-dessous représente la quantité $\mathcal{OAV}(b)$ en fonction de b. Il apparaît que les estimateurs $\hat{\gamma}_n^H(t)$ et $\hat{\gamma}_n^\pi(t)$ sont optimaux en ce sens qu'ils ont la plus petite variance possible compte tenu de leur biais respectif. Par contre, l'estimateur $\hat{\gamma}_n^Z(t)$ n'est pas optimal puisque l'on peut trouver un choix de a et λ conduisant à un estimateur ayant la même valeur de $\mathcal{MSB}$ mais avec une variance asymptotique plus faible.

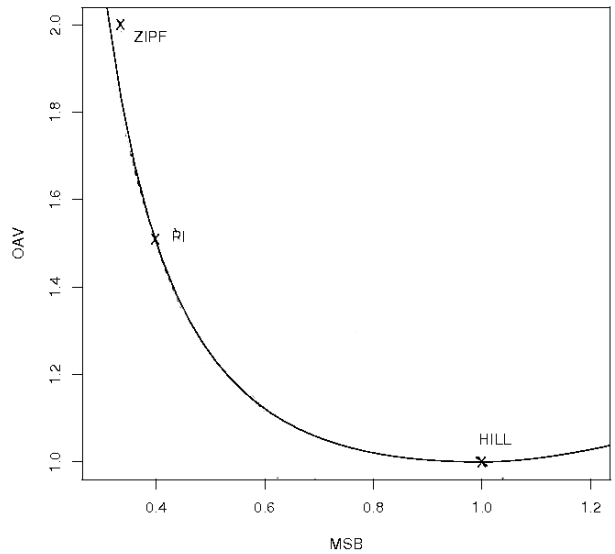

FIG. 2.1 – Variance optimale $\mathcal{OAV}$ en fonction de la valeur de $\mathcal{MSB}$. Les estimateurs $\hat{\gamma}_n^H(t)$ (HILL), $\hat{\gamma}_n^Z(t)$ (ZIPF) et $\hat{\gamma}_n^\pi(t)$ (PI) sont représentés par des croix ($\times$).

2.2.2 Estimation de quantiles extrêmes conditionnels

Nous nous intéressons à présent à l'estimation de quantiles extrêmes conditionnels au point $t \in E$ définis par

$$q(\alpha_{m_{n,t}}, t) = \alpha_{m_{n,t}}^{-\gamma(t)} \ell(1/\alpha_{m_{n,t}}, t),$$

où $\alpha_{m_{n,t}} \to 0$ lorsque $n \to \infty$. Rappelons que $m_{n,t}$ est le nombre d'observations sélectionnées pour l'estimation au point t et que $m_{n,t} \to \infty$ lorsque $n \to \infty$. Concernant la vitesse de convergence de $\alpha_{m_{n,t}}$ vers zéro, nous considérons trois situations :

(S.1) $\alpha_{m_{n,t}} \to 0$ et $m_{n,t}\alpha_{m_{n,t}} \to \infty$,
(S.2) $\alpha_{m_{n,t}} \to 0$ et $\lfloor m_{n,t}\alpha_{m_{n,t}} \rfloor \to r \in \{1, 2, \ldots\}$,
(S.3) $\alpha_{m_{n,t}} \to 0$ et $\lfloor m_{n,t}\alpha_{m_{n,t}} \rfloor \to 0$,

où $\lfloor x \rfloor$ est la partie entière de x. La situation **(S.1)** a été étudiée dans le cas non-conditionnel par Dekkers *et al.* [67] et le cas $m_{n,t}\alpha_{m_{n,t}} \to 0$ (cas particulier de la situation **(S.3)**) par de Haan [65] et de Haan *et al.* [66]. Dans la situation **(S.1)**, le quantile extrême conditionnel est situé dans l'intervalle $[Z_{1,m_{n,t}}, Z_{m_{n,t},m_{n,t}}]$ avec une probabilité tendant vers 1 (voir [79, Proposition 2]). On peut donc utiliser l'estimateur obtenu en inversant la fonction de répartition empirique calculée à partir des observations $\{Z_1, \ldots, Z_{m_{n,t}}\}$:

$$\hat{q}_1(\alpha_{m_{n,t}}, t) = Z_{m_{n,t} - \lfloor m_{n,t} \rfloor + 1, m_{n,t}}.$$

Dans la situation **(S.2)**, comme pour n assez grand, $\lfloor m_{n,t}\alpha_{m_{n,t}} \rfloor = r \geq 1$, on peut aussi utiliser l'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$. Enfin, dans la situation **(S.3)**, l'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$ n'est

plus utilisable. Le quantile extrême extrême conditionnel que l'on souhaite estimer est, avec une probabilité asymptotique non nulle, plus grand que l'observation maximale $Z_{m_{n,t},m_{n,t}}$. Nous proposons alors d'utiliser l'estimateur

$$\hat{q}_2(\alpha_{m_{n,t}}, t) = Z_{m_{n,t}-k_{n,t}+1,m_{n,t}} \left(\frac{k_{n,t}}{m_{n,t}\alpha_{m_{n,t}}} \right)^{\hat{\gamma}_n(t)}.$$

Cet estimateur est l'adaptation de l'estimateur proposé par Weissman [98] dans un cadre non-conditionnel. Nous donnons les résultats asymptotiques obtenus dans chacune des trois situations.

Situation (S.1) avec l'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$

Théorème 2.2 *Si la suite* $(\alpha_{m_{n,t}})$ *vérifie la condition* **(S.1)** *et s'il existe* $\delta_2 > 0$ *tel que* $(m_{n,t}\alpha_{m_{n,t}})^2 \omega_n(m_{n,t}^{-(1+\delta_2)}) \to 0$, *alors*

$$(m_{n,t}\alpha_{m_{n,t}})^{1/2} \left(\frac{\hat{q}_1(\alpha_{m_{n,t}}, t)}{q(\alpha_{m_{n,t}}, t)} - 1 \right) \xrightarrow{d} \mathcal{N}(0, \gamma^2(t)).$$

L'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$ est donc dans cette situation asymptotiquement normal. La variance asymptotique est proportionnelle à $\gamma^2(t)$ (la variance est d'autant plus importante que la queue de la distribution est lourde). Cette variance est aussi inversement proportionnelle à $\alpha_{m_{n,t}}$. Donc, plus le quantile conditionnel à estimer est extrême, plus la variance asymptotique est importante.

Situation (S.2) avec l'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$

Théorème 2.3 *Si la suite* $(\alpha_{m_{n,t}})$ *vérifie la condition* **(S.2)** *et s'il existe* $\delta_2 > 0$ *tel que* $(m_{n,t}\alpha_{m_{n,t}})^2 \omega_n(m_{n,t}^{-(1+\delta_2)}) \to 0$, *alors*

$$\left(\frac{\hat{q}_1(\alpha_{m_{n,t}}, t)}{q(\alpha_{m_{n,t}}, t)} - 1 \right) \xrightarrow{d} \mathcal{E}(r, \gamma(t)),$$

où $\mathcal{E}(r, \gamma(t))$ *est une loi non dégénérée.*

La loi asymptotique $\mathcal{E}(r, \gamma(t))$ admet une expression complexe que nous ne donnons pas ici pour ne pas alourdir la présentation. Il apparaît que dans la situation **(S.2)** l'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$ n'est pas asymptotiquement normal et surtout qu'il n'est pas consistant en ce sens que :

$$\frac{\hat{q}_1(\alpha_{m_{n,t}}, t)}{q(\alpha_{m_{n,t}}, t)} \text{ ne converge pas en probabilité vers 1.}$$

Situations (S.1), (S.2) et (S.3) avec l'estimateur $\hat{q}_2(\alpha_{m_{n,t}}, t)$

Selon la vitesse de convergence de $\hat{\gamma}_n(t)$ vers $\gamma(t)$, la loi asymptotique de $\hat{q}_2(\alpha_{m_{n,t}}, t)$ sera donnée soit par celle de la statistique d'ordre $Z_{m_{n,t}-k_{n,t}+1,m_{n,t}}$ soit par celle de l'estimateur $\hat{\gamma}_n(t)$.

Théorème 2.4 *Supposons que $k_{n,t} \to \infty$ et $k_{n,t}/m_{n,t} \to 0$. Soit $(\alpha_{m_{n,t}})$ une suite satisfaisant une des situations* **(S.1)**, **(S.2)** *ou* **(S.3)**. *Posons $\zeta_{m_{n,t}} = k_{n,t}^{1/2} \log(k_{n,t}/(m_{n,t}\alpha_{m_{n,t}}))$.*
S'il existe $\delta_2 > 0$ tel que $k_{n,t}^2 \omega_n(m_{n,t}^{-(1+\delta_2)}) \to 0$, une suite positive $\upsilon_n(t)$ et une loi $\mathcal{D}$ telles que

$$\upsilon_n(t)(\hat{\gamma}_n(t) - \gamma(t)) \xrightarrow{d} \mathcal{D}, \tag{2.4}$$

alors, deux situations peuvent se présenter :

(i) Sous la condition supplémentaire

$$\zeta_{m_{n,t}} \max\left\{\upsilon_n^{-1}(t), \Delta(k_{n,t}/m_{n,t}, t)\right\} \to 0, \tag{2.5}$$

on a

$$k_{n,t}^{1/2}\left(\frac{\hat{q}_2(\alpha_{m_{n,t}}, t)}{q(\alpha_{m_{n,t}}, t)} - 1\right) \xrightarrow{d} \mathcal{N}(0, \gamma^2(t)). \tag{2.6}$$

(ii) Sous la condition supplémentaire

$$\upsilon_n(t) \max\left\{\zeta_{m_{n,t}}^{-1}, \Delta(k_{n,t}/m_{n,t}, t)\right\} \to 0, \tag{2.7}$$

on a

$$\frac{\upsilon_n(t)}{\log\left(k_{n,t}/(m_{n,t}\alpha_{m_{n,t}})\right)}\left(\frac{\hat{q}_2(\alpha_{m_{n,t}}, t)}{q(\alpha_{m_{n,t}}, t)} - 1\right) \xrightarrow{d} \mathcal{D}. \tag{2.8}$$

Remarquons que dans la situation **(S.2)**, l'estimateur $\hat{q}_2(\alpha_{m_{n,t}}, t)$ est consistant ce qui n'était pas le cas pour l'estimateur $\hat{q}_1(\alpha_{m_{n,t}}, t)$. Dans le cas particulier où l'estimateur de $\gamma(t)$ est choisi dans la famille (2.2), nous avons

$$\zeta_{m_{n,t}} \upsilon_n^{-1}(t) = \log\left(\frac{k_{n,t}}{m_{n,t}\alpha_{m_{n,t}}}\right) \to \infty,$$

lorsque la suite $\alpha_{m_{n,t}}$ satisfait la condition **(S.2)** ou **(S.3)**. Ainsi, seul le cas (ii) du Théorème 2.4 est possible. Nous en déduisons le résultat suivant (voir [79, Corollaire 1]) sur la normalité asymptotique de l'estimateur

$$\hat{q}_2(\alpha_{m_{n,t}}, t, W) = Z_{m_{n,t}-k_{n,t}+1, m_{n,t}}\left(\frac{k_{n,t}}{m_{n,t}\alpha_{m_{n,t}}}\right)^{\hat{\gamma}_n(t,W)}.$$

Corollaire 2.1 *On se place sous le modèle (2.1). Sous les hypothèses* **(A.1)**, **(A.2)**, **(A.3)**, **(B.1)** *et* **(B.2)**, *si $k_{n,t} \to \infty$, $m_{n,t}/k_{n,t} \to \infty$ et s'il existe $\delta_2 > 0$ tel que*

$$k_{n,t}^2 \omega_n(m_{n,t}^{-(1+\delta_2)}) \to 0, \text{ et } k_{n,t}^{1/2}\Delta(m_{n,t}/k_{n,t}, t) \to 0,$$

alors, si $\alpha_{m_{n,t}}$ satisfait la condition **(S.2)** *ou* **(S.3)**,

$$\frac{k_{n,t}^{1/2}}{\log(k_{n,t}/(m_{n,t}\alpha_{m_{n,t}}))}\left(\frac{\hat{q}_2(\alpha_{m_{n,t}}, t, W)}{q(\alpha_{m_{n,t}}, t)} - 1\right) \xrightarrow{d} \mathcal{N}(0, \gamma^2(t)\mathcal{AV}(t, W)).$$

2.2.3 Application à l'estimation de niveaux de retour

Nous nous intéressons ici à l'estimation d'une carte de niveau de retour pour des quantités horaires de pluie. Pour ce faire, nous disposons de données fournies par le Laboratoire d'étude des Transferts en Hydrologie et Environnement de Grenoble dans le cadre d'un projet financé par l'Agence Nationale de la Recherche (programme VMC : Vulnérabilité ; Millieux, Climats). Les données sont des niveaux de pluie Y (l'unité étant le mm) mesurés toutes les heures sur 142 stations dans la région Cévennes-Vivarais (sud de la France, voir Figure 2.2). La covariable x est ici tri-dimensionnelle : la longitude, la latitude et l'altitude. Ces mesures ont été effectuées entre 1993 et 2000 et nous disposons de $n = 264056$ observations. Les pluies sur cette région ont aussi été étudiées par Bois *et al.* [55]. L'étude statistique de précipitations a intéressé de nombreux auteurs. Citons Coles *et al.* [58] et Cooley *et al.* [60] qui ont modélisé les précipitations par une loi de Pareto Généralisée.

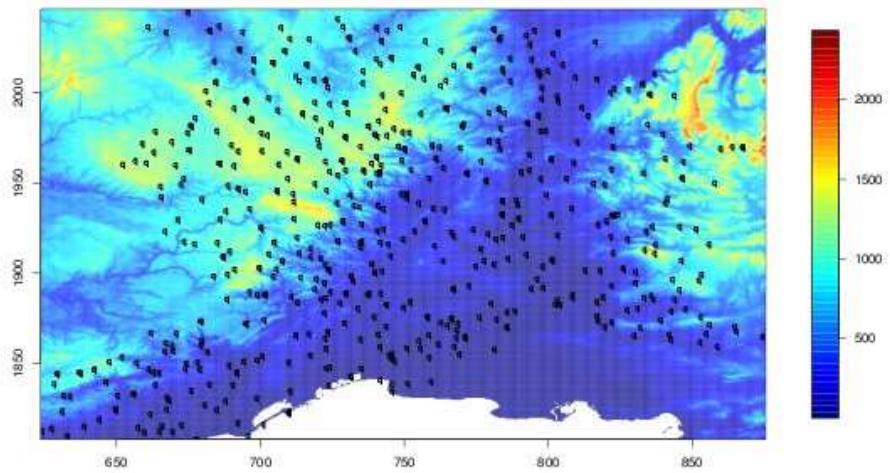

FIG. 2.2 – Carte des stations de mesure de la région Cévennes-Vivarais. En abscisse la longitude, en ordonnée la latitude. L'échelle de couleur représente l'altitude et les points noirs l'emplacement des stations.

Notre objectif est d'estimer le niveau de retour de 10 ans c'est à dire le niveau de pluie que l'on s'attend à voir dépasser tous les 10 ans et ceci pour tous les points t de la région étudiée. Il s'agit donc d'estimer un quantile extrême conditionnel d'ordre $1/(365.25 \times 24 \times 10)$. Ce quantile est bien extrême car nous disposons uniquement de 7 années d'observations. Avant d'appliquer l'estimateur de quantiles extrêmes conditionnels étudié dans le paragraphe précédent, il est important de remarquer que nous ne sommes pas ici dans le cadre d'application des résultats asymptotiques précédents. En effet, ces derniers ont été obtenus sous une hypothèse d'indépendance entre les observations. Cette hypothèse n'est clairement pas satisfaite pour nos données qui présentent deux types de dépendance : temporelle et spatiale. Nous avons donc étudié au préalable le comportement sur données simulées de nos estimateurs (en terme de biais et de variance) pour ces deux types de dépendance.

Effet de la dépendance temporelle Pour ne pas alourdir la présentation des résultats, nous nous plaçons ici dans un cadre non-conditionnel (sans covariable). L'utilisation d'une covariable sera par contre incontournable pour étudier la dépendance spatiale. Pour étudier l'effet de la dépendance temporelle sur l'estimation de l'indice des valeurs extrêmes conditionnels, nous simulons une série temporelle $\{y_1, \ldots, y_n\}$ avec $n = 500$ selon une méthode proposée par Fawcett *et al.* [71]. Plus précisément, nous générons dans un premier temps une série temporelle $\{f_1, \ldots, f_n\}$ dont les lois marginales sont Fréchet. La loi du couple (f_i, f_{i+1}), $i = 1, \ldots, n-1$ est une loi des valeurs extrêmes bivariée $G(u, v) = \exp\{-V(u, v)\}$ où la fonction de dépendance $V(.,.)$ est donnée par $V(u, v) = (u^{-1/\alpha} + v^{-1/\alpha})^\alpha$, $u > 0$, $v > 0$ et $\alpha \in]0, 1]$. Le paramètre α contrôle la dépendance entre deux variables consécutives : pour $\alpha = 1$ les variables sont indépendantes et pour $\alpha \to 0$ complètement dépendantes. L'algorithme de simulation utilisé pour générer la série temporelle $\{f_1, \ldots, f_n\}$ est décrit ci-dessous.

1. On simule la première observation f_1 selon une loi de Fréchet.
2. Pour $i = 1, \ldots, n-1$, on calcule la loi conditionnelle de f_{i+1} sachant f_i et on génère f_{i+1} selon cette loi.

Enfin, on effectue une transformation sur la série $\{f_1, \ldots, f_n\}$ pour obtenir des marges distribuées selon la loi de Burr. Rappelons que la fonction de répartition de la loi de Burr est donnée, pour $y \geq 0$, par $1 - (1 + y^{-\rho/\gamma})^{1/\rho}$ où ρ est le paramètre du second ordre introduit dans la condition **(A.2)**. Ici, nous prenons $\rho = -1$ et $\gamma = 0.2$. Nous simulons ainsi $N = 100$ séries temporelles dont les marges suivent une loi de Burr. L'estimateur $\hat{\gamma}_n^\pi$ est ensuite calculé sur ces N réplications. On obtient les valeurs :

$$\hat{\gamma}_{n,j}^\pi = \sum_{i=1}^{k} p\left(\frac{i}{k}, a_\pi, \lambda_\pi\right) i \log\left(\frac{y_{n-i+1,n}^{(j)}}{y_{n-i,n}^{(j)}}\right) \Bigg/ \sum_{i=1}^{k} p\left(\frac{i}{k}, a_\pi, \lambda_\pi\right), \; j = 1, \ldots, N,$$

où $y_{1,n}^{(j)} \leq \ldots \leq y_{n,n}^{(j)}$ est la $j-$ème réplication de la série temporelle rangée par orde croissant. Le biais au carré empirique ($\mathcal{ESB}$) et la variance empirique ($\mathcal{EV}$) définis par

$$\mathcal{ESB} = \frac{1}{N} \sum_{j=1}^{N} (\hat{\gamma}_{n,j}^\pi - \gamma)^2 \quad \text{et} \quad \mathcal{EV} = \frac{1}{N} \sum_{j=1}^{N} \left(\hat{\gamma}_{n,j}^\pi - \frac{1}{N} \sum_{j=1}^{N} \hat{\gamma}_{n,j}^\pi\right)^2$$

sont représentés en fonction de k sur les Figures 2.3 et 2.4 pour plusieurs valeurs du coefficient de dépendance $\alpha \in \{1, 0.8, 0.5, 0.2\}$. On remarque sur la Figure 2.3 que même pour une forte dépendance ($\alpha = 0.2$), un bon choix de k permet toujours d'obtenir un estimateur faiblement biaisé. Il apparaît donc que, concernant le biais, le choix du nombre de statistiques d'ordre k est plus critique que la présence ou non de dépendance dans les observations. Par contre, la variance de l'estimateur $\hat{\gamma}_n^\pi$ augmente avec le coefficient de dépendance temporelle (voir Figure 2.4). Ces remarques ont aussi été faites par Fawcett *et al.* [71].

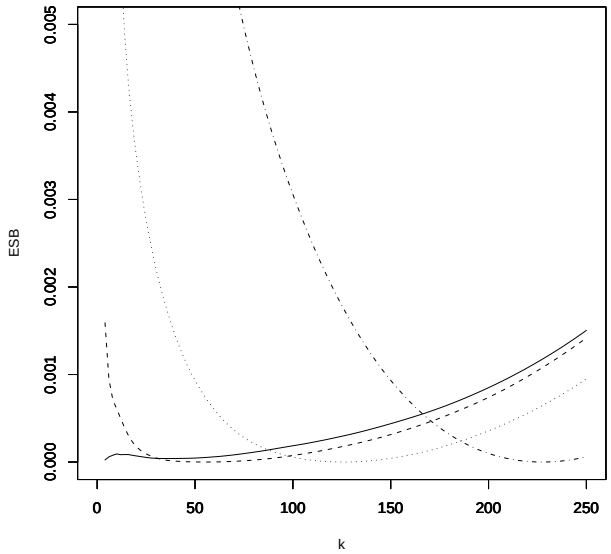

FIG. 2.3 – Biais au carré empirique ($\mathcal{ESB}$) de l'estimateur $\hat{\gamma}_n^{\pi}$ en fonction de k. Le paramètre de dépendance temporelle est égal à $\alpha = 1$ $(-)$, $\alpha = 0.8$ $(--)$, $\alpha = 0.5$ $(\ldots)$ et $\alpha = 0.2$ $(-.-)$.

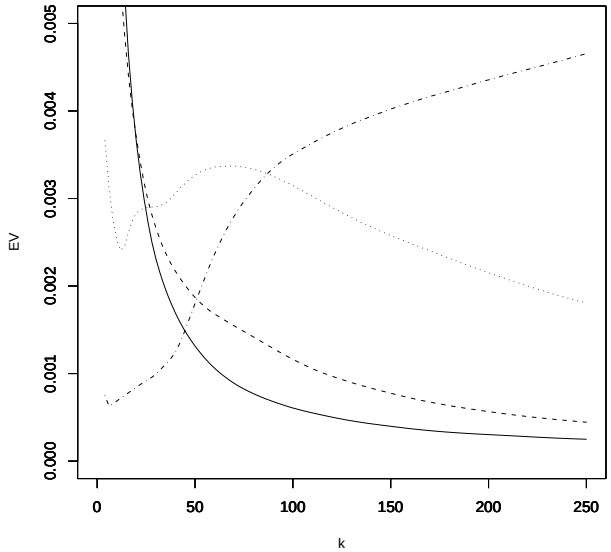

FIG. 2.4 – Variance empirique ($\mathcal{EV}$) de l'estimateur $\hat{\gamma}_n^{\pi}$ en fonction de k. Le paramètre de dépendance temporelle est égal à $\alpha = 1$ $(-)$, $\alpha = 0.8$ $(--)$, $\alpha = 0.5$ $(\ldots)$ et $\alpha = 0.2$ $(-.-)$.

Effet de la dépendance spatiale Nous nous intéressons à présent à l'influence de la dépendance spatiale. De la même façon que précédemment, nous simulons indépendamment $n_s = 10$ séries temporelles $\{s_1, \ldots, s_{n_s}\}$ de taille 500. Les marges sont ici de loi normale centrée et réduite. Le paramètre de dépendance temporelle est fixé à $\alpha = 0.5$. On peut par exemple interpréter n_s comme étant le nombre de stations de mesure. Nous disposons donc au total de $n = 5000$ observations. Nous introduisons de la dépendance spatiale en utilisant l'application linéaire suivante : $(s'_1, \ldots, s'_{n_s}) = (s_1, \ldots, s_{n_s})A(\theta)$ où $A(\theta)$ est une matrice circulante de dimension $n_s \times n_s$ définie pour tout $\theta \in [0, 1]$ et tout couple $(i, j) \in \{1, \ldots, n_s\}^2$ par :

$$A_{i,j}(\theta) = \begin{cases} 1/\delta & \text{si } \delta > (j - i) \text{ modulo } n_s, \\ 0 & \text{sinon} \end{cases}$$

avec $\delta = \lfloor n_s - (n_s - 1)\theta \rfloor$. Le paramètre θ contrôle la dépendance entre les séries temporelles $\{s'_1, \ldots, s'_{n_s}\}$. Plus précisément, chaque série temporelle s'_j est la moyenne de δ séries temporelles prises dans l'ensemble $\{s_1, \ldots, s_{n_s}\}$. Par exemple, si $\theta = 1$ alors $\delta = 1$ et $A(1)$ est la matrice identité. Les séries temporelles $s'_1, \ldots, s'_{n_s}$ sont indépendantes. Par contre, si $\theta = 0$ alors $\delta = n_s$ et $s'_1 = \ldots = s'_{n_s}$ qui correspond au cas de la dépendance complète. Un cas intermédiaire entre la dépendance complète et l'indépendance est par exemple obtenu en prenant $\theta = 4/5$ c'est à dire $\delta = 2$. Nous avons alors $s'_1 = (s_1 + s_2)/2$, $s'_2 = (s_2 + s_3)/2$, ... lorsque $n_s = 10$. Pour terminer, nous effectuons sur les séries temporelles s'_j, $j = 1, \ldots, n_s$ une transformation afin d'obtenir des marges de loi de Burr ayant un paramètre du second ordre $\rho = -1$ et un indice des valeurs extrêmes conditionnel $\gamma_j = 0.16 + j(0.26 - 0.16)/n_s$, $j = 1, \ldots, n_s$ ($\gamma_j \in]0.16, 0.26]$). Pour évaluer l'influence de la dépendance spatiale sur l'estimation de l'indice des valeurs extrêmes conditionnel, nous calculons (à partir de $N = 100$ réplications indépendantes des n_s séries temporelles) l'erreur moyenne quadratique empirique (définie par $\mathcal{EMSE} = \mathcal{EV} + \mathcal{ESB}$). Nous faisons varier le paramètre de dépendance spatiale θ dans l'ensemble $\{1, 0.8, 0.5, 0.2\}$. Le nombre d'observations sélectionnées $m_{n,t} = m$ varie dans l'ensemble $\{500, 500 \times 2, \ldots, 500 \times n_s\}$. Nous ne nous intéressons plus ici à l'influence du nombre $k_{n,t}$ de statistiques ordonnées utilisées pour construire l'estimateur de l'indice des valeurs extrêmes conditionnel. Nous prenons la valeur de $k_{n,t}$ minimisant l'erreur moyenne quadratique empirique $\mathcal{EMSE}$. Les résultats sont représentés sur la Figure 2.5. Le paramètre θ n'a visiblement que peu d'influence sur l'erreur moyenne quadratique de l'estimateur. Une étude plus précise montre qu'en fait la présence de dépendance spatiale augmente légèrement le biais mais diminue la variance de l'estimateur.

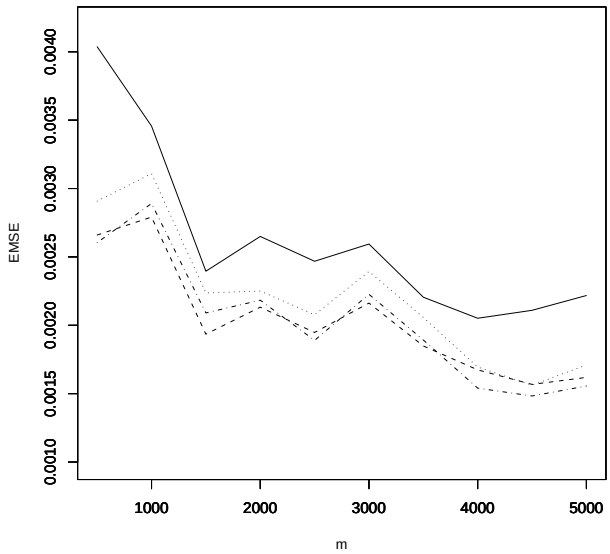

FIG. 2.5 – Erreur moyenne quadratique empirique ($\mathcal{EMSE}$) de l'estimateur $\hat{\gamma}_n^{\pi}$ en fonction du nombre m de points sélectionnés. Le nombre de séries temporelle est $n_s = 10$ avec un coefficient de dépendance temporelle $\alpha = 0.5$ et un coefficient de dépendance spatiale $\theta = 1$ $(-)$, $\theta = 0.8$ $(--)$, $\theta = 0.5$ $(\ldots)$ et $\theta = 0.2$ $(-.-)$.

Application aux données réelles Pour sélectionner les observations nous utilisons la méthode des plus proches voisins. En chaque point t de la région, nous conservons uniquement les $m_{n,t}$ observations associées aux covariables les plus proches de t. Pour choisir ce nombre $m_{n,t}$ ainsi que le nombre de plus grandes observations $k_{n,t}$ nous supposons dans un premier temps qu'ils ne dépendent pas de la covariable t. Nous proposons ensuite de minimiser une mesure de dissimilarité entre différents estimateurs de $\gamma(t)$. Plus précisément, nous prenons pour $m_{n,t}$ et $k_{n,t}$ les valeurs $\hat{m}$ et $\hat{k}$ définies par :

$$(\hat{k}, \hat{m}) = \arg\min_{k,m} \sum_{t \in S} \mathcal{DI}(\hat{\gamma}_n^H(t), \hat{\gamma}_n^Z(t), \hat{\gamma}_n^{\pi}(t)),$$

où S est l'ensemble des 142 stations de mesure et $\mathcal{DI}(u_1, u_2, u_3) = (u_1 - u_2)^2 + (u_2 - u_3)^2 + (u_3 - u_1)^2$. Cette méthode heuristique est utilisée notamment pour faire de l'estimation non paramétrique de densité. Nous trouvons ici $\hat{m} = 66000$ et $\hat{k} = 66$. Le résultat obtenu sur l'estimation de $\gamma(t)$ est représenté sur la Figure 2.6 en fonction de la latitude et de la longitude. La valeur estimée de $\gamma(t)$ varie entre 0.15 et 0.28 ce qui est en accord avec les résultats trouvés par Coles *et al.* [58]. La carte de l'estimation de la période de retour de 10 ans est donnée par la Figure 2.7. L'estimateur utilisé est $\hat{q}_2(\alpha, t, p(., a_\pi, \lambda_\pi))$ avec $\alpha = 1/(365.25 \times 24 \times 10)$. Les niveaux de retour semblent décroîtrent avec l'altitude. Ils sont plus importants dans la vallée que sur les plateaux. Cette remarque est cohérente avec les statistiques établient par Molinié *et al.* [89].

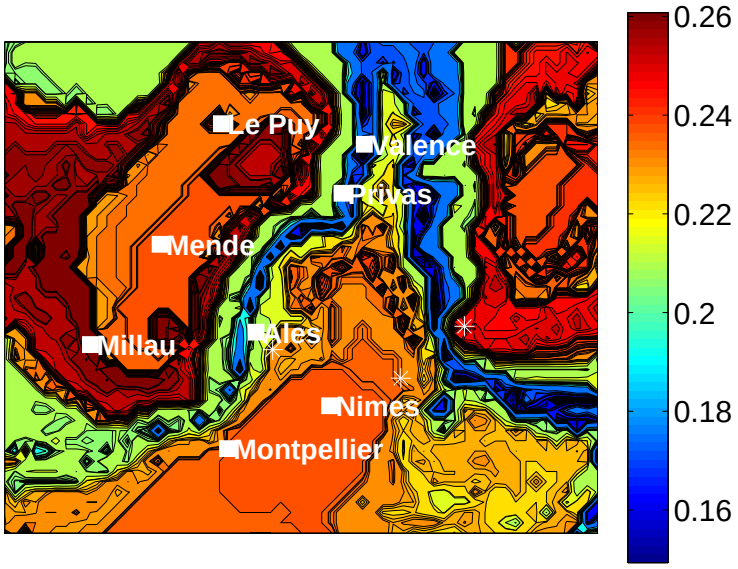

FIG. 2.6 – Estimation de l'indice des valeurs extrêmes conditionnel en fonction de la longitude et de la latitude.

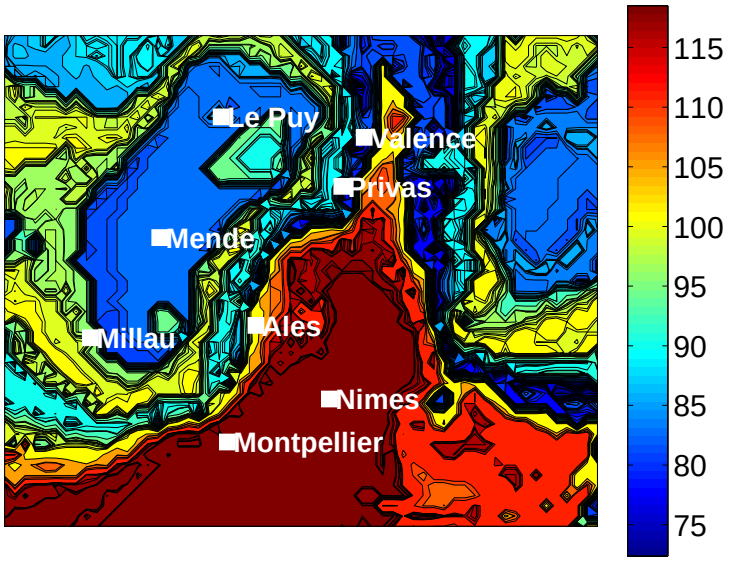

FIG. 2.7 – Niveaux de retour de 10 ans (en mm) en fonction de la longitude et de la latitude.

2.3 Cas d'une loi conditionnelle à queue lourde avec une covariable aléatoire

Nous nous intéressons dans ce paragraphe à l'estimation de quantiles extrêmes conditionnels lorsque la covariable est aléatoire. Plus précisément, nous disposons d'observations (X_i, Y_i), $i = 1, \ldots, n$ générées indépendamment à partir du vecteur aléatoire $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$. Nous souhaitons principalement proposer un estimateur ponctuel de la fonction $x \in \mathbb{R}^p \to q(\alpha_n, x)$ où

$$\bar{F}(q(\alpha_n, x), x) = \mathbb{P}(Y > q(\alpha_n, x)|X = x) = \alpha_n, \ \alpha_n \to 0.$$

L'estimation de quantiles classiques (c'est à dire lorsque l'ordre α ne dépend pas de la taille n de l'échantillon) a fait l'objet de nombreux travaux (voir par exemple Gannoun [74], Roussas [93], Stute [97] et Stone [96]). Le cas de quantiles extrêmes conditionnels n'a été considéré que récemment (voir Davison *et al.* [63]) et ceci malgré de nombreuses applications notamment en hydrologie. Comme dans le paragraphe précédent, nous considérons le cas où la loi conditionnelle de Y sachant X est à queue lourde. Autrement dit, la fonction de survie conditionnelle est donnée par :

$$\bar{F}(y, x) = y^{-1/\gamma(x)} L(y, x), \tag{2.9}$$

où pour tout $x \in \mathbb{R}^p$ fixé, $L(., x)$ est une fonction à variations lentes. La densité marginale de X est notée dans la suite $g(.)$. Pour tout $t \in \mathbb{R}^d$, nous proposons dans le paragraphe 2.3.1 un estimateur de *petites probabilités* c'est à dire des probabilités $\bar{F}(y_n, t)$ lorsque $y_n \to \infty$. Le paragraphe 2.3.2 est consacré à l'estimation de quantiles extrêmes conditionnels $q(\alpha_n, t)$. Ces estimateurs sont utilisés pour estimer l'indice des valeurs extrêmes conditionnel $\gamma(t)$ dans le paragraphe 2.3.3. Les résultats obtenus sont publiés dans la revue *Test* en collaboration avec A. Daouia, S. Girard et A. Lekina [62]. Pour établir la normalité asymptotique de ces estimateurs, nous introduisons les hypothèses suivantes :

(C.1) La fonction à variations lentes $L(., x)$ est normalisée.

Nous avons donc que

$$L(y, x) = c(x) \exp\left\{ \int_1^y \frac{\Delta(u, x)}{u} \right\},$$

où $\Delta(u, x) \to 0$ lorsque $u \to \infty$. Comme dans le paragraphe 2.2, la fonction $\Delta(., x)$ contrôle la vitesse de convergence du rapport $L(vy, x)/L(y, x)$ vers 1 ($v > 0$). Il est donc nécessaire de préciser la façon dont $\Delta(u, x)$ converge vers 0 lorsque $u \to \infty$. Ici nous ne supposons pas, comme dans la section 2.2, que la fonction est à variations régulières mais simplement qu'elle est asymptotiquement décroissante :

(C.2) La fonction $|\Delta(., x)|$ est asymptotiquement décroissante.

Les conditions suivantes sont des hypothèses de régularité sur les fonctions $\gamma(.)$, $L(y, .)$ et sur la densité marginale $g(.)$ de X. Soit $(x, x') \in \mathbb{R}^p \times \mathbb{R}^p$. On note par d la distance euclidienne de $\mathbb{R}^p$.

(D.1) Il existe une constante $c_\gamma > 0$ telle que

$$\left|\frac{1}{\gamma(x)} - \frac{1}{\gamma(x')}\right| \leq c_\gamma d(x, x').$$

(D.2) Il existe une constante $c_L > 0$ et $y_0 > 1$ tels que

$$\sup_{y \geq y_0} \left|\frac{\log L(y,x)}{\log y} - \frac{\log L(y,x')}{\log y}\right| \leq c_L d(x, x').$$

(D.3) Il existe une constante $c_g > 0$ telle que

$$|g(x) - g(x')| \leq c_g d(x, x').$$

2.3.1 Estimation de petites probabilités

Soit $t \in \mathbb{R}^p$. Nous proposons d'estimer la "petite probabilité" $\bar{F}(y_n, t)$ par un estimateur à noyau de la fonction de survie défini par :

$$\hat{\bar{F}}_n(y_n, t) = \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \mathbb{I}\{Y_i > y_n\} \Bigg/ \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right),$$

où $K(.)$ est une densité bornée sur $\mathbb{R}^p$ de support S inclus dans la boule unité de $\mathbb{R}^p$. La suite (h_n) est appelée suite de lissage. La normalité asymptotique de cet estimateur est établie dans le théorème ci-dessous (voir [62, Théorème 1]).

Théorème 2.5 *On se place sous le modèle (2.9) et on suppose les conditions* **(D.1)**, **(D.2)** *et* **(D.3)** *satisfaites. On pose :*

- $0 < a_1 < a_2 < \cdots < a_J$ *où* J *est un entier positif,*
- $y_n \to \infty$ *tel que* $nh_n^p \bar{F}(y_n, x) \to \infty$ *et* $nh_n^{p+2} \log^2(y_n) \bar{F}(y_n, x) \to 0$ *lorsque* $n \to \infty$,
- $y_{n,j} = a_j y_n$ *pour* $j = 1, \ldots, J$.

Pour tout $t \in \mathbb{R}^p$ *tel que* $g(t) > 0$, *le vecteur aléatoire*

$$\left\{\sqrt{nh_n^p \bar{F}(y_n, t)} \left(\frac{\hat{\bar{F}}_n(y_{n,j}, t)}{\bar{F}(y_{n,j}, t)} - 1\right)\right\}_{j=1,\ldots,J}$$

est asymptotiquement Gaussien de moyenne nulle et de matrice de variance-covariance $C(t)$ *avec, pour* $(j, j') \in \{1, \ldots, J\}^2$

$$C_{j,j'}(t) = \frac{\int_S K^2(u)du}{g(t)} a_{\min(j,j')}^{1/\gamma(t)}.$$

Un résultat similaire sur le comportement asymptotique de l'estimateur empirique de la fonction de survie a été obtenu par Einmahl [69] dans un cadre non conditionnel mais sans faire

d'hypothèses sur la distribution de l'échantillon. Nous montrons dans [62, Lemme 3] que la condition $nh_n^p \bar{F}(y_n, x) \to \infty$ est équivalente à

$$\mathbb{P}\big(\exists\, i \in \{1, \dots, n\} \text{ tel que } (X_i, Y_i) \in R_n(x)\big) \to 1 \text{ lorsque } n \to \infty,$$

avec $R_n(x) = B(x, h_n) \times]y_n, \infty[\subset \mathbb{R}^{p+1}$ où $B(x, h_n)$ est la boule de centre x et de rayon h_n. Ainsi cette condition signifie que la valeur y_n dont on souhaite estimer la probabilité de dépassement est située à l'intérieur de l'échantillon. Il n'y a donc pas d'extrapolation à faire et on peut ainsi utiliser la fonction de répartition empirique pour l'estimation.

2.3.2 Estimation de quantiles extrêmes conditionnels

Nous proposons ici un estimateur de quantiles extrêmes conditionnels. Pour ce faire, nous utilisons l'inverse généralisée de l'estimateur à noyau de la fonction de survie. Pour $\alpha \in]0, 1[$ et pour tout $t \in \mathbb{R}^p$, nous estimons le quantile conditionnel d'ordre α par :

$$\hat{q}_n(\alpha, t) = \hat{\bar{F}}_n^{\leftarrow}(\alpha, t) = \inf\{s,\ \hat{\bar{F}}_n(s, t) \leq \alpha\}.$$

La loi asymptotique de cet estimateur est donnée ci-dessous (voir [62, Théorème 2])

Théorème 2.6 *On se place sous le modèle (2.9) et on suppose les conditions* **(C.1)**, **(D.1)**, **(D.2)** *et* **(D.3)** *satisfaites. On pose :*

- $1 = \tau_1 > \tau_2 > \cdots > \tau_J > 0$ *où* J *est un entier positif,*
- $\alpha_n \to 0$ *tel que* $nh_n^p \alpha_n \to \infty$ *et* $nh_n^{p+2} \alpha_n \log^2(\alpha_n) \to 0$ *lorsque* $n \to \infty$,
- $\alpha_{n,j} = \tau_j \alpha_n$ *pour* $j = 1, \dots, J$.

Pour tout $t \in \mathbb{R}^p$ *tel que* $g(t) > 0$, *le vecteur aléatoire*

$$\left\{ \sqrt{nh_n^p \alpha_n} \left(\frac{\hat{q}_n(\alpha_{n,j}, t)}{q(\alpha_{n,j}, t)} - 1 \right) \right\}_{j=1,\dots,J}$$

est asymptotiquement Gaussien de moyenne nulle et de matrice de variance-covariance $\Sigma(t)$ *avec, pour* $(j, j') \in \{1, \dots, J\}^2$

$$\Sigma_{j,j'}(t) = \frac{\gamma^2(t) \int_S K^2(u) du}{g(t)} \tau_{\min(j,j')}^{-1}.$$

Contrairement à l'estimation de petites probabilités, l'indice des valeurs extrêmes conditionnel joue un rôle dans l'estimation des quantiles extrêmes conditionnels. En effet, une grande valeur de $\gamma(x)$ conduit à une variance importante pour l'estimateur $\hat{q}_n(\alpha_n, x)$. Ainsi, l'estimation de $q(\alpha_n, x)$ est d'autant plus difficile que la queue de la distribution est lourde. Nous montrons dans [62, Remarque 2] que si $n\alpha_n \log^2(\alpha_n) \to \infty$ alors la suite

$$h_n = \eta_n (n\alpha_n \log^2(\alpha_n))^{-1/(p+2)},$$

où (η_n) est une suite tendant vers zéro aussi lentement que l'on veut, satisfait les hypothèses du Théorème 2.6. Ce choix conduit à une vitesse de convergence proportionnelle à :

$$\eta_n^{p/2}\left(\frac{n\alpha_n}{\log^p(\alpha_n)}\right)^{1/(p+2)}.$$

Comme pour l'estimation de petites probabilités, la condition $nh_n^p\alpha_n \to \infty$ implique que les quantiles extrêmes conditionnels que l'on peut estimer sont à l'intérieur de l'échantillon. On ne peut donc pas faire d'extrapolation au delà du maximum des observations dans la boule $B(t, h_n)$. Pour pouvoir estimer des quantiles extrêmes d'ordre $\beta_n < \alpha_n$ où (β_n) est une suite aussi petite que l'on veut, nous proposons d'utiliser un estimateur adapté de celui proposé par Weissman [98]. Il est défini par :

$$\hat{q}_n^W(\beta_n, t) = \hat{q}_n(\alpha_n, t)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(t)},$$

où $\hat{\gamma}_n(t)$ est un estimateur de l'indice des valeurs extrêmes. La normalité asymptotique de $\hat{q}_n^W$ est donnée ci-dessous (voir [62, Théorème 3])

Théorème 2.7 *On se place sous le modèle (2.9) et on suppose les conditions* **(C.1)**, **(D.1)**, **(D.2)** *et* **(D.3)** *satisfaites. S'il existe une suite (σ_n) telle que $\sigma_n \to 0$, $\sigma_n^{-1}h_n \log\alpha_n \to 0$ et $\sigma_n^{-1}(\hat{\gamma}_n(t) - \gamma(t)) \xrightarrow{d} \mathcal{N}(0, v^2(t))$ avec $v^2(t) > 0$ alors, pour toutes suites (α_n) et (β_n) telles que $\alpha_n \to 0$, $\beta_n/\alpha_n \to 0$,*

$$\frac{\sigma_n^{-1}}{\sqrt{nh_n^p\alpha_n}\log(\alpha_n/\beta_n)} \to 0 \text{ et } \sigma_n^{-1}\Delta(q(\alpha_n, t), t) \to 0,$$

on a pour tout $x \in \mathbb{R}^p$,

$$\frac{\sigma_n^{-1}}{\log(\alpha_n/\beta_n)}\left(\frac{\hat{q}_n^W(\beta_n, t)}{q(\beta_n, t)} - 1\right) \xrightarrow{d} \mathcal{N}(0, v^2(t)).$$

2.3.3 Application à l'estimation de l'indice des valeurs extrêmes conditionnel

Nous utilisons le résultat du Théorème 2.6 pour proposer deux estimateurs de l'indice des valeurs extrêmes conditionnel. Le premier est une adaptation de l'estimateur proposé par Pickands [90]. Il est définit pour $t \in \mathbb{R}^p$ par :

$$\hat{\gamma}_n^P(t) = \frac{1}{\log 2}\log\left(\frac{\hat{q}_n(\alpha_n, t) - \hat{q}_n(2\alpha_n, t)}{\hat{q}_n(2\alpha_n, t) - \hat{q}_n(4\alpha_n, t)}\right),$$

où $\alpha_n \to 0$ lorsque $n \to \infty$. Notre deuxième estimateur est inspiré de l'estimateur proposé par Hill [84] :

$$\hat{\gamma}_n^H(t) = \sum_{j=1}^{J}\left[\log\hat{q}_n(\tau_j\alpha_n, t) - \log\hat{q}_n(\alpha_n, t)\right] \Bigg/ \sum_{j=1}^{J}\log(1/\tau_j),$$

où $\tau_1 > \tau_2 > \cdots > \tau_J > 0$. La normalité asymptotique de ces deux estimateurs est une conséquence du Théorème 2.6 (voir [62, Corollaires 1 et 2]).

Théorème 2.8 *Sous les hypothèses du Théorème 2.6, si la condition* **(C.2)** *est satisfaite et si*

$$\sqrt{nh_n^p\alpha_n}\max(\Delta(q(\alpha_n,t),t),\Delta(q(2\alpha_n,t),t)),$$

alors, pour tout $t \in \mathbb{R}^p$ *tel que* $g(t) > 0$, $\sqrt{nh_n^p\alpha_n}(\hat{\gamma}_n^P(t) - \gamma(t))$ *et* $\sqrt{nh_n^p\alpha_n}(\hat{\gamma}_n^H(t) - \gamma(t))$ *convergent vers une loi normale centrée de variance respective :*

$$\frac{\gamma^2(t)\int_S K^2(u)du}{g(t)}V_P(\gamma(t)) \quad \textit{et} \quad \frac{\gamma^2(t)\int_S K^2(u)du}{g(t)}V_H(J)$$

avec

$$V_P(\gamma(t)) = \frac{2^{2\gamma(t)+1}+1}{4(\log 2)^2(2^{\gamma(t)}-1)^2} \quad \textit{et} \quad V_H(J) = \left(\sum_{j=1}^{J}\frac{2(J-j)+1}{\tau_j} - J^2\right)\Bigg/\left(\sum_{j=1}^{J}\log(1/\tau_j)\right)^2.$$

La variance asymptotique de $\hat{\gamma}_n^H(t)$ dépend du choix des τ_j, $j = 1,\ldots,J$. En prenant par exemple $\tau_j = 1/j$ pour $j = 1,\ldots,J$, nous avons $V_H(J) = J(J-1)(2J-1)/(6\log^2(J!))$. Cette fonction de J est convexe et admet un unique minimum pour $J = 9$ correspondant à une valeur $V_H(J) \approx 1.25$.

2.4 Estimation de support

Dans ce dernier paragraphe, nous nous intéressons à l'estimation du support $\mathcal{S}$ de la loi d'un couple (X,Y) de la forme

$$\mathcal{S} = \{(x,y) \in \mathbb{R}^2 | 0 \le x \le 1 \text{ et } 0 \le y \le \xi(x)\}, \tag{2.10}$$

où $\xi(.)$ est une fonction inconnue et continue appelée *frontière du support*. Il est clair qu'ici l'estimation du support $\mathcal{S}$ est équivalente à l'estimation de la fonction $\xi(.)$. L'estimation de la fonction $\xi(.)$ est un cas particulier d'estimation de quantile extrême conditionnel. Elle est en effet équivalente à l'estimation du quantile extrême d'ordre $\alpha = 0$ lorsque la loi conditionnelle de Y sachant X est à support borné. On considère généralement que les premiers travaux relatifs à l'estimation de support sont dûs à Geffroy [80] et Rényi et Sulanke [91, 92]. Geffroy propose d'estimer la frontière du support avec un simple histogramme basé sur les plus grandes observations. Cet estimateur a été étudié et amélioré notamment par Chevalier [57], Bosq [56], et Jacob [85]. D'autres estimateurs de $\mathcal{S}$ ont été proposés par Deprins *et al.* [68] dans le cas où la fonction $\xi(.)$ est croissante et/ou concave. Une fonction polynomiale par morceaux est utilisée par Korostelev *et al.* [86] pour estimer la fonction support.

Cete section est un bref résumé des résultats sur l'estimation de $\xi(.)$ obtenus dans le cadre de ma thèse [76]. Pour effectuer cette estimation, nous disposons d'observations (X_i,Y_i), $i = 1,\ldots$ réparties dans l'ensemble $\mathcal{S}$. Nous avons considéré dans [76] deux situations.

Support d'un processus ponctuel de Poisson Dans la première situation, nous supposons que les observations sont issues de la superposition de n copies indépendantes d'un processus ponctuel de Poisson d'intensité constante. Ce travail a fait l'objet d'une publication dans les *Annales de l'Institut de l'Université de Paris* [75]. Une méthode classique pour estimer la fonction $\xi(.)$ consiste à tronquer son développement dans une base et d'en estimer les coefficients. Girard *et al.* [81, 82] proposent ainsi des estimateurs de la fonction support en utilisant les bases orthogonales de Harr et trigonométrique. Ils estiment ensuite les coefficients en utilisant des valeurs extrêmes. Nous avons utilisé un méthode d'estimation similaire en prenant la base de Faber-Schauder. Cette base, qui n'est pas une base Hilbertienne, est obtenue en prenant les primitives des fonctions définissant la base de Haar. L'utilisation de cette base a l'avantage de fournir un estimateur continu de la fonction $\xi(.)$ (au contraire de l'estimateur obtenu avec la base de Haar) et de ne plus avoir à supposer que $\xi(0) = \xi(1)$ (ce qui est le cas pour la base trigonométrique). La construction de notre estimateur s'effectue en deux étapes. Dans un premier temps, nous approchons la fonction $\xi(.)$ en tronquant à l'ordre r_n son développement dans la base de Faber-Schauder. Dans la deuxième étape, nous divisons le support $\mathcal{S}$ en s_n cellules et nous estimons les r_n premiers coefficients de la base en utilisant les s_n maxima des secondes coordonnées du processus sur ces cellules. Nous avons établi la convergence uniforme presque complète de l'estimateur ainsi obtenu (voir [75, Théorème 2]). Dans le cas où le rapport r_n/s_n reste constant, nous avons montré dans [75, Théorème 3] que la loi asymptotique de l'estimateur est une loi des valeurs extrêmes. La normalité asymptotique a été obtenue en supposant que le rapport r_n/s_n tendait vers zéro (voir [75, Théorème 6]). Notons que des résultats similaires ont été obtenus par Menneteau [88] dans une cadre plus général où les observations sont issues soit d'un processus ponctuel de Poisson, soit d'un échantillon uniformément distribué sur le support.

Estimation d'une fonction quantile extrême Dans la seconde situation, nous considérons le cas d'observations provenant d'un échantillon de n variables aléatoires indépendantes et identiquement distribuées. Nous supposons de plus que la loi conditionnelle de Y sachant $X = x$ appartient au domaine d'attraction de Weibull. Cette situation est évidemment moins favorable que celle du processus ponctuel de Poisson car les observations peuvent éventuellement être éloignées de la fonction support à estimer. En effet, dans le cas d'un processus ponctuel de Poisson d'intensité constante, la loi conditionnelle de Y sachant X est assimilable à une loi uniforme. Pour tout point $t \in [0, 1]$, nous nous sommes intéressés à l'estimation du quantile extrême conditionnel $\xi_{\alpha_n}(t)$ vérifiant :

$$\mathbb{P}(Y \leq \xi_{\alpha_n}(t)|X = t) = \alpha_n,$$

où $\alpha_n \to 0$. Evidemment, si $\alpha_n = 0$, $\xi_0(t) = \xi(t)$. Pour ce faire nous avons considéré les observations

$$Z_i = Y_i \mathbb{I}\{|X_i - t| \leq h_n\},\ i = 1, \ldots, n,$$

où (h_n) est une suite positive tendant vers zéro lorsque $n \to \infty$. Nous avons ensuite estimé la quantité $\xi_{\alpha_n}(t)$ en reprenant l'estimateur de quantile extrême défini dans le Chapitre 1, paragraphe 1.4.1 et en le calculant avec les observations Z_i, $i = 1, \ldots, n$. Nous avons montré dans [76, Théorème 4.5] que l'estimateur ainsi obtenu converge en probabilité vers $\xi_{\alpha_n}(t)$.

2.5 Conclusion et perspectives

Dans ce chapitre, nous nous sommes intéressés à l'estimation de quantiles extrêmes conditionnels. Ces quantiles sont des fonctions d'une covariable mesurée conjointement avec la variable d'intérêt. Nous avons considéré deux situations. Dans la première, nous supposons que la covariable est déterministe et pouvant être éventuellement de dimension infinie. Les estimateurs obtenus dans ce cadre ont été validés théoriquement par des résultats de convergence en loi mais aussi par une application à l'estimation d'une carte de niveaux de retour. Dans la deuxième situation, la covariable est aléatoire et nous nous sommes restreints au cas où elle appartient à un espace de dimension finie. Des résultats de normalité asymptotique ont également été obtenus sur les estimateurs. Enfin, un bref rappel des résultats sur l'estimation de support obtenus dans le cadre de ma thèse est présenté en fin de chapitre. Ces résultats sont en fait les premiers obtenus sur l'estimation de quantiles extrêmes conditionnels dans le cas où la loi conditionnelle appartient au domaine d'attraction de Weibull.

Les extensions possibles sur le thème de l'estimation de quantiles extrêmes conditionnels sont nombreuses. Dans un premier temps, nous souhaiterions étendre les résultats obtenus au cas d'une covariable aléatoire lorsque cette dernière est de dimension infinie.
Nous envisageons aussi (avec A. Daouia et S. Girard) d'étudier l'estimation d'un quantile extrême d'ordre α lorsque la covariable est aléatoire et la loi conditionnelle de Y sachant $X = x$ est à support borné. Dans ce cas, estimer le quantile extrême conditionnel d'ordre $\alpha = 0$ revient à estimer le support de la loi du couple (X, Y). Ces travaux seraient le prolongement direct de ceux effectués dans ma thèse. Un autre axe de recherche proche est l'estimation d'une frontière de production (voir Daouia *et al.* [61] pour plus de détails). Il s'agit en fait d'estimer les quantiles extrêmes de la loi conditionnelle de Y sachant que $X \leq x$ (et non $X = x$).
Lorsque la covariable est aléatoire, nous avons proposé des estimateurs de quantiles extrêmes basés sur l'estimateur empirique de la fonction de répartition conditionnelle. Pour l'instant une seule fonction noyau est utilisée pour lisser en la covariable et nous envisageons d'en utiliser une seconde pour lisser aussi en la variable d'intérêt. Ceci permettrait d'obtenir des estimateurs moins sensibles aux fluctuations des plus grandes observations.
A plus long terme, nous souhaiterions prouver des résultats de convergence uniforme sur les estimateurs de $\gamma(.)$ et $q(\alpha, .)$ en se basant par exemple sur des travaux de Einmahl *et al.* [70]. Une méthode pour choisir théoriquement les paramètres $h_{n,t}$ et $k_{n,t}$ reste aussi à définir. L'étude du comportement asymptotique de nos estimateurs lorsque les données ne sont plus supposées indépendantes doit aussi être effectuée.

Bibliographie du Chapitre 2

[52] J. Beirlant, G. Dierckx, A. Guillou, and C. Stărică. On exponential representations of log-spacings of extreme order statistics. *Extremes*, **5**, 157–180 (2002).

[53] J. Beirlant and Y. Goegebeur. Regression with response distributions of pareto type. *Computational Statistics and Data Analysis*, **42**, 595–619 (2003).

[54] J. Beirlant and Y. Goegebeur. Local polynomial maximum likelihood estimation for pareto-type distributions. *Journal of Multivariate Analysis*, **89**, 97–118 (2004).

[55] P. Bois, C. Obled, M. de Saintignon, and H. Mailloux. *Atlas expérimental des risques de pluies intenses : Cévennes-Vivarias.* Pôle grenoblois détudes et de recherche pour la prévention des risques naturels, 2nd Edition (1997).

[56] D. Bosq. Contribution à la théorie de l'estimation fonctionnelle. *Publications de l'Institut de Statistique de l'Université de Paris*, **XIX(2)**, 1–96 (1977).

[57] J. Chevalier. Estimation du support et du contenu du support d'une loi de probabilité. *Annales de l'Institut Henri Poincaré. Section B*, **12(4)**, 339–364 (1976).

[58] S. Coles and J. Tawn. A bayesian analysis of extreme rainfall data. *Journal of Applied Statistics*, **45**, 463–478 (1996).

[59] P. Consul and G. Jain. On the log-gamma distribution and its properties. *Statistische Hefe*, **12(2)**, 100–106 (1971).

[60] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, **102**, 824–840 (2007).

[61] A. Daouia, J. Florens, and L. Simar. Frontier estimation and extreme values theory. *Bernoulli* (2010). To appear.

[62] A. Daouia, L. Gardes, S. Girard, and A. Lekina. Kernel estimators of extreme level curves (2010). à paraître.

[63] A. Davison and N. Ramesh. Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, Series B*, **62**, 191–208 (2000).

[64] A. Davison and R. Smith. Models for exceedances over high threshols. *Journal of the Royal Statistical Society, Series B*, **52**, 393–442 (1990).

[65] L. de Haan. *Slow variation and characterization of domains of attraction.* Statistical Extremes and Application. Reidel, Dorchrecht, j. tiago de oliveira edition (1984).

[66] L. de Haan and H. Rootzén. On the estimation of high quantiles. *Journal of Statistical Planning and Inference*, **35**, 1–13 (1993).

[67] A. Dekkers and L. de Haan. On the estimation of the extreme value index and large quantile estimation. *The Annals of Statistics*, **17**, 1795–1832 (1989).

[68] D. Deprins, L. Simar, and H. Tulkens. Measuring labor efficiency in post offices. In M. Marchant, P. Pestiau, and H. Tulkens, editors, *The Performance of Public Enterprises : Concepts and Measurements*, pages 243–267. North-Holland, Amsterdam (1984).

[69] J. Einmahl. The empirical distribution function as a tail estimator. *Statistica Neerlandica*, **44**, 79–82 (1990).

[70] J. Einmahl and T. Lin. Asymptotic normality of extreme value estimators on C[0,1]. *The Annals of Statistics*, **34(1)**, 469–492 (2006).

[71] L. Fawcett and D. Walshaw. Improved estimation for temporally clustered extremes. *Environmetrics*, **18**, 173–188 (2007).

[72] F. Ferraty and P. Vieu. *Non parametric functional data analysis : Theory and practice*. Springer Series in statistics (2006).

[73] A. Feuerverger and P. Hall. Estimating a tail exponent by modelling departure from a Pareto distribution. *The Annals of Statistics*, **27**, 760–781 (1999).

[74] A. Gannoun. Estimation non paramétrique de la médiane conditionnelle, médianogramme et méthode du noyau. *Publications de l'Institut de Statistique de l'Université de Paris*, **XXXXVI**, 11–22 (1990).

[75] L. Gardes. Estimating the support of a poisson process via the faber-schauder basis and extreme values. *Annales de l'Institut de Statistique de l'Université de Paris*, **XXXXVI**, 43–72 (2002).

[76] L. Gardes. *Estimation d'une fonction quantile extrême*. Ph.D. thesis, Université Montpellier II (2003).

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

[78] L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions : an application to the estimation of extreme rainfall return levels. *Extremes*, **13(2)**, 177–204 (2010).

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

[80] J. Geffroy. Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, **XIII**, 191–200 (1964).

[81] S. Girard and P. Jacob. Extreme value and haar series estimates of point processes boundaries. *Scandinavian Journal of Statistics*, **30(2)**, 369–384 (2003).

[82] S. Girard and P. Jacob. Projection estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, **116(1)**, 1–15 (2003).

[83] P. Hall and N. Tajvidi. Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, **15**, 153–167 (2000).

[84] B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**, 1163–1174 (1975).

[85] P. Jacob. Estimation du contour discontinu d'un processus ponctuel sur le plan. *Publications de l'Institut de Statistique de l'Université de Paris*, **XXIX**, 1–25 (1984).

[86] A. Korostelev and A. Tsybakov. *Minimax Theory of Image Reconstruction.* Springer-Verlag, New-York / Berlin, lecture notes in statistics edition (1993).

[87] M. Kratz and S. Resnick. The qq-estimator and heavy tails. *Stochastic Models*, **12**, 699–724 (1996).

[88] L. Menneteau. Multidimensional limit theorems for smoothed extreme value estimates of point processes boundaries. *ESAIM : Probability and Statistics*, **12**, 273–307 (2008).

[89] G. Molinié, E. Yates, D. Ceresetti, S. Anquetin, B. Boudevillain, J. Creutin, and P. Bois. Rainfall regimes in a mountainous mediterranean region : statistical analysis at short time steps. Technical report (2010).

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

[91] A. Rényi and R. Sulanke. Über die konvexe hülle von n zufällig gewählten punkten. *Z. Wahrsch. Verw. gebiete*, **2**, 75–84 (1963).

[92] A. Rényi and R. Sulanke. Über die konvexe hülle von n zufällig gewählten punkten, ii. *Z. Wahrsch. Verw. gebiete*, **3**, 138–147 (1964).

[93] G. Roussas. Nonparametric estimation of the transition distribution function of a markov process. *The Annals of Mathematical Statistics*, **40**, 1386–1400 (1969).

[94] J. Schultze and J. Steinebach. On least squares estimates of an exponential tail coefficient. *Statistics and Decisions*, **14**, 353–372 (1996).

[95] R. Smith. Extreme value analysis of environmental time series : An application to trend detection in ground level ozone (with discussion). *Statistical Science*, **4**, 367–393 (1989).

[96] C. Stone. Consistent nonparametric regression (with discussion). *The Annals of Statistics*, **5**, 595–645 (1977).

[97] W. Stute. Conditional empirical processes. *The Annals of Statistics*, **14**, 638–647 (1986).

[98] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, **73**, 812–815 (1978).

Chapitre 3

Réduction de dimension pour la régression

3.1 Introduction et motivations

Les résultats présentés dans ce chapitre ont été obtenus dans le cadre d'un projet financé pour la période 2008-2011 par le programme ANR, Masse de Données et COnnaissances (MDCO). Ce projet intitulé "Visualisation et analyse d'images hyperspectrales multidimensionnelles en astrophysique" (VAHINE) a pour objectif l'étude des données hyperspectrales récoltées par l'instrument OMEGA (Observatoire pour la Minéralogie, l'Eau, la Glace et l'Activité) embarqué à bord de la mission Mars Express en orbite autour de la planète Mars depuis 2003. Plus précisément, cet instrument mesure, en plusieurs points (pixel) d'une surface donnée de Mars l'intensité lumineuse réfléchie par les différents matériaux présents sur le sol de la planète avec une résolution spatiale d'environ 2km par pixel. On dispose ainsi, pour tous les pixels de l'image de Mars étudiée, de spectres c'est à dire de vecteurs de taille $d \approx 184$, chaque dimension représentant une longueur d'onde (voir Figure 3.1).

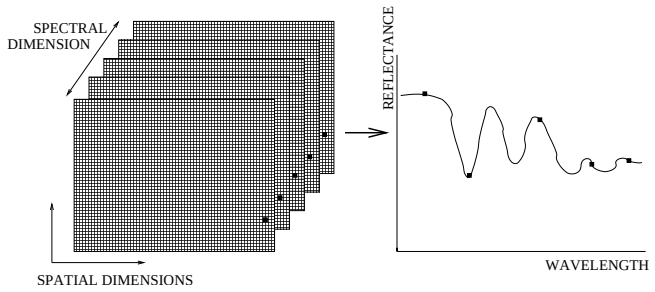

FIG. 3.1 – Schématisation d'une image hyperspectrale (source : BRGM).

Un des objectifs du projet est de retrouver, à partir de ces spectres, la structure physique de la planète (comme par exemple sa teneur en glace de dioxide de carbone (CO_2), en glace d'eau, en poussière, etc ...). Il s'agit donc de déterminer la relation liant le spectre à différents paramètres physiques. Pour ce faire, le Laboratoire de Planétologie de Grenoble (LPG) nous a fourni une base d'apprentissage. Cette dernière a été obtenue en utilisant un modèle de transfert radiatif simulant la propagation de la lumière sur un matériau dont on connaît les

paramètres physiques (voir Douté *et al.* [114]). Cette méthode permet donc, pour un jeu de paramètres donné, de connaître le spectre associé. Pour ce projet, nous nous sommes intéressés uniquement au pôle sud de Mars qui est principalement composé de glace d'eau, de glace de CO_2 et de poussière. En utilisant un modèle de transfert radiatif adapté à cette région, le LPG a ainsi mis à notre disposition un échantillon d'apprentissage $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$ où $X_i \in \mathbb{R}^d$ est le spectre associé au paramètre physique Y_i que l'on souhaite étudier. En pratique, on s'intéresse à 5 paramètres physiques : la proportion d'eau, la proportion de CO_2, la proportion de poussière, la taille des grains d'eau et enfin la taille des grains de CO_2. La taille n de l'échantillon d'apprentissage est d'environ 30000 spectres. Il s'agit donc d'estimer à partir de ces données la relation liant les spectres aux valeurs des paramètres. Ainsi, pour un nouveau spectre provenant de cette région de Mars, nous serons en mesure de déterminer les valeurs correspondantes de ses paramètres physiques. Plus formellement, on s'intéresse dans ce chapitre au problème de régression :

$$Y_i = g(X_i, \eta_i), \ i = 1, \ldots, n,$$

où $g : \mathbb{R}^{d+1} \to \mathbb{R}$ est la fonction de lien inconnue à estimer et η_i, $i = 1, \ldots, n$ est un terme d'erreur aléatoire. Pour résoudre ce problème de régression, on recense dans la littérature plusieurs méthodes parmi lesquelles :

La méthode des k plus proches voisins. Cette méthode est souvent utilisée par les physiciens pour l'étude d'images planétaires (voir par exemple Weiss *et al.* [129] et Carlson *et al.* [106]). Pour un vecteur $x \in \mathbb{R}^d$, on cherche dans une base de données simulées (X_i, Y_i), $i = 1, \ldots, n$ les k vecteurs les plus proches selon une certaine distance (par exemple la distance euclidienne). On estime alors la variable d'intérêt y associée au vecteur x par la moyenne (ou la médiane) des variables d'intérêt associées aux k plus proches voisins. Cette méthode est très simple à mettre en œuvre. Elle est cependant instable lorsque l'on travaille en grande dimension : une faible différence sur deux spectres peut correspondre à une erreur importante sur les paramètres physiques.

La méthode SVR (Support Vector Regression). Elle consiste à chercher dans la famille de fonctions

$$\mathcal{G} = \left\{ g : \mathbb{R}^d \to \mathbb{R} \text{ telles que } g(x) = \sum_{i=1}^{n} \alpha_i K(x, X_i) + b \right\},$$

(où $K(.,.)$ est une fonction noyau et b, α_i, $i = 1, \ldots, n$ sont des paramètres réels inconnus) celle minimisant le critère

$$\frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i) + \lambda \|f\|^2 \quad \text{avec} \quad \|f\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_j, X_i),$$

et

$$l(f(x), y) = \begin{cases} 0 & \text{si } |f(x) - y| \leq \varepsilon, \\ |f(x) - y| - \varepsilon & \text{sinon.} \end{cases}$$

Le paramètre ε contrôle le nombre de vecteurs supports c'est à dire le nombre de vecteurs X_i utilisés pour estimer la fonction de lien. Plus ε est grand, moins il y aura de vecteurs supports.

Le paramètre λ est un paramètre de régularité de la fonction de lien estimée. Elle sera d'autant plus régulière que le paramètre λ sera grand. Cette méthode est bien adaptée aux données de grande dimension mais les résultats sont difficilement interprétables et le temps de calcul assez important comme le font remarquer Durbha *et al.* [116]. Pour plus de détails sur la méthode SVR, voir par exemple Hastie *et al.* [120, Chapitre 12] et Cristianini *et al.* [113].

La régression PLS (Partial Least Square). Cette méthode est basée sur l'hypothèse que la fonction de lien est linéaire. Autrement dit, on considère un modèle de la forme :

$$Y_i = c_1 + c_2^t X_i + \eta_i,\ i = 1, \ldots, n,$$

où $c_1 \in \mathbb{R}$, $c_2 \in \mathbb{R}^p$ sont des paramètres inconnus et η_i, $i = 1, \ldots, n$ est un terme d'erreur aléatoire. Afin de pouvoir travailler avec des données de grande dimension, la régression PLS propose de remplacer les variables X_i par leur projection sur un sous-espace de plus petite dimension. Autrement dit, il s'agit de trouver une matrice A de dimension $m \times d$ où $m < d$ de telle sorte que les vecteurs $AX_i \in \mathbb{R}^m$ expliquent au mieux les variables Y_i. Nous donnons ci-dessous l'algorithme permettant de calculer la matrice A. On utilise les notations suivantes : soient u et v deux vecteurs de même dimension que l'on suppose centrés (*i.e.* dont la somme des composantes est nulle), on pose

$$\widehat{\mathrm{Cov}}(u, v) = u^t v \quad \text{et} \quad \widehat{\mathrm{Var}}(u) = u^t u.$$

Etape 0 : on centre et on réduit les variables $X_i = (X_{i,1}, \ldots, X_{i,d})^t$ et Y_i, $i = 1, \ldots, n$. On dénote les nouvelles variables ainsi obtenues par $\tilde{X}_i = (\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,d})^t$ et $\tilde{Y}_i$, $i = 1, \ldots, n$. On pose $\tilde{\mathcal{X}}_j = (\tilde{X}_{1,j}, \ldots, \tilde{X}_{n,j})^t$, $j = 1, \ldots, d$ et $\tilde{\mathcal{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^t$.

Etape 1 : initialisation : $l = 0$, $\tilde{\mathcal{X}}_{j,(0)} = \tilde{\mathcal{X}}_j$, $j = 1, \ldots, d$ et $\tilde{\mathcal{Y}}_{(0)} = \tilde{\mathcal{Y}}$.

Etape 2 :
i) $l = l + 1$.
ii) On pose $\mathbb{X}_{(l-1)}$ la matrice de dimension $n \times d$ dont les colonnes sont les vecteurs $\tilde{\mathcal{X}}_{j,(l-1)}$, $j = 1, \ldots, d$. On calcule ensuite $\hat{\beta}_{(l)} = \arg\max_\beta \widehat{\mathrm{Cov}}^2(\tilde{\mathcal{Y}}_{(l-1)}, \mathbb{X}_{(l-1)}\beta)$ sous la contrainte $\|\beta\|_2 = 1$. On pose $A_{(l)} = \mathbb{X}_{(l-1)}\hat{\beta}_{(l)}$.
iii) On calcule les résidus $\tilde{\mathcal{X}}_{(j,l)}$ $(j = 1, \ldots, d)$ et $\tilde{\mathcal{Y}}_{(l)}$ des régressions de $\tilde{\mathcal{X}}_{j,(l-1)}$ sur $A_{(l)}$ $(j = 1, \ldots, d)$ et de $\tilde{\mathcal{Y}}_{(l-1)}$ sur $A_{(l)}$, *i.e.*

$$\tilde{\mathcal{X}}_{(j,l)} = \tilde{\mathcal{X}}_{j,(l-1)} - A_{(l)} \frac{\widehat{\mathrm{Cov}}(\tilde{\mathcal{X}}_{j,(l-1)}, A_{(l)})}{\widehat{\mathrm{Var}}(A_{(l)})},$$

$$\tilde{\mathcal{Y}}_{(l)} = \tilde{\mathcal{Y}}_{(l-1)} - A_{(l)} \frac{\widehat{\mathrm{Cov}}(\tilde{\mathcal{Y}}_{(l-1)}, A_{(l)})}{\widehat{\mathrm{Var}}(A_{(l)})}.$$

Etape 3 : on répète l'étape 2 jusqu'à ce que $l = m$. La matrice A est la matrice dont les colonnes sont les vecteurs $A_{(l)}$, $l = 1, \ldots, m$.

La dimension m du sous-espace peut-être obtenue par validation croisée. La régression PLS est facile à utiliser et bien adaptée aux données de grande dimension. Elle présente cependant l'inconvénient de supposer l'existence d'une fonction de lien linéaire. Pour plus de détails sur la régression PLS, voir Hastie *et al.* [120, Chapitre 3].

La méthode SIR (Sliced Inverse Regression). Cette méthode a été proposée par Li [121]. Une présentation générale en est faite par Saracco *et al.* [125]. La méthode SIR est basée sur l'hypothèse suivante : les observations (X_i, Y_i), $i = 1, \ldots, n$ sont indépendantes et de même loi qu'un vecteur aléatoire (X, Y) pour lequel toute l'information sur Y fourni par la variable $X \in \mathbb{R}^d$ est en fait contenue dans la projection de X sur un sous-espace de plus petite dimension. Cette hypothèse peut se réecrire sous l'une des trois formes équivalentes suivantes : il existe des vecteurs de $\mathbb{R}^d$ $\beta_1 \ldots, \beta_m$ $(m < d)$ tels que :

1. Conditionnellement au vecteur $(\beta_1^t X, \ldots, \beta_m^t X)$, la variable aléatoire Y est indépendante de X.
2. La loi conditionnelle de Y sachant le vecteur $(\beta_1^t X, \ldots, \beta_m^t X)$ est la même que la loi de Y sachant X.
3. Il existe une fonction g de $\mathbb{R}^{m+1}$ dans $\mathbb{R}$ telle que $Y = g(\beta_1^t X, \ldots, \beta_m^t X, \eta)$ où η est un terme d'erreur indépendant de X.

La méthode SIR n'a pas pour objectif direct d'estimer la fonction de lien $g(.)$. Cependant, elle permet de réduire la dimension des observations facilitant ainsi l'estimation de cette fonction. En particulier, si $m = 1$, la fonction $g(.)$ est définie sur $\mathbb{R}$ et peut donc facilement être estimée à l'aide d'un estimateur à noyau par exemple. Les vecteurs $\beta_1, \ldots, \beta_m$ forment une base de l'espace de dimension m contenant la même information sur Y que l'espace complet. Ils ne sont pas définis de manière unique. Ce sous-espace de plus petite dimension est appelé espace *e.d.r.* ("effective dimension reduction"). La mise en œuvre de la méthode SIR permettant d'estimer une base de l'espace *e.d.r.* est décrite ci-dessous.

On découpe tout d'abord le support de la variable Y en H tranches $S_1, \ldots, S_H$. Pour $j = 1, \ldots, H$, on note n_j le nombre d'observations contenues dans la tranche S_j. On calcule ensuite l'estimateur de la matrice de covariance des X_i :

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^t, \text{ avec } \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

et l'estimateur de la matrice de covariance des moyennes conditionnelles des X_i sachant la tranche :

$$\hat{\Gamma} = \frac{1}{n} \sum_{j=1}^{H} n_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})^t, \text{ avec } \bar{X}_j = \frac{1}{n_j} \sum_{i : Y_i \in S_j} X_i. \tag{3.1}$$

Les vecteurs de la base *e.d.r.* sont solutions du problème d'optimisation ci-dessous :

$$\underset{\beta_1, \ldots, \beta_m}{\arg\max} \sum_{i=1}^{m} \beta_i^t \hat{\Gamma} \beta_i \text{ sous la contrainte } \beta_i^t \hat{\Sigma} \beta_i = 1 \ \forall \ i = 1, \ldots, m. \tag{3.2}$$

Autrement dit, la méthode SIR propose d'estimer la base de l'espace *e.d.r.* par les vecteurs maximisant la variance inter-tranches. Il est facile de montrer que la solution du problème (3.2) est $(\hat{\beta}_1, \ldots, \hat{\beta}_m)$ qui sont les m vecteurs propres de la matrice $\hat{\Sigma}^{-1}\hat{\Gamma}$ associés aux m plus grandes valeurs propres. De nombreuses extensions de SIR ont été proposées notamment par Barreda *et al.* [101]. Saracco [124] et Gannoun *et al.* [119] ont étudié les propriétés asymptotiques des estimateurs obtenus par SIR. Le cas de covariables fonctionnelles est considéré par Ferré *et al.* [118] et Amato *et al.* [99]. D'autres méthodes permettant d'estimer la base de l'espace *e.d.r.* ont été proposées, citons par exemple Cook et Weisberg [112] et Cook [108].

Remarquons que cette méthode de réduction de dimension est mieux adaptée pour estimer la fonction de lien g que la méthode plus classique de l'Analyse en Composantes Principales (ACP) qui consiste à résoudre le problème d'optimisation :

$$\underset{\beta_1,\ldots,\beta_m}{\arg\max} \sum_{i=1}^{m} \beta_i^t \hat{\Sigma} \beta_i \text{ sous la contrainte } \|\beta_i\|_2 = 1 \ \forall \ i = 1, \ldots, m.$$

En effet, la méthode SIR tient compte de l'information apportée par les Y_i ce qui n'est pas le cas de l'ACP. La méthode SIR est facile à implémenter et peu coûteuse en temps de calcul. Par contre, lorsque les données sont de grande dimension, la matrice $\hat{\Sigma}$ est souvent mal conditionnée ce qui conduit à une solution du problème (3.2) instable car elle est calculée en utilisant l'inverse de la matrice $\hat{\Sigma}$. Rappelons qu'une solution est instable si une petite variation sur les données engendre une erreur importante sur les estimateurs des vecteurs de la base *e.d.r.*

L'objectif de ce chapitre est de présenter une régularisation de la méthode SIR permettant d'obtenir des estimations plus stables des vecteurs de la base *e.d.r.* Dans la section 3.2, nous introduisons la méthode de régularisation GRSIR (Gaussian Regularized SIR) dans le cas où l'espace *e.d.r.* est de dimension 1. Une illustration sur simulation de la méthode GRSIR est présentée dans la section 3.3. La section 3.4 regroupe les résultats obtenus sur les images hyperspectrales du sol martien.

3.2 Régularisation de la méthode SIR

Dans cette section, nous proposons une régularisation de la méthode SIR dont le principe a été rappelé dans la section 3.1. L'intérêt de faire une régularisation est d'éviter les problèmes liés au mauvais conditionnement de la matrice de covariance $\hat{\Sigma}$ lorsque les données dont on dispose sont de grande dimension. Nous nous restreignons ici au cas où la dimension de l'espace *e.d.r.* est 1. Des régularisations de SIR ont été proposées dans la littérature, citons notamment Chiaromonte *et al.* [107] et Li *et al.* [122] qui proposent d'effectuer une ACP avant d'appliquer la méthode SIR et Antoniadis *et al.* [100] qui introduisent un a priori de Fisher-von Mises sur la direction *e.d.r.* Une autre méthode consiste à remplacer la matrice $\hat{\Sigma}$ par $\hat{\Sigma} + \tau I_d$, où $\tau > 0$ est le paramètre de régularisation et I_d est la matrice identité de dimension d (pour plus de détails, se référer à Draper *et al.* [115]).
Nous définissons dans le paragraphe 3.2.1 le modèle de régression inverse sur lequel repose notre méthode de régularisation. L'estimation par maximum de vraisemblance des paramètres de ce modèle est présentée dans le paragraphe 3.2.2. La méthode de régularisation est expliquée dans

le paragraphe 3.2.3. Elle s'appuie sur l'introduction d'un a priori Gaussien sur les paramètres du modèle de régression inverse. Ces résultats ont été obtenus en collaboration avec C. Bernard-Michel et S. Girard et ont été publiés dans la revue *Statistics and Computing* [105]. Enfin le paragraphe 3.2.4 est consacré à une analyse de la méthode de régularisation proposée par Li *et al.* [123]. Ce dernier paragraphe est issu d'une publication parue dans *Biometrics* [104] en collaboration avec C. Bernard-Michel et S. Girard.

3.2.1 Modèle de régression inverse

Le modèle de régression inverse présenté ici s'inspire de celui proposé par Cook [110]. Il s'écrit :

$$X = \mu + \tilde{g}(Y)b + \eta, \tag{3.3}$$

où μ et b sont des vecteurs non aléatoires et inconnus de $\mathbb{R}^d$, η est un vecteur aléatoire de loi multinormale $\mathcal{N}_d(0, I_d)$ indépendant de Y et $\tilde{g} : \mathbb{R} \to \mathbb{R}$ est une fonction inconnue. Cook [110, Proposition 1] montre que pour le modèle (3.3), l'espace *e.d.r.* est de dimension 1 et est engendré par le vecteur b. Pour mettre en place notre méthode de régularisation, nous définissons ci-dessous un modèle plus général où la matrice de covariance de l'erreur n'est plus nécessairement égale à la matrice identité :

$$X = \mu + \tilde{g}(Y)Vb + \zeta, \tag{3.4}$$

où V est une matrice définie positive de dimension $d \times d$ et ζ est un vecteur aléatoire de loi multinormale $\mathcal{N}_d(0, V)$. On montre aussi que le modèle (3.4) correspond à un espace *e.d.r.* de dimension 1 engendré par le vecteur b. Pour ce faire, il suffit de remarquer que (3.4) peut se réécrire sous la forme du modele de Cook : $X^* = \mu^* + c(Y)b^* + \zeta^*$ avec $X^* = V^{-1/2}X$, $\mu^* = V^{-1/2}X$, $b^* = V^{1/2}b$ et $\zeta^* = V^{1/2}\zeta \sim \mathcal{N}_d(0, I_d)$. On peut aussi définir le rapport signal sur bruit selon la direction *e.d.r.* b du modèle (3.4) par :

$$\rho = \frac{\mathrm{Var}(b^t X)}{\mathrm{Var}(b^t \zeta)} = \frac{b^t \Sigma b}{b^t V b}. \tag{3.5}$$

Une valeur de ρ proche de 1 conduit généralement à une mauvaise estimation des paramètres du modèle. Dans le paragraphe 3.2.2, nous proposons des estimateurs des paramètres inconnus du modèle (3.4). Afin d'estimer la fonction $\tilde{g}(.)$, nous supposons qu'elle se décompose dans une base de $H-1$ fonctions $s_j(.)$, $j = 1, \ldots, H-1$. Autrement dit,

$$\tilde{g}(.) = \sum_{j=1}^{H-1} c_j s_j(.),$$

où les coefficients $c_j \in \mathbb{R}$, $j = 1, \ldots, H-1$ sont inconnus. Les fonctions de base $s_j(.)$, $j = 1, \ldots, H-1$ et le paramètre H sont par contre supposés connus. Le modèle (3.4) se réécrit donc :

$$X = \mu + c^t s(Y)Vb + \zeta, \tag{3.6}$$

où $c = (c_1, \ldots, c_{H-1})^t$ et $s(.) = (s_1(.), \ldots, s_{H-1}(.))^t$.

Avant de définir les estimateurs des paramètres (μ, V, b, c), remarquons que le vecteur b défini par le modèle (3.6) est aussi l'unique vecteur propre d'une certaine matrice de rang 1. Pour ce faire, on introduit les notations suivantes :

$$W = \mathbb{E}\big\{\big(s(Y) - \mathbb{E}(s(Y))\big)\big(s(Y) - \mathbb{E}(s(Y))\big)^t\big\},\ \Sigma = \mathbb{E}\big\{\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^t\big\}$$

$$M = \mathbb{E}\big\{\big(s(Y) - \mathbb{E}(s(Y))\big)\big(X - \mathbb{E}(X)\big)^t\big\}.$$

Théorème 3.1 *Si les matrices Σ et W sont inversibles alors le vecteur b défini par le modèle (3.6) est le vecteur propre de la matrice $\Sigma^{-1}M^tW^{-1}M$ associé à l'unique valeur propre non nulle $\lambda = 1 - b^tVb/b^t\Sigma b$.*

Ce théorème fait partie d'un travail en cours (non soumis) avec Anne-Françoise Yao (Université d'Aix-Marseille II). On peut montrer que $M^tW^{-1}M = c^tWc(Vb)(Vb)^t$. Donc la matrice $\Sigma^{-1}M^tW^{-1}M$ est de rang 1 et possède ainsi une unique valeur propre non nulle.

3.2.2 Estimation des paramètres par maximum de vraisemblance

Nous allons à présent estimer les paramètres (μ, V, b, c) du modèle de régression inverse (3.6). Pour ce faire, on dispose d'un échantillon (X_i, Y_i), $i = 1, \ldots, n$ de variables aléatoires indépendantes et de même loi que le couple (X, Y) défini par le modèle (3.6). Sans perte de généralités, nous supposons dans la suite que

$$\bar{s}_j = \frac{1}{n}\sum_{i=1}^{n} s_j(Y_i) = 0,\ \forall j = 1, \ldots, H-1.$$

En utilisant le Théorème 3.1, l'idée la plus simple pour estimer la direction *e.d.r.* b est de calculer le vecteur propre $\hat{b}$ associé à la plus grande valeur propre de la matrice $\hat{\Sigma}^{-1}\hat{M}^t\hat{W}^{-1}\hat{M}$ avec

$$\hat{W} = \frac{1}{n}\sum_{i=1}^{n} s(Y_i)s^t(Y_i) \text{ et } \hat{M} = \frac{1}{n}\sum_{i=1}^{n} s(Y_i)(X_i - \bar{X})^t.$$

L'estimateur $\hat{b}$ ainsi défini maximise aussi la vraisemblance $L(\mu, V, b, c)$ des paramètres du modèle (3.6) au vu des observations (X_i, Y_i), $i = 1, \ldots, n$ avec

$$L(\mu, V, b, c) = \prod_{i=1}^{n} f_{(X,Y)}(X_i, Y_i) = \prod_{i=1}^{n} f_{X|Y=Y_i}(X_i) f_Y(Y_i) \propto \prod_{i=1}^{n} f_{X|Y=Y_i}(X_i),$$

où $f_{(X,Y)}(.,.)$ est la densité du couple (X, Y), $f_Y(.)$ la densité marginale de Y et $f_{X|Y=y}(.)$ la densité conditionnelle de X sachant $Y = y$. D'après le modèle (3.6), $f_{X|Y=y}(.)$ est la densité associée à une loi multinormale $\mathcal{N}_d(\mu + s^t(y)cVb, V)$. Nous avons montré dans [105, Lemme 1] que maximiser la vraisemblance $L(\mu, V, b, c)$ revient à minimiser en (μ, V, b, c) la quantité

$$G(\mu, V, b, c) = \log\det V + \mathrm{tr}(\hat{\Sigma}V^{-1}) + (\mu - \bar{X})^tV^{-1}(\mu - \bar{X}) + (c^tWc)(b^tVb) - 2c^tMb.$$

Les estimateurs du maximum de vraisemblance des paramètres (μ, V, b, c) sont explicites et données par le résultat ci-dessous (voir [105, Proposition 1]).

Théorème 3.2 *Sous le modèle (3.6), si les matrices $\hat{W}$ et $\hat{\Sigma}$ sont inversibles alors les estimateurs du maximum de vraisemblance des paramètres (μ, V, b, c) sont donnés par :*

- *$\hat{b}$ est le vecteur propre associé à la plus grande valeur propre $\hat{\lambda}$ de la matrice $\hat{\Sigma}^{-1}\hat{M}^t\hat{W}^{-1}\hat{M}$,*
- *$\hat{c} = \hat{W}^{-1}\hat{M}\hat{b}/(\hat{b}^t\hat{V}\hat{b})$,*
- *$\hat{\mu} = \bar{X}$,*
- *$\hat{V} = \hat{\Sigma} - \hat{\lambda}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}/(\hat{b}^t\hat{\Sigma}\hat{b})$.*

On déduit de la dernière équation du Théorème 3.2 que $\hat{\lambda} = 1 - (\hat{b}^t\hat{V}\hat{b})/(\hat{b}^t\hat{\Sigma}\hat{b}) \in]0,1[$. On dispose ainsi d'un estimateur du rapport signal sur bruit du modèle (3.6) :

$$\hat{\rho} = \frac{1}{1-\hat{\lambda}}.$$

L'estimateur $\hat{\lambda}$ peut donc être vu comme une mesure de la qualité des estimateurs. Une valeur de $\hat{\lambda}$ proche de zéro correspond à un rapport signal proche de 1 et donc à des estimateurs de qualité médiocre.

Lien avec la méthode SIR Comme expliqué dans le paragraphe d'introduction 3.1, la méthode SIR suppose un découpage en H tranches $S_1, \ldots, S_H$ du support de la variable Y. Plaçons nous dans le cas particulier où les fonctions de base $s_j(.)$, $j = 1, \ldots, H-1$ sont constantes sur les $H-1$ premières tranches *i.e.*

$$s_j(.) = \mathbb{I}\{. \in S_j\} - \frac{n_j}{n},\ j = 1, \ldots, H-1, \tag{3.7}$$

où n_j est le nombre d'observations Y_i appartenant à la tranche S_j, $j = 1, \ldots, H-1$. Nous montrons alors (voir [105], Corollaire 1) que $\hat{M}^t\hat{W}^{-1}\hat{M} = \hat{\Gamma}$, la matrice $\hat{\Gamma}$ étant la matrice de covariance inter-tranches définie par (3.1). Ainsi, le vecteur $\hat{b}$ coïncide avec l'estimateur de la direction *e.d.r.* fourni par la méthode SIR.

3.2.3 Régularisation par introduction d'un a priori Gaussien

Nous présentons dans un premier temps l'a priori Gaussien introduit dans le modèle de régression inverse (3.6) afin de régulariser la méthode SIR. La direction *e.d.r.* b et le vecteur c sont à présent supposés aléatoires. On pose $\tilde{\rho} = (b^t\hat{\Sigma}b)/(b^tVb)$. Cette variable aléatoire est égale (à l'estimation de la matrice de covariance Σ près) au rapport signal sur bruit défini en (3.5). Nous introduisons un a priori sur la variable aléatoire

$$\Theta = \tilde{\rho}^{-1/2}s^t(Y)cb. \tag{3.8}$$

Nous supposons que conditionnellement à $(\tilde{\rho}, Y)$,

$$\Theta \sim \mathcal{N}_d(0, \Omega). \tag{3.9}$$

La matrice de covariance a priori Ω est supposée connue. Elle décrit les directions de $\mathbb{R}^d$ que nous privilégions pour l'axe b. De plus, on remarque à partir de (3.8) et (3.9) que plus grande est la valeur $\tilde{\rho}$ (c'est à dire le rapport signal sur bruit) moins l'a priori sur les paramètres b et c sera informatif. Pour estimer (μ, V, b, c) en utilisant l'information a priori introduite précédemment, on maximise la vraisemblance de la loi conditionnelle du couple (X, Θ) sachant $(Y, \tilde{\rho})$. Plus précisément, on maximise en (μ, V, b, c) la quantité

$$\tilde{L}(\mu, V, b, c) = \prod_{i=1}^{n} f_{(X,\Theta)|(Y=Y_i,\tilde{\rho})}(X_i, \Theta) = \prod_{i=1}^{n} f_{X|(\Theta,Y=Y_i,\tilde{\rho})}(X_i) f_{\Theta|(Y=Y_i,\tilde{\rho})}(\Theta),$$

où $f_{(X,\Theta)|(Y=Y_i,\tilde{\rho})}(.,.)$ est la densité conditionnelle du couple (X, Θ) sachant $(Y = Y_i, \tilde{\rho})$. La densité conditionnelle de X sachant $(\Theta, Y = Y_i, \tilde{\rho})$ notée $f_{X|(\Theta,Y=Y_i,\tilde{\rho})}(.)$ est la densité associée à une loi multinormale $\mathcal{N}_d(\mu + s^t(Y_i)cVb, V)$ d'après le modèle (3.6). Enfin $f_{\Theta|(Y=Y_i,\tilde{\rho})}(.)$ est la densité conditionnelle de la variable Θ sachant $(Y = Y_i, \tilde{\rho})$ qui est associée à une loi multinormale $\mathcal{N}_d(0, \Omega)$ d'après (3.9). Nous avons montré dans [105, Lemme 2] que maximiser $\tilde{L}(\mu, V, b, c)$ revient à minimiser en (μ, V, b, c) la quantité

$$G_\Omega(\mu, V, b, c) = G(\mu, V, b, c) + \frac{(b^t\Omega^+ b)(c^t W c)}{\tilde{\rho}},$$

où Ω^+ est l'inverse de Moore-Penrose de Ω. La minimisation de la fonction $G_\Omega(\mu, V, b, c)$ fournit des estimateurs explicites des paramètres. Nous donnons leurs expressions dans le résultat ci-dessous (voir [105, Proposition 2]).

Théorème 3.3 *Sous le modèle (3.6) avec l'a priori (3.9), si les matrices $\hat{W}$ et $\Omega\hat{\Sigma} + I_d$ sont inversibles alors les estimateurs obtenus en maximisant la fonction $G_\Omega(\mu, V, b, c)$ en (μ, V, b, c) sont les suivants :*

- *$\hat{b}$ est le vecteur propre associé à la plus grande valeur propre $\hat{\lambda}$ de la matrice $(\Omega\hat{\Sigma} + I_d)^{-1}\Omega\hat{M}^t\hat{W}^{-1}\hat{M}$,*
- *$\hat{c} = \hat{\rho}\hat{W}^{-1}\hat{M}\hat{b}/(\hat{b}^t(\Omega^+ + \hat{\Sigma})\hat{b})$, où $\hat{\rho} = (\hat{b}^t\hat{\Sigma}\hat{b})/(\hat{b}^t\hat{V}\hat{b})$,*
- *$\hat{\mu} = \bar{X}$,*
- *$\hat{V} = \hat{\Sigma} - \hat{\lambda}\hat{\Sigma}\hat{b}\hat{b}^t\hat{\Sigma}/(\hat{b}^t\hat{\Sigma}\hat{b})$.*

On remarque que le calcul de la direction *e.d.r.* estimée $\hat{b}$ ne requiert plus l'inversion de la matrice $\hat{\Sigma}$ comme dans le Théorème 3.2 mais l'inversion de la matrice $\Omega\hat{\Sigma} + I_d$. Ainsi, on peut choisir une matrice Ω pour obtenir une matrice $\Omega\hat{\Sigma} + I_d$ ayant un meilleur conditionnement que la matrice $\hat{\Sigma}$. Comme pour le Théorème 3.2, on a $\hat{\lambda} = 1 - 1/\hat{\rho}$. La valeur propre peut donc être vu comme une mesure de la qualité des estimateurs obtenus.

Méthode GRSIR En prenant comme fonctions de bases $s_j(.)$, $j = 1, \ldots, H-1$ celles définies par (3.7), on montre (voir [105, Corollaire 2]) que $\hat{b}$ est le vecteur propre associé à la plus

grande valeur propre de la matrice $(\Omega\hat{\Sigma} + I_d)^{-1}\Omega\hat{\Gamma}$, $\hat{\Gamma}$ étant comme précédemment la matrice de covariance inter-tranches. L'estimateur ainsi obtenu est appelé *estimateur GRSIR* de la direction *e.d.r.* b (GRSIR pour Gaussian Regularized Sliced Inverse Regression).

Nous définissons à présent une famille de matrices de covariance Ω pouvant être utilisées pour calculer l'estimateur GRSIR de b. Posons,

$$\mathcal{F}(\varphi, m) = \left\{ \Omega = \sum_{j=1}^{m} \varphi(\hat{\delta}_j)\hat{q}_j\hat{q}_j^t \right\},$$

où m est un entier positif tel que $m \leq d$, $(\hat{q}_1, \ldots, \hat{q}_m)$ sont les vecteurs propres de la matrice $\hat{\Sigma}$ associés aux m plus grandes valeurs propres $\hat{\delta}_1 \geq \ldots \geq \hat{\delta}_m > 0$. Nous montrons (voir [105, Proposition 4]) que l'estimateur GRSIR de b obtenu en utilisant une matrice $\Omega \in \mathcal{F}(\varphi, m)$ peut-être aussi obtenu en :

1) projetant les variables $X_i, i = 1, \ldots, n$ dans le sous-espace de dimension m engendré par les vecteurs $(\hat{q}_1, \ldots, \hat{q}_m)$,
2) en calculant sur ces variables projetées l'estimateur GRSIR obtenu en prenant comme matrice de covariance a priori

$$\sum_{j=1}^{d} \varphi(\hat{\delta}_j)\hat{q}_j\hat{q}_j^t.$$

L'étape 1) correspond à une pré-réduction de la dimension sur les m premiers axes de l'ACP. Evidemment, si $m = d$, cette étape devient inutile. La fonction φ donne la forme de la matrice de covariance a priori. Nous donnons ci-dessous plusieurs choix de φ et m conduisant à des régularisations connues de SIR ainsi qu'à de nouvelles méthodes.

Liens avec des méthodes de régularisations existantes

- **Méthode SIR sans régularisation**. Il est facile de remarquer que si l'on prend

  $$\varphi(t) = 1/t \text{ et } m = d,$$

  alors la matrice de covariance a priori est donnée par $\hat{\Sigma}^{-1}$. Ainsi, l'estimateur GRSIR de la direction b correspond à l'estimateur classique obtenu par la méthode SIR. Ce choix de matrice de covariance a priori conduit à privilégier pour b les directions où la variance de X est faible. Ceci conduit en pratique à des instabilités lorsque la covariance empirique $\hat{\Sigma}$ est proche de la singularité.

- **Méthode Ridge**. Cette méthode de régularisation proposée par Zhong *et al.* [130] et Scrucca [126, 127] consiste à remplacer la matrice $\hat{\Sigma}$ par $\hat{\Sigma}+\tau I_d$ où $\tau > 0$ est un paramètre de régularisation. Le rôle de τ est de contrôler l'importance de la régularisation : une petite valeur de τ implique une faible transformation de la matrice $\hat{\Sigma}$ et donc une régularisation

peu importante. Cette méthode revient à choisir notre matrice de covariance a priori dans la famille $\mathcal{F}(\varphi, m)$ avec

$$\varphi(t) = 1/\tau \text{ pour tout } t > 0 \text{ et } m = d,$$

autrement dit, prendre pour matrice de covariance a priori la matrice I_d/τ. La méthode Ridge ne favorise aucune direction particulière pour l'axe b. Remarquons enfin que si τ tend vers zéro, la méthode Ridge revient à effectuer la méthode SIR sans régularisation. Par contre, si τ tend vers l'infini, il est facile de montrer que la méthode Ridge correspond alors à calculer le vecteur propre de la matrice $\hat{\Gamma}$ associé à la plus grande valeur propre.

- **Méthode ACP + SIR**. Cette méthode proposée par Chiaromonte *et al.* [107] et Li *et al.* [122] consiste à projeter les variables X_i, $i = 1, \ldots, n$ sur le sous-espace de dimension $m < d$ engendré par les d premiers vecteurs propres issus de l'ACP de $\hat{\Sigma}$ et à appliquer ensuite la méthode SIR classique sur ces variables projetées. Comme mentionné ci-dessus cela revient à prendre

$$\varphi(t) = 1/t \text{ et } m < d,$$

dans notre famille $\mathcal{F}(\varphi, m)$ de matrice de covariance a priori.

Nouvelles méthodes de régularisation

- **Méthode de Tikhonov**. Elle correspond à une matrice de covariance a priori dans la famille $\mathcal{F}(\varphi, m)$ avec

$$\varphi(t) = t/\tau \text{ et } m = d,$$

où $\tau > 0$ est un paramètre de régularisation. L'estimateur $\hat{b}$ est alors obtenu en prenant le vecteur propre de la matrice $(\hat{\Sigma}^2 + \tau I_d)^{-1}\hat{\Sigma}\hat{\Gamma}$ associé à la plus grande valeur propre. Cette régularisation privilégie pour b les directions où la variance de X est grande (au contraire de la méthode SIR). Elle est donc moins sensible au mauvais conditionnement de $\hat{\Sigma}$. Si le paramètre de régularisation τ tend vers l'infini, on peut montrer que l'estimateur obtenu converge vers le vecteur propre de la matrice $\hat{\Sigma}\hat{\Gamma}$ associé à la plus grande valeur propre. Cette méthode de régularisation peut être rapprochée de la régularisation de Tikhonov introduite dans un cadre d'estimation par moindres carrés (voir par exemple [128, Eq. (1.34)]).

- **Méthode ACP + Tikhonov**. Cette régularisation consiste à faire une étape préliminaire où l'on projette les variables X_i, $i = 1, \ldots, n$ sur le sous-espace engendré par les m premiers vecteurs propres de $\hat{\Sigma}$ et ensuite appliquer la méthode de Tikhonov présentée ci-dessus. Elle correspond au choix suivant pour φ et m :

$$\varphi(t) = t/\tau \text{ et } m < d.$$

- **Méthode ACP + Ridge**. Il s'agit dans un premier temps de projeter les variables X_i, $i = 1, \ldots, n$ sur le sous-espace engendré par les m premiers vecteurs propres de $\hat{\Sigma}$ et ensuite appliquer la méthode Ridge. Elle correspond au choix suivant pour φ et m :

$$\varphi(t) = 1/\tau \text{ pour tout } t > 0 \text{ et } m < d.$$

3.2.4 Discussion sur une autre méthode de régularisation

Cette autre méthode de régularisation présentée par Li *et al.* [123] est basée sur une définition alternative de l'estimateur SIR de la direction b. On découpe tout d'abord le support de Y en H tranches $S_1, \ldots, S_H$ et on définit la fonction :

$$G(a, \tilde{c}) = \sum_{j=1}^{H} \frac{n_j}{n} \big((\bar{X}_j - \bar{X}) - \tilde{c}_j \hat{\Sigma} a\big)^t \hat{\Sigma}^{-1} \big((\bar{X}_j - \bar{X}) - \tilde{c}_j \hat{\Sigma} a\big),$$

où $a \in R^d$, $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_H) \in \mathbb{R}^H$, les autres notations étant identiques à celles employées pour la présentation de la méthode SIR dans la section 3.1. Cook *et al.* [109, 111] montrent que le vecteur a minimisant la fonction $G(a, \tilde{c})$ correspond à l'estimateur SIR de la direction b. Par soucis de comparaison avec notre méthode de régularisation, nous supposons ici que l'espace *e.d.r.* est de dimension 1. Le résultat montré par Cook *et al.* [109, 111] est en fait valable dans le cadre plus général d'un espace *e.d.r.* de dimension supérieure ou égale à 1. Pour régulariser la méthode SIR, Li *et al.* [123] proposent de minimiser en a et $\tilde{c}$ la fonction :

$$G_\tau(a, \tilde{c}) = \sum_{j=1}^{H} \frac{n_j}{n} \|(\bar{X}_j - \bar{X}) - \tilde{c}_j \hat{\Sigma} a\|_2^2 + \tau \|a\|_2^2,$$

où $\tau > 0$ est un paramètre de régularisation. Nous montrons dans [104] que cette méthode n'est pas justifiable théoriquement bien qu'elle semble fournir de bons résultats sur les simulations et l'application sur données réelles présentées dans [123]. Plus précisément, nous remarquons dans un premier temps que pour tout réel $\lambda > 0$, $G_\tau(a, \tilde{c}) \neq G_\tau(\lambda a, \tilde{c})$. Ainsi, les vecteurs a et λa (engendrant pourtant le même sous-espace) ne seront pas estimés de la même façon par cette méthode de régularisation. Nous montrons de plus (voir [104, Proposition 1]) que s'il existe, le vecteur $\hat{a}$ minimisant la fonction $G_\tau(a, \tilde{c})$ est le vecteur nul de $\mathbb{R}^d$. Nous proposons dans [104, Eq. (5)] une version corrigée de la fonction $G_\tau(a, \tilde{c})$ permettant d'obtenir un estimateur de a acceptable d'un point de vue théorique :

$$H(a, \tilde{c}) = \sum_{j=1}^{H} \frac{n_j}{n} \big((\bar{X}_j - \bar{X}) - \tilde{c}_j \hat{\Sigma} a\big)^t \hat{\Sigma}^{-1} \big((\bar{X}_j - \bar{X}) - \tilde{c}_j \hat{\Sigma} a\big) + \tau \sum_{j=1}^{H} \frac{n_j}{n} \tilde{c}_j \|a\|_2^2.$$

Il est facile alors de montrer que la valeur $\hat{a}$ minimisant la fonction $H(a, \tilde{c})$ est le vecteur propre de la matrice $(\hat{\Sigma} + \tau I_d)^{-1} \hat{\Gamma}$ associé à la plus grande valeur propre. Autrement dit, $\hat{a}$ est l'estimateur de la direction *e.d.r.* obtenu par la méthode ridge présentée précédemment.

3.3 Illustration sur simulation

Dans ce paragraphe, nous présentons quelques simulations nous permettant de comparer les différentes méthodes de régularisation présentées précédemment. On génère pour ce faire des observations selon la loi d'un couple (X, Y) où X est de loi multinormale $\mathcal{N}(0, I_d)$ et la variable Y est déterminée par le modèle de régression :

$$Y = \sin\left(\frac{\pi}{2} b^t X\right) + \xi,$$

où $\xi \sim \mathcal{N}(0, 0.009)$ et la direction b engendrant l'espace *e.d.r.* de dimension 1 est donnée par $b = (1, \ldots, 1)/\sqrt{d}$. Comme nous l'illustrons dans [105, Paragraphe 4], la fonction de lien n'a que peu d'influence sur la qualité de l'estimateur de la direction *e.d.r.* b. Nous nous restreignons donc ici à une seule fonction de lien. Afin d'apprécier le comportement des différents estimateurs de b, nous générons indépendamment $N = 100$ échantillons de taille $n = 100$ selon la loi du couple (X, Y). Nous disposons donc des observations $(X_i^{(r)}, Y_i^{(r)})$, $i = 1, \ldots, 100$, $r = 1, \ldots, 100$ à l'aide desquelles nous calculons les estimateurs $\hat{b}^{(r)}$ de la direction b (les vecteurs $\hat{b}^{(r)}$ sont normalisés de telle sorte que $\|\hat{b}^{(r)}\|_2 = 1$). Le nombre de tranches utilisées pour calculer la matrice $\hat{\Gamma}$ est fixé à $H = 10$. La qualité de l'estimateur est mesurée par la moyenne des cosinus carrés entre l'axe b et les axes estimés $\hat{b}^{(r)}$, $r = 1, \ldots, 100$. Plus précisément, nous calculons le critère :

$$\mathrm{CC} = \frac{1}{N} \sum_{r=1}^{N} (b^t \hat{b}^{(r)})^2 \in [0, 1].$$

Une valeur du critère CC proche de 0 correspond à une mauvaise estimation de b (les vecteurs $\hat{b}^{(r)}$, $r = 1, \ldots, 100$ sont presque orthogonaux à b). Inversement, si CC est proche de 1, les vecteurs $\hat{b}^{(r)}$, $r = 1, \ldots, 100$ sont à peu de choses près colinéaires. Une adaptation de ce critère dans le cas d'un espace *e.d.r.* de dimension supérieure à 1 est proposée dans [117]. Nous étudions ici le comportement des estimateurs de b en fonction de 3 paramètres : la dimension d du vecteur alátoire X, le paramètre de régularisation τ et la dimension m du sous-espace utilisé dans les méthodes ACP + SIR, ACP + Ridge et ACP + Tikhonov. Remarquons tout d'abord que la dimension d est directement liée au conditionnement de la matrice de covariance empirique $\hat{\Sigma}$ (voir Figure 3.2) : une grande dimension d correspond à un mauvais conditionnement (*i.e.* forte valeur) de $\hat{\Sigma}$.

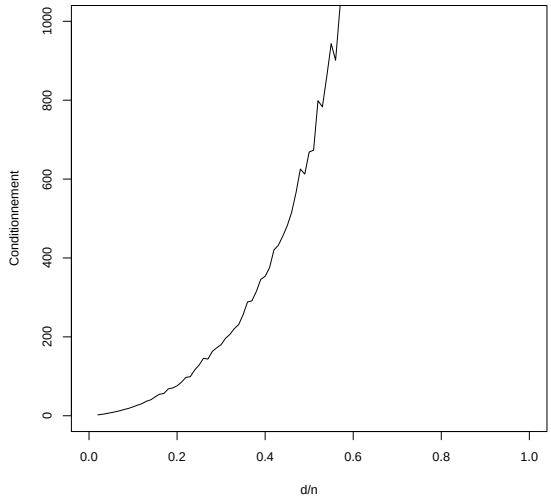

FIG. 3.2 – Lien entre la dimension d du vecteur aléatoire X et le conditionnement de la matrice de covariance empirique $\hat{\Sigma}$. En abscisse, le rapport $d/n \in]0, 1]$ et en ordonnée le conditionnement de $\hat{\Sigma}$.

Influence du paramètre de régularisation. Nous calculons le critère CC en fonction du paramètre τ. Afin d'obtenir des graphiques plus lisibles, nous utilisons une échelle logarithmique et nous faisons varier $\log(\tau)$ entre -10 et 10. Les paramètres d et m sont fixés respectivement à 70 et 20. Les résultats sont représentés sur la figure 3.3. On remarque que la méthode SIR donne de mauvais résultats selon le critère CC. Pour une valeur de τ correctement choisie, les méthodes Ridge et Tikhonov améliorent la qualité des estimateurs. L'influence du paramètre τ semble moins importante pour la méthode Ridge. La méthode ACP + SIR est plus performante que la méthode SIR classique. Pour un bon choix de τ, les méthodes ACP + Ridge et ACP + Tikhonov donnent de meilleurs résultats que ACP + SIR.

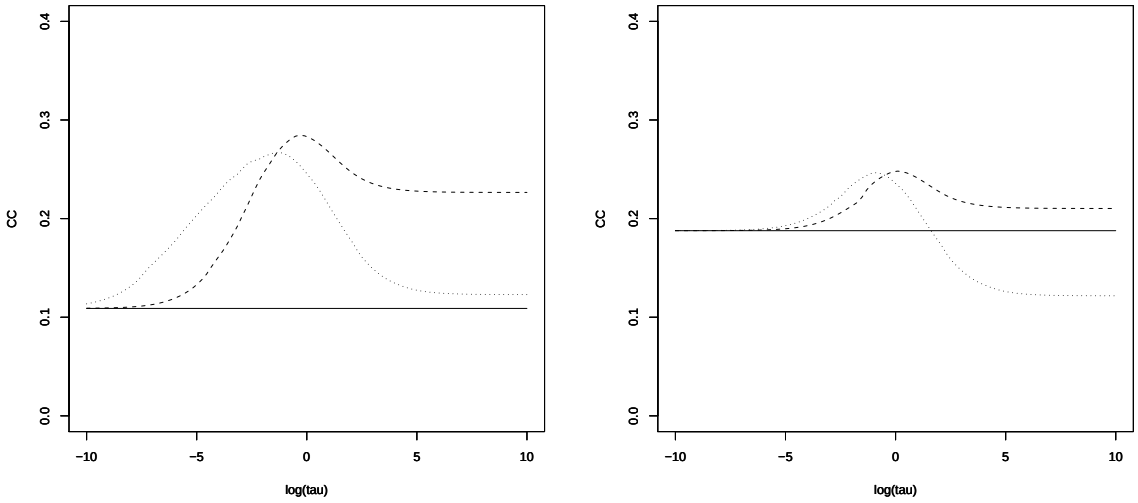

FIG. 3.3 – Influence du paramètre de régularisation sur l'estimation de b avec : figure de gauche les méthodes SIR (trait continu), Ridge (tirets) et Tikhonov (pointillé), figure de droite les méthodes ACP + SIR (trait continu), ACP + Ridge (tirets) et ACP + Tikhonov (pointillés). En abscisse, le logarithme de τ et en ordonnée le critère CC. La dimension d est fixé à 70 et la dimension m à 20.

Influence de la dimension m du sous-espace pour les méthodes utilisant une ACP. Le critère CC est ici calculé en fonction de m qui est le nombre de directions fournies par l'ACP que nous utilisons pour effectuer les méthodes ACP + SIR, ACP + Ridge et ACP + Tikhonov. Le paramètre de régularisation est fixé à $\tau = 0.5$ et le vecteur X est de dimension $d = 70$. Les résultats représentés dans la figure 3.4 montrent que les trois méthodes donnent des résultats très similaires pour $m \leq 40$. Les méthodes ACP + Ridge et ACP + Tikhonov continuent à fournir de bons résultats pour des dimensions $m > 40$. Il semble, au vu de cette simulation, qu'il est inutile d'effectuer une ACP avant d'appliquer les méthodes de régularisation Ridge et Tikhonov. Par contre, la méthode ACP + SIR améliore nettement les résultats de SIR pour peu que l'on choisisse correctement la dimension m.

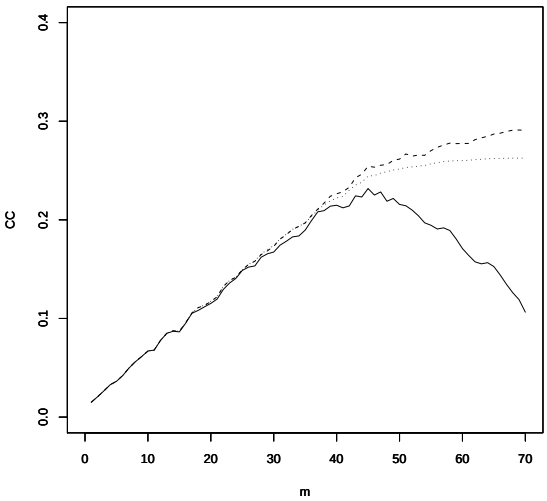

FIG. 3.4 – Critère CC en fonction de la dimension m du sous-espace utilisé pour les méthodes ACP + SIR (trait continu), ACP + Ridge (tirets) et ACP + Tikhonov (pointillés). En abscisse, la dimension m et en ordonnée le critère CC.

Influence de la dimension d du vecteur aléatoire X. Nous calculons ici le critère CC en fonction de la dimension d de la variable explicative X. Le paramètre τ est fixé à la valeur 0,5. Nous comparons uniquement les méthodes SIR, Ridge et Tikhonov.

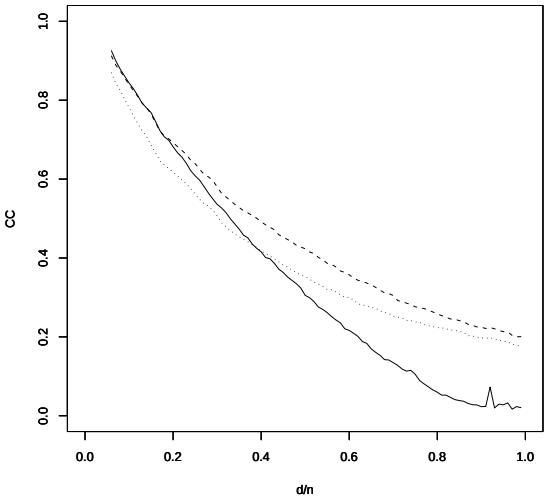

FIG. 3.5 – Critère CC en fonction de la dimension d de la variable X pour les méthodes SIR (trait continu), Ridge (tirets) et Tikhonov (pointillés). En abscisse, le rapport d/n et en ordonnée le critère CC.

La figure 3.5 montre que les méthodes Ridge et Tikhonov ont une valeur du critère CC très proche de celle obtenue avec SIR pour un rapport $d/n < 0,3$. Pour de plus grandes valeurs de d, les méthodes Ridge et Tikhonov fournissent de meilleurs résultats que SIR.

3.4 Application à l'étude d'images hyperspectrales du sol martien

L'objectif de cette section est d'utiliser la méthode de régularisation GRSIR présentée précédemment pour inverser une image hyperspectrale de la planète Mars. Cette application a été effectuée en collaboration avec C. Bernard-Michel, S. Douté, M. Fauvel et S. Girard et publiée dans *Journal of Geophysical Research* [102]. Inverser une image signifie en fait estimer les paramètres physiques associés aux spectres dont on dispose. Les résultats obtenus sont comparés à ceux fournis par la méthode des plus proches voisins et les régressions SVR et PLS (voir la section 3.1 pour une description rapide de ces méthodes). La nature des images hyperspectrales de Mars est décrite dans la section 3.1 (voir aussi [102, Section 2]). Nous allons ici utiliser trois types de données :

- **L'image hyperspectrale observée**. Nous disposons d'une image comprenant environ 15000 spectres acquise durant l'été martien par l'instrument OMEGA. Les spectres sont en dimension $d = 184$. Cette image couvre une grande partie du pôle sud de la planète. L'objectif est d'estimer pour ces spectres les paramètres physiques correspondants inconnus car nous ne disposons pas de vérité terrain. On se focalise ici sur 5 paramètres : les tailles des grains de CO_2 et d'eau, les proportions de poussière, d'eau et de glace de CO_2.

- **Une base d'apprentissage simulée**. Afin d'apprendre la liaison liant un spectre à ces paramètres physiques, nous disposons d'une base d'apprentissage simulée par un modèle de transfert radiatif correspondant au pôle sud martien. Cette base contient $n = 31500$ spectres de dimension $d = 184$ obtenus pour différentes valeurs des 5 paramètres physiques prises sur une grille régulière (voir Tableau 3.1). On dispose ainsi d'un échantillon

$$(X_i, Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}, Y_{i,5}),\ i = 1, \ldots, n,\ X_i \in \mathbb{R}^{184} \text{ et } Y_{i,j} \in \mathbb{R},\ j = 1, \ldots, 5,$$

  où $Y_{i,1}$ et $Y_{i,2}$ sont les tailles des grains de CO_2 et des grains d'eau associés au spectre X_i, $Y_{i,3}$, $Y_{i,4}$ et $Y_{i,5}$ les proportions de poussière, d'eau et de glace de CO_2. Les 5 paramètres sont traités indépendamment les uns des autres.

Paramètres	# de valeurs distinctes	Intervalle de variation
Proportion d'eau	15	[0.0006 0.002]
Proportion de CO_2	29	[0.996 0.9988]
Proportion de poussière	15	[0.0006 0.002]
Taille des grains d'eau	5	[100 400]
taille des grains de CO_2	28	[40000 105000]

TAB. 3.1 – Tableau donnant pour chacun des 5 paramètres le nombre de valeurs différentes selectionnées ainsi que les intervalles de variation.

- **Une base de test**. Ne disposant pas de vérité terrain sur Mars, nous proposons de valider les résultats d'estimation des paramètres obtenus par les méthodes GRSIR, plus proches voisins, SVR et PLS sur une base test :

$$(X_i^*, Y_{i,1}^*, Y_{i,2}^*, Y_{i,3}^*, Y_{i,4}^*, Y_{i,5}^*),\ i = 1, \ldots, n^* = 3500,\ X_i^* \in \mathbb{R}^{184} \text{ et } Y_{i,j}^* \in \mathbb{R},\ j = 1, \ldots, 5.$$

  Cette base a été obtenue en utilisant le même modèle de transfert radiatif que celui utilisé pour construire la base d'apprentissage. Pour simuler le bruit du aux effets de l'atmosphère de Mars, au mauvais fonctionnement de l'appareil de mesure, etc ... nous avons rajouté aux spectres simulés par le modèle une variable aléatoire de loi multinormale centrée et dont la matrice de covariance a été déterminée expérimentalement par le LPG.

Tout d'abord, nous estimons la fonction de lien entre les spectres et les paramètres avec la base d'apprentissage. Les résultats obtenus sont ensuite validés sur la base de test. Enfin, nous appliquons la méthode GRSIR pour inverser l'image hyperspectrale du pôle sud martien.

Estimation de la fonction de lien et validation des méthodes. Pour estimer la fonction de lien à partir de la base d'apprentissage, nous utilisons les méthodes suivantes : GRSIR-Tikhonov, plus proches voisins, SVR et PLS. On note par $\hat{Y}_{i,j}^*$ l'estimation par l'une de ces méthodes du $j-$ème paramètre physique associé au spectre X_i^*. Pour valider les estimations des paramètres sur la base test, nous proposons d'utiliser le critère suivant :

$$NRMSE_j = \sqrt{\frac{\sum_{i=1}^{n^*}(\hat{Y}_{i,j}^* - Y_{i,j}^*)^2}{\sum_{i=1}^{n^*}(Y_{i,j}^* - \bar{Y}_j^*)^2}},\ j = 1, \ldots, 5,\ \text{avec } \bar{Y}_j^* = \frac{1}{n^*}\sum_{i=1}^{n^*} Y_{i,j}^*.$$

Ce critère mesure l'écart entre le paramètre estimé et sa vraie valeur. Une valeur proche de 0 pour le $NRMSE_j$ signifie donc une bonne prédiction. Le dénominateur a pour but de normaliser le critère afin de pouvoir comparer sa valeur pour différents paramètres n'ayant pas forcément le même intervalle de variation. Nous donnons ci-dessous des détails sur l'implémentation des différentes méthodes d'estimation.

- **GRSIR-Tikhonov**. Nous avons choisi d'utiliser uniquement la méthode de régularisation Tikhonov. Nous avons remarqué (voir [102, Paragraphe 5.4] que l'on pouvait supposer que l'espace *e.d.r.* était de dimension 1. En effet, le premier axe donné par GRSIR-Tikhonov explique à lui seuil 98% de la variance totale. Nous devons donc calculer avec la base d'apprentissage le vecteur propre de la matrice $(\hat{\Sigma}^2 + \tau I_d)^{-1}\hat{\Sigma}\hat{\Gamma}$ associé à la plus grande valeur propre. Pour ce faire, nous devons au préalable choisir une valeur pour le paramètre de régularisation τ. Pour ce faire, nous avons retenu la valeur de τ minimisant le critère $NRMSE_j$. Le nombre H de tranches utilisées pour calculer $\hat{\Gamma}$ est égal au nombre de valeurs différentes prises par le paramètre étudié (voir Tableau 3.1). Une fois la direction *e.d.r.* $\hat{b}$ obtenue, nous estimons le paramètre Y (par exemple la taille des grains de CO_2) associé à un nouveau spectre x par :

$$\hat{Y} = \begin{cases} m_1^{param} & \text{si } t \in \left]-\infty, m_1^{proj}\right] \\ m_h^{param} + (t - m_h^{proj}) \left(\frac{m_{h+1}^{param} - m_h^{param}}{m_{h+1}^{proj} - m_h^{proj}} \right) & \text{si } t \in \left]m_h^{proj}, m_{h+1}^{proj}\right], \ h = 1, \ldots, H-1 \\ m_H^{param} & \text{si } t \in \left]m_H^{proj}, +\infty\right[\end{cases} \quad (3.10)$$

où, si j dénote le numéro du paramètre étudié,

$$m_h^{proj} = \frac{1}{n_h} \sum_{Y_{i,j} \in S_h} \hat{b}_j^t X_i, \text{ et } m_h^{param} = \frac{1}{n_h} \sum_{Y_{i,j} \in S_h} \hat{b}^t X_i, \ h = 1, \ldots, H,$$

les notations n_h et S_h ayant été introduites dans la section 3.1. Cette méthode d'estimation est illustrée par la Figure 3.6.

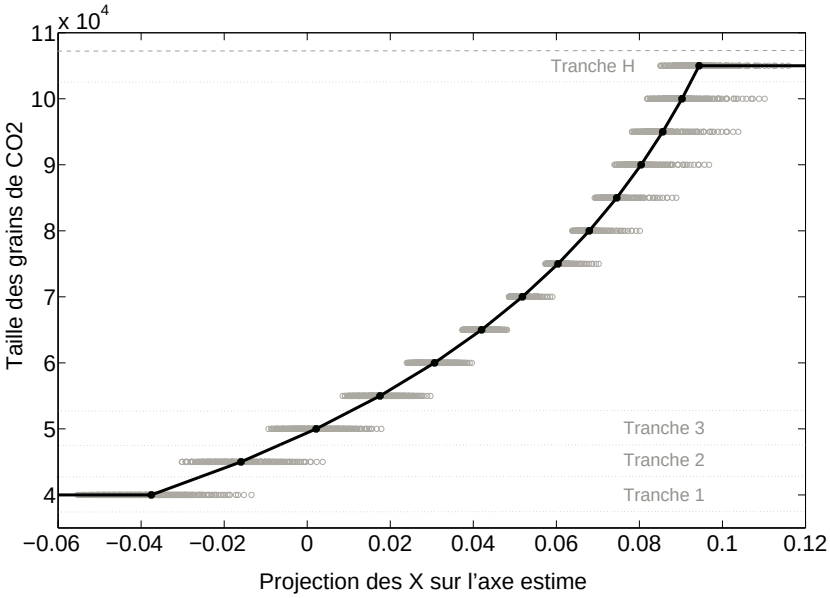

FIG. 3.6 – Figure illustrant l'estimation de la fonction de lien (en noir). En abscisse : les projections des spectres sur la direction $\hat{b}$. En ordonnée : la taille des grains de CO_2. Les points gris représentent les données de la base de test.

- **Plus proches voisins**. On conserve ici uniquement un voisin. Autrement dit, pour un nouveau spectre x, on estime son paramètre par celui associé au spectre le plus proche de x (pour la distance euclidienne de $\mathbb{R}^{184}$) dans la base d'apprentissage.

- **Régression PLS**. En reprenant les mêmes notations que dans la section 3.1, on projette les spectres X_i sur un sous-espace de dimension $m < d = 184$ engendré par la matrice A. On estime le paramètre associé au spectre x par $\hat{c}_1 + \hat{c}_2^t A x$ où $\hat{c}_1$ et $\hat{c}_2$ sont les estimateurs des moindres carrés des paramètres du modèle de régression linéaire multiple $Y_i = c_1 + c_2^t A X_i + \eta_i$, $i = 1, \ldots, n$.

- **Régression SVR**. On utilise le noyau gaussien

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|_2^2).$$

  La méthode SVR dépend alors de trois paramètres : ϵ, λ et γ. Nous fixons la valeur de ϵ à 0,01. Les paramètres λ et γ sont choisis en minimisant le critère $NRMSE_j$.

Les résultats de l'estimation des paramètres sur la base de test sont présentés dans le Tableau 3.2 et la Figure 3.7. Le Tableau 3.2 fournit les valeurs du critère $NRMSE$ pour les 5 paramètres. Comme attendu, la méthode des plus proches voisins donne les moins bons résultats. La méthode GSIR est préférable (sauf pour le paramètre "taille des grains d'eau") à la méthode des plus proches voisins et à la régression PLS. Dans la base d'apprentissage, nous disposons uniquement de 5 valeurs différentes pour le paramètre "taille des grains d'eau" ce qui peut expliquer le mauvais comportement de la méthode GRSIR. La régression SVR fournit les meilleurs résultats mais au prix d'un temps de calcul important (environ 15 heures contre seulement 1 minute pour GRSIR). De plus, la méthode GRSIR présente l'avantage de fournir une direction *e.d.r.* estimée dont les composantes sont interprétables par un planétologue (voir [102, Paragraphe 5.5]).
Dans la Figure 3.7, nous avons représenté les estimations du paramètre "proportion de CO_2" en fonction des valeurs observées des paramètres sur la base de test. Ici encore, nous remarquons le mauvais comportement de la méthode des plus proches voisins et la très bonne estimation obtenue avec la régression SVR. La méthode GRSIR présente un comportement satisfaisant. La régression PLS fournit un nuage de points légèrement incurvé par rapport à la première bissectrice. Ce comportement est dû au fait que la méthode PLS suppose une relation linéaire entre les spectres et les paramètres.

Estimation de la proportion de poussière de l'image hyperspectrale de Mars. Nous allons utiliser les estimateurs de la fonction de lien trouvés à l'aide de la base d'apprentissage par les méthodes des plus proches voisins, GRSIR-Tikhonov, PLS et SVR. La base d'apprentissage a été construite sans tenir compte des spectres observés sur Mars. Ainsi, certains spectres de la base d'apprentissage sont très éloignés des spectres observés et sont donc inutiles (voire même nuisibles) pour inverser l'image hyperspectrale. A l'inverse, certains spectres observés sont très différents des spectres de la base d'apprentissage et ne peuvent donc pas être inversés par la fonction de lien estimée. Il faut donc au préalable faire coïncider au mieux les spectres observés et ceux de la base d'apprentissage. Pour ce faire nous proposons dans [102, Paragraphe 4] de projeter ces spectres sur les 2 premiers axes de l'ACP des spectres de la base d'apprentissage. Nous conservons pour l'estimation uniquement les spectres suffisamment proches selon la distance euclidienne de $\mathbb{R}^2$. Avec la nouvelle base d'apprentissage ainsi obtenue (contenant 15407 spectres), nous estimons la liaison entre les spectres et les paramètres par les différentes méthodes considérées. Nous estimons ensuite pour les spectres observés sur Mars la proportion de poussière. Les résultats sont représentés sur la Figure 3.8. Les méthodes GRSIR et SVR donnent des images similaires à celles obtenues avec les plus proches voisins bien que cette dernière semble moins lisse. De nombreuses autres estimations de paramètres physiques à partir d'images hyperspectrales de Mars sont consultables dans un rapport de recherche [103]

rédigé durant le post-doctorat de C. Bernard-Michel.

Paramètres	Plus proches voisins	PLS	SVR	GRSIR
Proportion d'eau	0.86	0.52	0.17	0.40
Proportion de CO_2	0.88	0.56	0.18	0.30
Proportion de poussière	0.44	0.36	0.11	0.17
Taille des grains d'eau	0.43	0.44	0.17	0.54
Taille des grains de CO_2	0.53	0.47	0.14	0.22

TAB. 3.2 – Valeur du $NRMSE$ pour les 5 paramètres obtenue sur la base de test avec les méthodes GRSIR-Tikhonov, plus proches voisins, PLS et SVR.

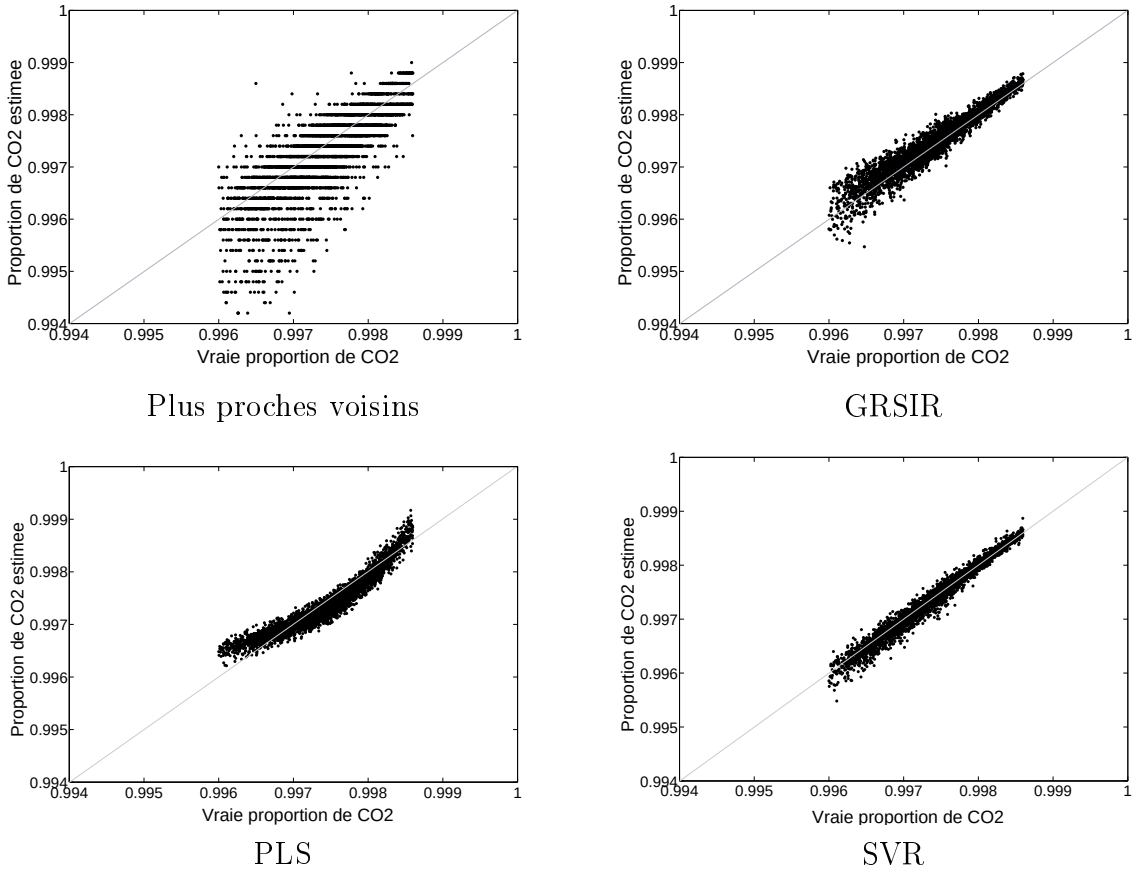

FIG. 3.7 – En abscisse : la vraie valeur de la proportion de CO_2. En ordonnée : la proportion de CO_2 estimée par les méthodes des plus proches voisins, GRSIR, PLS et SVR. Le trait en pointillé représente la première bissectrice.

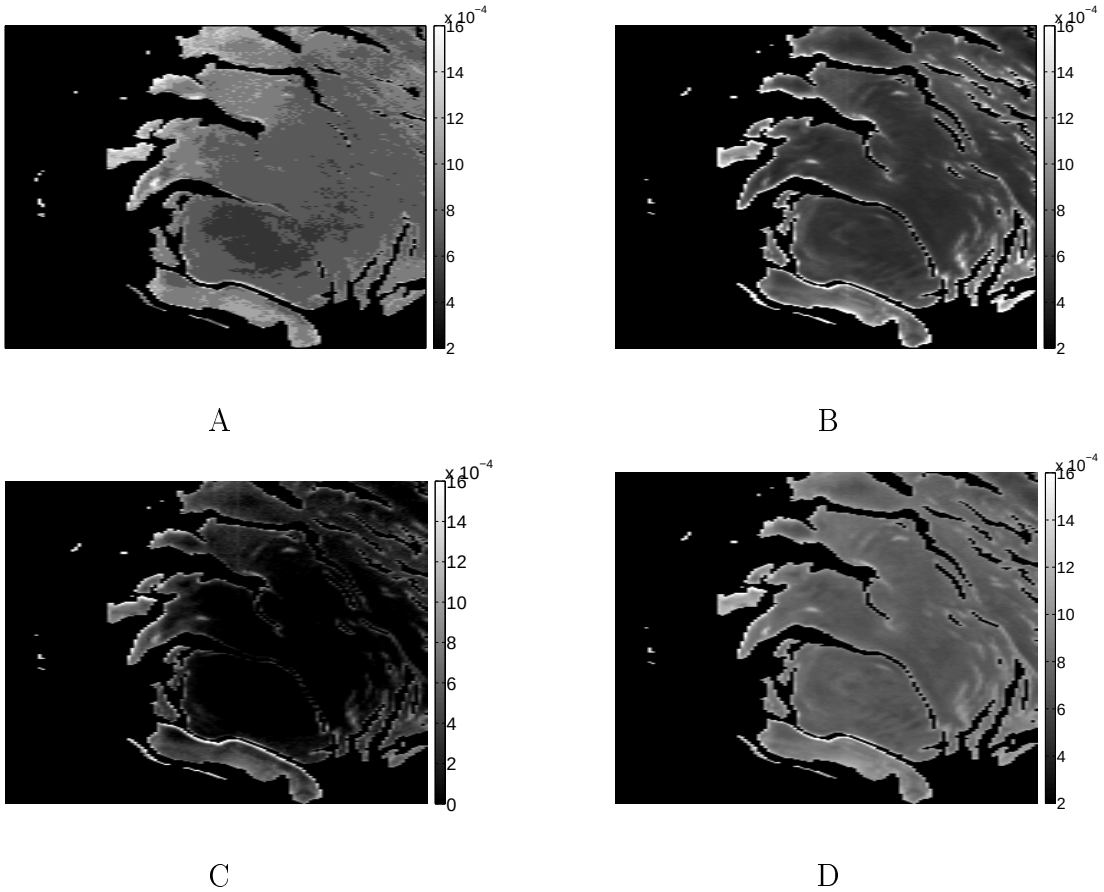

FIG. 3.8 – Estimation de la proportion de poussière par la méthode des plus proches voisins (A), GRSIR (B), PLS (C) et SVR (D).

3.5 Conclusion et perspectives

Dans ce chapitre, nous avons proposé une régularisation de la méthode SIR basé sur l'introduction d'un a priori Gaussien. Différents choix pour la matrice de variance-covariance de la loi a priori conduisent à des méthodes de régularisation connues ainsi qu'à de nouvelles approches. La méthode GRSIR s'est avérée efficace et compétitive face à d'autres méthodes de régression en grande dimension (SVR, PLS, etc ...). Elle semble aussi fournir de bons résultats pour l'application sur des images hyperspectrales du sol martien. Dans le cadre du projet ANR *Visualisation et analyse d'images hyperspectrales multi-dimensionnelles en astrophysique*, un logiciel regroupant de nombreuses méthodes d'analyse d'images hyperspectrales est actuellement développé par L. Léau-Mercier. La méthode GRSIR sera implémentée dans ce logiciel.

La méthode GRSIR a pour l'instant été validée uniquement sur données simulées et réelles. En collaboration avec S. Girard et A.F. Yao, nous tentons de déterminer la vitesse de convergence théorique de l'estimateur $\hat{b}$ obtenu avec la méthode GRSIR vers la vraie direction *e.d.r.* Nous pourrons ainsi la comparer à celle obtenue lorsque la méthode SIR classique (sans régularisation) est utilisée.

Il serait également intéressant d'appliquer la méthode de régularisation proposée pour SIR à d'autres méthodes (par exemple celle des plus proches voisins).
Dans de nombreuses applications (et notamment en planétologie), la possibilité de pouvoir affecter une nouvelle observation à une classe est souvent appréciable. Nous souhaiterions donc coupler la méthode SIR à une méthode de classification (par exemple l'algorithme E-M) afin de pouvoir (en une seule étape) classer et estimer les paramètres physiques.
Enfin, concernant l'application en Planétologie, la méthode GRSIR est pour l'instant effectuée indépendamment sur chaque pixel de l'image. Il serait souhaitable de mettre en place une méthode de régularisation spatiale afin d'obtenir une carte d'estimation plus lisse.

Bibliographie du Chapitre 3

[99] U. Amato, A. Antoniadis, and I. D. Feiss. Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, **50**, 2422–2446 (2006).

[100] A. Antoniadis, G. Grégoire, and I. McKeague. Bayesian estimation in single-index models. *Statistica Sinica*, **14**, 1147–1164 (2004).

[101] L. Barreda, A. Gannoun, and J. Saracco. Some extensions of multivariate SIR. *Journal of Statistical Computation and Simulation*, **77(1-2)**, 1–17 (2007).

[102] C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes, and S. Girard. Retrieval of Mars surface physical properties from omega hyperspectral images using regularized sliced inverse regression. *Journal of Geophysical Research - Planets*, **114**, E06005 (2009).

[103] C. Bernard-Michel, S. Douté, L. Gardes, and S. Girard. Estimation of Mars surface physical properties from hyperspectral images using Sliced Inverse Regression (2007). http ://hal.inria.fr/inria-00187444/fr/.

[104] C. Bernard-Michel, L. Gardes, and S. Girard. A note on Sliced Inverse Regression with regularizations. *Biometrics*, **64 (3)**, 982–984 (2008).

[105] C. Bernard-Michel, L. Gardes, and S. Girard. Gaussian regularized sliced inverse regression. *Statistics and Computing*, **19**, 85–98 (2009).

[106] R. Carlson, M. Anderson, R. Mehlman, and R. Johnson. Distribution of hydrate on europa : Further evidence for sulfuric acid hydrate. *Icarus*, **177 (2)**, 451–471 (2005).

[107] F. Chiaromonte and J. Martinelli. Dimension reduction strategies for analysing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123–144 (2002).

[108] R. Cook. *Regression graphics. Ideas for studying regressions through graphics.* Wiley Series in Probability and Statistics, New York (1998).

[109] R. Cook. Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, **32**, 1062–1092 (2004).

[110] R. Cook. Fisher lecture : Dimension reduction in regression. *Statistical Science*, **22 (1)**, 1–26 (2007).

[111] R. Cook and L. Ni. Sufficient dimension reduction via inverse regression : A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410–428 (2005).

[112] R. Cook and S. Weisberg. Discussion of "sliced inverse regression for dimension reduction". *K.C. Li, Journal of the American Statistical Association*, **86**, 328–332 (1991).

[113] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press (2000).

[114] S. Douté and B. Schmitt. A multilayer bidirectional reflectance model for the analysis of planetary surface hyperspectral images at visible and near-infrared wavelengths. *Journal of Geophysical Research (Planets)*, **103 (12)**, 31367–31390 (1998).

[115] N. Draper and R. Smith. *Applied regression analysis (3rd edition)*. Wiley (1998).

[116] S. Durbha, R. King, and N. Younan. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sensing of Environment*, **107**, 348–361 (2007).

[117] L. Ferré. Determining the dimension in Sliced Inverse Regression and related method. *Journal of the American Statistical Association*, **93**, 132–140 (1998).

[118] L. Ferré and A. Yao. Smoothed functional inverse regression. *Statistica Sinica*, **15**, 665–683 (2005).

[119] A. Gannoun and J. Saracco. An asymptotic theory for SIR_α method. *Statistica Sinica*, **13**, 297–310 (2003).

[120] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer (2003).

[121] K. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327 (1991).

[122] L. Li and H. Li. Dimension reduction methods for micro-arrays with application to censored survival data. *Bioinformatics*, **20 (18)**, 3406–3412 (2004).

[123] L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrics*, **64 (1)**, 124–131 (2008).

[124] J. Saracco. An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and Methods*, **26(9)**, 2141–2171 (1997).

[125] J. Saracco, I. Larramendy, and Y. Aragon. La régression inverse par tranches ou méthodes SIR : présentation générale. *La Revue Modulad*, **22**, 21–39 (1999).

[126] L. Scrucca. *Regularized Sliced Inverse Regression with applications in classification*. Springer-Verlag, Berlin (2006).

[127] L. Scrucca. Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. *Computational Statistics and Data Analysis*, **52**, 438–451 (2007).

[128] C. Vogel. *Computational methods for inverse problems*. Society for Industrial and Applied Mathematics, Philadelphia (2002).

[129] M. Weiss, F. Baret, R. Myneni, A. Pragnère, and Y. Knyazikhin. Investigation of a model inversion technique to estimate canopy biophysical variables from spectral and directional reflectance data. *Agronomie*, **20**, 3–22 (2000).

[130] W. Zhong, P. Zeng, P. Ma, J. Liu, and Y. Zhu. Rsir : Regularized sliced inverse regression for motif discovery. *Bioinformatics*, **21 (22)**, 4169–4175 (2005).

Conclusion générale

Ce mémoire d'Habilitation à Diriger des Recherches présente l'état actuel d'avancement de mes travaux dans les trois thématiques suivantes : inférence sur les lois à queue de type Weibull, lois des valeurs extrêmes conditionnelles et réduction de dimension pour la régression. Le premier thème a fait l'objet de 5 publications dans des revus internationales, les deux autres à 3 publications chacun. Avant de donner quelques directions de recherche possibles permettant de relier ces trois thèmes, je donne un bref résumé des résultats obtenus.

Dans le chapitre 1, nous nous sommes intéressés aux lois à queue de type Weibull. Plus précisément, à partir d'un échantillon de variables aléatoires indépendantes issues d'une loi à queue légère, nous avons proposé plusieurs estimateurs de quantiles extrêmes. Leur validité théorique a été établie par le biais de résultats de normalité asymptotique. De nombreuses simulations (qui ne sont pas toutes présentées dans ce mémoire) ont permis de valider le comportement des estimateurs sur des échantillons de taille finie.

Dans le chapitre 2, nous avons aussi proposé plusieurs estimateurs de quantiles extrêmes mais dans un cadre conditionnel c'est à dire lorsque la variable d'intérêt est mesurée conjointement avec une covariable. Les quantiles extrêmes sont alors fonction de la covariable. Nous nous sommes restreint au cas de lois conditionnelles à queue lourde. Comme dans le chapitre précédent, des résultats de normalité asymptotique ont été établis montrant ainsi le bien fondé théorique de ces estimateurs. Leur utilisation pour l'estimation d'une carte de période de retour a permis de mettre en évidence l'intérêt applicatif de ces estimateurs.

Enfin, dans le chapitre 3, nous avons proposé une régularisation de la méthode SIR en introduisant un a priori Gaussien. Différents choix pour la matrice de variance-covariance de la loi a priori conduisent à des méthodes de régularisation connues ainsi qu'à de nouvelles approches. Le bon comportement des estimateurs régularisés obtenus a été mis en évidence sur des simulations ainsi que sur une application sur données réelles. La méthode SIR régularisée s'est avérée compétitive face à d'autres méthodes de régression comme la méthode SVR, celle des plus proches voisins, etc ...

Voici à présent quelques futurs axes de recherche possibles. Tout d'abord, il serait intéressant d'étendre les définitions des estimateurs de quantiles extrêmes conditionnels aux cas de loi à queue de type Weibull conditionnelle. Ceci permettrait d'appliquer les méthodes développées dans le Chapitre 2 aux estimateurs introduits dans Chapitre 1.
En remarquant que l'indice des valeurs extrêmes conditionnel introduit dans le Chapitre 2 peut

aussi être interprété comme une fonction de régression, il semble envisageable d'utiliser les résultats sur la réduction de dimension obtenus dans le Chapitre 3 pour en faire l'estimation. En effet, dans le cas où la covariable considérée est de grande dimension, il serait pertinent d'en réduire la dimension avant d'étudier le comportement de la queue de distribution de la loi conditionnelle.
Enfin, d'autres domaines de la Statistique pourraient être utilisés pour améliorer les résultats présentés dans ce mémoire. Par exemple, des techniques d'estimation fonctionnelle pourraient être appliquées pour estimer la fonction à variations lentes présente dans l'expression des fonctions de répartition des lois à queue lourde et légère. Ceci permettrait d'utiliser davantage d'observations pour estimer les quantiles extrêmes puisque nous ne serions plus alors contraints d'utiliser uniquement les plus grandes observations pour lesquelles la fonction à variations lentes peut être considérée comme étant constante.