UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I
U.F.R. D'INFORMATIQUE
ET DE MATHÉMATIQUES APPLIQUÉES

Mémoire d'habilitation

présenté par

Stéphane GIRARD

en vue de l'obtention du diplôme d'Habilitation à Diriger des Recherches de l'

UNIVERSITÉ JOSEPH FOURIER

Spécialité Informatique et Mathématiques Appliquées

Contributions à l'inférence statistique semi- et non-paramétrique

Soutenance le 6 juillet 2004 devant le jury composé de

Jean-Noël BACRO	Professeur, Université Montpellier 2	Examinateur
Philippe BESSE	Professeur, Université Université Paul Sabatier	Rapporteur
Gilles CELEUX	Directeur de Recherches, INRIA futurs	Examinateur
Irène GIJBELS	Professeur, Université Catholique de Louvain	Rapporteur
Ivette GOMES	Professeur, Université de Lisbonne	Rapporteur
Anatoli IOUDITSKI	Professeur, Université Joseph Fourier	Examinateur

Habilitation préparée au sein du Laboratoire de Modélisation et Calcul de l'IMAG

Remerciements

Je tiens à remercier Philippe Besse, Irène Gijbels et Ivette Gomes pour avoir bien voulu consacrer une part de leur temps à la lecture critique de ce mémoire et à la rédaction de leur rapport. Mes remerciements s'adressent aussi à Jean-Noël Bacro, Gilles Celeux et Anatoli Iouditski qui m'ont fait l'honneur de participer au jury de cette habilitation.

J'aimerais remercier ici les membres des équipes de recherche qui m'ont accueilli depuis 1993 : le département Systèmes du LETI/CEA, le département Image de l'ENST Paris, le projet IS2 de l'INRIA Rhône-Alpes, le Laboratoire de Probabilités et Statistique de l'Université Montpellier 2 et le Laboratoire de Modélisation et Calcul de l'Université Grenoble 1. Plus particulièrement, j'adresse ma sincère reconnaissance à mes co-auteurs cités tout au long de ce mémoire.

Enfin, je voudrais exprimer ma gratitude à Fabrice Bellet pour m'avoir fourni l'équipement et l'assistance informatiques nécessaires à la rédaction de ce mémoire.

Table des matières

Introduction

Ce mémoire est une synthèse de mon activité de recherche depuis ma thèse débutée en octobre 1993. Les travaux présentés s'inscrivent dans le cadre de l'inférence statistique au sens large. Plus précisément, ils s'articulent autour des thèmes suivants :

- estimation de quantiles extrêmes,
- estimation de frontière,
- réduction de dimension en analyse d'images,
- estimation de courbes de référence.

Mes contributions à ces quatre domaines sont décrites en autant de chapitres, numérotés de 1 à 4, pouvant être lus indépendamment.

Le Chapitre 1 est consacré à l'estimation de quantiles extrêmes. Le quantile x_{p_n} d'ordre p_n d'une variable aléatoire X est le nombre qui a probabilité p_n d'être dépassée : $P(X > x_{p_n}) = p_n$. Dans le cas où $p_n < 1/n$, ce quantile est dit extrême car il est "en général" supérieur à l'observation maximale. L'estimation de tels quantiles nécessite des méthodes semi-paramétriques d'extrapolation au-delà de l'observation maximale faisant le minimum d'hypothèses sur la loi de X.

Le problème abordé dans le Chapitre 2 est l'estimation d'un ensemble D à partir de points disposés aléatoirement dans celui-ci. Le problème n'est pas traité ici dans toute sa généralité mais on se restreint au cas d'ensembles de la forme $D = \{(x,y) : x \in E\ ;\ 0 \leq y \leq f(x)\}$, E étant un sous-ensemble de $\mathbb{R}^d$ connu, et f une fonction de E dans $\mathbb{R}^+$ inconnue, si bien que l'estimation de D se ramène à celle de la fonction frontière f.

Les méthodes de réduction de dimension non-linéaires introduites Chapitre 3 ont été motivées par des applications à l'analyse d'images. Une image peut en effet être représentée par un vecteur de grande dimension et les méthodes de réduction de dimension linéaires sont souvent mal adaptées à représenter les déformations même simples d'une image.

Dans le Chapitre 4 de nouvelles méthodologies pour l'estimation de courbes de références sont présentées. Ces approches sont basées sur une estimation non-paramétrique de quantiles conditionnels précédée si besoin est d'une étape de réduction de dimension de la covariable.

Enfin, je montre dans le dernier chapitre de ce document comment la confrontation de ces domaines de recherche avec d'autres thématiques fait naitre de nouvelles perspectives. Avant cela j'aimerais souligner les liens qui unissent ces recherches.

Tout d'abord, les problématiques des chapitres 1, 3 et 4 sont issues de collaborations avec des industriels, respectivement EDF (électricité de France), le CEA (commissariat à l'énergie atomique) et le CERIES (centre de recherche et d'investigation épidermiques et sensorielles de Chanel). En règle générale, les travaux décrits dans ce mémoire s'étendent de l'étude théorique d'une méthode statistique au développement d'outils logiciels l'implémentant. Dans le cas de l'estimation de quantiles extrêmes et de courbes de référence, les logiciels sont mis à la disposition de tous [32, 50].

Ensuite, les quatre domaines de recherche abordés font appel à des thèmes communs. Ainsi, l'estimation de frontière, de courbes de référence et la réduction de dimension en analyse d'images relèvent toutes trois de l'estimation fonctionnelle, et estimation de quantiles extrêmes ou de courbes de références font appel à des méthodes semi-paramétriques. Notons que ces deux dernières problématiques associées à l'estimation de frontière peuvent être considérées comme différents aspects de l'estimation de quantiles, conditionnels ou non. D'autre part le souci de la réduction de dimension se retrouve à la fois en analyse d'images et dans le cadre de l'estimation de courbes de référence.

Conventions

Les notations suivantes seront utilisées dans ce mémoire :

– Si (A_n) et (B_n) sont deux suites réelles positives, on écrit $A_n \asymp B_n$ lorsque

$$0 < \lim_{n\to\infty} \inf A_n/B_n \leq \lim_{n\to\infty} \sup A_n/B_n < +\infty.$$

$A_n \asymp 0$ s'interprète par $A_n \to 0$. On note $A_n \sim B_n$ si $A_n/B_n \to 1$ quand $n \to \infty$.

– La convergence en loi est notée $\xrightarrow{d}$ et la convergence en probabilité est notée $\xrightarrow{P}$.

Enfin, dans les chapitres suivants, les résultats sont parfois énoncés sous des hypothèses simplifiées afin de ne pas alourdir la présentation.

Chapitre 1

Estimation de quantiles extrêmes

Supposons que l'on dispose de n observations $x_1,\ldots,x_n$ d'une grandeur physique modélisée par une variable aléatoire X. Le quantile x_{p_n} d'ordre p_n de X est la quantité qui a la probabilité p_n d'être dépassée: $P(X > x_{p_n}) = p_n$. Dans le cas où $p_n < 1/n$, ce quantile est dit extrême car il est "en général" supérieur à l'observation maximale. Plus précisément, si $X_{n,n}$ désigne la maximum de n variables aléatoires indépendantes et de même loi que X, alors $np_n \to 0$ implique $P(x_{p_n} > X_{n,n}) \to 1$ quand $n \to \infty$.
Les problèmes d'estimation de quantiles extrêmes se trouvent typiquement en hydrologie: à partir de mesures $x_1,\ldots,x_n$ de débit d'une rivière sur 50 ans, estimer le débit de la crue du siècle. Les problèmes abordés dans ce chapitre trouvent cependant leur motivation dans un domaine différent, la fiabilité. A l'exception du paragraphe 1.4, les résultats obtenus ici ont en effet été acquis dans le cadre d'une collaboration de 7 années entre le projet IS2 de l'INRIA Rhône-Alpes et la direction des études et recherche de EDF. La conclusion de cette collaboration a été le développement d'un logiciel d'étude des queues de distribution.
Nous présentons au paragraphe 1.1 la théorie des valeurs extrêmes, et la notion de domaine d'attraction, qui sont à la base des méthodes développées ici. Les paragraphes 1.2 à 1.4 présentent les méthodes d'estimation des quantiles extrêmes, classées par domaine d'attraction, que nous avons proposées. Un test d'adéquation dédié aux queues de distribution et utilisant la notion de quantile extrême est introduit paragraphe 1.5. Le logiciel EXTREMES où sont implantées quelques unes de ces méthodes est brièvement présenté paragraphe 1.6. Enfin, quelques perspectives sont proposées paragraphe 1.7.

1.1 La théorie des valeurs extrêmes

Soit X une variable aléatoire réelle de fonction de répartition F et de fonction de survie $\bar{F} = 1 - F$. On note $x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}$ le point terminal de F. On introduit également $u \leq x_F$ un réel appelé seuil. L'excès Y de X au delà du seuil u est la variable aléatoire définie par $Y = X - u$ quand $X > u$. Soit $\{X_1,\ldots,X_n\}$ un échantillon de n variables aléatoires indépendantes et de même loi que X et soit $X_{1,n} \leq \cdots \leq X_{n,n}$ les statistiques d'ordre associées.
La théorie des valeurs extrêmes établit deux types de comportement asymptotique. D'une part, le théorème des valeurs extrêmes donne la loi asymptotique de $X_{n,n}$, le maximum de l'échantillon, lorsque n tend vers l'infini. Ce résultat est présenté dans le paragraphe 1.1.1. D'autre part, le

théorème de Pickands donne la loi asymptotique de l'excès Y quand le seuil u tend vers le point terminal x_F. Ce résultat est présenté dans le paragraphe 1.1.2.

1.1.1 Le théorème des valeurs extrêmes

Théorème 1.1.1 *Sous certaines conditions de régularité sur F, il existe $\xi \in \mathbb{R}$ et deux suites réelles $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ $(\beta_n > 0)$ tels que $\forall x \in \mathbb{R}$,*

$$\lim_{n\to\infty} P(\beta_n^{-1}(X_{n,n} - \alpha_n) \leq x) = \lim_{n\to\infty} F^n(\alpha_n + \beta_n x) = H_\xi(x),$$

où H_ξ est la fonction de répartition de la loi des valeurs extrêmes (EVD) :

$$H_\xi(x) = \begin{cases} \exp\left[-(1+\xi x)_+^{-1/\xi}\right] & si\ \xi \neq 0, \quad où\ y_+ = \max(0, y). \\ \exp(-\exp(-x)) & si\ \xi = 0. \end{cases} \tag{1.1}$$

Les conditions de régularité sur F sont décrites dans [19], page 108 et [46], page 54. Elles sont vérifiées pour la plupart des lois usuelles. La loi de fonction de répartition H_ξ est appelée loi des valeurs extrêmes, et le paramètre ξ est appelé indice des valeurs extrêmes. Si F vérifie le Théorème 1.1.1, on dit que F appartient au domaine d'attraction de H_ξ. On distingue alors trois cas :

- si $\xi < 0$, F appartient au domaine d'attraction de Weibull, et l'on note $F \in \mathrm{DA}(\text{Weibull})$,
- si $\xi = 0$, F appartient au domaine d'attraction de Gumbel, et l'on note $F \in \mathrm{DA}(\text{Gumbel})$,
- si $\xi > 0$, F appartient au domaine d'attraction de Fréchet, et l'on note $F \in \mathrm{DA}(\text{Fréchet})$.

Des descriptions de ces trois domaines d'attraction sont proposées dans le paragraphe 1.1.3.

1.1.2 Le théorème de Pickands

La fonction de répartition de l'excès Y au-delà du seuil u est notée F_u, la fonction de survie associée s'écrit pour $x \geq 0$:

$$\bar{F}_u(x) = P(X - u > x | X > u) = \bar{F}(u + x)/\bar{F}(u).$$

Le théorème de Pickands [107] donne une approximation de cette fonction de survie (ou de façon équivalente de la fonction de répartition) lorsque le seuil u est proche du point terminal x_F.

Théorème 1.1.2 *F appartient au domaine d'attraction de H_ξ si et seulement si il existe une fonction σ telle que*

$$\lim_{u\to x_F} \sup_{0<x<x_F - u} \left| \bar{F}_u(x) - \bar{F}^{GPD}_{\xi,\sigma(u)}(x) \right| = 0, \tag{1.2}$$

où $\bar{F}^{GPD}_{\xi,\sigma}$ est la fonction de survie de la loi de Pareto généralisée (GPD) :

$$\bar{F}^{GPD}_{\xi,\sigma}(x) = \begin{cases} (1 + \xi x/\sigma)^{-1/\xi} & si\ \xi \neq 0, \\ \exp(-x/\sigma) & si\ \xi = 0, \end{cases} \tag{1.3}$$

définie pour $x \geq 0$ si $\xi \geq 0$ et $0 \leq x \leq -\sigma/\xi$ sinon.

Remarquons que si $F \in \mathrm{DA}(\text{Gumbel})$, alors $\bar{F}^{\mathrm{GPD}}_{0,\sigma}$, la fonction de survie associée par le théorème de Pickands, est la fonction de survie d'une loi exponentielle d'espérance σ. De plus, si F est elle-même la fonction de répartition d'une loi exponentielle alors l'approximation de Pickands est exacte dans le sens où la loi de l'excès est exponentielle quel que soit le seuil.

1.1.3 Description des domaines d'attraction

Nous rappelons les conditions nécessaires et suffisantes sur une fonction de répartition pour qu'elle appartienne à un domaine d'attraction. Ces caractérisations font appel aux classes de fonctions à variations régulières $\mathcal{RV}_\rho$ et de fonctions à variations régulières lisses $\mathcal{SR}_\rho$ [11], paragraphes 1.4, 1.5, 1.8, 2.3 et 2.4. Dans les deux cas, $\rho \in \mathbb{R}$ est appelé indice de variations régulières, et si $\rho = 0$, on parle de variations lentes (lisses). Les domaines d'attraction de Fréchet et Weibull se caractérisent alors aisément à partir de $\mathcal{RV}_0$.

Théorème 1.1.3

(i) F appartient à DA(Fréchet) avec un indice de valeurs extrêmes $\xi > 0$ *si et seulement si (a)* $x_F = +\infty$ *et (b) il existe* $\ell \in \mathcal{RV}_0$ *tel que* $\bar{F}(x) = x^{-1/\xi}\ell(x)$.

(ii) F appartient à DA(Weibull) avec un indice de valeurs extrêmes $\xi < 0$ *si et seulement si (a)* $x_F < +\infty$ *et (b) il existe* $\ell \in \mathcal{RV}_0$ *tel que* $\bar{F}(x) = (x_F - x)^{-1/\xi}\ell((x_F - x)^{-1})$.

Le domaine d'attraction de Gumbel peut, quant à lui, être décrit à partir des fonctions de type Von-Mises [42], Théorème 3.3.26. Cependant, il n'en existe pas de caractérisation simple. Nous proposons ci-dessous un exemple de classe de lois représentatif de la diversité des lois de ce domaine d'attraction. Ce dernier contient des fonctions de répartition F de point terminal fini ou infini. Ici, nous nous limitons au second cas. Plus précisément, nous définissons la famille suivante de fonctions de répartition, introduite dans [37].

Définition 1.1.1 *Une fonction de répartition F appartient à* $\mathcal{C} \subset$ *DA(Gumbel) si*

(i) F est inversible.

(ii) La fonction $V : x \in \mathbb{R}_*^+ \to V(x) = \bar{F}^{-1}(\exp(-x)) \in \mathbb{R}$ *(fonction de hasard cumulée inverse) appartient à* $\mathcal{C}_\theta^1 \cup \mathcal{C}^2 \cup \mathcal{C}_\theta^3$ *avec*

$\mathcal{C}_\theta^1 = \mathcal{SR}_\theta,\ \theta > 0,\ \theta \neq 1,$

$\mathcal{C}_1^1 = \mathcal{C}_{1,\infty}^1 \cup \mathcal{C}_{1,\tau}^1 = \{V \in \mathcal{SR}_1\ :\ V'' = 0\} \cup \{V \in \mathcal{SR}_1\ :\ |V''| \in \mathcal{SR}_{-1-\tau}\},\ \tau \geq 0,$

$\mathcal{C}^2 = \{V \in \mathcal{SR}_0,\ V' \in \mathcal{SR}_{-1}\},$

$\mathcal{C}_\theta^3 = \{V = \exp g,\ g \in \mathcal{SR}_\theta,\ 0 < \theta < 1\}.$

Pour conclure ce paragraphe les domaines d'attraction associés à quelques lois usuelles sont précisés.

Exemple 1.1.1

- *Domaine d'attraction de Fréchet* $(\xi > 0)$ *: Loi de Burr, loi de Fréchet, loi de Pareto, loi de Student.*
- *Domaine d'attraction de Weibull* $(\xi < 0)$ *: Loi uniforme* $(\xi = -1)$*, loi inverse de Burr.*
- *Domaine d'attraction de Gumbel* $(\xi = 0)$ *: Loi exponentielle* $(\mathcal{C}_{1,\infty}^1)$*, loi gamma* $(\mathcal{C}_{1,1}^1)$*, loi normale* $(\mathcal{C}_{1/2}^1)$*, loi de Weibull* $(\mathcal{C}_{1/\beta}^1$*,* β *étant le paramètre de forme), loi double-exponentielle* $(\mathcal{C}^2)$ *de fonction de survie* $\bar{F}(x) = \exp(-\exp(x))$*, loi lognormale* $(\mathcal{C}_{1/2}^3)$.

1.1.4 Méthode des excès pour l'estimation des quantiles extrêmes

Nous donnons ici les grandes lignes de la méthode des excès, définie dans [15] et basée sur le théorème de Pickands. Rappelons que l'on cherche à estimer le quantile extrême x_{p_n} défini par $\bar{F}(x_{p_n}) = p_n$ avec $0 < p_n < 1/n$. Pour cela, on introduit un second quantile u_n, classique celui-ci, défini par $\bar{F}(u_n) = c_n$, avec $1/n \leq c_n < 1$. L'heuristique de la méthode des excès consiste alors à appliquer l'approximation (1.2) avec $u = u_n$ et $x = x_{p_n} - u_n$ pour obtenir l'approximation $\tilde{x}_{p_n}^{\mathrm{GPD}}$ de x_{p_n} :

$$\tilde{x}_{p_n}^{\mathrm{GPD}} = u_n + (\bar{F}_{\xi,\sigma(u_n)}^{\mathrm{GPD}})^{-1}(p_n/c_n) = u_n - \frac{\sigma(u_n)}{\xi}\left(1 - (c_n/p_n)^{\xi}\right). \tag{1.4}$$

L'estimateur correspondant est obtenu en estimant u_n par $X_{n-k_n+1,n}$ où $k_n = nc_n$ et en remplaçant $\sigma(u_n)$ et ξ par des estimateurs appropriés $\hat{\sigma}_n$ et $\hat{\xi}_n$ construits sur la base des excès ordonnés $\{X_{n-i+1,n} - X_{n-k_n+1,n},\ i = 1,\ldots,k_n - 1\}$. Des exemples sont donnés au paragraphe 1.1.5. On obtient alors :

$$\hat{x}_{p_n}^{\mathrm{GPD}} = X_{n-k_n+1,n} - \frac{\hat{\sigma}_n}{\hat{\xi}_n}\left(1 - (c_n/p_n)^{\hat{\xi}_n}\right). \tag{1.5}$$

Un cas particulier important de cet estimateur est l'estimateur ET (Exponential Tail) introduit dans [15]. Il consiste à supposer que la fonction de répartition F est dans le domaine d'attraction de Gumbel. Dès lors, il suffit de poser $\hat{\xi}_n = 0$ dans (1.5) et d'estimer $\sigma(u_n)$ par la moyenne empirique des excès,

$$\hat{\sigma}_n = \frac{1}{k_n - 1}\sum_{i=1}^{k_n-1}(X_{n-i+1,n} - X_{n-k_n+1,n}), \tag{1.6}$$

pour obtenir l'estimateur ET de x_{p_n} :

$$\hat{x}_{p_n}^{\mathrm{ET}} = X_{n-k_n+1,n} + \hat{\sigma}_n \log(c_n/p_n). \tag{1.7}$$

1.1.5 Estimation des paramètres de la loi GPD

L'estimation des paramètres ξ et σ de la loi GPD à partir d'un échantillon d'excès est un problème délicat en pratique. La principale difficulté provient du fait que, sur un échantillon de n observations, on ne peut disposer que de k_n excès pour réaliser l'estimation (le choix du nombre k_n d'excès est également un problème en lui-même). Dans ce contexte, l'estimateur du maximum de vraisemblance [115] est peu utilisé car d'une part, il pose des problèmes numériques et d'autre part il est peu performant pour des nombres d'excès inférieurs à 500 [28, 81]. Parmi les nombreuses propositions existantes pour pallier ces limitations citons les estimateurs des moments pondérés [87].

1.2 Estimation dans DA(Gumbel)

Notre contribution à l'estimation des quantiles extrêmes dans le domaine d'attraction de Gumbel porte sur deux points. Dans un premier temps, nous avons étudié les propriétés asymptotiques de l'estimateur ET défini en (1.7). Dans un second temps, nous avons proposé un estimateur des quantiles extrêmes dédié à la classe $\mathcal{C}_\theta^1$, $\theta > 0$.

1.2.1 Propriétés asymptotiques de l'estimateur ET

Ce travail a été réalisé en collaboration avec Jean Diebolt (CNRS, Université de Marne-la-Vallée). L'étude des propriétés asymptotiques de l'estimateur ET a été négligée dans la littérature. On trouve cependant dans [29] un résultat donnant la loi asymptotique de la différence $\hat{x}_{p_n}^{\mathrm{ET}} - x_{p_n}$, mais nous avons montré [37] que ce résultat est en général faux. Pour notre part, nous avons tout d'abord étudié le comportement du terme déterministe $\tilde{x}^{\mathrm{ET}}_{p_n} - x_{p_n}$ puis du terme stochastique $\hat{x}^{\mathrm{ET}}_{p_n} - \tilde{x}^{\mathrm{ET}}_{p_n}$, où $\tilde{x}^{\mathrm{ET}}_{p_n}$ est l'approximation du quantile extrême définie par analogie avec (1.4) par

$$\tilde{x}^{\mathrm{ET}}_{p_n} = u_n + (\bar{F}^{\mathrm{GPD}}_{0,\sigma(u_n)})^{-1}(p_n/c_n) = u_n + \sigma(u_n)\log(c_n/p_n). \tag{1.8}$$

1.2.1.1 Etude du terme déterministe

Les détails de cette étude sont parus dans [72]. Nous avons montré que, dans la classe $\mathcal{C}$, l'approximation (1.8) est un cas particulier de

$$\tilde{x}^{\mathrm{ET},k}_{p_n} = u_n + \sum_{i=1}^{k} \frac{\sigma_n^{[i]}}{i!} \log^i(c_n/p_n),$$

où $\sigma_n^{[i]} = V^{(i)}(-\log c_n)$, appelée approximation d'ordre k du quantile x_{p_n}. On vérifie en effet que $\tilde{x}^{\mathrm{ET},1}_{p_n} = \tilde{x}^{\mathrm{ET}}_{p_n}$. On note $\varepsilon_n^{\mathrm{app},k} = (x_{p_n} - \tilde{x}^{\mathrm{ET},k}_{p_n})/x_{p_n}$, l'erreur d'approximation d'ordre k et on introduit la suite de fonctions $K_k(x) = x^k V^{(k)}(x)/V(x)$, $x > 0$, $k \geq 0$. On suppose que les ordres des quantiles peuvent s'écrire $p_n = 1/n^{q+\eta_n}$, $c_n = 1/n^{q'+\eta'_n}$ avec $0 < q' \leq 1 \leq q$, $\eta_n \to 0$, $\eta'_n \to 0$ et $\eta_n \asymp \eta'_n$. De plus, si $q = q' = 1$ alors on suppose que $\eta'_n < 0 < \eta_n$. Enfin, on note $\mathbb{N}_k = \{1, \ldots, k\}$. Le résultat suivant donne des conditions nécessaires et suffisantes pour que l'erreur d'approximation d'ordre k converge vers 0.

Théorème 1.2.1 *Soit $F \in \mathcal{C}$.*

(i) Si $V \in \mathcal{C}^1_{1/\theta}$ et $\theta \in \mathbb{N}_k$ alors $\varepsilon_n^{app,k} \to 0$ pour tout $0 < q' \leq 1 \leq q$.
De plus, si $q \neq q'$ alors $\varepsilon_n^{app,k} \asymp K_{k+1}(\log n)$.

(ii) Si $V \in \mathcal{C}^1_{1/\theta}$ et $\theta \notin \mathbb{N}_k$ alors [$\varepsilon_n^{app,k} \to 0 \Leftrightarrow q = q' = 1$].
De plus, si $q = q' = 1$ alors $\varepsilon_n^{app,k} \sim \dfrac{\theta(\theta-1)\ldots(\theta-k)}{(k+1)!}(\eta_n - \eta'_n)^{k+1}$.

(iii) Si $V \in \mathcal{C}^2$ alors $\varepsilon_n^{app,k} \to 0$ pour tout $0 < q' \leq 1 \leq q$.
De plus, si $q \neq q'$ alors $\varepsilon_n^{app,k} \asymp K_{k+1}(\log n)$.

(iv) Si $V \in \mathcal{C}^3_\theta$ alors [$\varepsilon_n^{app,k} \to 0 \Rightarrow q = q' = 1$].
Inversement, si $q = q' = 1$ et s'il existe $s > 0$ tel que $\eta_n \log^{\theta+s}(n) \to 0$ alors $\varepsilon_n^{app,k} \to 0$ et $\varepsilon_n^{app,k} \asymp (\eta_n - \eta'_n)^{k+1} K_{k+1}(\log n)$.

Ce résultat établit le lien entre les ordres p_n et c_n des quantiles à estimer et la classe à laquelle la loi appartient. Dans les cas (ii) et (iv) la convergence de l'erreur d'approximation vers 0 impose de choisir $p = p' = 1$, ce qui implique $\log(1/p_n)/\log(n) \to 1$. Dans de telles situations, les approximations $\tilde{x}^{\mathrm{ET},k}_{p_n}$ ne sont proches de x_{p_n} que pour des quantiles "peu" extrêmes, c'est à dire proches de l'observation maximale. Considérons le cas $k = 1$, correspondant à la méthode ET. Le Théorème 1.2.1 montre que l'approximation ET est de bonne qualité pour les lois dont la fonction de survie décroit très vite vers 0 (classe $\mathcal{C}^2$), le quantile extrême x_{p_n} peut être approché sans condition sur son ordre p_n. A l'inverse, les lois dont la fonction de survie décroit relativement lentement vers 0 (classe $\mathcal{C}^3_\theta$) demandent de fortes conditions sur l'ordre du quantile p_n afin d'obtenir des approximations acceptables. La classe $\mathcal{C}^1_\theta$, $\theta \neq 1$ représente un cas intermédiaire. Enfin, la classe $\mathcal{C}^1_1$ conduit à des approximations de bonne qualité car les lois considérées sont proches de la loi exponentielle.

Il est intéressant de remarquer que la convergence ou non vers 0 de l'erreur d'approximation $\varepsilon_n^{\mathrm{app},k}$ ne dépend pas de l'ordre de l'approximation k dans les classes $\mathcal{C}^1_{1/\theta}$, $\theta \notin \mathbb{N}$, $\mathcal{C}^2$ et $\mathcal{C}^3_\theta$. Par exemple l'erreur correspondant à l'approximation naïve $\tilde{x}_{p_n}^{\mathrm{ET},0} = u_n$ converge vers 0 sous les mêmes conditions que l'erreur associée à l'approximation ET : $\tilde{x}_{p_n}^{\mathrm{ET},1} = \tilde{x}_{p_n}^{\mathrm{ET}}$. De ce point de vue, l'approximation de Pickands ne modifie pas la nature de la convergence.

1.2.1.2 Etude du terme stochastique

Les détails de cette étude sont publiés dans [37]. Les résultats sont obtenus ici pour une classe de fonctions de répartition plus générale que la précédente.

Définition 1.2.1 *Une fonction de répartition F appartient à $\mathcal{D} \subset DA(Gumbel)$ si*

(i) F est inversible et deux fois dérivable.

(ii) La fonction définie par $A : x \in \mathbb{R}_*^+ \to A(x) = V''/V'(\log x) \in \mathbb{R}$,
où $V(x) = \bar{F}^{-1}(\exp(-x))$ vérifie les conditions suivantes :
$A(x) \to 0$ quand $x \to +\infty$,
A est asymptotiquement de signe constant,
Il existe $\rho \leq 0$ tel que $|A| \in \mathcal{RV}_\rho$.

On a comme annoncé $\mathcal{C} \subset \mathcal{D} \subset$ DA(Gumbel), et le résultat est le suivant:

Théorème 1.2.2 *Soit $F \in \mathcal{D}$ et soit $a : x \in \mathbb{R} \to V''/V'(V^{-1}(x)) \in \mathbb{R}$. Si $k_n \to +\infty$, $c_n \to 0$, $p_n/c_n \to 0$ et $k_n^{1/2}a(u_n) \to 0$ alors*

$$\frac{k_n^{1/2}}{\sigma(u_n)\log(c_n/p_n)}\left(\hat{x}_{p_n}^{ET} - \tilde{x}_{p_n}^{ET}\right) \xrightarrow{d} \mathcal{N}(0{,}1) \text{ quand } n \to \infty.$$

L'application ce théorème aux lois de la classe $\mathcal{C}$ permet d'obtenir des formes plus explicites pour la condition $k_n^{1/2}a(u_n) \to 0$.

Corollaire 1.2.1 *Soit $F \in \mathcal{C}$. Si $k_n \to +\infty$ et $p_n/c_n \to 0$ alors dans les cas suivants :*

(i) $V \in \mathcal{C}^1_\theta \cup \mathcal{C}^2$, $\theta \neq 1$ et $k_n = o((\log n)^2)$,

(ii) $V \in \mathcal{C}^1_{1,\infty}$ et $k_n = o(n)$,

(iii) $V \in \mathcal{C}^1_{1,\tau}$ et $k_n = O((\log n)^{2(1+\tau)-\delta})$ $\forall\delta > 0$ arbitrairement petit,

(iv) $V \in \mathcal{C}^3_\theta$ et $k_n = O((\log n)^{2(1-\theta)-\delta})$ $\forall\delta > 0$ arbitrairement petit,

on a,

$$\frac{k_n^{1/2}}{\sigma(u_n)\log(c_n/p_n)}\left(\hat{x}_{p_n}^{ET} - \tilde{x}_{p_n}^{ET}\right) \xrightarrow{d} \mathcal{N}(0{,}1) \text{ quand } n \to \infty.$$

1.2.2 Cas des lois à queue de type Weibull

Parmi les familles de lois de $\mathcal{C}$, la classe $\mathcal{C}^1_\theta$ est la plus intéressante dans le sens où elle englobe la majorité des lois de DA(Gumbel): Weibull, gamma, normale ... De ce fait, des estimateurs des quantiles extrêmes dédiés à cette famille de lois ont été introduits, par exemple [9, 5, 16, 94]. Plus précisémment, ces estimateurs s'adressent aux lois dont la fonction de survie satisfait l'hypothèse suivante :

(A.1) : $\bar{F}(x) = \exp(-H(x))$, avec $V(t) = H^{-1}(t) = t^\theta \ell(t)$ et $\ell \in \mathcal{RV}_0$.

De telles lois sont appelées lois à queue de type Weibull. Ce sont essentiellement les lois de la classe $\mathcal{C}_\theta^1$ dont les conditions de régularité sur V, l'inverse de la fonction de hasard cumulée, sont assouplies. Le paramètre θ est appelé indice de queue de type Weibull. L'estimation des quantiles extrêmes pour les lois à queue de type Weibull passe alors par l'estimation de θ.

1.2.2.1 Estimation de l'indice de queue de Weibull

Dans [67], nous proposons l'estimateur suivant du paramètre θ :

$$\hat{\theta}_n = \sum_{i=1}^{k_n-1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Bigg/ \sum_{i=1}^{k_n-1} (\log_2(n/i) - \log_2(n/k_n)) \,, \tag{1.9}$$

où $\log_2(t) = \log(\log(t))$, $t > 1$ et (k_n) est une suite d'entiers tels que $1 \leq k_n < n$. Cet estimateur apparaît naturellement si l'on considère la fonction quantile définie par

$$q(t) = \bar{F}^{-1}(t) = V(\log(1/t)) = (\log(1/t))^\theta \ell(\log(1/t)), \tag{1.10}$$

et si l'on remarque que pour s et t proches de 0 que

$$\begin{aligned} \log(q(t)) - \log(q(s)) &= \theta(\log_2(1/t) - \log_2(1/s)) + \log\left(\frac{\ell(\log(1/t))}{\ell(\log(1/s))}\right) \\ &\simeq \theta(\log_2(1/t) - \log_2(1/s)). \end{aligned} \tag{1.11}$$

Cette dernière approximation est justifiée par le fait que ℓ est une fonction à variations lentes. Il est important de remarquer que (1.11) est exacte dans le cas des lois de Weibull où ℓ est constante. Cette propriété n'est en général pas vérifiée pour les autres estimateurs de θ (par exemple [16] ou [5]) ce qui se révèle pénalisant dans la pratique, voir [67] pour une comparaison des différents estimateurs sur simulations. La consistance de $\hat{\theta}_n$ y est établie :

Théorème 1.2.3 *Sous* **(A1)**, *si* $k_n \to \infty$ *et* $k_n/n \to 0$ *alors* $\hat{\theta}_n \xrightarrow{P} \theta$.

La normalité asymptotique requiert l'hypothèse de second ordre habituelle sur la fonction à variations lentes ℓ : il existe $\rho \leq 0$ et $b(x) \to 0$ tels que, uniformément localement en $\lambda \geq 1$ quand $x \to \infty$,

(A.2) : $\log\left(\dfrac{\ell(\lambda x)}{\ell(x)}\right) \sim b(x) K_\rho(\lambda)$,

où $K_\rho(\lambda) = \int_1^\lambda u^{\rho-1} du$. Le paramètre $\rho \leq 0$ contrôle la vitesse de convergence du rapport $\ell(\lambda x)/\ell(x)$ vers 1. La condition **(A2)** est la clé de voute des preuves de normalité asymptotique pour les estimateurs basés sur les valeurs extrêmes. Nous renvoyons à [85] ou [7] pour d'autres utilisations dans des contextes identiques. Notre résultat est alors le suivant :

Théorème 1.2.4 *Sous* **(A1)** *et* **(A2)**, $k_n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0,\theta^2)$, *pour toute suite* (k_n) *telle que*

$$k_n \to \infty,\ k_n^{1/2} b(\log(n/k_n)) \to 0 \text{ et } k_n^{1/2}/\log(n/k_n) \to 0. \tag{1.12}$$

1.2.2.2 Estimation des quantiles extrêmes

Disposant d'un estimateur $\hat{\theta}_n$ de l'indice de queue de Weibull θ, nous proposons dans [57] d'estimer le quantile x_{p_n} par l'estimateur WT (Weibull Tail) :

$$\hat{x}_{p_n}^{\mathrm{WT}} = X_{n-k_n+1,n} \left(\frac{\log(1/p_n)}{\log(1/c_n)} \right)^{\hat{\theta}_n}, \tag{1.13}$$

où, comme précédemment, $c_n = k_n/n$. Là encore, cet estimateur bénéficie d'une justification intuitive. Pour s et t proches de 0, la fonction quantile définie en (1.10) vérifie

$$\frac{q(t)}{q(s)} = \frac{V(\log(1/t))}{V(\log(1/s))} \simeq \left(\frac{\log(1/t)}{\log(1/s)} \right)^{\theta},$$

ce qui permet, de façon similaire à l'approche des excès décrite paragraphe 1.1.4, de remplacer l'estimation d'une quantile extrême par l'estimation d'un quantile classique.
En introduisant $\tau_n = \log(1/p_n)/\log(1/c_n)$, on a les théorèmes asymptotiques suivants.

Théorème 1.2.5 *Sous* **(A1)**, *si* $k_n \to \infty$, $k_n/n \to 0$ *et* $\tau_n \asymp 1$ *alors* $\hat{x}_{p_n}^{WT}/x_{p_n} \xrightarrow{P} 1$ *quand* $n \to \infty$.

Théorème 1.2.6 *Sous* **(A1)** *et* **(A2)**, *si (1.12) est vérifiée et si* $\tau_n \to 1$ *alors,*

$$\frac{\log(1/c_n) k_n^{1/2}}{\log(c_n/p_n)} \left(\frac{\hat{x}_{p_n}^{WT}}{x_{p_n}} - 1 \right) \xrightarrow{d} \mathcal{N}(0,\theta^2).$$

L'étude du comportement de cet estimateur sur simulations est basée sur la notion de pouvoir d'extrapolation introduite dans [41].

1.3 Estimation dans DA(Fréchet)

L'estimation des quantiles extrêmes dans le domaine d'attraction de Fréchet par la méthode des excès (1.5) nécessite l'estimation des deux paramètres ξ et σ de la loi GPD. Les difficultés liées à cette estimation ont été évoquées dans le paragraphe 1.1.5. Nous proposons ici une méthode d'inférence bayésienne pour tenter de dépasser ces difficultés.
Ce travail a été réalisé en collaboration avec Jean Diebolt (CNRS, Université de Marne-la-Vallée), Mhamed-Ali El-Aroui (ISG de Tunis) et Myriam Garrido (ENAC, Université Toulouse 3) dans le cadre de la thèse de cette dernière ([58], Chapitre 3). Une synthèse de la méthode est présentée dans [33].

1.3.1 Estimation bayésienne des paramètres de la loi GPD

Reparamétrisation de la loi GPD. La paramétrisation standard de la fonction de survie $\bar{F}_{\xi,\sigma}^{\mathrm{GPD}}$ de loi GPD décrite en (1.3) est remplacée par une nouvelle paramétrisation mieux adaptée dans le cas $\xi > 0$ qui nous intéresse ici. Deux nouveaux paramètres positifs $\alpha = 1/\xi$ et $\beta = \sigma/\xi$ sont introduits et leur couple est noté $\theta = (\alpha,\beta)$. La fonction de survie ainsi reparamétrée est notée $\bar{F}_{\mathrm{GPD}}(.\,|\,\theta)$ et s'écrit

$$\bar{F}_{\mathrm{GPD}}(y\,|\,\theta) = \left(1 + \frac{y}{\beta}\right)^{-\alpha}, \quad y \geq 0, \tag{1.14}$$

et la densité associée est

$$f_{\text{GPD}}(y\,|\,\theta) = \frac{\alpha}{\beta}\left(1 + \frac{y}{\beta}\right)^{-\alpha-1}, \quad y \geq 0\,. \tag{1.15}$$

Nous supposons disposer de réalisations $\mathbf{y} = (y_1,\ldots,y_k)$ de variables aléatoires indépendantes et identiquement distribuées $Y_1,\ldots,Y_k$ selon (1.14)–(1.15). En pratique, il s'agit d'excès au-delà d'un seuil u de variables aléatoires $X_1,\ldots,X_n$ indépendantes et identiquement distribuées selon une loi appartenant à DA(Fréchet). Le point de départ de cette étude est la représentation de la densité (1.15) par un mélange introduite dans [110], page 157 :

$$f_{\text{GPD}}(y\,|\,\theta) = \int p(y\,|\,z) g(z\,|\,\theta)\,dz\,, \tag{1.16}$$

où $p(.|z)$ est la densité de la loi exponentielle d'espérance $1/z$ définie par $p(y|z) = z\,e^{-yz}\mathbb{I}_{\{y>0\}}$ pour tout $z > 0$ et où $g(.\,|\,\theta)$ est la densité de la loi gamma de couple de paramètres $\theta = (\alpha,\beta)$: $g(z\,|\,\theta) = \beta^{\alpha}(\Gamma(\alpha))^{-1}\,z^{\alpha-1}e^{-\beta z}\,\mathbb{I}_{\{z>0\}}$. La représentation sous forme de mélange (1.16) permet de pallier l'absence de classe conjuguée pour la loi GPD en construisant une classe quasi-conjuguée pour la loi GPD à partir de la classe conjuguée pour la loi gamma.

Classe conjuguée pour la loi gamma. La définition de la classe conjuguée pour la loi gamma repose sur la loi Gamcon de type II [27]. Elle est notée Gamcon II(c, d) où $c > 1$ et $d > 0$ sont deux paramètres. Sa densité s'écrit :

$$\xi_{c,\,d}(x) = I_{c,\,d}^{-1}\,\Gamma(dx + 1)\,(\Gamma(x))^{-d}\,(c\,d)^{-dx}\mathbb{I}_{\{x>0\}}\,, \tag{1.17}$$

où $I_{c,\,d}$ est un coefficient de normalisation. Soit $\mathbf{z} = (z_1,\ldots,z_k)$ un ensemble de k réalisations de variables aléatoires $Z_1,\ldots,Z_k$ indépendantes et de même loi gamma de couple de paramètres $\theta = (\alpha,\beta)$. D'après [27], Théorème 2, la densité a priori conjuguée sur θ avec comme hyper-paramètres $\delta > 0$ et $\eta > \mu > 0$ est donnée par $\pi(\theta) = \pi(\alpha)\pi(\beta\,|\,\alpha)$ où $\pi(\alpha)$ est la densité de la loi Gamcon II de paramètres $c = \eta/\mu$ et $d = \delta$ et $\pi(\beta\,|\,\alpha)$ est la densité de la loi gamma de couple de paramètres $(\delta\alpha + 1, \delta\eta)$. Les densités a posteriori correspondantes sont alors

$$\pi(\theta\,|\,\mathbf{z}) = \pi(\alpha\,|\,\mathbf{z})\pi(\beta\,|\,\alpha, \mathbf{z}) \tag{1.18}$$

avec $\pi(\alpha\,|\,\mathbf{z})$ la densité de la loi Gamcon II de paramètres $c' = \eta'/\mu'$ et $d' = \delta'$, où

$$\delta' = \delta + k\,, \quad \eta' = \frac{\delta\eta + \sum_{i=1}^{k} z_i}{\delta + k} \quad \text{et} \quad \mu' = \mu^{\delta/(\delta+k)}\left(\prod_{i=1}^{k} z_i\right)^{1/(\delta+k)}\,; \tag{1.19}$$

et $\pi(\beta\,|\,\alpha, \mathbf{z})$ la densité de la loi gamma de couple de paramètres $(\delta'\alpha + 1, \delta'\eta')$.

Application à la loi GPD. Nous avons remarqué [33], paragraphe 2.1, que la loi a posteriori de θ sachant $\mathbf{y}$ s'écrit comme un mélange dont la densité est

$$\pi(\theta\,|\,\mathbf{y}) = \int q_{\pi}(\mathbf{z}\,|\,\mathbf{y})\,\pi(\theta\,|\,\mathbf{z})\,d\mathbf{z} \tag{1.20}$$

où $q_\pi(.\,|\,\mathbf{y})$ est la densité définie par

$$q_\pi(\mathbf{z}\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{z})g_\pi(\mathbf{z})}{\int p(\mathbf{y}\,|\,\mathbf{z}')g_\pi(\mathbf{z}')\,d\mathbf{z}'} \text{ et } g_\pi(\mathbf{z}) = \int g(\mathbf{z}\,|\,\theta')\pi(\theta')d\theta',$$

avec les notations

$$g(\mathbf{z}\,|\theta) = \prod_{i=1}^{k} g(z_i\,|\,\theta) \text{ et } p(\mathbf{y}\,|\,\mathbf{z}) = \prod_{i=1}^{k} p(y_i\,|\,z_i).$$

Il apparaît que la densité (1.20) ne peut être calculée explicitement. On a cependant accès aux densités conditionnelles $\pi(\theta\,|\,\mathbf{y},\mathbf{z})$ par (1.18) et $\pi(\mathbf{z}\,|\,\mathbf{y},\theta)$ en remarquant que

$$\pi(\mathbf{z}\,|\,\mathbf{y},\theta) \propto p(\mathbf{y}\,|\,\mathbf{z})g(\mathbf{z}\,|\theta) = \prod_{i=1}^{k} z_i^{\alpha} e^{-(\beta+y_i)z_i}\mathbb{I}_{\{z_i>0\}},$$

et donc que pour $i = 1,\ldots,k$, conditionnellement à θ et y_i, Z_i suit une loi gamma de couple de paramètres $(\alpha+1,\beta+y_i)$. Il est alors possible (voir par exemple [111], Chapitre 5) de simuler des réalisations a posteriori de θ sachant $\mathbf{y}$ grâce à un échantillonneur de Gibbs. Le principe de la $(m+1)$ème itération est le suivant :

1. Simulation des $z_i^{(m+1)}$ de loi Gamma($\alpha^{(m)}+1,\beta^{(m)}+y_i$) ;
2. Simulation de $\alpha^{(m+1)}$ de loi Gamcon II($\eta'/\mu',\delta'$), avec $\delta' = \delta+k$, η' et μ' calculés à partir des $\mathbf{z}^{(m+1)}$ par l'équation (1.19) ;
3. Simulation de $\beta^{(m+1)}$ de loi Gamma($\delta'\alpha^{(m+1)}+1,\delta'\eta'$).

Les nombreux problèmes de mise en œuvre de cet algorithme sont abordés dans [33] : simulation d'une loi Gamcon II, choix des hyperparamètres δ, η et μ, détection du régime stationnaire ...
Soit (α_j,β_j), $j = 1,\ldots,K$ un échantillon de K couples simulés par l'algorithme précédent, une fois le régime stationnaire atteint. Il est alors possible d'estimer le couple (α,β) par la moyenne, le mode ou la médiane empirique de la distribution précédente.

1.3.2 Application à l'estimation des quantiles extrêmes

A partir de l'échantillon (α_j,β_j), $j = 1,\ldots,K$ précédent, on calcule K estimations de x_{p_n} sur la base de la méthode des excès (1.5) :

$$\hat{x}_{p_n}^{\text{GPD},j} = X_{n-k_n+1,n} - \beta_j\left(1-(c_n/p_n)^{1/\alpha_j}\right).$$

De même, de nombreux estimateurs peuvent être calculés à partir de cet ensemble de valeurs [33].

1.4 Estimation dans DA(Weibull)

Les travaux présentés ici sont issus de la thèse de Laurent Gardes [55] co-encadrée par Pierre Jacob (Université de Montpellier 2) et moi-même. L'estimation des quantiles extrêmes dans DA(Weibull) est basée sur la caractérisation du Théorème 1.1.3. En remarquant que $\bar{F}^{-\xi}$ est approximativement linéaire au voisinage du point terminal x_F, on en déduit l'approximation

$$\frac{\bar{F}^{-\xi}(a)-\bar{F}^{-\xi}(b)}{\bar{F}^{-\xi}(c)-\bar{F}^{-\xi}(b)} \simeq \frac{a-b}{c-b}, \tag{1.21}$$

valable pour a, b, et c proches de x_F. En donnant des valeurs bien choisies à ces paramètres, il est alors possible d'estimer l'indice des valeurs extrêmes ξ (paragraphe 1.4.1) puis des quantiles extrêmes (paragraphe 1.4.2).

1.4.1 Estimation de l'indice des valeurs extrêmes

Nous donnons dans ce paragraphe deux exemples d'estimateurs de ξ qu'il est possible d'obtenir sur la base de l'approximation (1.21). Dans les deux cas, on pose $a = u_n X_{n,n}$, $b = X_{n,n}$ et $c = v_n X_{n,n}$ où (u_n) et (v_n) sont deux suites de l'intervalle]0,1[. On approche alors $\bar{F}(u_n)$ par la variable aléatoire τ_{u_n}/n avec

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

De même, on approche $\bar{F}(v_n)$ par τ_{v_n}/n où τ_{v_n} est défini similairement. Nous proposons deux approximations possibles de $\bar{F}(X_{n,n})$ donnant lieu à deux estimateurs de ξ.

1.4.1.1 Un estimateur à double seuil explicite

Dans ce paragraphe, on approche $\bar{F}(X_{n,n})$ par 0. La validité de l'approximation déduite de (1.21) qui en résulte est garantie par [55], Théorème 3.2. Sous certaines conditions sur (u_n) et (v_n), on a en effet

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

L'estimateur qui en résulte est le suivant :

$$\hat{\xi}_{1,n} = -\frac{\log(1 - u_n) - \log(1 - v_n)}{\log(\tau_{u_n}) - \log(\tau_{v_n})}.$$

Ses propriétés théoriques sont établies dans [55], Chapitre 3. En particulier, il est montré que $\hat{\xi}_{1,n}$ converge en probabilité vers ξ pour tout $\xi < 0$ et converge en loi si $\xi < -1/2$. Le résultat (publié dans [54], Théorème 2.2) est le suivant :

$$(n\bar{F}(u_n x_F))^{1/2}(\hat{\xi}_{1,n} - \xi) \xrightarrow{d} \mathcal{N}\left(0, \frac{(1 - c^{-1/\xi})\xi^4}{\log^2(c)}\right), \tag{1.22}$$

où c est une constante appartenant à l'intervalle]0,1[. Ce résultat est obtenu au prix d'une condition de second ordre de type **(A2)** et de contraintes sur les suites (u_n) et (v_n). Cependant, il est remarquable que la vitesse de la convergence (1.22) soit en puissance de n. De plus la convergence presque sûre est établie pour $\xi < -1$. Néanmoins, l'intérêt de cet estimateur est principalement théorique car ses performances sur simulations sont médiocres. L'estimateur $\hat{\xi}_{1,n}$ souffre en effet d'un important biais systématique dû sans doute à l'approximation grossière de $\bar{F}(X_{n,n})$.

1.4.1.2 Un estimateur à double seuil implicite

Cette limitation pratique est surmontée en approchant $\bar{F}(X_{n,n})$ par $1/n$. La validité de l'approximation déduite de (1.21) qui en résulte est garantie par [55], Théorème 3.4. Sous certaines conditions sur (u_n) et (v_n), on a en effet

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

L'estimateur $\hat{\xi}_{3,n}$ résultant de cette approximation est défini comme la racine en θ de l'équation :

$$\left(\frac{1-v_n}{1-u_n}\right)\left(\frac{\tau_{u_n}^{-\theta}-1}{\tau_{v_n}^{-\theta}-1}\right) = 1. \tag{1.23}$$

Il n'est pas possible de tirer de cette équation une formulation explicite pour $\hat{\xi}_{3,n}$. Néanmoins, on prouve que, sous certaines conditions sur les suites (u_n) et (v_n), l'équation (1.23) admet une unique solution avec une probabilité qui tend vers 1 quand n tend vers l'infini. La consistance faible de $\hat{\xi}_{3,n}$ est également établie, voir [55], Théorème 3.7. L'obtention d'autres propriétés asymptotiques pour cet estimateur est actuellement à l'étude. Son comportement sur simulations est satisfaisant.

1.4.2 Estimation des quantiles extrêmes

Laurent Gardes [55], paragraphe 4.2, montre que sur la base de l'approximation (1.21) et d'un estimateur $\hat{\xi}_n$ consistant de ξ, il est possible de construire un estimateur du quantile extrême x_{p_n}. Pour cela, en considérant $\xi = \hat{\xi}_n$, $a = u_n X_{n,n}$, $b = x_{p_n}$ et $c = v_n X_{n,n}$ dans (1.21), on dispose d'une nouvelle approximation impliquant le quantile recherché. La validité de cette approximation est prouvée dans [55], Lemme 4.1 avec la convergence

$$\left(\frac{x_{p_n}-v_nX_{n,n}}{x_{p_n}-u_nX_{n,n}}\right)\left(\frac{\tau_{u_n}^{-\hat{\xi}_n}-(np_n)^{-\hat{\xi}_n}}{\tau_{v_n}^{-\hat{\xi}_n}-(np_n)^{-\hat{\xi}_n}}\right) \xrightarrow{P} 1.$$

On déduit alors de cette approximation l'estimateur suivant de x_{p_n} :

$$\hat{x}_{p_n}^{\mathrm{DAW}}(\hat{\xi}_n) = u_nX_{n,n} - \frac{\lambda_n(\hat{\xi}_n)}{\hat{\xi}_n}\left(1-(\tau_{u_n}/np_n)^{\hat{\xi}_n}\right),$$

où l'on a défini

$$\lambda_n(\hat{\xi}_n) = \hat{\xi}_n X_{n,n}\frac{u_n-v_n}{\tau_{u_n}^{-\hat{\xi}_n}-\tau_{v_n}^{-\hat{\xi}_n}}.$$

L'estimateur du quantile extrême ainsi défini est similaire à l'estimateur GPD (1.5). Il fait intervenir un seuil aléatoire $u_nX_{n,n}$ et le pourcentage de points τ_{u_n}/n dépassant ce seuil. Notons qu'ici ce pourcentage est aléatoire alors qu'il est déterministe (c_n) dans (1.5). On montre dans [55], Théorème 4.2, que $\hat{x}_{p_n}^{\mathrm{DAW}}(\hat{\xi}_n)$ est consistant dès que $\hat{\xi}_n$ est un estimateur consistant de ξ. Il apparaît sur simulations que l'estimateur $\hat{x}_{p_n}^{\mathrm{DAW}}(\hat{\xi}_{3,n})$ bénéficie de bonnes performances.

1.5 Tests de queues de distribution

La théorie des valeurs extrêmes, telle qu'elle est exposée dans le paragraphe 1.1.1, permet, grâce à un modèle semi-paramétrique, d'extrapoler la comportement de la queue de distribution au-delà de l'observation maximale à partir des plus grandes valeurs de l'échantillon. De ce fait, les estimateurs qui en découlent n'utilisent qu'une petite partie de l'information présente dans l'échantillon (par exemple les k_n excès) et sont peu efficaces lorsque le nombre de données est petit ou modéré. Pour cette raison, dans des situations concrètes, issues de la fiabilité par exemple, il est plus intéressant de disposer d'un modèle paramétrique construit sur l'ensemble des données. Un tel modèle présente de plus l'intérêt d'être interprétable pour les ingénieurs et d'être disponible dans la majorité des logiciels de fiabilité. Les tests d'adéquation classiques (Cramer von Mises, Anderson Darling, ...) permettent de sélectionner un tel modèle. Cependant, ces procédures testent essentiellement l'adéquation d'un modèle à la partie centrale des données. Les dangers causés par l'extrapolation des résultats de ces tests aux queues de distribution sont décrits dans [38] et [82]. Pour cela, nous avons développé une procédure permettant de vérifier l'adéquation d'un modèle paramétrique aux valeurs extrêmes de l'échantillon. Son principe général est décrit paragraphe 1.5.1 et plusieurs variantes en sont présentées paragraphes 1.5.2 et 1.5.3.

1.5.1 Principe des tests de queue de distribution

Supposons qu'un test d'adéquation usuel ne rejette pas l'hypothèse nulle $\mathcal{H}_0$ selon laquelle F appartient à la famille de modèles paramétriques $\{F_\theta : \theta \in \Theta\}$. Le but du test est de contrôler l'adéquation de la queue de $F_{\widehat{\theta}_n}$, où $\widehat{\theta}_n$ est par exemple l'estimateur du maximum de vraisemblance de θ, aux plus grandes données et de vérifier si cette queue de distribution permet des extrapolations raisonnables au-delà de l'observation maximale. Il s'agit donc de tester $\mathcal{H}_0 : F \in \{F_\theta : \theta \in \Theta\}$ contre $\mathcal{H}_1 : F \notin \{F_\theta : \theta \in \Theta\}$ en queue de distribution. Le principe du test est de comparer deux estimateurs du quantile extrême x_{p_n} sous $\mathcal{H}_0$. Le premier est l'estimateur paramétrique du quantile $\hat{x}_{p_n}^{\text{param}} = \bar{F}_{\widehat{\theta}_n}^{-1}(p_n)$. Le second est l'estimateur GPD introduit en (1.5) :

$$\hat{x}_{p_n}^{\text{GPD}} = X_{n-k_n+1,n} - \frac{\hat{\sigma}_n}{\hat{\xi}_n}\left(1 - (c_n/p_n)^{\hat{\xi}_n}\right).$$

Plus précisément, nous construisons sous $\mathcal{H}_0$ un intervalle de confiance IC_α de niveau α pour la différence entre les deux quantiles estimés et nous rejetons l'hypothèse nulle si $\hat{x}_{p_n}^{\text{GPD}} - \hat{x}_{p_n}^{\text{param}} \notin IC_\alpha$. Plusieurs versions du test peuvent être déclinées selon le principe de construction de l'intervalle IC_α et les estimateurs $\hat{\sigma}_n$ et $\hat{\xi}_n$ utilisés pour calculer $\hat{x}_{p_n}^{\text{GPD}}$.

1.5.2 Le test ET

Dans le cas où les lois F_θ considérées dans l'hypothèse nulle appartiennent à DA(Gumbel), il suffit de choisir $\hat{\xi}_n = 0$ et $\hat{\sigma}_n$ défini par (1.6). On a alors $\hat{x}_{p_n}^{\text{GPD}} = \hat{x}_{p_n}^{\text{ET}}$ (voir équation (1.7)), et le test obtenu est appelé test ET. Chronologiquement, il s'agit du premier test que nous avons mis en œuvre. Son principe a été défini en collaboration avec Jean Diebolt [36] et des développements ont été proposés dans la thèse de Myriam Garrido [58], Chapitre 1. En particulier, plusieurs versions du test ET ont été introduites.

Le test ET asymptotique. Sur la base du Théorème 1.2.2, nous avons construit un intervalle de confiance IC_α asymptotique. Le comportement asymptotique du niveau et de la puissance de ce test a été établi [35] dans le cas d'hypothèses simples et de lois appartenant à la classe $\mathcal{C}$. Néanmoins, l'intérêt de cette version du test est uniquement théorique. En effet, la convergence en loi énoncée dans le théorème est très lente, et de ce fait le test ET asymptotique est peu puissant pour des échantillons de petite taille.

Le test ET “bootstrap” paramétrique. Pour pallier cette limitation, nous avons développé une version du test ET basée sur une évaluation par bootstrap paramétrique de l'intervalle IC_α. Nous avons vérifié sur simulations que le test ainsi construit bénéficie d'une bonne puissance [34]. En contrepartie, cette version du test est très coûteuse en temps de calcul. Pour cette raison, nous avons également proposé une version “bootstrap” paramétrique simplifiée. Constatant que les fluctuations de $\hat{x}_{p_n}^{\mathrm{ET}}$ sont de l'ordre de $k_n^{-1/2}$ et celles de $\hat{x}_{p_n}^{\mathrm{param}}$ de l'ordre de $n^{-1/2}$, nous avons négligé ces dernières en ne “bootstrappant” pas l'estimateur paramétrique. Le gain de temps est appréciable lorsque l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ est lourd à calculer, et la perte de puissance peu importante.

1.5.3 Le test GPD

Le principe du test GPD est celui décrit dans le paragraphe 1.5.1. Le choix des estimateurs $\hat{\xi}_n$ et $\hat{\sigma}_n$ pour le calcul de $\hat{x}_{p_n}^{\mathrm{GPD}}$ a été étudié dans le cadre du stage de DEA de Abdelhak Imoussaten [89]. Pour des raisons aussi bien théoriques (invariance des estimateurs par rapport aux paramètres de position et d'échelle) que pratiques (bonne puissance expérimentale), il est apparu que les estimateurs des moments pondérés (décrits au paragraphe 1.1.5) étaient bien adaptés. Nous avons également choisi de ne pas développer de version asymptotique du test GPD, toujours pour des raisons de puissance. Seules les versions bootstrap et boostrap simplifiée ont été proposées.

1.6 Le logiciel EXTREMES

Le logiciel EXTREMES a été écrit par Jérôme Ecarnot, sous la direction conjointe de Jean Diebolt et moi-même et dans le cadre d'un contrat entre EDF et le projet IS2 de l'INRIA Rhône-Alpes. Ce logiciel regroupe quelques unes des différentes méthodes de modélisation de queues de distribution et d'estimation de quantiles extrêmes décrites dans ce chapitre. Par exemple, on y trouve plusieurs fonctions pour estimer l'indice des valeurs extrêmes ξ ou les paramètres de la loi GPD (voir paragraphe 1.1.5). Les procédures développées dans le cadre de la thèse de Myriam Garrido y sont également présentes. Par exemple, les tests de queue de distribution décrits au paragraphe 1.5 sont implémentés. Ce logiciel, écrit en C++ avec une interface Matlab se veut un outil convivial pour expérimenter graphiquement les procédures de statistique des extrêmes. Il est disponible librement à l'adresse suivante : `http://www.inrialpes.fr/is2/pub/software/EXTREMES` accompagné d'une documentation complète. Un descriptif résumé du logiciel est également disponible [32].

1.7 Perspectives

Je travaille actuellement à l'extension à tous les domaines d'attraction de l'estimateur à double seuil implicite $\hat{\xi}_{3,n}$, dédié au DA(Weibull). Le travail mené en collaboration avec Laurent Gardes consiste à remplacer les seuils déterministes u_n et v_n par des seuils aléatoires $X_{n-k'_n+1,n}/X_{n,n}$ et $X_{n-k_n+1,n}/X_{n,n}$ où (k_n) et (k'_n) sont des suites déterministes tendant vers l'infini moins vite que n. Le nouvel estimateur de ξ est obtenu en résolvant en θ l'équation similaire à (1.23) :

$$\left(\frac{X_{n,n} - X_{n-k_n+1,n}}{X_{n,n} - X_{n-k'_n+1,n}}\right)\left(\frac{{k'_n}^{-\theta} - 1}{{k_n}^{-\theta} - 1}\right) = 1.$$

La consistance faible de cet estimateur est établie dans [56], Théorème 1, quel que soit le signe de ξ, c'est à dire quel que soit le domaine d'attraction de la loi des observations. Nous avons également montré sous certaines conditions (voir [56], Théorème 2) que la loi limite de l'estimateur convenablement renormalisé est gaussienne si $\xi < -1/2$ et une loi des valeurs extrêmes si $\xi > -1/2$. Ce résultat laisse à penser que la limitation du résultat (1.22) à $\xi < -1/2$ est bien une nécessité et non une facilité de calcul.
A court terme, nous projetons de construire un nouvel estimateur des quantiles extrêmes basé sur cet estimateur de ξ.
D'autre part, le test GPD présenté paragraphe 1.5.3 mérite d'être amélioré. Il est en effet souhaitable d'adapter l'estimateur de ξ utilisé suivant le domaine d'attraction où se situent les lois F_θ considérées dans l'hypothèse nulle. Par exemple, si F_θ est un ensemble de lois de type Weibull, il semble préférable d'utiliser l'estimateur ad hoc présenté paragraphe 1.2.2 plutôt qu'un estimateur "généraliste". Cette intuition doit être validée sur simulations et, pour cela, le logiciel EXTREMES est un outil précieux.
A plus long terme, je souhaite développer une correction du biais pour les estimateurs proposés dans ce chapitre. Sous une hypothèse de second ordre de type **(A2)**, le terme dominant du biais des estimateurs est dû à la fonction b. Cette fonction peut alors être estimée [7, 102] dans le cadre d'une modèle de régression pour corriger l'estimateur. Cette démarche doit permettre également de simplifier la sélection du paramètre k_n optimal. Il s'agit d'une seconde direction de recherche que je souhaite développer en m'appuyant sur les travaux existants pour l'estimateur de Hill [39, 80].
Enfin, l'estimation de frontière présentée au chapitre suivant est un prolongement naturel de la problématique d'estimation des quantiles extrêmes par l'introduction d'une covariable.

Chapitre 2

Estimation de frontière

Le problème abordé dans ce chapitre est l'estimation d'un ensemble D borné de $\mathbb{R}^{d+1}$, $d \geq 1$, à partir d'un ensemble N_n de points disposés aléatoirement dans celui-ci. Nous ne traitons pas ici le problème dans toute sa généralité mais nous nous plaçons dans le cas particulier où D est de la forme $D = \{(x,y) : x \in E \; ; \; 0 \leq y \leq f(x)\}$, E étant un sous-ensemble de $\mathbb{R}^d$ connu, et f une fonction de E dans $\mathbb{R}^+$. L'estimation de D se ramène donc à l'estimation de f appelée fonction frontière. Le problème ainsi posé a été introduit initialement par Geffroy [59]. Depuis, des applications se sont développées en traitement d'images [96, 101] et en économétrie. Dans ce dernier cas, f est appelée la frontière de production optimale. Des estimateurs sont proposés selon que f est supposée croissante (estimateur FDH, Free Disposal Hull [31]) ou concave et croissante (estimateur DEA, Data Envelopment Analysis [43, 63]). Les estimateurs étudiés ici ne nécessitent pas d'hypothèse de monocité sur f. Nous étudions leurs propriétés dans les deux cadres suivants :

(a) N_n est un échantillon de points $(X_i,Y_i) \in E \times \mathbb{R}^+$, $i = 1,\dots,n$ indépendants et identiquement distribués sur D selon une densité ϕ.

(b) N_n est un processus de Poisson de mesure moyenne $nc\nu \otimes \lambda \mathbb{I}_D$, c est une constante strictement positive, λ désigne la mesure de Lebesgue sur $\mathbb{R}$, et ν une mesure absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. On note dans ce cas $\{(X_i,Y_i),\; i \geq 1\} \subset E \times \mathbb{R}^+$ l'ensemble des points associés.

Nous nous focalisons ici sur la description de la loi asymptotique d'estimateurs non-paramétriques de f lorsque n tend vers l'infini. Par commodité, nous considérons parfois les cas particuliers $E = [0,1]$, ϕ constante sur D dans le cas (a) et $E = [0,1]$, $\nu = \lambda$ dans le cas (b).
Nous présentons dans le paragraphe 2.1 les estimateurs de la littérature qui ont inspirés nos travaux. Le paragraphe 2.2 est dédidé à la présentation des estimateurs que nous avons proposés demandant une partition de D. Cette partition préalable est abandonnée dans le paragraphe 2.3 avec l'introduction des estimateurs par programmation linéaire. Les perspectives sont évoquées au paragraphe 2.4.

2.1 Nos points de départ

Nos travaux se sont appuyés sur deux contributions à l'estimation de frontière. Outre le fait de se placer dans le cas $E = [0,1]$, leur point commun est l'introduction d'une partition du support D en k_n cellules $D_{n,r} = \{(x,y), x \in I_{n,r}, 0 \leq y \leq f(x)\}$ où $\{I_{n,r}, r = 1,\dots,k_n\}$ est une partition régulière de $E = [0,1]$. La suite (k_n), supposée tendre vers l'infini avec n, est un paramètre commun aux deux méthodes d'estimation. La première d'entre elles, qui est aussi la première à notre connaissance dans le domaine, est due à Geffroy [59]. L'estimateur proposé est construit à partir des k_n points les plus hauts dans chaque cellule $D_{n,r}$. Ses propriétés, détaillées dans le paragraphe 2.1.1, relèvent de la théorie des valeurs extrêmes. La seconde méthode que nous avons considérée a été proposée par Jacob & Suquet [90]. La fonction f inconnue est décomposée en séries orthogonales dont les coefficients sont estimés à partir des nombres de points dans chaque cellule (paragraphe 2.1.2).

2.1.1 L'estimateur de Geffroy

Ce paragraphe entre dans le cadre (a) avec $E = [0,1]$, l'ensemble N_n considéré est un échantillon de points (X_i,Y_i), $i = 1,\ldots,n$ indépendants et identiquement distribués selon une densité ϕ. Les valeurs extrêmes de l'échantillon sont définies par $Y^*_{n,r} = \max\{Y_i, (X_i,Y_i) \in D_{n,r}\}$, où l'on a posé $\max\{\emptyset\} = 0$. L'estimateur de Geffroy [59] est alors la fonction constante par morceaux définie par

$$\hat{f}_n^{\mathrm{G}}(x) = \sum_{r=1}^{k_n} \mathbb{I}\{x \in I_{n,r}\} Y^*_{n,r}. \tag{2.1}$$

Dans ce même article, il est montré sous diverses hypothèses sur les fonctions f et ϕ et sur la suite (k_n) que la distance L_∞ définie par

$$d_\infty(\hat{f}_n^{\mathrm{G}},f) = \sup_{x\in[0,1]} \left|\hat{f}_n^{\mathrm{G}}(x) - f(x)\right| \tag{2.2}$$

convenablement normalisée converge en loi vers une loi de Gumbel de fonction de répartition H_0 définie en (1.1). Plus récemment, Korostelev & Tsybakov [96] ont inclu l'estimateur (2.1) dans la famille plus générale des estimateurs polynômiaux par morceaux

$$\hat{f}_n^{\mathrm{KT}}(x;\theta) = \sum_{r=1}^{k_n} \mathbb{I}\{x \in I_{n,r}\} P^*_{n,r}(x;\theta), \tag{2.3}$$

où, sur chaque intervalle $I_{n,r}$, $P^*_{n,r}(.;\theta)$ est le polynôme de degré θ englobant tous les points et d'aire minimum. En d'autres termes, il s'agit de résoudre le problème d'optimisation sous contraintes suivant :

$$\min \int_{I_{n,r}} P_{n,r}(x;\theta)dx \ \text{ s.c. } \ P_{n,r}(X_i;\theta) \geq Y_i, \ X_i \in I_{n,r}. \tag{2.4}$$

En particulier, on a $\hat{f}_n^{\mathrm{G}}(x) = \hat{f}_n^{\mathrm{KT}}(x;0)$. On remarque alors que d'après [96], Théorème 4.1.1, l'estimateur $\hat{f}_n^{\mathrm{G}}$ est minimax si ϕ est constante et si f est 1-lipschitzienne dans les cadres suivants

(i) $k_n = (n/\log n)^{1/2}$ et pour la distance L_∞ définie en (2.2),

(ii) $k_n = n^{1/2}$ et pour la distance L_1 définie par

$$d_1(\hat{f}_n^{\mathrm{G}},f) = \int_0^1 \left|\hat{f}_n^{\mathrm{G}}(x) - f(x)\right| dx. \tag{2.5}$$

Afin de compléter ces résultats, nous avons montré [61] avec Pierre Jacob (Université Montpellier 2) et Jean Geffroy la normalité asymptotique de l'erreur L_1 convenablement normalisée :

Théorème 2.1.1 *Si ϕ est constante et si f est α-lipschitzienne ($0 < \alpha \leq 1$), $k_n = o(n/\log n)$, $n = o(k_n^{1+\alpha})$, alors*

$$\frac{nc}{k_n^{1/2}} \left(d_1(\hat{f}_n^G,f) - \mathbb{E}\left[d_1(\hat{f}_n^G,f)\right]\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{2.6}$$

Pour prouver (2.6), nous avons déjà établi ce résultat lorsque N_n est un processus de Poisson [77], puis nous avons utilisé la technique de Geffroy [60] pour étendre le résultat au cas de l'échantillon. En renforçant les conditions sur la suite (k_n), la convergence (2.6) est conservée si on remplace l'espérance $\mathbb{E}[d_1(\hat{f}_n^{\mathrm{G}},f)]$ par $k_n/(nc)$, voir [61], Corollaire 2. De même, la constante c peut être remplacée par l'estimateur

$$\hat{c}_n^{\mathrm{G}} = k_n \Bigg/ \sum_{r=1}^{k_n} Y_{n,r}^* \ ,$$

sans perturber la convergence en loi, [61], Corollaire 3.

2.1.2 L'estimateur de Jacob & Suquet

Jacob & Suquet [90] ont adapté la méthode de projection sur séries orthogonales, déjà connue en estimation de la densité ou en régression, au cas de l'estimation de frontière. Ces travaux entrent dans le cadre (b) simplifié où $E = [0,1]$ et $\nu = \lambda$ de sorte que N_n est un processus de Poisson homogène de mesure moyenne $nc\lambda_2 \mathbb{I}_D$, c étant une constante strictement positive et λ_2 désignant la mesure de Lebesgue sur $\mathbb{R}^2$. Le principe est le suivant. Soit (b_n) une suite d'entiers qui tend vers l'infini et soit $(e_\ell)_{\ell\in\mathbb{N}}$ une base orthonormée de L^2. Le développement de f sur la base est tronqué au (b_n+1)ème terme et chaque coefficient

$$a_\ell = \int_0^1 f(t)e_\ell(t)dt = \sum_{r=1}^{k_n} \int_{I_{n,r}} f(t)e_\ell(t)dt \tag{2.7}$$

est approché par

$$a_{\ell,k_n} = \sum_{r=1}^{k_n} e_\ell(x_r) \int_{I_{n,r}} f(t)dt = \sum_{r=1}^{k_n} e_\ell(x_{n,r})\lambda_2(D_{n,r}), \tag{2.8}$$

où $x_{n,r}$ est le centre de $I_{n,r}$. En introduisant $N_{n,r} = \sharp\{Y_i, (X_i,Y_i) \in D_{n,r}\}$, et en remarquant que $\mathbb{E}[N_{n,r}] = nc\lambda_2(D_{n,r})$, les auteurs proposent d'estimer a_{ℓ,k_n} par

$$\hat{a}_{\ell,k_n}^{\mathrm{JS}} = \sum_{r=1}^{k_n} e_\ell(x_{n,r}) \frac{N_{n,r}}{nc}. \tag{2.9}$$

L'estimateur de la frontière s'écrit donc

$$\hat{f}_n^{\mathrm{JS}}(x) = \sum_{r=1}^{k_n} K_n(x,x_{n,r}) \frac{N_{n,r}}{nc}, \tag{2.10}$$

avec K_n désignant le noyau de Dirichlet d'ordre b_n associé à la base (e_ℓ) défini par

$$K_n(x,y) = \sum_{\ell=0}^{b_n} e_\ell(x)e_\ell(y),\ (x,y) \in [0,1]^2.$$

Les auteurs étudient alors différents types de convergence de $\hat{f}_n^{\mathrm{JS}}$ vers f sous des hypothèses générales sur le noyau de Dirichlet associé à une base C^1. La normalité asymptotique ponctuelle de l'estimateur centré sur son espérance $\hat{f}_n^{\mathrm{JS}}(x) - \mathbb{E}[\hat{f}_n^{\mathrm{JS}}(x)]$ et convenablement normalisé est également établie. Ces résultats sont illustrés sur l'exemple de la base trigonométrique. Notons que les auteurs établissent en parallèle les mêmes résultats pour la base de Haar. Néanmoins, les estimateurs de type (2.10) présentent l'inconvénient de nécessiter la connaissance du coefficient c pour leur mise en œuvre, ce qui n'est pas le cas de l'estimateur de Geffroy (2.1). Cette constatation est à la base des travaux décrits dans le paragraphe suivant.

2.2 Estimation à partir de partitions

L'ensemble ce paragraphe se situe dans le cadre (b), c'est à dire que N_n est un processus de Poisson.

2.2.1 Estimation par projection

Les travaux décrits dans ce paragraphe ont été menés en collaboration avec Pierre Jacob. Nous montrons tout d'abord dans le paragraphe 2.2.1.1 comment adapter l'estimateur $\hat{f}_n^{\mathrm{JS}}$ au cas c inconnu en utilisant les valeurs extrêmes du processus de façon similaire à $\hat{f}_n^{\mathrm{G}}$. Le cas des bases non orthogonales est abordé avec l'exemple de la base de Faber-Shauder dans le cadre de la thèse de Laurent Gardes, le principe d'estimation étant décrit ici paragraphe 2.2.1.2. L'ensemble de ces travaux suppose $E = [0,1]$ et $\nu = \lambda$.

2.2.1.1 Projection sur une base orthogonale

Le principe d'estimation des coefficients (a_ℓ) consiste à reprendre l'approximation (2.8) dans laquelle $\lambda_2(D_{n,r})$ est estimée par l'aire d'un rectangle de base $I_{n,r}$ et de hauteur $[0,Y^*_{n,r}]$:

$$\hat{a}^{\mathrm{GJ}}_{\ell,k_n} = \sum_{r=1}^{k_n} e_\ell(x_{n,r})\frac{Y^*_{n,r}}{k_n}. \tag{2.11}$$

L'estimateur de la frontière s'écrit donc

$$\hat{f}_n^{\mathrm{Pr}}(x) = \sum_{r=1}^{k_n} K_n(x,x_{n,r})\frac{Y^*_{n,r}}{k_n}, \tag{2.12}$$

et ne nécessite pas la connaissance de c. L'inconvénient de ce type d'estimateurs provient du fait que $Y^*_{n,r}/k_n$ est un estimateur de $\lambda_2(D_{n,r})$ biaisé inférieurement comme le montre le développement suivant :

$$\mathbb{E}\left[\frac{Y^*_{n,r}}{k_n}\right] = \lambda_2(D_{n,r}) - \frac{1}{nc} + o\left(\frac{n}{k_n^4}\right),$$

établi dans [76], Lemme 2, dans le cas où f est une fonction C^1, et sous les conditions $n = o(k_n^2)$ et $k_n = o(n/\log n)$. Il est cependant possible de réduire ce biais en modifiant l'estimateur (2.11) de façon à éliminer le facteur $1/(nc)$:

$$\hat{a}^{\mathrm{GJ},2}_{\ell,k_n} = \sum_{r=1}^{k_n} e_\ell(x_{n,r})\frac{Y^*_{n,r} + Z_n}{k_n}, \tag{2.13}$$

où Z_n est la variable aléatoire définie par

$$Z_n = \frac{1}{k_n} \sum_{r=1}^{k_n} Z^*_{n,r}, \tag{2.14}$$

avec $Z^*_{n,r} = \min\{Y_i, (X_i,Y_i) \in D_{n,r}\}$. Cette correction de biais est justifiée par le fait que $Z^*_{n,r}/k_n$ a un comportement symétrique à $Y^*_{n,r}/k_n$. L'estimateur de la frontière correspondant est alors

$$\hat{f}_n^{\mathrm{Pr},2}(x) = \sum_{r=1}^{k_n} K_n(x,x_{n,r}) \frac{Y^*_{n,r} + Z_n}{k_n}. \tag{2.15}$$

Les estimateurs (2.12) et (2.15) s'écrivent comme des combinaisons linéaires des valeurs extrêmes de N_n. Leurs propriétés asymptotiques dépendent en grande partie du comportement des coefficients de cette combinaison linéaire et donc du noyau de Dirichlet de la base utilisée. Deux cas très différents ont été envisagés : les bases C^1 et la base de Haar. Dans ces deux situations on pose

$$\sigma_n^{\mathrm{Pr}}(x) = \frac{k_n^{1/2}}{nc} K_n^{1/2}(x,x). \tag{2.16}$$

Bases C^1. Cette situation est étudiée en détails dans [76]. En particulier, la normalité asymptotique de $(\sigma_n^{\mathrm{Pr}}(x))^{-1}(\hat{f}_n^{\mathrm{Pr}}(x) - \mathbb{E}[\hat{f}_n^{\mathrm{Pr}}(x)])$ et $(\sigma_n^{\mathrm{Pr}}(x))^{-1}(\hat{f}_n^{\mathrm{Pr},2}(x) - f(x))$ est établie à $x \in [0,1]$ fixé sous diverses conditions sur le noyau de Dirichlet ([76], théorèmes 2 & 3). Cette situation est illustrée avec la base trigonométrique dont le noyau de Dirichlet est donné par

$$K_n(x,y) = \left| \begin{array}{ll} \dfrac{\sin\,(1+b_n)\pi(x-y)}{\sin \pi(x-y)} & \text{si } x \neq y, \\ 1 + b_n & \text{sinon.} \end{array} \right.$$

Les conditions de convergence se réécrivent alors simplement en fonction des suites (k_n) et (b_n).

Base de Haar. Ce cas est présenté dans [75]. La base de Haar est définie à partir d'une subdivision dyadique $\{J_\ell\}_{\ell \geq 1}$ de $[0,1]$. Pour chaque entier naturel ℓ, l'intervalle J_ℓ est défini par

$$J_\ell = \left[\frac{p_\ell}{2^{q_\ell - 1}}, \frac{p_\ell + 1}{2^{q_\ell - 1}} \right),$$

où p_ℓ et q_ℓ sont les entiers déterminés de manière unique par $\ell = 2^{q_\ell - 1} + p_\ell$ et $0 \leq p_\ell < 2^{q_\ell - 1}$. La base de Haar est définie par :

$$e_0 = \mathbb{I}\{[0,1]\}, \qquad e_\ell = 2^{\frac{q_\ell - 1}{2}} \left(\mathbb{I}\{J_{2\ell}\} - \mathbb{I}\{J_{2\ell+1}\} \right), \ell \geq 1.$$

Le nombre de termes dans le développement de f sur la base est nécessairement une puissance de deux : $h_n + 1 = 2^{b'_n}$, $b'_n \in \mathbb{N}$, et on impose de plus $k_n = \gamma_n(b_n+1)$, $\gamma_n \in \mathbb{N}^*$ de sorte que la partition définie par les $\{I_{n,r}\}$ soit un raffinement de celle associée aux $\{J_\ell\}$. Ainsi, pour tout $\ell \leq h_n$, J_ℓ est exactement l'union de γ_n sous-intervalles $I_{n,r}$. On introduit alors $R(n,\ell) = \{r = 1, \ldots, k_n,\ I_{n,r} \subset J_\ell\}$, et on a card $R(n,\ell) = \gamma_n$. Sous ces conditions et si $x \in J_\ell$, les coefficients pondérant les valeurs extrêmes dans les estimateurs (2.12) et (2.15) sont donnés par

$$K_n(x,x_{n,r}) = \left| \begin{array}{ll} b_n + 1 & \text{si } r \in R_{n,\ell}, \\ 0 & \text{sinon.} \end{array} \right.$$

L'estimateur $\hat{f}_n^{\mathrm{Pr}}$ est alors tout simplement la moyenne arithmétique de γ_n valeurs extrêmes. Lorsque $\gamma_n = 1$, l'estimateur $\hat{f}_n^{\mathrm{Pr}}$ se réduit à l'estimateur de Geffroy $\hat{f}_n^{\mathrm{G}}$ rappelé (2.1). Dans ce cadre, et sous certaines conditions sur la suite (k_n), nous avons montré [75], Théorème 4 que $\hat{f}_n^{\mathrm{Pr}}(x)$ convenablement normalisé converge en loi à $x \in [0,1]$ fixé vers une loi de Weibull des valeurs extrêmes de fonction de répartition H_{-1} définie en (1.1). Par contre, lorsque $\gamma_n \to \infty$, on retrouve une comportement analogue à celui apparaissant avec les bases C^1. La normalité asymptotique de $(\sigma_n^{\mathrm{Pr}}(x))^{-1}(\hat{f}_n^{\mathrm{Pr}}(x) - \mathbb{E}[\hat{f}_n^{\mathrm{Pr}}(x)])$ et $(\sigma_n^{\mathrm{Pr}}(x))^{-1}(\hat{f}_n^{\mathrm{Pr},2}(x) - f(x))$ est établie à $x \in [0,1]$ fixé dans [75], théorèmes 5 & 7 avec $\sigma_n^{\mathrm{Pr}}(x)$ introduit en (2.16). Enfin, nous avons montré la normalité asymptotique de l'erreur L^1 entre l'estimateur de Haar et la vraie frontière $d_1(\hat{f}_n^{\mathrm{Pr}},f)$ dans le cas où N_n est un processus de Poisson [68] :

Théorème 2.2.1 *Supposons f α-lipschitzienne ($0 < \alpha \leq 1$). Si $k_n = o(n/\log n)$, $n = o(h_n^{1+\alpha})$ et $k_n = o(h_n^{4/3})$ alors*

$$\frac{nc}{k_n^{1/2}} \left(d_1(\hat{f}_n^{Pr},f) - \mathbb{E}\left[d_1(\hat{f}_n^{Pr},f) \right] \right) \xrightarrow{d} \mathcal{N}(0,1). \tag{2.17}$$

Notons que ce résultat englobe les deux cas opposés $\gamma_n = 1$ (estimateur de Geffroy) et $\gamma_n \to \infty$. En renforçant les conditions sur la suite (h_n), la convergence (2.17) est conservée si on remplace l'espérance $\mathbb{E}\left[d_1(\hat{f}_n^{\mathrm{Pr}},f)\right]$ par $k_n/(nc)$, voir [68], Théorème 1.

2.2.1.2 Cas d'une base non-orthogonale

L'exemple de la base de Faber-Shauder est étudié dans la thèse de Laurent Gardes [55], Chapitre 1. La base de Faber-Shauder est définie à partir de la subdivision dyadique $\{J_\ell\}_{\ell \geq 1}$ de $[0,1]$ introduite dans le paragraphe précédent. Les fonctions qui la composent sont continues et sont données par

$$e_{-1}(x) = \mathbb{I}\{x \in [0,1]\}, \qquad e_0(x) = x\mathbb{I}\{x \in [0,1]\},$$

et pour $\ell \geq 1$ par

$$e_\ell(x) = 2^{q_\ell} \left[\left(x - \frac{p_\ell}{2^{q_\ell - 1}} \right) \mathbb{I}\{x \in J_{2\ell}\} - \left(x - \frac{p_\ell + 1}{2^{q_\ell - 1}} \right) \mathbb{I}\{x \in J_{2\ell+1}\} \right].$$

Les suites (b_n) et (k_n) sont définies de la même manière que pour l'estimateur de Haar. L'estimateur $\hat{f}_n^{\mathrm{LG}}$ obtenu s'écrit encore sous la forme (2.15) où l'expression du noyau est donnée par [53], Proposition 1. Suivant les ordres de grandeur respectifs de (b_n) et (k_n), la différence $\hat{f}_n^{\mathrm{LG}}(x) - f(x)$ convenablement normalisée converge en loi vers une loi des valeurs extrêmes (voir [53], Théorème 5) ou vers une loi normale centrée réduite (voir [53], Théorème 7).

2.2.2 Estimation par la méthode du noyau

Les travaux décrits dans ce paragraphe ont été menés en collaboration avec Pierre Jacob et sont décrits dans [78]. Par souci de comparaison avec le cadre (a), nous avons posé $c = 1/\lambda_2(D)$ de sorte que $E(N_n(D)) = n$. Formellement, l'estimateur de la frontière basé sur la méthode du noyau est semblable à (2.12). Il s'écrit

$$\hat{f}_n^{\mathrm{Ke}}(x) = \sum_{r=1}^{k_n} K_n(x - x_{n,r}) \frac{Y_{n,r}^*}{k_n}, \tag{2.18}$$

où K_n est défini par

$$K_n(t) = \frac{1}{h_n} K\left(\frac{t}{h_n}\right), \ t \in \mathbb{R}.$$

(h_n) désigne une suite de réels positifs tendant vers 0 parfois appelée fenêtre de lissage et K est un noyau de Parzen-Rosenblatt. Nous avons également introduit $\hat{f}_n^{\mathrm{Ke},2}$, une version de cet estimateur corrigée du biais,

$$\hat{f}_n^{\mathrm{Ke},2}(x) = \sum_{r=1}^{k_n} K_n(x - x_{n,r}) \frac{Y_{n,r}^* + Z_n'}{k_n},$$

où, contrairement à (2.14), la correction de biais définie par

$$Z_n' = \frac{1}{n - k_n} \sum_{r=1}^{k_n} Y_{n,r}^*,$$

ne fait appel qu'à des observations situées au voisinage de la frontière. La normalité asymptotique à $x \in]0,1[$ fixé de $(\sigma_n^{\mathrm{Ke}})^{-1}(\hat{f}_n^{\mathrm{Ke}}(x) - \mathbb{E}[\hat{f}_n^{\mathrm{Ke}}(x))]$ et $(\sigma_n^{\mathrm{Ke}})^{-1}(\hat{f}_n^{\mathrm{Ke},2}(x) - f(x))$ est établie pour

$$\sigma_n^{\mathrm{Ke}} = \frac{k_n^{1/2}}{nch_n^{1/2}} \left(\int_{\mathbb{R}} K^2(t)dt\right)^{1/2}, \tag{2.19}$$

et sous diverses conditions sur le noyau K, la régularité de la fonction f et les suites (k_n) et (h_n), voir théorèmes 3 & 5. La convergence en loi n'a lieu ici que sur l'intérieur de l'intervalle [0,1]. De même, $\hat{f}_n^{\mathrm{Ke}}$ ne converge uniformément vers f que sur les compacts de]0,1[, voir par exemple [78], Théorème 1 ou Théorème 2. Ce mauvais comportement des estimateurs à noyau sur les bords de l'intervalle d'estimation est bien connu, il a été également constaté dans la pratique. Des méthodes de symétrisation des données ont été proposées afin de pallier ces limitations. Par exemple, l'application de la technique décrite dans [25] à $\hat{f}_n^{\mathrm{Ke},2}$ conduit à l'estimateur suivant:

$$\hat{f}_n^{\mathrm{Ke},3}(x) = \sum_{r=1}^{k_n} \left(K_n(x - x_{n,r}) + K_n(x + x_{n,r}) + K_n(x + x_{n,r} - 2)\right) \left(\frac{Y_{n,r}^* + Z_n}{k_n}\right).$$

Les propriétés asymptotiques de $\hat{f}_n^{\mathrm{Ke},2}$ sur les compacts de]0,1[peuvent alors être étendues à $\hat{f}_n^{\mathrm{Ke},3}$ sur l'intervalle [0,1].

2.2.3 Estimation par la méthode du noyau généralisé

Les travaux présentés ici sont issus d'une collaboration avec Ludovic Menneteau (Université Montpellier 2). La famille d'estimateurs proposée [79] englobe les estimateurs $\hat{f}_n^{\mathrm{Pr}}$ et $\hat{f}_n^{\mathrm{Ke}}$, tout en offrant un cadre général pour construire de nouveaux estimateurs de la frontière f.

2.2.3.1 Cadre de l'étude

Nous nous plaçons dans le cas (b) le plus général où le support s'écrit

$$D = \{(x,y) : x \in E \; ; \; 0 \leq \; y \; \leq \; f(x)\}, \tag{2.20}$$

l'ensemble E étant un sous-ensemble de $\mathbb{R}^d$. Dans ce contexte, l'ensemble des $(I_{n,r})$, $r = 1,\dots,k_n$ est une partition quelconque de E. D'autre part, N_n est un processus de Poisson de mesure moyenne $nc\nu \otimes \lambda \mathbb{I}_D$, λ désignant la mesure de Lebesgue sur $\mathbb{R}$, et ν une mesure absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. Nous introduisons la famille d'estimateurs

$$\hat{f}_n^{\mathrm{GK}}(x) = \sum_{r=1}^{k_n} \nu_{n,r} \kappa_{n,r}(x) Y_{n,r}^*, \tag{2.21}$$

où $\nu_{n,r} = \nu(I_{n,r})$ et $\kappa_{n,r} : E \to \mathbb{R}$ est un noyau généralisé sur lequel quelques conditions seront imposées par la suite. On note $m_{n,r} = \inf\{f(x), x \in I_{n,r}\}$ et $M_{n,r} = \sup\{f(x), x \in I_{n,r}\}$. La version de (2.21) corrigée du biais est

$$\hat{f}_n^{\mathrm{GK},2}(x) = \sum_{r=1}^{k_n} \nu_{n,r} \kappa_{n,r}(x) \left(1 + \frac{1}{N_{n,r}}\right) Y_{n,r}^*. \tag{2.22}$$

Cette correction est motivée par la remarque que, conditionnellement à $N_{n,r}$, $Y_{n,r}^*$ a approximativement la même loi que le maximum de $N_{n,r}$ variables aléatoires indépendantes et distribuées uniformément sur $[0,m_{n,r}]$, voir [79], Lemme 1. La correction de biais est effectuée indépendamment sur chaque cellule $D_{n,r}$ de la partition de façon à être adaptée à des mesures ν quelconques.

2.2.3.2 Comportement asymptotique

On introduit

$$\kappa_n(x) = \left(\sum_{r=1}^{k_n} \kappa_{n,r}^2(x) \right)^{1/2}, \; x \in E,$$

le noyau généralisé normalisé $w_{n,r}(x) = \kappa_{n,r}(x)/\kappa_n(x)$, ainsi que $\nu_n = \min\{\nu_{n,r},\ 1 \leq r \leq k_n\}$ et $\Delta_n = \max\{M_{n,r} - m_{n,r},\ 1 \leq r \leq k_n\}$. Les hypothèses sont alors les suivantes :

(H.1) $k_n \to \infty$ et $n\nu_n \to \infty$ quand $n \to \infty$.

(H.2) $0 < m \leq M < +\infty$ et $\delta_n := \max_{1 \leq r \leq k_n} \nu_{n,r}(M_{n,r} - m_{n,r}) = o(1/n)$.

Il existe $F \subset E$ tel que :

(H.3) Pour tout $(x_1,\dots,x_p) \subset F$, il existe une matrice de covariance $\Sigma_{(x_1,\dots,x_p)} = [\sigma(x_i,x_j)]_{1 \leq i,j \leq p}$ de $\mathbb{R}^p$ telle que pour tout $1 \leq i,\ j \leq p$,

$$\sum_{r=1}^{k_n} w_{n,r}(x_i) w_{n,r}(x_j) \to \sigma(x_i,x_j) \text{ quand } n \to \infty.$$

(H.4) Pour tout $x \in F$,

$$\max_{1 \leq r \leq k_n} |w_{n,r}(x)| \to 0 \text{ quand } n \to \infty.$$

(H.5) Pour tout $x \in F$,

$$\left|\sum_{r=1}^{k_n} \nu_{n,r}\kappa_{n,r}(x)m_{n,r} - f(x)\right| = o\left(\frac{\kappa_n(x)}{n}\right) \text{ quand } n \to \infty.$$

(H.6) Pour tout $x \in F$,

$$\sum_{r=1}^{k_n} |w_{n,r}(x)| \max\left((n\delta_n)^2, \Delta_n, n\nu_n \exp(-mnc\nu_n)\right) \to 0 \text{ quand } n \to \infty.$$

Les hypothèses (H.1)–(H.4) sont dédiées au contrôle de l'estimateur centré $\hat{f}_n^{\mathrm{GK},2}(x) - \mathbb{E}[\hat{f}_n^{\mathrm{GK},2}(x)]$. La condition (H.1) impose que le nombre moyen de points dans chaque cellule $D_{n,r}$ tende vers l'infini. (H.2) assure que la fonction f ne s'annule pas et que le nombre moyen de points dans $D_{n,r}$ au-dessus de $m_{n,r}$ converge vers 0. Le couple de conditions (H.1) et (H.2) implique la convergence uniforme de l'oscillation de f sur $I_{n,r}$ vers 0 : $\max\{(M_{n,r} - m_{n,r}), 1 \le r \le k_n\} \to 0$ quand $n \to \infty$. L'hypothèse (H.3) est dédiée aux aspects multidimensionnels de la convergence en loi. (H.4) impose aux coefficients $\kappa_{n,r}(x)$ de la combinaison linéaire (2.22) d'être tous approximativement du même ordre afin d'obtenir un comportement asymptotique gaussien. Le couple de conditions (H.5) et (H.6) est dédié au contrôle du biais $\mathbb{E}[\hat{f}_n^{\mathrm{GK},2}(x)] - f(x)$ de façon à ce qu'il reste négligeable devant l'écart-type de l'estimateur

$$\sigma_n^{\mathrm{GK}}(x) = \frac{\kappa_n(x)}{nc}.$$

Sous ces hypothèses, on a la convergence en loi suivante :

Théorème 2.2.2 *Sous* (H.1)-(H.6) *et pour tout* $(x_1, \dots, x_p) \subset F$,

$$\left\{(\sigma_n^{GK}(x_j))^{-1}\left(\hat{f}_n^{GK,2}(x_j) - f(x_j)\right) : 1 \le j \le p\right\} \xrightarrow{d} \mathcal{N}(0, \Sigma_{(x_1,\dots,x_p)}),$$

où $\mathcal{N}(0, \Sigma_{(x_1,\dots,x_p)})$ *est la loi normale sur* $\mathbb{R}^p$ *centrée et de matrice de covariance* $\Sigma_{(x_1,\dots,x_p)}$.

Dans ce résultat, le coefficient de normalisation σ_n^{GK} dépend de la constante c supposée inconnue. Nous avons montré qu'il est possible de la remplacer par l'estimateur suivant

$$\hat{c}_n^{\mathrm{GK}} = \sum_{r=1}^{k_n} N_{n,r} \Big/ \sum_{r=1}^{k_n} n\nu_{n,r}\left(1 + \frac{1}{N_{n,r}}\right) Y_{n,r}^*$$

sans modifier la convergence du Théorème 2.2.2.

2.2.3.3 Exemples

Lien avec les estimateurs existants.

– On se place dans le cas où $d = 1$, $\mu = \lambda$, $E = [0,1]$ et $\{I_{n,r}\}$ est une partition régulière de $[0,1]$ de sorte que $\nu_{n,r} = 1/k_n$. En choisissant $\kappa_{n,r}(x) = K_n(x, x_{n,r})$ où $x_{n,r}$ est le centre de $I_{n,r}$ et K_n est un noyau de Dirichlet associé à une base orthogonale, l'estimateur $\hat{f}_n^{\mathrm{GK},2}$ est un estimateur de type projection similaire à $\hat{f}_n^{\mathrm{Pr},2}$ mais avec une correction de biais différente. L'application du Théorème 2.2.2 à l'estimateur ainsi obtenu dans le cas des bases C^1 et de la base de Haar permet de retrouver les résultats de normalité asymptotique établis dans [76], Théorème 3 et [75], Théorème 7, avec le même coefficient de normalisation $\sigma_n^{\mathrm{GK}} = \sigma_n^{\mathrm{Pr}}$.

– Dans le même cadre que précédemment, mais en considérant

$$\kappa_{n,r}(x) = \frac{1}{h_n} K\left(\frac{x - x_{n,r}}{h_n}\right)$$

avec K un noyau de Parzen-Rosenblatt et h_n une fenêtre de lissage, l'estimateur $\hat{f}_n^{\mathrm{GK},2}$ est un estimateur de type noyau similaire à $\hat{f}_n^{\mathrm{Ke},2}$ mais avec une correction de biais différente. L'application du Théorème 2.2.2 à l'estimateur ainsi obtenu permet de retrouver les résultats de normalité asymptotique établis dans [78], Théorème 5 avec le même coefficient de normalisation $\sigma_n^{\mathrm{GK}} = \sigma_n^{\mathrm{Ke}}$.

– Enfin, dans le cadre décrit paragraphe 2.2.3.1, et quand la mesure ν est inconnue, l'estimateur suivant de la frontière peut être considéré :

$$\hat{f}_n^{\mathrm{GK},3}(x) = \sum_{r=1}^{k_n} \hat{\nu}_{n,r} \kappa_{n,r}(x) \left(1 + \frac{1}{N_{n,r}}\right) Y_{n,r}^*,$$

$\hat{\nu}_{n,r}$ étant un estimateur de $\nu_{n,r}$. On remarque alors que, si $\kappa_{n,r}$ est le noyau de Dirichlet associé à une base orthogonale, le choix

$$\hat{\nu}_{n,r} = N_{n,r} \Big/ nc \left(1 + \frac{1}{N_{n,r}}\right) Y_{n,r}^*$$

conduit à $\hat{f}_n^{\mathrm{GK},3} = \hat{f}_n^{\mathrm{JS}}$ introduit initialement dans [90] lorsque $d = 1$, $\mu = \lambda$, $E = [0,1]$, $\{I_{n,r}\}$ est une partition régulière de $[0,1]$ et dont l'expression est rappelée en (2.10).

Introduction de nouveaux estimateurs.

– Par souci de simplicité, on se place dans le cas où $d = 1$, $\mu = \lambda$, $E = [0,1]$ et $\{I_{n,r}\}$ est une partition régulière de $[0,1]$ de sorte que $\nu_{n,r} = 1/k_n$. Il est possible d'exhiber dans la famille (2.22) un nouveau type d'estimateurs par projection. Pour cela, il suffit dans la définition (2.7) des coefficients de projection sur la base, de remplacer $f(t)$ sur l'intervalle $I_{n,r}$ par l'estimateur $Y_{n,r}^*(1 + N_{n,r}^{-1})$. On obtient alors

$$\hat{a}_{\ell,k_n}^{\mathrm{GM}} = \sum_{r=1}^{k_n} \left(\int_{I_{n,r}} e_\ell(t)dt\right) \left(1 + \frac{1}{N_{n,r}}\right) Y_{n,r}^*. \tag{2.23}$$

La différence entre les estimations (2.13) et (2.23), outre la correction de biais, réside dans le fait que la seconde d'entre elles ne repose pas sur l'approximation de la valeur moyenne de e_ℓ sur $I_{n,r}$ par $e_\ell(x_{n,r})$. Ce nouvel estimateur s'écrit alors

$$\hat{f}_n^{\mathrm{Pr},3}(x) = \sum_{r=1}^{k_n} \left(\int_{I_{n,r}} K_n(t,x)dt\right) \left(1 + \frac{1}{N_{n,r}}\right) Y_{n,r}^*,$$

il entre dans la famille (2.22) en posant

$$\kappa_{n,r}(x) = \frac{1}{\nu_{n,r}} \int_{I_{n,r}} K_n(t,x)dt = k_n \int_{I_{n,r}} K_n(t,x)dt,$$

où K_n désigne le noyau de Dirichlet associé à la base orthogonale. Nous avons établi dans [79], Corollaire 3, que dans le cas du noyau de Dirichlet associé à la base trigonométrique, la vitesse dans la convergence en loi de $\hat{f}_n^{\mathrm{Pr},3}$ était supérieure à celle de de $\hat{f}_n^{\mathrm{Pr},2}$.

– La construction de nouveaux estimateurs à noyau multidimensionnels entre également aisément dans le cadre du paragraphe 2.2.3.1. Posons $E = [0,1]^d$, $d \in \mathbb{N}^*$, $\nu = \lambda_d$ la mesure de Lebesgue sur E, et introduisons $\{I_{n,r} : 1 \le r \le k_n\}$ la partition régulière de E définie par $I_{n,r} = \prod_{j=1}^d J_{n,r,j}$ avec $\lambda(J_{n,r,j}) = 1/k_n^{1/d}$, de sorte que $\nu_{n,r} = 1/k_n$ pour tout $1 \le r \le k_n$. On note toujours $x_{n,r}$ le centre de $I_{n,r}$, $r = 1,\dots,k_n$. De façon similaire à l'étude précédente, il apparaît qu'il est souhaitable de considérer l'estimateur à noyau multidimensionnel défini par

$$\kappa_{n,r}(x) = \frac{1}{\nu_{n,r}} \int_{I_{n,r}} \frac{1}{h_n^d} K\left(\frac{x-t}{h_n}\right) dt,$$

où $K : \mathbb{R}^d \to \mathbb{R}^+$ est un noyau de Parzen-Rosenblatt multidimensionnel et (h_n) est une suite de réels positifs tendant vers 0. Les hypothèses (H.1)–(H.6) sont vérifiées sous certaines hypothèses sur le noyau K, les suites (k_n) et (h_n), voir [79], Corollaire 2. On obtient alors une vitesse de l'ordre de $n^{-\frac{\alpha}{\alpha+d}}$ dans la convergence en loi en dimension d pour une frontière α-Lipschitzienne.

2.2.4 Illustration sur simulations

A titre d'illustration, nous avons simulé dans le cadre (b) simplifié un processus de Poisson de paramètre d'intensité $nc = 800$ et de support défini par la fonction frontière

$$f(x) = [0.1 + \sin(\pi x)] \left[1.1 - 0.5 \exp\left(-64(x-0.5)^2\right)\right],$$

pour $x \in [0,1]$. La partition de l'intervalle $[0,1]$ comporte $k_n = 32$ de même longueur. Dans le cas de l'estimateur à noyau (paragraphe 2.2.2), le paramètre de lissage est $h_n = 0.025$. Dans le cas des estimateurs de Haar, trigonométriques (paragraphe 2.2.1.1) et de Faber-Shauder (paragraphe 2.2.1.2) $b_n = 15$ termes sont utilisés dans le développement de f. Les résultats sont présentés Figure 2.1 où les estimateurs sont superposés à la vraie frontière. Les théorèmes de normalité asymptotiques sont utilisés pour tracer des intervalles de confiance ponctuels à 90% pour $f(x)$ en 50 points différents.

2.3 Estimation par programmation linéaire

Les travaux décrits ici sont menés en collaboration avec Guillaume Bouchard (INRIA Rhône-Alpes), Anatoli Iouditski (Université Grenoble 1) et Alexander Nazin (Institute of Control Sciences, Moscou) [13, 14].
Le principe d'estimation proposé dans ce paragraphe ne requiert pas de partition du support D. Ce point est important dans la pratique car le choix de la partition $I_{n,r}$, $r = 1,\dots,k_n$ et donc de la suite (k_n) est un problème ouvert en ce qui concerne tous les estimateurs du paragraphe 2.2. Dans la suite, on se place dans le cadre (a) simplifié, N_n est un échantillon de points (X_i,Y_i), $i = 1,\dots,n$ indépendants et uniformément distribués sur le support $D = \{(x,y) : x \in [0,1] \,;\, 0 \le y \le f(x)\}$. Par commodité, nous considérons l'extension de f sur tout $\mathbb{R}$ en posant $f(x) = 0$ si $x \notin [0,1]$.

2.3.1 Construction de l'estimateur

L'estimateur proposé s'insère dans la famille suivante

$$\begin{cases} \hat{f}_n^{\text{PL}}(x) = \displaystyle\sum_{i=1}^n K_n(x - X_i)\alpha_i\,, & K_n(t) = h_n^{-1} K(t/h_n), \\ \alpha_i \ge 0, \quad i = 1,\dots,n, & \end{cases} \tag{2.24}$$

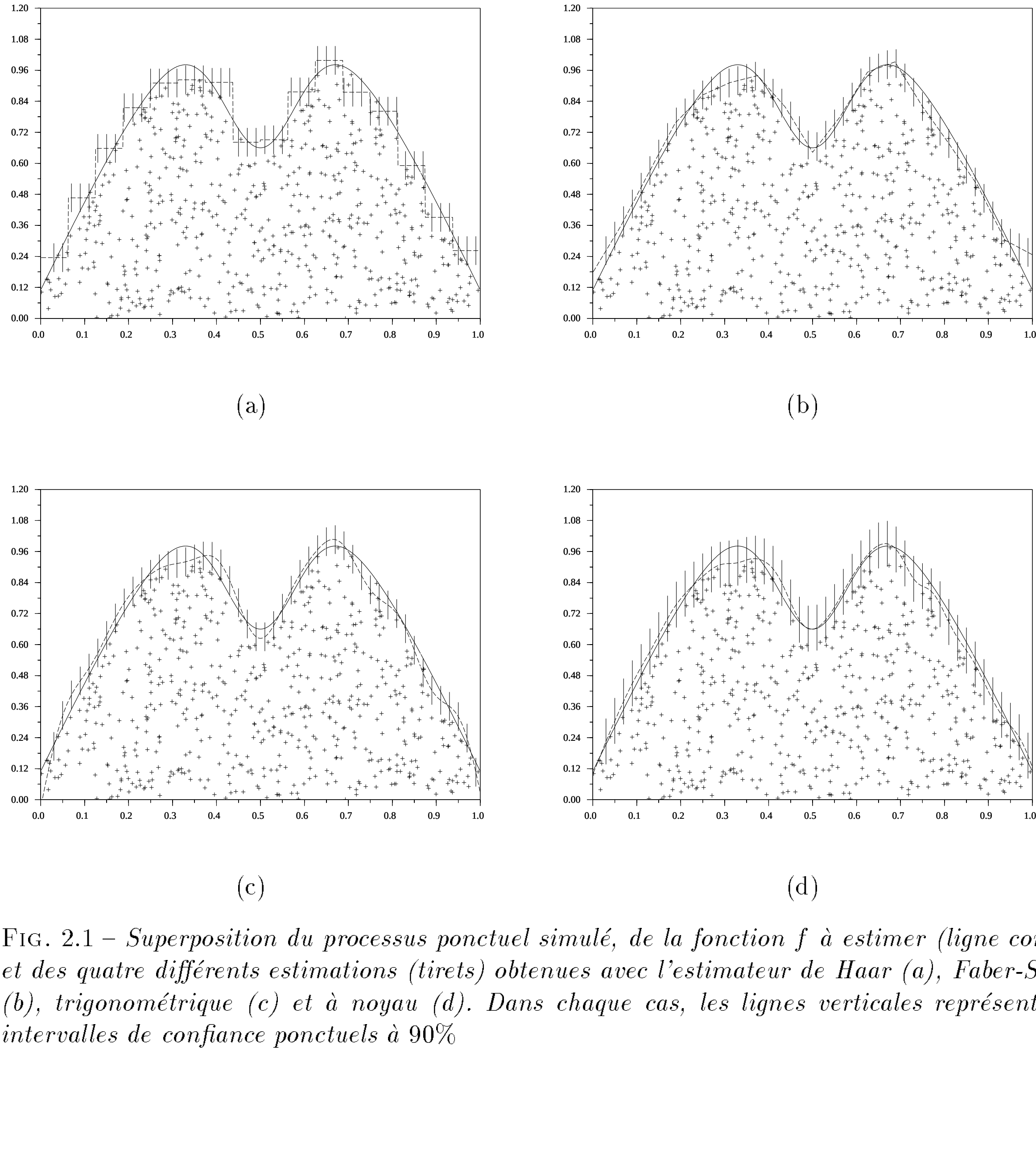


FIG. 2.1 – *Superposition du processus ponctuel simulé, de la fonction f à estimer (ligne continue) et des quatre différents estimations (tirets) obtenues avec l'estimateur de Haar (a), Faber-Shauder (b), trigonométrique (c) et à noyau (d). Dans chaque cas, les lignes verticales représentent les intervalles de confiance ponctuels à* 90%

où K est un noyau de Parzen-Rosenblatt réel et h_n est une fenêtre de lissage. Formellement, cette famille est similaire à celle des estimateurs à noyau (2.18). Cependant, l'expression définissant $\hat{f}_n^{\text{PL}}$ fait *a priori* intervenir tous les points de l'échantillon. En pratique, seuls les points (X_i,Y_i) pour lesquels $\alpha_i \neq 0$ interviennent dans l'estimation. De tels points sont appelés "vecteurs support" par analogie avec les *Support Vector Machines* (SVM). Nous renvoyons à [26] pour une synthèse sur ce sujet et à [113], Chapitre 8 pour des exemples d'application des SVM à l'estimation de quantiles. La contrainte $\alpha_i \geq 0$ pour $i = 1,\dots,n$ implique $\hat{f}_n^{\text{PL}}(x) \geq 0$ pour tout $x \in \mathbb{R}$ et assure une certaine régularité à l'estimateur. Nous reviendrons sur ce point plus tard. En remarquant que la surface du support estimé

$$\hat{D}_n = \{(x,y) : x \in [0,1]\ ;\ 0 \leq\ y\ \leq\ \hat{f}_n^{\text{PL}}(x)\}, \tag{2.25}$$

est donnée par

$$\int_{\mathbb{R}} \hat{f}_n^{\text{PL}}(x)\, dx = \sum_{i=1}^{n} \alpha_i, \tag{2.26}$$

il est naturel de déterminer le vecteur de paramètres $\alpha =^t (\alpha_1,\dots,\alpha_n)$ par le problème de programmation linéaire suivant :

$$\min_{\alpha}\ {}^t\mathbf{1}\alpha \tag{2.27}$$

sous les contraintes

$$A\alpha \geq Y \tag{2.28}$$

$$\alpha \geq 0. \tag{2.29}$$

Les notations suivantes sont adoptées: $\mathbf{1} =^t (1,1,\dots,1) \in \mathbb{R}^n$, $A = (K_n(X_i - X_j))_{1\leq i,j\leq n}$ et $Y =^t (Y_1,\dots,Y_n)$. Ainsi $A\alpha =^t (\hat{f}_n^{\text{PL}}(X_1),\dots,\hat{f}_n^{\text{PL}}(X_n))$, et la contrainte (2.28) se traduit par $\hat{f}_n^{\text{PL}}(X_i) \geq Y_i$, $i = 1,\dots,n$. L'estimateur ainsi obtenu $\hat{f}_n^{\text{PL}}$ est la fonction de la famille (2.24) associée au support (2.25) contenant tous les points de N_n et de surface minimale. En général, la solution du problème de programmation linéaire est parsimonieuse dans le sens où le nombre de coefficients α_i non nuls est faible pour des valeurs modérées de h_n.

2.3.2 Lien avec d'autres méthodes

Estimateur du maximum de vraisemblance. L'estimateur obtenu comme solution du problème de programmation linéaire (2.27)–(2.29) peut être considéré comme l'estimateur du maximum de vraisemblance dans la famille de fonctions (2.24). La densité jointe de l'ensemble des observations N_n connaissant la frontière f s'écrit :

$$P(N_n \mid f) = \prod_{i=1}^{n} \frac{f(X_i)}{C_f} \cdot \frac{1}{f(X_i)} \mathbb{I}\{0 \leq Y_i \leq f(X_i)\},$$

où C_f est l'aire du support D: $C_f = \int_0^1 f(x)dx = \int_{\mathbb{R}} f(x)dx$. En remarquant que d'après (2.26): $C_{f|f=\hat{f}_n^{\text{PL}}} = \sum_{i=1}^n \alpha_i$, la fonction de log-vraisemblance est donnée par

$$\log P(N_n \mid \hat{f}_n^{\text{PL}}) = -n \log \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \log \mathbb{I}\{Y_i \leq \hat{f}_n^{\text{PL}}(X_i)\},$$

et par conséquent sa maximisation sur l'ensemble des paramètres α positifs est équivalente au problème (2.27)–(2.29).

Estimateurs polynomiaux par morceaux. La famille d'estimateurs (2.3) introduite dans [96] est à la fois basée sur une partition du support et la résolution de problèmes de programmation linéaire. Sur chaque intervalle $I_{n,r}$, les coefficients du polynôme $P_{n,r}$ sont estimés par résolution d'un problème d'optimisation linéaire (2.4) similaire à (2.27)–(2.29).

Estimation par la méthode du noyau. Comme nous l'avons remarqué, la famille d'estimateurs (2.24) est similaire à celle des estimateurs à noyau (2.18). Ces derniers peuvent être interprétés comme reposant sur un estimation des coefficients α_i à partir de la partition $I_{n,r}$ par

$$\widehat{\alpha}_i = \left| \begin{array}{ll} Y^*_{n,r}/k_n & \text{si } \exists\, r \in \{1,\ldots,k_n\}\,;\, Y_i = Y^*_{n,r} \\ 0 & \text{sinon,} \end{array} \right.$$

alors que pour les estimateurs de type programmation linéaire, les coefficients α_i sont déterminés automatiquement, c'est à dire sans choix d'une partition au préalable par l'utilisateur.

2.3.3 Propriétés asymptotiques

Nous avons établi la convergence presque sure de la norme L^1 des estimateurs définis par le problème de programmation linéaire (2.27)–(2.29).

Théorème 2.3.1 *Si f est 1-lipschitzienne et sous quelques conditions sur le noyau K (voir [12], Théorème 1), si $nh_n^2/\log n \to \infty$ quand $n \to \infty$, alors*

$$\limsup_{n\to\infty}\ \varepsilon_n^{-1} d_1(\hat{f}_n^{PL},f) \le C < \infty \quad \text{p.s.}$$

avec $\varepsilon_n = \max\left\{h_n,(\log n)^{1/2}/(n^{1/2}h_n)\right\}$.

Dans ces conditions, la vitesse maximum de convergence est

$$d_1(\hat{f}_n^{\mathrm{PL}},f) = O_P\left((\log n/n)^{1/4}\right). \tag{2.30}$$

Il faut remarquer que cette vitesse est relativement faible en regard de la vitesse minimax $n^{-1/2}$ atteinte par l'estimateur de Geffroy (2.1). Le résultat (2.30) peut cependant être amélioré légérement en choisissant des noyaux K particuliers. Nous donnons dans [12], paragraphe 5.6, un exemple de noyau permettant d'obtenir une vitesse de l'ordre de $(\log n/n)^{1/3}$. L'obtention de vitesses supérieures passe par une modification de l'estimateur $\hat{f}_n^{\mathrm{PL}}$ envisagée au paragraphe 2.4 dans le cadre de nos perspectives de recherche.

2.4 Perspectives

Les deux types d'estimateurs décrits dans ce chapitre, estimateurs basés sur des partitions et estimateurs basés sur une optimisation, ouvrent des perspectives d'ordres différents.

– L'estimateur à noyau généralisé multidimensionnel défini au paragraphe 2.2.3.3 est l'estimateur le plus prometteur dans la famille des estimateurs basés sur des partitions car il bénéficie d'une vitesse de l'ordre de $n^{-\frac{\alpha}{\alpha+d}}$ en dimension d pour une frontière α-Lipschitzienne. Sur un plan théorique, il serait intéressant d'établir ses vitesses de convergence en norme L_1 et L_∞ pour les comparer aux vitesses minimax. Il est probable que les vitesses optimales soient atteintes à un facteur logarithmique près. Il serait alors nécessaire d'étudier dans quelle mesure la technique de preuve utilisée dans le cas de l'estimateur de Haar [68] peut être adaptée, puis d'employer la méthode de Geffroy [60] afin de passer du cadre des processus de Poisson à celui de l'échantillon. Sur un plan pratique, il est nécessaire d'étudier les performances de cet estimateur sur simulations.

– L'estimateur $\hat{f}_n^{\mathrm{PL}}$ ne constitue que le premier pas vers la définition d'estimateurs de frontière performants sur la base de méthodes d'optimisation. La première modification envisagée est l'incorportation dans le problème d'optimisation linéaire (2.27)–(2.29) d'une contrainte de Lispschitz sur l'estimateur. Cela doit permettre, au prix d'une complexité algorithmique plus importante, d'améliorer sensiblement les vitesses obtenues au paragraphe 2.3.3. L'extension de cet estimateur au cas multivarié doit également être réalisée. Enfin, il me parait également intéressant d'étudier la loi asymptotique de ce type d'estimateurs définis par des problèmes d'optimisation. Pour cela, l'étude de travaux connexes [95, 83] est indispensable.

Les deux familles d'estimateurs nécessitent également des recherches communes. Dans les deux cas, il est nécessaire de définir une procédure de choix adaptative du paramètre de lissage h_n ne dégradant pas les performances asymptotiques de l'estimateur. Une piste pourrait être l'utilisation de méthodes de type Lepski [98]. Ces deux familles d'estimateurs doivent aussi être adaptées au cas où la densité de points n'est pas uniforme et s'annule au voisinage de la fonction frontière. Les méthodes à développer dans ce cas sont issues de la théorie des valeurs extrêmes et passent par l'estimation d'un indice des valeurs extrêmes décrivant la vitesse de décroissance de la densité vers 0 [64, 6]. Cette direction de recherche "Estimation de quantiles extrêmes conditionnels" rejoint alors les travaux de thèse de Laurent Gardes [55], Chapitre 4, ainsi que les méthodes d'estimation de quantiles extrêmes et de quantiles conditionnels exposées aux chapitres 1 & 4 de ce mémoire.

Chapitre 3

Réduction de dimension et analyse d'images

L'analyse d'images est un champ d'application privilégié des méthodes de réduction de dimension. Une image de $M \times M$ pixels en niveaux de gris peut en effet être représentée par un vecteur de $\mathbb{R}^p$ avec $p = M^2$. Même pour des tailles d'images raisonnables, on obtient des données dans des espaces de très grande dimension. L'analyse en composantes principales (ACP) est alors un outil généralement efficace pour réduire la dimension de ces données [103, 119]. Toutefois, même des déformations très simples entre images peuvent se traduire par des fortes non linéarités dans $\mathbb{R}^p$, diminuant ainsi de beaucoup l'efficacité de l'ACP. Cette remarque nous a conduit à introduire les modèles auto-associatifs permettant de construire de nouvelles méthodes de réduction de dimension non-linéaires. Ces modèles sont présentés dans un cadre général paragraphe 3.1 et leur application à l'analyse d'images est illustrée paragraphe 3.2.

3.1 Les modèles auto-associatifs

L'ACP [91] est une méthode couramment utilisée pour la réduction de dimension en analyse des données. Elle bénéficie de plusieurs interprétations :

– Etant donné un ensemble de points de $\mathbb{R}^p$, et pour un entier $0 \leq d \leq p$, l'ACP construit le sous-espace affine de dimension d approchant au mieux les points au sens de la distance euclidienne. Partant de ce point de vue géométrique, de nombreux auteurs ont proposé des extensions non-linéaires de cette méthode. Les approches de type courbes ou surfaces principales [84] font partie de cette famille de méthodes.

– L'ACP peut également être interprétée en termes de poursuite de projection [88, 92]. Elle recherche le sous-espace affine de dimension d maximisant la variance projetée. Un algorithme de type poursuite de projection permettant de réaliser l'ACP itérativement est présenté dans le paragraphe 3.1.1. L'introduction de critères autres que la variance permet de définir autant de méthodes d'exploration des données [44, 106]. Dans les approches de type ACPVI-Spline (analyse en composantes principales de variables instrumentales [40]) et ACP curvilinéaire [10], l'introduction de transformations non-linéaires des coordonnées permet de conserver un critère de variance projetée sur les données transformées.

– Enfin, il est également possible d'associer un modèle probabiliste gaussien à l'ACP [118], le

sous-espace affine étant alors obtenu par maximisation d'une vraisemblance. Cette approche permet d'obtenir d'autres méthodes de réduction de dimension en considérant des modèles non-gaussiens, par exemple des modèles de mélange.
Construire une l'ACP non-linéaire est donc un problème difficile si l'on ne souhaite pas perdre ses trois interprétations. Ainsi, la construction d'un modèle probabiliste satisfaisant est souvent impossible sans spécifier la loi des observations. La méthode ainsi construite est alors ad hoc et de peu d'utilité en pratique. De plus, l'introduction d'une non-linéarité peut fait perdre l'interprétation géométrique du modèle construit. Les notions de variables principales, de directions principales, d'inertie expliquée et résiduelle se généralisent alors malaisément. La non-linéarité du modèle pose également les problèmes de l'existence, unicité et calculabilité d'un estimateur du modèle.
Dans le paragraphe 3.1.2, nous définissons les modèles auto-associatifs (AA), candidats à la généralisation de l'ACP. Les modèles AA ont été introduits initialement du point de vue géométrique dans le cadre de ma thèse [65] dirigée par Bernard Chalmond (Université de Cergy-Pontoise) et Jean-Marc Dinten (LETI/CEA). Ces modèles reposent sur l'approximation du nuage des observations par une variété différentiable [71]. Nous montrons dans le paragraphe 3.1.3 que ces modèles peuvent également être interprétés comme des modèles de poursuite de projection en régression [45] adaptés au cas auto-associatif. De ce fait, nous proposons un algorithme de mise en œuvre aisée. Deux exemples de modèles particuliers sont présentés paragraphe 3.1.4. Les aspects de mise en œuvre sont évoqués paragraphe 3.1.5. Ces travaux ont été réalisés en collaboration avec Serge Iovleff (Université Lille 1) et sont publiés dans [74].

3.1.1 Exemple de l'analyse en composantes principales

Soit X est un vecteur aléatoire de $\mathbb{R}^p$ possèdant un moment du second ordre. Il peut se décomposer en une somme de d variables aléatoires non corrélées (variables principales) et d'un résidu en appliquant de manière itérative les étapes [A] (recherche d'Axes), [P] (Projection), [R] (Régression) et [M] (Mise à jour) suivantes :

Algorithme 3.1.1

- *Pour* $j = 0$, *on pose* $R^0 = X - \mathbb{E}[X]$.
- *Pour* $j = 1, \dots, d$ *:*

  [A] *Déterminer* $a^j = \arg\max_{x \in \mathbb{R}^p} \mathbb{E}\left[\langle x, R^{j-1}\rangle^2\right]$ *s.c.* $\|x\| = 1$ *et* $\langle x, a^k\rangle = 0$, $1 \le k < j$.

  [P] *Calculer* $Y_j = \langle a^j, R^{j-1}\rangle$.

  [R] *Déterminer* $b^j = \arg\min_{x \in \mathbb{R}^p} \mathbb{E}\left[\|R^{j-1} - Y_j x\|^2\right]$ *s.c.* $\langle x, a^j\rangle = 1$,
  (on trouve $b^j = a^j$*) puis poser* $s^j(Y_j) = Y_j b^j$.

  [M] *Calculer* $R^j = R^{j-1} - s^j(Y_j)$.

Les vecteurs a^j sont appelées directions révélatrices, les variables aléatoires Y_j variables principales, les fonctions s^j fonctions de régression, et les vecteurs aléatoires R^j résidus. L'étape [A] consiste à rechercher un axe privilégié perpendiculaire aux précédents, qui maximise un certain critère : ici la variance projetée. L'étape [P] est une projection des résidus sur l'axe trouvé pour déterminer les variables principales, l'étape [R] consiste à chercher la meilleure fonction linéaire des variables principales qui approche les résidus. L'étape [M] est une mise à jour des résidus.

Les modèles AA étendent l'algorithme précédent en considérant des étapes [A] et [R] plus générales. Pour l'étape [A], nous considérons un certain nombre d'autres critères issus de la poursuite de projection correspondant à des objectifs différents. Nous envisageons l'étape [R] comme un problème de régression pouvant être abordé par des outils de type splines ou estimateurs à noyaux. Nous montrons que ce type de généralisation de l'ACP permet de conserver ses principales propriétés théoriques (construction d'un modèle exact, décroissance des résidus, ...) ou de les étendre (approximation des réalisations de X non plus par un sous-espace affine mais par une variété différentiable).

3.1.2 Définition des modèles auto-associatifs

Définition 3.1.1 *Une application $F : \mathbb{R}^p \to \mathbb{R}^p$ est dite auto-associative de dimension d, s'il existe d vecteurs orthonormés a^j et d fonctions $s^j : \mathbb{R} \to \mathbb{R}^p$ vérifiant $P_{a^j} \circ s^j = \mathbb{I}_{\mathbb{R}^p}$ et $P_{a^k} \circ s^j = 0$, $1 \le k < j \le d$, où $P_{a^j}(x) = \langle a^j, x \rangle$, tels que*

$$F = \left(\mathbb{I}_{\mathbb{R}^p} - s^d \circ P_{a^d}\right) \circ \ldots \circ \left(\mathbb{I}_{\mathbb{R}^p} - s^1 \circ P_{a^1}\right) = \prod_{k=d}^{1} \left(\mathbb{I}_{\mathbb{R}^p} - s^k \circ P_{a^k}\right).$$

Les vecteurs a^j sont appelés les directions révélatrices, les fonctions s^j sont appelées les fonctions de régression et on écrit $F \in \mathcal{A}^d_{a,s}$.

Par la suite, le signe produit représentera le produit de composition. Un modèle auto-associatif de dimension d est donc déterminé par d fonctions de régression et d directions révélatrices. On montre (voir [70], Lemme 2) que les modèles auto-associatifs additifs définis ci-dessous sont un cas particulier des modèles auto-associatifs.

Définition 3.1.2 *Une application $F : \mathbb{R}^p \to \mathbb{R}^p$ est dite auto-associative additive de dimension d, s'il existe d vecteurs orthonormés a^j et d fonctions $s^j : \mathbb{R} \to \mathbb{R}^p$ vérifiant $P_{a^j} \circ s^j = \mathbb{I}_{\mathbb{R}^p}$ et $P_{a^k} \circ s^j = 0$, $1 \le k < j \le d$, où $P_{a^j}(x) = \langle a^j, x \rangle$, tels que*

$$F = \mathbb{I}_{\mathbb{R}^p} - \sum_{k=1}^{d} s^k \circ P_{a^k}.$$

D'autre part, un modèle auto-associatif est dit linéaire si les fonctions de régressions sont linéaires. Le lemme suivant est établi dans [66].

Lemme 3.1.1 *Soit $F \in \mathcal{A}^d_{a,s}$, et supposons que les s^j, $j = 1, \ldots, d$ soient C^1 de $\mathbb{R}$ dans $\mathbb{R}^p$. L'équation $F(x) = 0$ définit alors une sous-variété différentiable de dimension d.*

Ce résultat est essentiel puisqu'il permet de généraliser l'interprétation géométrique de l'ACP : l'approximation de nuages de points par un sous-espace vectoriel.

Définition 3.1.3 *Soit X un vecteur aléatoire de $\mathbb{R}^p$ de carré intégrable. On dit que X vérifie un modèle auto-associatif de dimension d de directions révélatrices $(a^1, \ldots, a^d)$, de fonctions de régression $(s^1, \ldots, s^d)$ et de résidu ε, si X vérifie $F(X - \mu) = \varepsilon$ où $F \in \mathcal{A}^d_{a,s}$, $\mu \in \mathbb{R}^p$ et où ε est un vecteur aléatoire centré.*

Donnons deux exemples de modèles auto-associatifs triviaux :

1. Tout X satisfait un modèle AA de dimension 0. Il suffit de choisir $F = \mathbb{I}_{\mathbb{R}^p}$, $\mu = \mathbb{E}[X]$ et $\varepsilon = X - \mathbb{E}[X]$. On a alors $\mathrm{Var}[\|\varepsilon\|^2] = \mathrm{Var}[\|X\|^2]$.

2. X satisfait toujours également un modèle AA de dimension p. Dans ce cas $F = 0$, $\mu = 0$ et $\varepsilon = 0$ conduisent à $\mathrm{Var}\left[\|\varepsilon\|^2\right] = 0$.

Dans la pratique, il s'agit de réaliser un compromis entre ces deux extrêmes en construisant un modèle de dimension $d \ll p$ et tel que $\mathrm{Var}\left[\|\varepsilon\|^2\right] \ll \mathrm{Var}\left[\|X\|^2\right]$. Par exemple, dans le cas où X possède une matrice de variance-covariance Σ de rang d, il satisfait un modèle AA additif linéaire de dimension d de résidu nul. En effet, en notant a^j, $j = 1,\dots,d$ les vecteurs propres de Σ associés aux valeurs propres non nulles. Les choix

$$F(x) = \coprod_{k=d}^{1} \left(\mathbb{I}_{\mathbb{R}^p} - P_{a^k} a^k\right)(x),$$

$\mu = \mathbb{E}[X]$ et $\varepsilon = 0$ p.s. définissent un modèle auto-associatif additif linéaire pour X :

$$X = \mathbb{E}[X] + \sum_{k=1}^{d} \left\langle a^k, X - \mathbb{E}[X] \right\rangle a^k \quad p.s. \tag{3.1}$$

Il s'agit de la décomposition de X obtenue par ACP, comme il sera vu dans le Corollaire 3.1.2. Enfin, si X vérifie un modèle auto-associatif additif, on a

$$X = \mu + \sum_{k=1}^{d} s^k \left(\left\langle a^k, X \right\rangle\right) + \varepsilon, \tag{3.2}$$

ce qui est une forme parfois proposée dans les approches de type réseaux de neurones (voir par exemple [93] ou le Perceptron Multicouches [24]) afin de généraliser le modèle (3.1). Les limites de cette extension sont évoquées au paragraphe 3.1.4.1.

Définition 3.1.4 *On dit qu'un ensemble $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$ de fonctions mesurables de $\mathbb{R}$ dans $\mathbb{R}^p$ est admissible s'il est un sous-ensemble fermé de L_2 et s'il vérifie la condition suivante :*

$$(\mathcal{R}) : \left\{ \begin{array}{ll} \forall a \in \mathbb{R}^p \text{ tel que } \|a\| = 1, & s \in \mathcal{S}(\mathbb{R},\mathbb{R}^p) \Rightarrow s - \langle a,s \rangle\, a \in \mathcal{S}(\mathbb{R},\mathbb{R}^p) \\ \forall b \in \mathbb{R}^p & s \in \mathcal{S}(\mathbb{R},\mathbb{R}^p) \Rightarrow s + b \in \mathcal{S}(\mathbb{R},\mathbb{R}^p) \\ & \mathbb{I}_{\mathbb{R}} b \in \mathcal{S}(\mathbb{R},\mathbb{R}^p). \end{array} \right.$$

La condition $(\mathcal{R})$ s'interprète comme une invariance par projection et par translation. Un choix possible de $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$ est l'ensemble des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}^p$. Cet exemple est présenté dans le paragraphe 3.1.4.1.

Définition 3.1.5 *Soit a un vecteur unitaire de $\mathbb{R}^p$. Un index I est une fonctionnelle qui associe à la projection du vecteur aléatoire X sur a (i.e. $\langle a,X \rangle$) un réel positif.*

Un choix de I possible est $I(\langle a,X \rangle) = \mathrm{Var}[\langle a,X \rangle]$, la variance projetée. D'autres exemples sont présentés au paragraphe 3.1.5.2.

3.1.3 Construction et propriétés

Soient $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$ un ensemble de fonctions admissibles et $d \in \{0,\dots,p\}$. On introduit l'algorithme suivant:

Algorithme 3.1.2

- *Pour* $j = 0$, *on pose* $\mu = \mathbb{E}[X]$ *et* $R^0 = X - \mu$.
- *Pour* $j = 1,\dots,d$ *:*
  - [A] *Déterminer* $a^j = \arg\max_{x \in \mathbb{R}^p} I(\langle x,R^{j-1}\rangle)$ *s.c.* $\|x\| = 1$, $\langle a^k,x\rangle = 0$, $1 \le k < j$.
  - [P] *Calculer* $Y_j = \langle a^j,R^{j-1}\rangle$.
  - [R] *Choisir* $s^j \in \arg\min_{s \in \mathcal{S}(\mathbb{R},\mathbb{R}^p)} \mathbb{E}\left[\left\|R^{j-1} - s(Y_j)\right\|^2\right]$ *s.c.* $P_{a^j} \circ s^j = \mathbb{I}$.
  - [M] *Calculer* $R^j = R^{j-1} - s^j(Y_j)$.

Nous vérifions Théorème 3.1.1 que cet algorithme construit un modèle auto-associatif de dimension d. De plus, à l'issue de p itérations il construit une représentation exacte de X.
L'étape [R] dépend fortement du choix de $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$. Deux cas extrêmes sont étudiés: le cas où $\mathcal{S}(\mathbb{R},\mathbb{R}^p) = \mathcal{A}(\mathbb{R},\mathbb{R}^p)$ l'ensemble des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}^p$, dans le paragraphe 3.1.4.1 et le cas où $\mathcal{S}(\mathbb{R},\mathbb{R}^p) = L_2$, dans le paragraphe 3.1.4.2. Le choix de l'index I est discuté paragraphe 3.1.5.2.

Théorème 3.1.1 *L'algorithme 3.1.2 construit un modèle auto-associatif de dimension d de directions révélatrices $(a^1,\dots,a^d)$, de fonctions de régression $(s^1,\dots,s^d)$ et de résidu $\varepsilon = R^d$. De plus, si $d = p$ alors $\varepsilon = R^p = 0$ et on a la décomposition exacte :*

$$X = \mathbb{E}[X] + \sum_{k=1}^{p} s^k(Y_k) \quad p.s.$$

Ces propriétés sont générales dans le sens où elles ne dépendent ni de l'index I choisi, ni de l'ensemble de fonctions admissible $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$. Dans le paragraphe 3.1.4 quelques propriétés complémentaires correspondant à des choix de I et $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$ particuliers sont établies. Le corollaire suivant est utile en pratique pour choisir la dimension d'un modèle.

Corollaire 3.1.1 *Soit Q_d la fraction d'information représentée par un modèle AA de dimension d :*

$$Q_d = 1 - \mathbb{E}\left[\left\|R^d\right\|^2\right] / \mathrm{Var}\left[\|X\|\right].$$

On a alors $Q_0 = 0$, $Q_p = 1$ et la suite (Q_d) est croissante.

3.1.4 Deux modèles particuliers

Nous considérons deux cas importants en pratique où il est possible de fournir une solution explicite à l'étape [R]: les modèles auto-associatifs linéaires et les modèles auto-associatifs de régression. Ces modèles héritent des propriétés établies dans le paragraphe précédent. Dans chaque cas, nous complétons ces propriétés générales par quelques caractéristiques propres à ces deux cas particuliers.

3.1.4.1 Les modèles auto-associatifs linéaires

Nous nous intéressons au cas où $\mathcal{S}(\mathbb{R},\mathbb{R}^p) = \mathcal{A}(\mathbb{R},\mathbb{R}^p)$. L'étape [R] s'écrit

[R] Trouver $b^j = \arg\min_{x\in\mathbb{R}^p} \mathbb{E}\left[\|R^{j-1} - Y_j x\|^2\right]$, s.c. $\langle a^j,x\rangle = 1$,

et on obtient comme solution du problème d'optimisation sous contrainte

$$b^j = \Sigma^{j-1}a^j/({}^t a^j \Sigma^{j-1} a^j) \tag{3.3}$$

où Σ^j est la matrice de variance-covariance de R^j. On a l'analogue du Théorème 3.1.1 suivant :

Théorème 3.1.2 *L'algorithme 3.1.2 construit un modèle auto-associatif linéaire de dimension d pour X de fonctions de régression $s^j(t) = tb^j$. De plus pour $d = p$, on a la décomposition :*

$$X = \mathbb{E}[X] + \sum_{k=1}^{p} Y_k b^k, \ \ p.s. \tag{3.4}$$

où les variables aléatoires Y_k, $k = 1,\dots,p$ sont non corrélées.

Sur la base de ce résultat, il est possible de construire des modèles auto-associatifs linéaires qui ne soient pas additifs à partir de l'étape [R] ainsi définie. On montre également dans le corollaire suivant que pour une étape [A] bien choisie, on retrouve le modèle linéaire et additif de l'ACP.
Corollaire 3.1.2 *Si de plus $I(\langle x,R^{j-1}\rangle) = \mathrm{Var}\left[\langle x,R^{j-1}\rangle\right]$, alors l'algorithme 3.1.2 effectue une ACP de X. En particulier, pour $d = p$, on a*

$$X = \mathbb{E}[X] + \sum_{k=1}^{p} Y_k a^k, \ \ p.s.$$

avec $Y_k = \langle X - \mathbb{E}[X],a^k\rangle$ et a^k est le vecteur propre de la matrice de covariance de X associé à la kème plus grande valeur propre.
Nous avons également prouvé un résultat réciproque ([70], Théorème 1) : le seul modèle auto-associatif additif est celui de l'ACP. Ce résultat limite considérablement l'intérêt des modèles auto-associatifs additifs tels que (3.2).

3.1.4.2 Les modèles auto-associatifs de régression

Nous considérons désormais le cas où $\mathcal{S}(\mathbb{R},\mathbb{R}^p) = L_2$. Dans ce cas, l'étape [R] possède une solution explicite : $s^j(Y_j) = \mathbb{E}\left[R^{j-1}|Y_j\right]$. En effet, l'espérance conditionnelle est un projecteur orthogonal et vérifie la contrainte du problème d'optimisation puisque $\langle a^j,R^{j-1}\rangle = Y_j$. On a alors le résultat suivant :

Théorème 3.1.3 *L'algorithme 3.1.2 construit un modèle auto-associatif de dimension d. De plus si $d = p$ alors on a la décomposition exacte :*

$$X = \mathbb{E}[X] + \sum_{j=1}^{p} s^j(Y_j), \ \ p.s. \tag{3.5}$$

où Y_j et Y_{j+1} sont non-corrélées, $j = 1,\dots,p-1$.

Dans la suite ces modèles seront désignés par modèles auto-associatifs de régression (AAR).

3.1.5 Mise en œuvre

Le paramètre μ est estimé par la moyenne empirique. Les deux étapes cruciales dans l'algorithme 3.1.2 sont [A] et [R] : la détermination de la direction révélatrice et l'estimation de la fonction de régression. Le choix de l'index I et de la classe de fonctions $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$ déterminent en effet à la fois la nature du modèle obtenu et la complexité de calcul associée aux problèmes d'optimisation [A] et [R].

3.1.5.1 Estimation de la fonction de régression

Pour ce problème on peut remarquer que si $\mathcal{S}(\mathbb{R},\mathbb{R}^p)$ est l'ensemble des fonctions linéaires de $\mathbb{R}$ dans $\mathbb{R}^p$ alors il existe une unique solution donnée par (3.3). Il suffit donc de remplacer dans cette expression Σ^{j-1} par son estimateur empirique et a^j par son estimation obtenue à l'étape [A].
Dans le cas des modèles auto-associatifs de régression, il s'agit d'estimer l'espérance conditionnelle de R^{j-1} sachant Y_j. Ce problème standard peut être résolu (par exemple) par une régression par noyau ou une régression spline. Par rapport au problème classique de régression, nous avons une contrainte supplémentaire sur la fonction à estimer. A l'itération j, elle doit vérifier $P_{a^j} \circ s^j = \mathbb{I}$. En se plaçant dans la base orthonormée B^j obtenue en complétant $\{a^1,\dots,a^j\}$, il suffit de réaliser $p-j$ régressions, composantes par composantes. Pour la composante numéro $k \in \{j+1,\dots,p\}$ l'estimateur à noyau s'écrit dans la base B^j :

$$\tilde{s}_k^j(u) = \sum_{i=1}^{n} \tilde{R}_{i,k}^{j-1} K_h(u - Y_{i,j}) \Bigg/ \sum_{i=1}^{n} K_h(u - Y_{i,j}) \ , \qquad (3.6)$$

où $\tilde{R}_{i,k}^{j-1}$ représente la $k$ème coordonnée du résidu de l'individu i à la $(j-1)$ème itération dans la base B^j. $Y_{i,j}$ représente la valeur de la $j$ème variable principale pour l'individu i.

3.1.5.2 Détermination des directions révélatrices

Le choix de l'index I est à la base de tout problème de poursuite de projection (PP) où il s'agit de trouver des directions "intéressantes". Nous renvoyons à [88, 92] pour des articles de synthèse sur ce type de problème. Le sens de l'adjectif "intéressantes" dépend du problème d'analyse des données considéré. Par exemple, Friedman *et al* [44] ont proposé un index pour faire apparaître des groupes ou utilisent une distance à la normalité pour mettre en évidence d'éventuelles structures plus complexes du nuage de points. Citons également les index dédiés à la recherche de points aberrants [106].
Dans le cadre des modèles auto-associatifs de régression, il s'agit de rechercher des directions qui paramétrent au mieux la variété que l'on cherche à estimer. Dans ce cadre, Demartines [30] propose un index qui favorise les directions dans lesquelles la projection conserve approximativement les distances. Nous donnons deux exemples d'index favorisant les directions dans lesquelles la structure de voisinage est respectée par projection.

Premier index. L'index suivant comptabilise le nombre de points qui sont plus proches voisins dans $\mathbb{R}^p$ et le sont encore après projection. Il est introduit dans [66]. A l'itération j, il s'écrit :

$$I\left(\langle x,R^{j-1}\rangle\right) = \sum_{i=k}^{n} \sum_{\ell \neq k} \mathbb{I}\left\{R_k^{j-1} \text{ ppv } R_\ell^{j-1}\right\} \mathbb{I}\left\{\left\langle x,R_k^{j-1}\right\rangle \text{ ppv } \left\langle x,R_\ell^{j-1}\right\rangle\right\}, \qquad (3.7)$$

où "ppv" est l'abréviation de "plus proche voisin de". Cet index peut être réécrit sous la forme

$$I\left(\langle x, R^{j-1}\rangle\right) = \sum_{i=k}^{n}\sum_{\ell\neq k} m_{k\ell}^{j}\mathbb{I}\left\{\left\langle x, R_k^{j-1}\right\rangle \text{ ppv } \left\langle x, R_\ell^{j-1}\right\rangle\right\},$$

où $m_{k\ell}^{j} = \mathbb{I}\{R_k^{j-1} \text{ ppv } R_\ell^{j-1}\}$ sont les coefficients de la matrice de contiguïté du premier ordre : $M_j = (m_{k\ell}^{j})$. Il apparaît que l'index bénéficie des propriétés d'invariance par translation et échelle ([66], Lemme 3.1) permettant de le placer dans la classe III définie par Huber [88]. La maximisation de cet index est un problème délicat. Une solution basée sur un algorithme de recuit simulé est introduite dans [21]. Néanmoins, cette solution est très lourde à mettre en œuvre. Une version simplifiée de l'index (3.7) est alors introduite.

Second index. Nous proposons ici une approche similaire à celle de Lebart [97] qui consiste à définir un coefficient de contiguïté dont la minimisation permet de déplier les stuctures non linéaires. En terme d'index, il s'agit de maximiser en x à l'itération j le rapport de formes quadratiques :

$$I\left(\langle x, R^{j-1}\rangle\right) = \sum_{i=1}^{n}\left\langle x, R_i^{j-1}\right\rangle^2 \Big/ \sum_{k=1}^{n}\sum_{\ell=1}^{n} m_{k\ell}^{j}\left\langle x, R_k^{j-1} - R_\ell^{j-1}\right\rangle^2. \tag{3.8}$$

Le numérateur de (3.8) est la variance projetée des résidus, son dénominateur représente la distance entre la projection de résidus plus proches voisins dans $\mathbb{R}^p$. La maximisation de (3.8) doit alors révéler les directions dans lesquelles la projection préserve la structure de voisinage du premier ordre. La direction révélatrice a^j est donnée par le vecteur propre associé à la plus petite valeur propre de la matrice $V_j^\star V_j^{-1}$ où

$$V_j^\star = \sum_{k=1}^{n}\sum_{\ell=1}^{n} m_{k\ell}^{j}\,{}^t(R_k^{j-1} - R_\ell^{j-1})(R_k^{j-1} - R_\ell^{j-1})$$

est proportionnelle à la matrice de covariance locale. La matrice

$$V_j = \sum_{k=1}^{n} {}^tR_k^{j-1}R_k^{j-1}$$

est proportionnelle à la matrice de covariance empirique de R^{j-1}. V_j^{-1} désigne son inverse généralisé, V_j n'étant pas inversible puisque R^j est orthogonal à $\{a^1, \ldots, a^j\}$ d'après [74], Proposition 2.1(ii). Notons que cette approche est équivalente à celle de Lebart lorsque la matrice de contiguïté M_j est symétrique. Dans la pratique, les approches (3.7) et (3.8) donnent des résultats comparables.

3.2 Application en analyse d'images

Nous présentons deux exemples d'application des méthodes de réduction de dimension précédentes en analyse d'images.

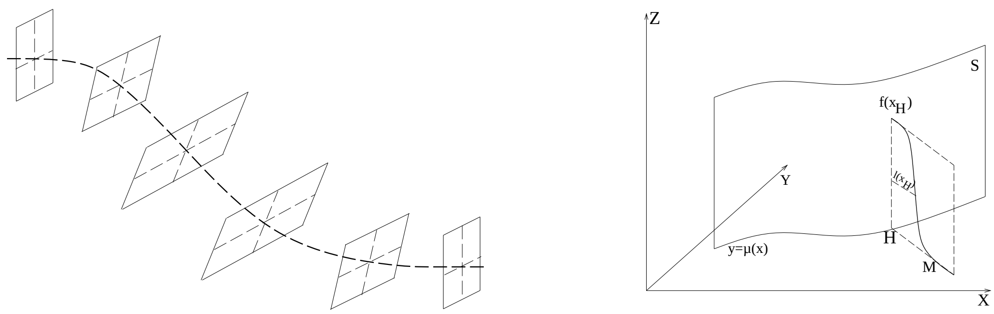
3.2.1 Reconstruction à partir d'une seule projection radiographique

L'objectif général de la reconstruction d'images est de déterminer une image tridimensionnelle de la structure interne d'un objet à partir de projections radiographiques bidimensionnelles de celui-ci. Le champ d'application le plus connu de ce type de problématique est la médecine, mais l'industrie en fait également partie avec le contrôle non destructif. C'est dans ce contexte que se situent nos travaux en collaboration avec Bernard Chalmond (Université de Cergy-Pontoise) et Jean-Marc Dinten (LETI/CEA). Plus précisément, nous nous plaçons dans le cas où les cadences de contrôle en ligne imposent la prise d'une seule radiographie, la reconstruction devant alors être menée à partir d'une seule projection. Il s'agit d'un problème inverse mal posé qui ne peut être résolu sans l'introduction de fortes informations a priori sur la géométrie de l'objet à reconstruire. Nous présentons dans la suite une démarche permettant de construire un modèle a priori d'objet flexible (dans le sens ou un objet n'est pas nécessairement de géométrie figée) à partir d'un jeu d'exemples de ses variations de forme. Ce modèle est destiné à être utilisé dans une procédure d'estimation bayésienne de l'objet.

Modèle de la projection de l'objet. L'objet à reconstruire est modélisé par un cylindre généralisé, c'est à dire un volume caractérisé par une courbe axiale et une courbe fermée décrivant le contour des sections transverses [4]. Nous nous restreignons à des sections transverses rectangulaires. Ce modèle très simple d'objet (voir figure 3.1a) conduit à une représentation de sa projection radiographique par une surface d'équation

$$\phi(M) = f(x_H) \exp\left\{ -K_q \left(\frac{\|HM\|}{\ell(x_H)} \right)^q \right\}, \tag{3.9}$$

avec $K_q = 2^q \log 2$ et où M est un point dans le plan (Oxy) de l'image, H est le point le plus proche de M sur la surface S d'équation $y = \mu(x)$ dans $\mathbb{R}^3$ et x_H son abscisse sur cette surface (voir figure 3.1b). La fonction μ décrit la projection de la courbe axiale du cylindre généralisé. La fonction f décrit la section de la surface (3.9) par S, et la fonction ℓ la largeur à mi-hauteur des sections transverses. Le paramètre $q > 0$ quantifie le flou induit par le système d'acquisition radiographique. En résumé, la projection radiographique est entièrement décrite par les trois fonctions μ, ℓ et f. Leurs graphes sont appelés courbes caractéristiques.



(a) Modèle de l'objet (b) Modèle de la projection

FIG. 3.1 – *Modélisation de l'objet par un cylindre généralisé et de sa projection radiographique par une surface.*

Apprentissage des déformations de la projection de l'objet. Il s'agit de modéliser les déformations de la projection de l'objet à partir d'un ensemble d'observations de celles-ci. Compte-tenu du formalisme du paragraphe précédent, il suffit de modéliser les déformations des trois courbes caractéristiques. Considérons par exemple la courbe décrite par la fonction f. On suppose que l'on dispose d'un échantillon $\{f_1, \dots, f_n\}$ de ses déformations, chacune d'elles étant discrétisée en p points $f_i = ^t (f_i^1, \dots f_i^p)$. On obtient alors un ensemble de n vecteurs de $\mathbb{R}^p$, pouvant être interprété comme la réalisation d'un vecteur aléatoire X lorsque les courbes sont discrétisées sur un même intervalle. L'apprentissage des déformations de la projection est alors ramené à la construction d'un modèle auto-associatif décrite au paragraphe 3.1. Dans la suite, nous supposons que le vecteur aléatoire X est gaussien justifiant l'utilisation de modèles auto-associatifs linéaires construits par ACP. Les limitations de cette hypothèse sont illustrées dans [69] par l'introduction d'un décalage simulé par translation dans l'échantillon formé par les fonctions discrétisées. D'après le Corollaire 3.1.2, on obtient un modèle linéaire à d paramètres des déformations de la courbe caractéristique :

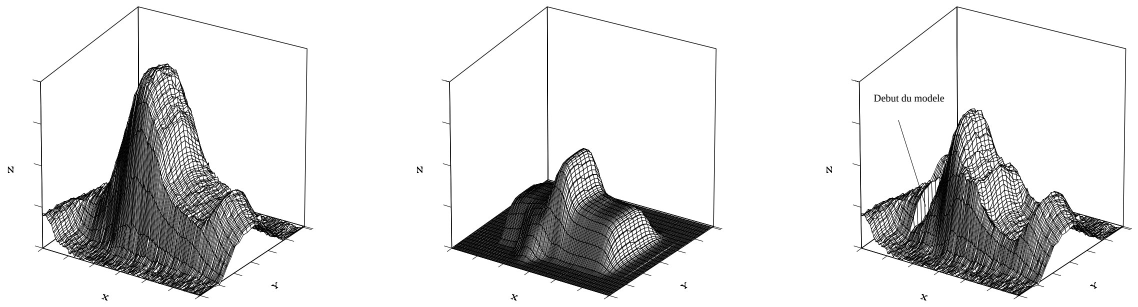
$$\hat{f}(Y_1, \dots, Y_d) = \bar{f} + \sum_{k=1}^{d} Y_k \bar{f}^k$$

où $\bar{f}$ est une estimation de $\mathbb{E}[X]$ obtenue par la moyenne empirique des $\{f_1, \dots, f_n\}$ et $\bar{f}^k$ est l'estimation de la direction principale a^k par le vecteur propre associé à la $k$ème plus grande valeur propre λ_k de la matrice de covariance empirique des $\{f_1, \dots, f_n\}$. Dans ce contexte, $\bar{f}^k$ est appelé le $k$ème mode de déformation (ou mode propre) de la courbe caractéristique et peut bénéficier d'une interprétation physique (torsion, élongation, ...). Le paramètre d est choisi à l'aide de la fraction d'information représentée, voir Corollaire 3.1.1. En accord avec l'hypothèse de normalité de X, les lois des paramètres de déformation sont modélisées par des lois normales centrées et indépendantes : $Y_k \sim \mathcal{N}(0, \lambda_k)$.

Exemple en contrôle non destructif. Le principe de modélisation décrit ci-dessus a été mis en œuvre dans le cadre du contrôle non destructif de soudures de circuits imprimés [73]. A partir d'une image radiographique de circuits imprimés soudés, il est nécessaire de construire une image de la soudure seule. Le problème est donc le suivant : en tout point M de la projection radiographique ϕ_{ps} d'une patte de composant soudé, on cherche à distinguer la projection radiogrpahique de la soudure ϕ_s de celle de la patte du composant ϕ_p sachant que $\phi_{ps}(M) = \phi_p(M) + \phi_s(M)$. Dans cet objectif, on modélise la projection radiographique d'une patte de composant par (3.9). Cette modélisation est valide pour les pattes de type Gull-Wing considérées. Les variations de géométrie de cette projection (dues aux tolérances de fabrication et aux déformations possibles d'une patte de composant) sont apprises sur une base d'images de composants non-soudés. Le modèle paramétrique ainsi obtenu fournit un modèle a priori pour ϕ_p permettant son identification à partir de la seule observation de ϕ_{ps} grâce à une démarche bayésienne non décrite ici. On pourra se reporter à [20], Chapitre 12 pour plus de détails. Disposant alors de l'estimation $\hat{\phi}_p$, on en déduit immédiatement une estimation $\hat{\phi}_s$ de ϕ_s par soustraction, et il est également possible de reconstruire la géométrie tridimensionnelle de la patte de composant. Un exemple de résultats est présenté figure 3.2.

3.2.2 Représentation de bases d'images

Nous présentons deux exemples de représentation de bases d'images extraits de [21].

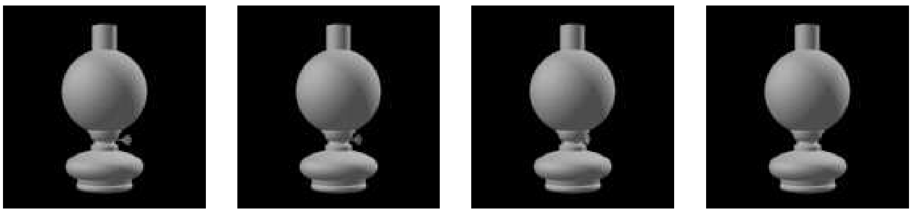


(a) (b) (c)

FIG. 3.2 – *Représentation des images en surfaces de niveaux de gris: (a) patte soudée ϕ_{ps}, (b) estimation $\hat{\phi}_p$ correspondant à la patte seule, (c) estimation $\hat{\phi}_s$ associée à la soudure.*

Images de synthèse. Nous étudions ici une base de 45 images de taille 256×256 issue de l'archive du Centre For Intelligent Systems, Faculty of Human Sciences and Faculty of Technology, University of Plymouth. Il s'agit d'images d'un objet de synthèse vu sous différents angles d'élévation et d'azimuth. Un extrait de la base est représenté figure 3.3.

(a) (b) (c) (d) (e)



(f) (g) (h) (i)

FIG. 3.3 – *Extrait de la base d'images de synthèse. (a) image référence, (b-e) rotation utilisant l'angle d'élévation, (f-i) rotation utilisant l'angle d'azimuth.*

Chaque image est représentée par un vecteur de dimension $M^2 = 256^2$. On obtient donc un nuage de $n = 45$ points en dimension 65536. Cependant, par un simple changement de repère, on se ramène à un ensemble de points en dimension $p = 44$. Notre but est de comparer les résultats de modélisation obtenus par ACP et par les modèles AAR introduits au paragraphe 3.1.4.2. Il apparaît qu'un modèle AAR de dimension $d = 1$ permet de représenter plus de $Q_1 = 96\%$ de l'information. A titre de comparaison, un modèle linéaire construit par ACP doit être de dimension 4 pour atteindre ce pourcentage. De plus, le coude dans la courbe représentant la suite $(Q_d)_{d\geq 0}$ associée aux modèles

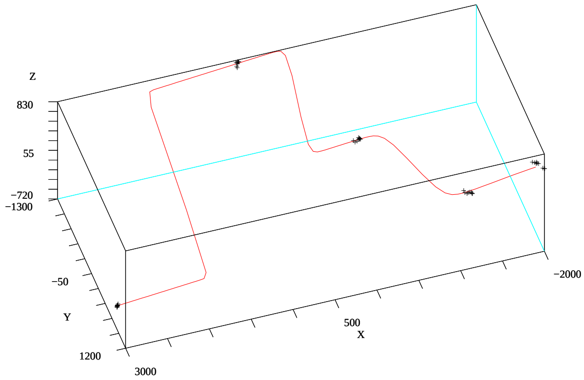


FIG. 3.4 – *Représentation de la projection de la variété de dimension 1 et du nuage de points dans le repère formé par les trois premiers axes principaux.*

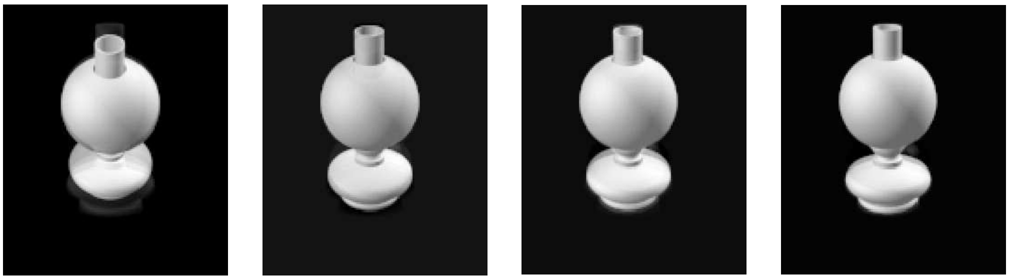


FIG. 3.5 – *Simulation de 4 images par le modèle AAR de dimension 1. La variable Y_1 est simulée uniformément sur l'intervalle* $[\min_i Y_{1,i}, \max_i Y_{1,i}]$.

AAR semble indiquer que $d = 1$ est un choix correct de dimension de modèle. La projection de la variété correspondante dans l'espace formé par les trois premiers axes principaux est représentée figure 3.4 où elle est superposée à la projection du nuage de points. Modéliser ce nuage de points par une variété de bidimensionnelle a également un sens puisque les images sont générées par rotation de l'objet dans deux directions orthogonales.
Il est intéressant de noter que la variable principale Y_1 associée au modèle AAR de dimension 1 est interprétable. Elle correspond à l'angle de rotation d'élévation. Pour s'en convaincre, il suffit de simuler des réalisations uniformes de cette variable et de représenter les images ainsi simulées avec le modèle AAR de dimension 1 (figure 3.5). La variable Y_2 n'est pas aussi aisément interprétable. Pour cette raison, le modèle AAR de dimension 1 semble préférable.

Images réelles. Une expérimentation similaire est menée à partir d'une base de 400 images de taille 112×92 issue du Olivetti and Oracle Research Laboratory. Il s'agit d'images de visages de 40 individus. L'ACP demande un modèle de dimension 210 pour obtenir les mêmes performances d'approximation d'un modèle AAR de dimension 89. A dimension égale (à 89), l'ACP conduit à une erreur relative par pixel de 35% contre 20% les modèles AAR. Ceci ce traduit visuellement sur les visages (figure 3.6). Par ACP, on obtient uniquement la forme globale du visage avec disparition des principaux traits caractéristiques. Avec les modèles AAR, les yeux et la bouche, primitives souvent utilisées par les méthodes de reconnaissance de visage, sont beaucoup mieux reconstruits.

FIG. 3.6 – *Représentation de 5 images différentes du même individu (première ligne) par le modèle AAR (seconde ligne) et par ACP (dernière ligne).*

3.3 Perspectives

Les travaux décrits dans ce chapitre trouvent une partie de leurs prolongements dans le cadre d'une ACI "masse de données" impliquant en particulier le projet LEAR (apprentissage et reconnaissance en vision par ordinateur) de l'INRIA Rhône-Alpes. Dans ce même contexte, je co-encadre depuis octobre 2003 avec Cordelia Schmid (LEAR) la thèse de Charles Bouveyron sur le thème "Modèles statistiques pour la sélection et l'organisation de descripteurs d'images". Nous nous proposons d'étudier les problèmes particuliers posés par l'utilisation des méthodes de réduction de dimension en analyse d'images. Ainsi, la modélisation d'une image par un vecteur de $\mathbb{R}^p$ proposée en introduction n'est pas satisfaisante car ne prenant pas en compte la notion de proximité entre pixels. Une solution envisageable est de ne pas travailler sur les pixels eux-mêmes, mais sur des descripteurs extraits de l'image. Dans les deux cas, il est indispensable d'adapter la méthode de réduction de dimension à la nature spatiale des données.
Indépendamment du contexte de l'analyse d'image, l'étude des modèles auto-associatifs demande à être poursuivie. Je projette d'étudier les propriétés asymptotiques des estimateurs mis en œuvre et ainsi de définir un test permettant le choix de la dimension du modèle. D'autre part, les modèles

auto-associatifs tels qu'ils sont définis ici supposent l'existence de directions dans lesquelles la variété est paramétrable. Cette hypothèse limite le champ d'application de ces modèles. Afin de lever cette limitation, j'envisage d'étendre la définition de ces modèles aux variétés paramétrables par des couples de directions. Enfin, j'aimerais également étendre ces modèles à la représentation de réunion de variétés afin de pouvoir prendre en compte les distributions de mélange.

Chapitre 4

Estimation de courbes de référence

De nombreuses expérimentations, en particulier dans le cadre d'études biomédicales, sont conduites pour établir des intervalles de valeurs qui sont prises “normalement” par une variable d'intérêt Y, dans une population cible. Ici, le terme “normalement” fait référence aux valeurs que l'on est susceptible d'observer avec une probabilité donnée, dans des conditions normales et pour des individus types présumés en bonne santé. Ces intervalles sont appelés intervalles de référence et les valeurs correspondantes valeurs de référence. Par exemple, on peut s'intéresser à un intervalle excluant les 5% d'observations les plus grandes et les 5% d'observations les plus petites. Ainsi, la construction d'intervalles de référence repose sur le calcul de quantiles. Par ailleurs, il arrive régulièrement que, sur la population cible, l'on dispose simultanément, avec la variable d'intérêt Y, d'une information complémentaire sous la forme d'une covariable X. Très souvent, X représente l'âge du sujet. Pour une valeur donnée x de X, on peut construire un intervalle de référence. Lorsque x varie, on obtient alors des “courbes de référence” ou plus précisément des hypersurfaces. Dans ce cadre, il est nécessaire de travailler avec les quantiles conditionnels de Y sachant X. Le tracé de courbes de référence sur le nuage des valeurs prises par le couple (X,Y) pour les sujets de référence donne un résumé graphique très utile et interprétable lorsque la covariable X est unidimensionnelle. Ainsi, un individu i représenté par le point (x_i,y_i) pourra être comparé à la population de référence. En d'autres termes, une “anormalité” de cet individu sera suspectée si ce point se situe en dessous de la courbe de référence inférieure ou au dessus de la courbe de référence supérieure. L'estimation des courbes de références se pose aussi bien dans les domaines biomédical et biométrique, que dans le domaine industriel. Nos travaux dans le cadre unidimensionnel sont présentés paragraphe 4.1. L'extension au cas multidimensionnel fait l'objet du paragraphe 4.2. L'ensemble de l'étude résumée dans ce chapitre est le fruit d'une collaboration avec Ali Gannoun, Jérôme Saracco (Université Montpellier 2) et Christiane Guinot (CERIES).

4.1 Covariable unidimensionnelle

Nous présentons dans le paragraphe 4.1.1 un extrait de la littérature concernant l'estimation des quantiles conditionnels. Le paragraphe 4.1.2 est consacré à la présentation des estimateurs non paramétriques des quantiles conditionnels que nous avons retenus. Un extrait de l'étude comparative de ces estimateurs sur simulations menée dans [49] est présenté dans le paragraphe 4.1.3. Enfin, nous exposons dans le paragraphe 4.1.4 une application visant à établir des courbes de référence, en fonction de l'âge, des propriétés biophysiques de la peau de femmes sur deux zones du visage et une zone de l'avant-bras. L'étude complète peut être trouvée dans [48]. Les logiciels développés pour mener cette étude sont décrits dans [50].

4.1.1 Quantiles conditionnels et courbes de référence

Le quantile conditionnel d'ordre $\alpha \in]0.5,1[$ de Y sachant $X = x$ est défini par $q_\alpha(x) = F^{-1}(\alpha|x)$, où $F(.|x)$ désigne la fonction de répartition conditionnelle de Y sachant $X = x$. Une caractérisation alternative de $q_\alpha(x)$ est obtenue sous forme d'un problème d'optimisation :

$$q_\alpha(x) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}[\rho_\alpha(Y-\theta)|X=x], \tag{4.1}$$

où ρ_α est la fonction définie par $\rho_\alpha(z) = \alpha z \mathbb{I}_{[0,\infty)}(z) - (1-\alpha) z \mathbb{I}_{(-\infty,0)}(z)$. Pour une valeur x donnée, l'intervalle de référence contenant $100(2\,\alpha\ -\ 1)\%$ des sujets de référence est ensuite défini par $I_\alpha(x) = [q_{1-\alpha}(x),\, q_\alpha(x)]$. Les courbes de référence sont alors les ensembles de points $\{(x,q_{1-\alpha}(x)\}$ et $\{(x,q_\alpha(x)\}$ lorsque x varie. Soit $q_{n,\alpha}(x)$ un estimateur de $q_\alpha(x)$ obtenu à partir de l'échantillon $\{(X_i,Y_i),\ i = 1,\ldots,n\}$ de n réalisations indépendantes du couple de variables aléatoires (X,Y). L'estimateur correspondant de $I_\alpha(x)$ est défini par $I_{n,\alpha}(x) = [q_{n,1-\alpha}(x),\, q_{n,\alpha}(x)]$. En pratique, pour obtenir les courbes de référence à 90%, α est choisi égal à 0.95. Trois types d'approches existent pour l'estimation des quantiles conditionnels.

Approche paramétrique. Quand l'échantillon est de petite taille, des hypothèses paramétriques sont habituellement imposées afin de réduire le nombre de paramètres à estimer. En particulier, lorsque la fonction de répartition conditionnelle est supposée gaussienne, un estimateur du quantile conditionnel est obtenu à partir d'estimateurs de l'espérance et de la variance conditionnelles. Un modèle linéaire ou polynomial [112] associé à la méthode des moindres carrés est généralement utilisé pour estimer ces deux quantités. Cependant ces estimations sont très sensibles aux valeurs aberrantes de Y. De plus, il peut être nécessaire de transformer les données de départ dans l'espoir d'obtenir des résidus normalement distribués avec cette nouvelle échelle. L'existence d'une telle transformation n'est nullement garantie. Il est aussi connu que les hypothèses paramétriques sont restrictives et peuvent rarement être faites avec certitude [22]. Ainsi, l'approche paramétrique peut être mal adaptée à la réalité des données en particulier biologiques. Des approches semi ou non paramétriques du problème ont alors été développées afin de pallier ces problèmes d'hypothèses et de modélisation paramétriques.

Approche semi paramétrique. Les méthodes semi paramétriques consistent à rechercher une transformation des valeurs observées de la variable Y de façon à les rendre normalement distribuées. Par exemple, la méthode LMS [22] effectue la transformation suivante

$$Z_i = \frac{(Y_i/M(X_i))^{L(X_i)} - 1}{L(X_i)S(X_i)}.$$

Les fonctions L, M et S sont estimées sous forme de fonctions splines respectivement par $\hat{L}$, $\hat{M}$ et $\hat{S}$ grâce à une méthode de maximum de vraisemblance pénalisée. L'estimateur correspondant du quantile $q_\alpha(x)$ est alors $\hat{M}(x)(1+\hat{L}(x)\hat{S}(x)z_\alpha)^{1/\hat{L}(x)}$ où z_α est le quantile correspondant de la loi normale centrée réduite.

Approche non paramétrique. De nombreux travaux récents ont été menés pour l'estimation non paramétrique des quantiles conditionnels. Le point commum à ces méthodes est de ne pas nécessiter d'hypothèse sur la nature de la distribution. A titre d'exemple, une approche robuste permettant d'obtenir des courbes de référence est introduite dans [86]. Cette méthode est basée sur une partition du support de X en intervalles. Pour chaque intervalle, les deux quantiles $q_\alpha(x)$ et $q_{1-\alpha}(x)$ sont estimés de façon usuelle. L'ensemble des quantiles estimés est alors lissé selon x pour obtenir les courbes de référence.
Les approches non paramétrique que nous considérons ici ne nécessitent pas de partition du support de X. De plus, elles sont robustes dans le sens où les courbes de références calculées sont peu sensibles à la présence de points aberrants. Dans la suite, trois méthodes non paramétriques d'estimation des quantiles conditionnels sont présentées.

4.1.2 Méthodes non paramétriques d'estimation des quantiles conditionnels

1– Méthode d'estimation par noyau. Définissons tout d'abord un estimateur non paramétrique de la fonction de répartition conditionnelle de Y sachant $X = x$, pour $y \in \mathbb{R}$:

$$\tilde{F}_n(y|x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \mathbb{I}\{Y_i \leq y\} \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{4.2}$$

La fonction K, appelée noyau, est d'intégrale égale à un. Le paramètre h_n permet de contrôler le lissage appliqué aux données. Il est alors naturel d'estimer le quantile conditionnel $q_\alpha(x)$ par $\tilde{q}_{\alpha,n}(x) = \tilde{F}_n^{-1}(\alpha|x)$. Les propriétés asymptotiques de cet estimateur sont établies en particulier dans [47].

2– Méthode de la constante locale. La méthode dite de la constante locale consiste à estimer la solution du problème (4.1) par

$$\bar{q}_{n,\alpha}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \rho_\alpha(Y_i - a) K\left(\frac{x - X_i}{h_n}\right),$$

où h_n et K désignent la fenêtre et le noyau mentionnés précédemment. Cette méthode directe d'estimation présente en particulier l'avantage d'un bon comportement face aux effets de bords De plus, sous des conditions générales, la convergence de cet estimateur est obtenue sans étudier préalablement la convergence de l'estimateur de la fonction de répartition conditionnelle [120].

3– Méthode d'estimation par noyau produit. Une version plus "lisse" de l'estimateur de la fonction de répartition conditionnelle définie en (4.2) peut être introduite en remplaçant la fonction indicatrice par une nouvelle densité symétrique ω. L'estimateur correspondant, appelé estimateur par noyau produit est défini comme suit :

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_{1,n}}\right) \Omega\left(\frac{y - Y_i}{h_{2,n}}\right) \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_{1,n}}\right),$$

où Ω est la fonction de répartition associée à ω. Cet estimateur peut également être vu comme une primitive de l'estimateur à noyau de la densité conditionnelle. Il en découle l'estimateur suivant

$\hat{q}_{\alpha,n}(x) = \hat{F}_n^{-1}(\alpha|x)$. Cette approche est attractive mais nécessite le choix de deux paramètres de lissage $h_{1,n}$ et $h_{2,n}$. Il apparaît en pratique que cet estimateur est extrêmement sensible au choix de ces deux paramètres. Les propriétés asymptotiques de cet estimateur sont établies par exemple dans [8].

Choix des paramètres de lissage. La qualité des estimateurs non paramétriques basés sur les noyaux est étroitement liée aux choix des paramètres de lissage. Une importante littérature est consacrée à ce sujet, et en particulier aux méthodes de sélection automatique par minimisation d'un critère, telles que la validation croisée. Nous avons retenu les choix suivants pour les différentes fenêtres intervenant dans chacun des estimateurs. Pour l'estimateur $\tilde{q}_{\alpha,n}(x)$, une approche dérivée du critère de validation croisée est utilisée :

$$h_n = \arg\min_{h>0} \sum_{j=1}^{n} \int_{\mathbb{R}} \{\mathbb{I}\{Y_j \leq y\} - \tilde{F}_{n,-j}(y|x)\}^2 \omega(y) dy, \tag{4.3}$$

où $\tilde{F}_{n,-j}(.|x)$ est l'estimateur de $F(.|x)$ défini au paragraphe 4.1.2 mais calculé à partir de l'échantillon $\{(X_i,Y_i),\ i = 1,\ldots,n\}$ privé de la j-ème observation.
Pour les estimateurs $\bar{q}_{n,\alpha}(x)$ et $\bar{q}_{n,\alpha}(x)$, une règle empirique reposant sur l'hypothèse de normalité de la loi conditionnelle de Y sachant X et proposée dans [120] est retenue. L'utilisation de telles règles empiriques présente l'avantage d'une mise en œuvre simple et rapide. Cependant cet avantage est acquis au prix d'une perte de généralité due à l'ajout d'une hypothèse de normalité.

4.1.3 Comparaison sur simulations des trois méthodes non paramétriques

Les trois méthodes non paramétriques décrites précédemment sont évaluées dans [49] sur des données simulées mais réalistes au vu de l'application sur données réelles considérée paragraphe 4.1.4. Plus précisément, lorsque le choix de la fenêtre de l'estimateur est basé sur des considérations de normalité, son comportement vis à vis de cette hypothèse est examiné.

Description du modèle simulé. Nous considérons le modèle suivant :

$$y = 16(x - 0.5)^2 + \varepsilon, \tag{4.4}$$

où le terme d'erreur ε est égal à $\varepsilon^* - a\lambda$ avec ε^* suivant une loi gamma de paramètres d'échelle λ et de paramètre de forme a. Rappelons que $\mathbb{E}[\varepsilon^*] = a\lambda$, $\mathrm{Var}[\varepsilon^*] = a\lambda^2$ et $S(\varepsilon^*) = 2/\sqrt{a}$, S désignant le coefficient d'asymétrie, ou *skewness*. Dans la suite de la simulation, on choisit $a = 1/\lambda^2$ afin de travailler avec un terme d'erreur de variance unité. Dans ce cadre particulier, $\varepsilon \in [-1/\lambda, +\infty[$, l'espérance de ε^* étant égale à $1/\lambda$, le terme d'erreur ε est centré. De plus, le coefficient d'asymétrie est alors égal à 2λ. Ainsi, plus λ est grand, plus la distribution de l'erreur est dissymétrique et donc s'éloignera de la normalité. Les données réelles étudiées dans la suite, présentent, pour la plupart des variables d'intérêt, une répartition du nuage de points du même type que $\lambda = 1$ ou 1.5 dans le modèle (4.4). Dans le cadre de ce modèle, le quantile conditionnel d'ordre α s'écrit ici explicitement sous la forme $q_\alpha(x) = 16(x - 0.5)^2 + q_\alpha^{(g)} - 1/\lambda$, où $q_\alpha^{(g)}$ est le quantile d'ordre α de la loi gamma de paramètres $(\lambda,1/\lambda^2)$.
Pour chacune des valeurs de λ considérées, $N = 50$ échantillons de taille $n = 200$ sont simulés selon le modèle (4.4) avec la covariable X générée selon la loi uniforme discrète sur $\{j/30, j = 0,1,\ldots,30\}$.

Le choix de l'uniforme discrète se justifie par le fait que, sur les données réelles étudiées, la covariable est l'âge (arrondi à l'année) des individus. Pour chaque échantillon simulé, nous avons ensuite estimé $q_\alpha(z_j)$, pour $z_j = j/50$, $j = 0,1,\ldots,50$, par chacune des trois méthodes non paramétriques. Afin d'évaluer les différentes méthodes, pour chaque échantillon ($k = 1,\ldots,50$) l'erreur relative médiane, notée $ERM(k)$ est calculée :

$$ERM(k) = m\acute{e}diane \left\{ \left| \frac{q_{\alpha,n}^{(k)}(z_j) - q_\alpha(z_j)}{q_\alpha(z_j)} \right|, \ j = 0,1,\ldots,50 \right\},$$

où $q_{\alpha,n}^{(k)}(z_j)$ correspond à l'estimation de $q_\alpha(z_j)$ obtenue par la méthode non paramétrique considérée sur l'échantillon k.

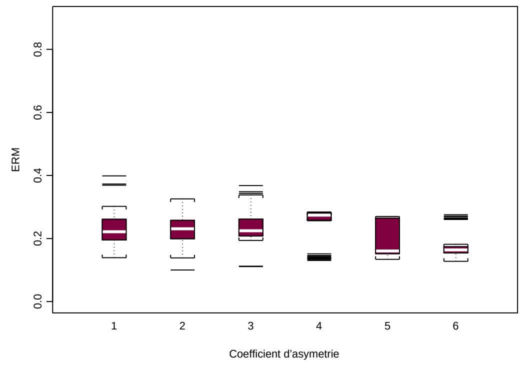
Exemple de résultats. Les distributions des ERM sont présentées sous forme de boîtes à moustaches en fonction du degré d'asymétrie du bruit en figure 4.1. La dégradation des performances des estimateurs 2 et 3 avec l'augmentation de l'asymétrie du bruit n'est pas surprenante, le choix des fenêtres étant basé sur une hypothèse de normalité. La détérioration des résultats est surtout sensible pour l'estimation des quantiles d'ordre 5%, c'est à dire pour l'estimation des quantiles proches de l'extrémité finie ($-1/\lambda$) du support de la loi du bruit. Ceux-ci sont globalement sous-estimés, le plus souvent situés en deçà du support ($q_{200,\,0.05} < -1/\lambda$). Pour éliminer ce problème, la solution consisterait à changer de règle de choix de largeur de fenêtre.

4.1.4 Application à des données réelles

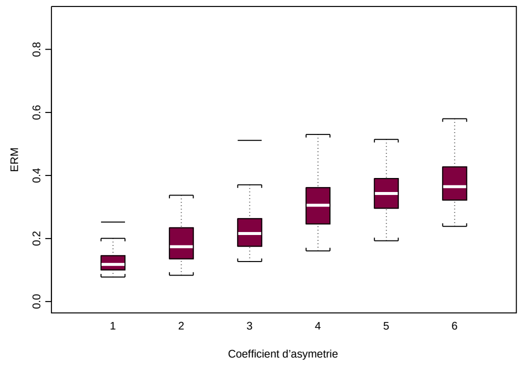
Une étude réalisée par le CE.R.I.E.S a été conduite entre novembre 1998 et décembre 1999 sur $n = 322$ femmes de type caucasien agées de 20 à 80 ans présentant une peau apparemment saine. Chaque volontaire a été examinée en atmosphère contrôlée. Cette étude comportait des questionnaires sur les habitudes de vie, un interrogatoire et un examen médical cutané, ainsi qu'une évaluation des propriétés biophysiques cutanées. Les propriétés biophysiques de la peau incluaient : le taux de sécrétion de sébum (taux instantané de lipides ($\mu g/m^2$), la température cutanée (exprimée en degrés Celsius), la perte insensible en eau (g/m^2h), le pH cutané, l'hydratation de la peau estimée par la capacitance et la conductance, et la couleur de la peau. La couleur a été exprimée à l'aide des trois paramètres luminosité, coordonnées de chromacité rouge/vert et coordonnées de chromacité jaune/bleu. L'évaluation des paramètres biophysiques a été effectuée sur deux zones du visage (front et joue) et sur la face antérieure de l'avant-bras gauche, sauf pour le taux de sébum mesuré uniquement sur les deux zones du visage.
En résumé, les variables d'intérêt sont au nombre de 12 sur la face antérieure de l'avant-bras gauche et de 13 sur le front et la joue. La covariable est l'âge des volontaires. Les détails de cette étude sont publiés dans [48].

4.1.4.1 Les méthodes d'estimation

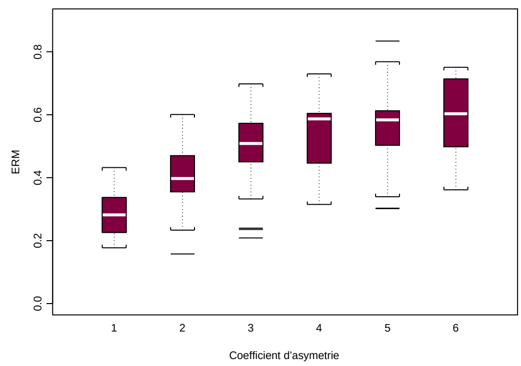
Méthode paramétrique. La méthodologie utilisée pour l'établissement de valeurs de référence des mesures biophysiques cutanées en fonction de l'âge des sujets découle de celle proposée dans [112]. Elle comporte six étapes légèrement modifiées afin de l'adapter aux données.



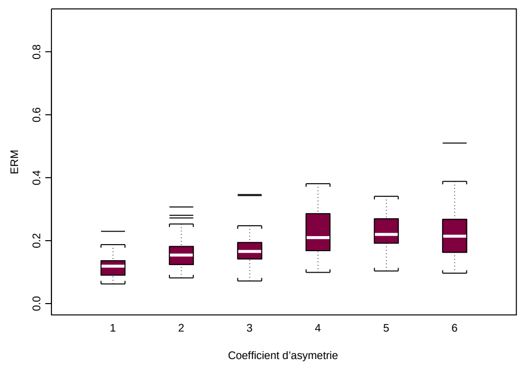
(a) Estimateur 1, $\alpha = 5\%$



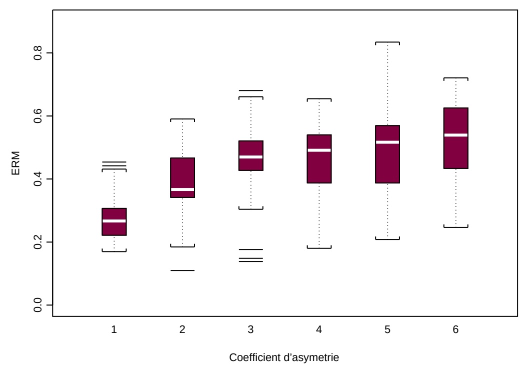
(b) Estimateur 1, $\alpha = 95\%$



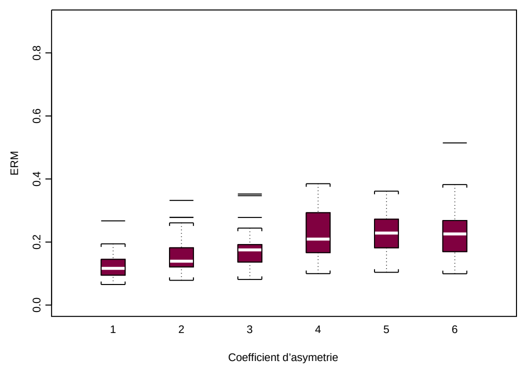
(c) Estimateur 2, $\alpha = 5\%$



(d) Estimateur 2, $\alpha = 95\%$



(e) Estimateur 3, $\alpha = 5\%$



(f) Estimateur 3, $\alpha = 95\%$

FIG. 4.1 – *Boîtes à moustaches pour les différentes erreurs relatives médianes obtenues avec les estimateurs des quantiles conditionnels unidimensionnels.*

Méthode semi paramétrique. L'estimateur LMS présenté au paragraphe 4.1.1 est calculé grâce à un logiciel spécifique écrit par T. J. Cole et P. Green. Ce logiciel nous a été fourni par les auteurs.

Méthodes non paramétriques. La covariable ne prend que des valeurs entières allant de 20 à 80 ans. Notons $\{z_t,\ t = 1,\ldots,T\}$ ces T valeurs distinctes. Pour chacun des estimateurs non paramétriques, le quantile conditionnel $q_\alpha(x)$ a été évalué en ces valeurs. On obtient donc l'ensemble des points $\{(z_t,\widehat{q_\alpha}(z_t)),\ t = 1,\ldots,T\}$. Pour la représentation graphique des courbes de référence, une approche basique consiste à réaliser une interpolation linéaire entre ces différents points. Cependant les courbes obtenues avec cette approche présentent un aspect visuel non "lisse". Ainsi, pour pallier ce défaut, nous avons opté pour un lissage par la méthode du noyau de ces points, le noyau choisi étant le noyau gaussien et la fenêtre utilisée étant obtenue par validation croisée.

4.1.4.2 Résultats

Nous précisons tout d'abord les critères utilisés pour accepter ou rejeter a posteriori les courbes de référence obtenues. Puis, nous résumons les résultats obtenus par les différentes méthodes. Nous remarquons enfin que, même lorsqu'il existe un modèle paramétrique et que les courbes de références sont acceptées, l'approche non paramétrique fonctionne bien.

Critère d'acceptabilité des courbes de références. Les courbes de référence (obtenues par les méthodes paramétrique, semi paramétrique ou non paramétriques) sont considérées comme acceptables si elles satisfont les trois conditions suivantes :

– Elles n'incluent pas de valeurs impossibles pour Y (par exemple des valeurs nulles ou négatives alors que la variable Y ne peut prendre en réalité que des valeurs strictement positives).

– Elles contiennent le pourcentage désiré d'individus à savoir $100(2\alpha - 1)\%$.

– Les valeurs individuelles qui se trouvent en dehors des limites des courbes de référence sont réparties de façon uniforme en fonction de la covariable et aucun regroupement de valeurs individuelles n'apparaît.

Synthèse des résultats obtenus par la méthode paramétrique. Pour un certain nombre des variables étudiées, nous avons pu établir des courbes de référence en fonction de l'âge par la méthode paramétrique. Le tableau 4.1(a) résume les résultats ainsi obtenus. Précisons que le modèle paramétrique est acceptable lorsque les résidus sont normalement distribués (après transformation au préalable des données si nécessaire). Si le modèle paramétrique n'est pas accepté, les courbes de référence ne peuvent alors être construites. Il apparaît que la modélisation polynomiale est appropriée pour 25 variables sur 38. Seules 21 courbes de références sont ensuite acceptables.

Synthèse des résultats obtenus par la méthode semi paramétrique. Les résultats obtenus par la méthode LMS sont résumés dans le tableau 4.1(b). Pour la plupart des variables (34 sur 38) les courbes de référence sont acceptées. Les 4 courbes restantes ne répondent pas au premier critère.

Synthèse des résultats obtenus par les méthodes non paramétriques. Les courbes de référence en fonction de l'âge ont été établies avec succès pour tous les paramètres biophysiques analysés par l'estimateur à noyau (méthode 1). Pour la presque totalité des paramètres, la méthode de la constante locale (méthode 2) donne des courbes de référence acceptables. Par contre, les courbes de référence obtenues par la méthode du noyau produit (méthode 3) sont moins satisfaisantes. Il apparaît clairement que la règle empirique pour le choix de $h_{2,n}$ est mal adaptée et entraîne des courbes de référence sur-lissées. Pour remédier à ce problème déjà constaté sur simulations, d'autres méthodes de choix ne reposant pas sur une hypothèse de normalité doivent être envisagées. Le tableau 4.1 résume l'ensemble des résultats obtenus avec les trois méthodes non paramétriques.

	Joue	Front	Avant-bras
Nombre de variables	13	13	12
Méthode paramétrique			
Nombre de modèles acceptés	7	10	8
Nombre de courbes de référence acceptées	6	9	6
Méthode semi paramétrique			
Nombre de courbes de référence acceptées	11	11	12
Méthodes non paramétriques			
Méthode 1 : nombre de courbes de référence acceptées	13	13	12
Méthode 2 : nombre de courbes de référence acceptées	12	12	12
Méthode 3 : nombre de courbes de référence acceptées	6	5	3

TAB. 4.1 – *Récapitulatif des résultats obtenus avec les différentes méthodes (méthode 1 : méthode d'estimation par noyau, méthode 2 : méthode de la constante locale, méthode 3 : méthode d'estimation par noyau produit).*

Illustration graphique. Lorsqu'un modèle paramétrique a pu être construit et que les courbes de référence ont été déclarées acceptables, les approches semi paramétrique et non paramétrique donnent des résultats semblables. Par exemple, sur la figure 4.2 toutes les courbes de référence obtenues avec les diverses méthodes évoluent de façon comparable sur la fourchette d'âges étudiée. D'autres situations sont examinées en détail dans [48].

4.2 Covariable multidimensionnelle

L'extension formelle des estimateurs de quantiles conditionnels à des covariables de dimension $p > 1$ (notée $\mathbf{X}$ dans ce cadre multidimensionnel) ne pose pas de problème. Cependant, la mise en œuvre des estimateurs souffre du fléau de la dimension ou *curse of dimensionality*. Plus généralement, la rareté des observations en grande dimension est un problème récurrent en estimation non paramétrique. Ceci apparaît également sur le plan théorique, la vitesse de convergence des estimateurs non paramétriques diminuant lorsque p augmente. Pour plus de détails sur cette question, on pourra se reporter à [116]. De plus, dans ce cas les courbes de référence sont en fait un couple d'hypersurfaces de dimension p, leur visualisation est difficile, et elles sont alors moins utiles pour les analyses exploratoires. Il est par exemple délicat de détecter graphiquement si un individu est normal ou non.

Notre but est de réduire la dimension du vecteur $\mathbf{X}$ sans perte d'information sur la loi conditionnelle de Y sachant $\mathbf{X}$ et sans utiliser de modèle paramétrique. Des études similaires existent dans le cadre

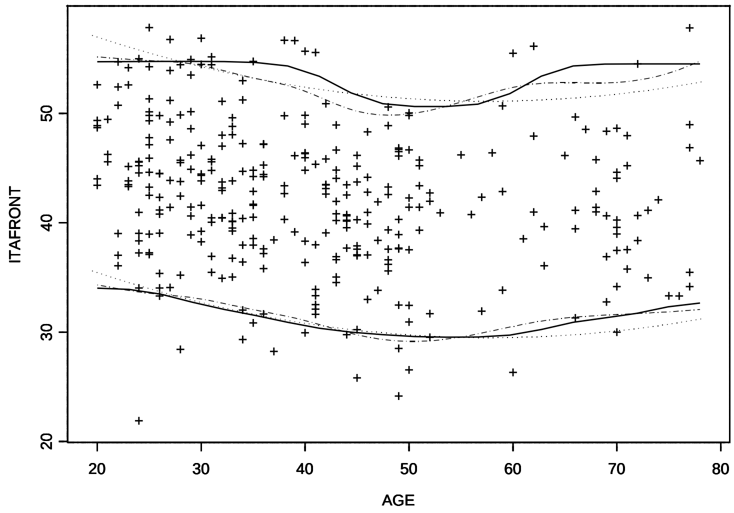


FIG. 4.2 – *Exemples de courbes de référence à 90% obtenues avec la méthode paramétrique (pointillés), la méthode LMS (tirets) et l'estimateur à noyau, méthode 1 (ligne continue).*

de la régression. Par exemple, dans [117], des modèles additifs sont utilisés pour pallier le fléau de la dimension. Ici, une méthode de projection linéaire est considérée afin de réduire la dimension des covariables et obtenir un estimateur efficace des quantiles conditionnels. La réduction de dimension proprement dite est basée sur la méthode SIR (*slice inverse regression*) introduite dans [99].
Les aspects théoriques liés à la réduction de dimension en régression ainsi que la méthode SIR sont présentés paragraphe 4.2.1. Un estimateur semi paramétrique des quantiles conditionnels basé sur cette méthode est introduit paragraphe 4.2.2. Quelques résultats asymptotiques sont établis paragraphe 4.2.3. Les performances de l'estimateur proposé sont brièvement illustrées sur simulations paragraphe 4.2.4. Enfin, dans le paragraphe 4.2.5, la méthode est mise en œuvre dans le cadre de l'application décrite paragraphe 4.1. L'étude complète est publiée dans [51].

4.2.1 Aspects théoriques de la réduction de dimension en régression

Dans ce paragraphe, nous introduisons la notion de sous-espace de réduction de dimension ainsi que le principe de construction d'une base de sous-espace par la méthode SIR. Nous insistons sur les conséquences concernant l'estimation des quantiles conditionnels.

Sous-espaces de réduction de dimension. On suppose l'existence d'une matrice B de taille $p \times r$ telle que

$$F(y|\mathbf{x}) = F(y|\,^tB\mathbf{x}), \tag{4.5}$$

où $F(.|.)$ est la fonction de répartition conditionnelle de Y. Une telle matrice existe toujours puisque (4.5) est trivialement vraie avec $B = I_p$ la matrice identité de taille $p \times p$. L'hypothèse (4.5) implique que le vecteur des prédicteurs $\mathbf{X}$ de taille $p \times 1$ peut être remplacé par le vecteur $^tB\mathbf{X}$

de taille $r \times 1$ sans perte d'information de régression. Si $r < p$ alors on réalisé une réduction de dimension. La définition suivante [99] est introduite :

Définition 4.2.1 *Le sous-espace vectoriel $S(B)$ engendré par les colonnes de B est appelé un sous-espace de réduction de dimension.*

La dimension de $S(B)$ révèle le nombre de composantes linéaires de $\mathbf{X}$ nécessaires pour expliquer Y. Lorsque (4.5) est vraie, alors cette propriété est encore vraie en remplaçant B par tout matrice dont les colonnes forment une base de $S(B)$. Il est donc clair que la connaissance du sous-espace de réduction de dimension de plus petite dimension fournit la caractérisation de Y sachant $\mathbf{X}$ la plus parcimonieuse. On a alors la définition suivante [23] :

Définition 4.2.2 *On note $S_{Y|\mathbf{X}}$ l'unique sous-espace de réduction de dimension de dimension minimale, appelé sous-espace central de réduction de dimension parfois abrégé espace EDR pour effective dimension reduction subspace.*

Soit $d = \dim(S_{Y|\mathbf{X}})$, $d \leq r$, la dimension de ce sous-espace et β la matrice de taille $p \times d$ dont les colonnes forment une base de $S_{Y|\mathbf{X}}$. D'après (4.5), on a alors $q_\alpha(\mathbf{x}) = q_\alpha({}^t\beta\mathbf{x})$.

Définition 4.2.3 *La courbe de régression inverse centrée est l'ensemble*

$$S_{\mathbb{E}[\mathbf{X}|Y]} = \{\mathbb{E}[\mathbf{X}|Y] - \mathbb{E}[\mathbf{X}] : \ Y \in \Omega_Y\}$$

où $\Omega_Y \in \mathbb{R}$ est l'ensemble des valeurs de Y.

Sous (4.5), on suppose que la loi marginale de $\mathbf{X}$ satisfait la condition de linéarité suivante :

(LC) : Pour tout $b \in \mathbb{R}^p$, $\mathbb{E}\left[{}^t b\mathbf{X}|\, {}^t\beta\mathbf{X}\right]$ est linéaire en ${}^t\beta\mathbf{X}$.

Soit Σ la matrice de covariance de $\mathbf{X}$, supposée définie positive. Sous **(LC)**, il est prouvé [99] que la courbe de régression inverse centrée appartient au sous-espace engendré par les colonnes de $\Sigma\beta$ noté $S(\Sigma\beta)$. On a alors

$$S_{\mathbb{E}[\mathbf{X}|Y]} \subseteq S(\Sigma\beta) = \Sigma S_{Y|\mathbf{X}}. \tag{4.6}$$

Soit $\mathbf{Z}$ le prédicteur $\mathbf{X}$ normalisé défini par $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \mathbb{E}[\mathbf{X}])$. On montre [23] qu'il est possible de se restreindre à travailler sur $\mathbf{Z}$ sans perte de généralité puisque tout base de $S_{Y|\mathbf{Z}}$ peut être transformée en une base de $S_{Y|\mathbf{X}}$ selon le principe $S_{Y|\mathbf{X}} = \Sigma^{-1/2}S_{Y|\mathbf{Z}}$. Comme conséquence de (4.6), on a la propriété [23] :

Proposition 4.2.1 *Sous **(LC)**, on a $S_{\mathbb{E}[\mathbf{Z}|Y]} \subseteq S(\eta) = S_{Y|\mathbf{Z}}$, avec $\eta = \Sigma^{1/2}\beta$.*
De plus $S(\mathrm{Var}\,[\mathbb{E}[\mathbf{Z}|Y]]) = S_{\mathbb{E}[\mathbf{Z}|Y]}$, sauf sur un ensemble de mesure nulle.

Le sous-espace central de réduction de dimension peut donc être estimé à partir de la courbe de régression inverse $S_{\mathbb{E}[\mathbf{Z}|Y]}$ via l'estimation de la matrice $\mathrm{Var}\,[\mathbb{E}[\mathbf{Z}|Y]]$. Les méthodes basées sur ce principe n'estiment en fait que des parties du sous-espace central de réduction de dimension, l'égalité entre $S_{\mathbb{E}[\mathbf{Z}|Y]}$ et $S_{Y|\mathbf{Z}}$ n'étant pas garantie. Le paragraphe suivant est dédié à la méthode SIR introduite dans [99] et qui repose sur une estimation non lisse et non paramétrique de $S_{Y|\mathbf{Z}}$.

La méthode SIR. La méthode SIR est basée sur un partitionnement des valeurs de Y en un nombre fixé H de tranches notées $\mathcal{S}_1, \ldots, \mathcal{S}_H$. Les p composantes de $\mathbf{Z}$ sont alors régressées sur $\widetilde{Y}$ la version discrétisée de Y par le découpage en tranches. Cette régression inverse se traduit simplement par p régressions unidimensionnelles. Soit la matrice $M = \mathrm{Var}[\mathbb{E}[\mathbf{Z}|\widetilde{Y}]]$. On déduit de la Proposition 4.2.1 la série d'inclusions suivante : $S(M) = S_{\mathbb{E}[\mathbf{Z}|\widetilde{Y}]} \subseteq S_{\widetilde{Y}|\mathbf{Z}} \subseteq S_{Y|\mathbf{Z}}$, la dernière inclusion étant une conséquence du fait que $\widetilde{Y}$ est une fonction de Y, et donc que $S_{Y|\mathbf{Z}}$ est un sous-espace de réduction de dimension pour la régression de $\widetilde{Y}$ sur $\mathbf{Z}$. A partir du découpage $\mathcal{S}_1,\dots,\mathcal{S}_H$, on a la décomposition :

$$M = \sum_{h=1}^{H} p_h m^h \, {}^t m^h, \tag{4.7}$$

avec $p_h = \mathrm{P}(Y \in \mathcal{S}_h)$ et $m^h = \mathbb{E}[\mathbf{Z}|Y \in \mathcal{S}_h]$. En supposant $d = \dim(S(M))$, on note $s_1 \geq \dots \geq s_d$ les d valeurs propres non nulles de M, $u^1,\dots,u^d$ les vecteurs propres correspondants et $b^k = \Sigma^{-1/2}u^k$, $k = 1,\dots,d$. On a alors $S(M) = S(u^1,\dots,u^d)$, et $b^1,\dots,b^d$ forment une base de $S(\beta)$ et sont parfois appelées directions EDR (cf Définition 4.2.2).

4.2.2 Procédure d'estimation

Dans la suite, on note y_i la $i$ème observation de la variable réponse unidimensionnelle Y et $\mathbf{x}_i$ le vecteur de taille $p \times 1$ des valeurs observées de la covariable, $i = 1,\dots,n$.

Etape d'estimation liée à SIR. Soient $\overline{\mathbf{x}}$ et $\widehat{\Sigma}$ la moyenne et la matrice de covariance empiriques des $\mathbf{x}_i$. On note $\widehat{\mathbf{z}}_i$ le $i$ème prédicteur normalisé défini par $\widehat{\mathbf{z}}_i = \widehat{\Sigma}^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}})$, $i = 1,\dots,n$. L'estimation de M définie en (4.7) par la méthode SIR est donnée par

$$\widehat{M} = \sum_{h=1}^{H} \widehat{p}_h \widehat{m}^h \, {}^t \widehat{m}^h,$$

avec $\widehat{p}_h = n_h/n$ où n_h est le nombre d'observations dans la $h$ème tranche et $\widehat{m}^h$ est le vecteur obtenu en moyennant les $\widehat{\mathbf{z}}_\mathbf{i}$ de la tranche h. On introduit $\hat{s}_1 \geq \dots \geq \hat{s}_p$ les valeurs propres de $\widehat{M}$ et $\hat{u}^1,\dots,\hat{u}^p$ les vecteurs propres associés. Si la dimension d de $S(M)$ est connue, $S(\widehat{M}) = S(\hat{u}^1,\dots,\hat{u}^d)$ est alors un estimateur consistant de $S(M)$. Dans la pratique, la dimension d est estimée par $\hat{d}$, le nombre de valeurs propres que l'on estime non nulles [99].
Lorsque $\dim(S(\eta)) = d$, $\widehat{M}$ fournit une estimation d'une base de $S(\eta)$ et les directions EDR estimées $\hat{b}^k = \widehat{\Sigma}^{-1/2}\hat{u}^k$, $k = 1,\dots,d$ sont une estimation d'une base de l'espace EDR $S(\beta)$.

Estimation des quantiles conditionnels. Par souci de simplicité, on suppose pour l'instant $d = 1$. On note $\hat{b} = \hat{b}^1$ la direction estimée de $S_{Y|\mathbf{X}} = S(\beta)$ et $\hat{v} = {}^t\hat{b}\mathbf{x}$ la projection des observations de la covariable sur cette direction, appelée indice. De façon similaire à (4.2), on introduit l'estimateur à noyau de $F(y|\mathbf{x})$ à partir des observations $(y_i,\hat{v}_i)$, $i = 1,\dots,n$:

$$F_n(y|\,{}^t\hat{b}\mathbf{x}) = F_n(y|\hat{v}) = \sum_{i=1}^{n} K\left(\frac{\hat{v} - \widehat{v}_i}{h_n}\right) \mathbb{I}\{y_i \leq y\} \Bigg/ \sum_{i=1}^{n} K\left(\frac{\hat{v} - \widehat{v}_i}{h_n}\right). \tag{4.8}$$

On déduit de (4.8) un estimateur de $q_\alpha(\mathbf{x})$ par

$$q_{n,\alpha}({}^t\hat{b}\mathbf{x}) = q_{n,\alpha}(\hat{v}) = F_n^{-1}(\alpha \mid \hat{v}), \tag{4.9}$$

et les courbes de références à $100 \times (2\alpha - 1)\%$ sont estimées pour $\alpha > 0.5$ par

$$I_{n,\alpha}(\mathbf{x}) = [q_{n,1-\alpha}(\hat{v}), q_{n,\alpha}(\hat{v})] = [q_{n,1-\alpha}({}^t\hat{b}\mathbf{x}), q_{n,\alpha}({}^t\hat{b}\mathbf{x})]. \tag{4.10}$$

L'extension de ce principe d'estimation au cas $d > 1$ ne pose pas de problème, on travaille alors avec plusieurs directions EDR $\hat{b}^j = \widehat{\Sigma}^{-1/2}\hat{u}^j$, $j = 1,\dots,d$, plusieurs indices $({}^t\hat{b}^1\mathbf{x}_i,\dots,{}^t\hat{b}^d\mathbf{x}_i)$, $i = 1,\dots,n$ et on utilise un noyau multidimensionnel dans (4.8) pour obtenir $q_{n,\alpha}({}^t\hat{b}^1\mathbf{x},\dots,{}^t\hat{b}^d\mathbf{x})$.

4.2.3 Propriétés asymptotiques

Dans le cas $d = 1$, nous avons établi la consistance de $q_{n,\alpha}({}^t\hat{b}\mathbf{x})$ dans [51] sous les hypothèses suivantes :

(A1) Les vecteurs aléatoires $(\mathbf{X}_i, Y_i)$, $i = 1,\dots,n$ sont indépendants et de même loi.

(A2) Le noyau $K : \mathbb{R} \longrightarrow \mathbb{R}$ est une densité de probabilité bornée telle que $|v|K(v) \to 0$ quand $|v| \to \infty$, $\int vK(v)dv = 0$ et $\int v^2K(v)dv < \infty$.

(A3) $h_n \to 0$ et $nh_n/\log n \to \infty$ quand $n \to \infty$.

(A4) La densité de $\mathbf{X}$ est continue.

(A5) Pour tout $\mathbf{x} \in \mathbb{R}^p$ et $y \in \mathbb{R}$, $F(. \mid {}^tb\mathbf{x})$ et $F(y \mid .)$ sont continues.

(A6) Quels que soient $\alpha \in]0,1[$ et $\mathbf{x} \in \mathbb{R}^p$, $F(. \mid {}^tb\mathbf{x})$ a un unique quantile d'ordre α.

Cette dernière hypothèse peut être évitée en utilisant l'inverse généralisé pour définir le quantile conditionnel par : $q_\alpha(\mathbf{x}) = \inf\{y : F(y \mid {}^tb\mathbf{x}) \geq \alpha\}$. On a les deux résultats de consistance suivants.

Théorème 4.2.1 *Sous les hypothèses (4.5), **(LC)**, **(A1)-(A5)**, et pour $\mathbf{x}$ fixé dans $\mathbb{R}^p$, on a*

$$\sup_{y\in\mathbb{R}} \left| F_n(y \mid {}^t\hat{b}\mathbf{x}) - F(y \mid \mathbf{x}) \right| \xrightarrow{P} 0.$$

*Si de plus **(A6)** est vérifiée alors $q_{n,\alpha}({}^t\hat{b}\mathbf{x}) \xrightarrow{P} q_\alpha(\mathbf{x})$.*

4.2.4 Validation sur simulations

Dans ce paragraphe, les performances de la méthode d'estimation proposée sont évaluées sur simulations. En particulier, nous comparons notre estimateur à l'estimateur non paramétrique classique n'incluant pas d'étape de réduction de dimension.

Méthodes d'estimation. Nous considérons trois estimateurs du quantile conditionnel d'ordre α au point $\mathbf{x}$:

(a) $q^{(a)}_{n,\alpha}(\mathbf{x}) = q_{n,\alpha}({}^t\hat{b}\mathbf{x})$ est l'estimateur défini en (4.9). La direction $\widehat{b}$ est estimée par la méthode SIR et le quantile conditionnel par inversion numérique de la fonction de répartition conditionnelle estimée (4.8). Le noyau choisi est le noyau gaussien et la largeur de fenêtre est calculée par validation croisée.

(b) $q^{(b)}_{n,\alpha}(\mathbf{x}) = q_{n,\alpha}({}^t\beta\mathbf{x})$ est introduit uniquement pour comparaison avec l'estimateur précédent. La direction de projection n'est pas estimée mais fixée à sa valeur théorique.

(c) $q^{(c)}_{n,\alpha}(\mathbf{x}) = q_{n,\alpha}(\mathbf{x})$ est l'estimateur non paramétrique classique des quantiles conditionnels. Il est calculé en inversant numériquement la fonction de répartition conditionnelle estimée par (4.2) en utilisant un noyau multidimensionnel. Nous avons retenu un noyau gaussien, et la largeur de fenêtre est, là encore, calculée par validation croisée.

Modèles simulés. Nous considérons deux modèles.

(M1) $Y = f(\beta^T \mathbf{X}) + \varepsilon$,

est un modèle de régression avec comme fonction de lien $f(t) = 1 + \exp(2t/3)$. Le vecteur aléatoire $\mathbf{X}$ suit la loi normale multidimensionnelle $N_p(0,I_p)$ et la variable aléatoire ε est normale centrée-réduite $\varepsilon \sim \mathcal{N}(0,1)$ et indépendante de $\mathbf{X}$. Ce modèle permet d'évaluer le comportement des estimateurs en fonction de la dimension $p \in \{3,5,\dots,13\}$. La direction β est définie par ${}^t\beta = (p-1)^{-1/2}[c,-c,0]$ avec $c = [1,\dots,1]$ le vecteur ligne de longueur $(p-1)/2$, Le quantile conditionnel théorique s'écrit $q_\alpha(\mathbf{x}) = f({}^t\beta\mathbf{x}) + z_\alpha$, où z_α est le quantile d'ordre α de la loi normale centrée-réduite.

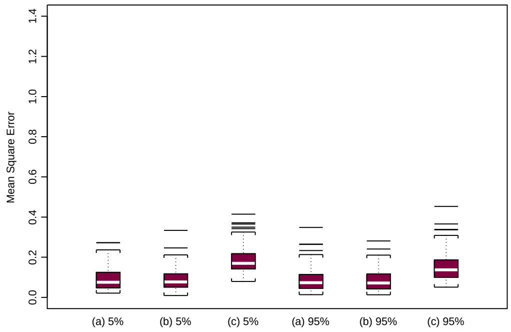
(M2) $Y = (1-\theta)g({}^t\beta\mathbf{X}) + \theta h(\mathbf{X}) + \varepsilon$,

est une modèle de mélange en dimension $p = 3$ avec comme fonction de lien $g(t) = 1 + 2t/3$. Le paramètre θ détermine l'importance de la contamination du modèle de régression par la fonction $h(x,y,z) = 2xyz/3$. Ce type de modèle permet d'évaluer la robustesse des méthodes d'estimation vis à vis du degré de contamination $\theta \in \{0,0.2,\dots,1\}$, et donc de l'hypothèse (4.5). De même que précédemment ${}^t\beta = 2^{-1/2}[1,-1,0]$, $\mathbf{X}$ suit la normale multidimensionnelle $N_3(0,I_3)$, ε est normal centré-réduit $\varepsilon \sim N(0,1)$ et indépendant de $\mathbf{X}$. De plus on a $\mathrm{Var}\left[g({}^t\beta\mathbf{X})\right] = \mathrm{Var}\left[h(\mathbf{X})\right] = 1$ et le quantile conditionnel théorique s'écrit $q_\alpha(\mathbf{x}) = (1-\theta)g({}^t\beta\mathbf{x}) + \theta h(\mathbf{x}) + z_\alpha$.

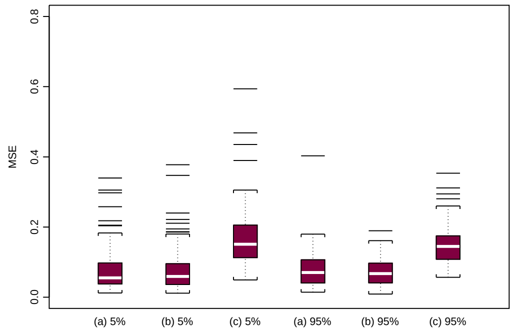
Evaluation des résultats. Les trois estimateurs **(a)**, **(b)** et **(c)** sont comparés au quantile théorique dans les situations **(M1)** et **(M2)**. Pour cela, $N = 100$ échantillons de taille $n = 200$ sont simulés dans chaque situation. Les quantiles conditionnels sont estimés pour $\alpha = 5\%$ et $\alpha = 95\%$ sur une grille en dimension p. Cette grille est constituée de 125 points $\{z_\ell,\ \ell = 1,\dots,125\}$ tirés au hasard suivant une loi uniforme sur $[-3/2,3/2]^p$. La performance des estimateurs est mesurée sur chacun des N échantillons par l'erreur quadratique moyenne :

$$E_{n,\alpha}^{(\Theta)} = \frac{1}{125}\sum_{\ell=1}^{125}\left(q_{n,\alpha}^{(\Theta)}(z_\ell) - q_\alpha(z_\ell)\right)^2, \quad \text{où } \Theta \in \{\mathbf{a},\mathbf{b},\mathbf{c}\}.$$

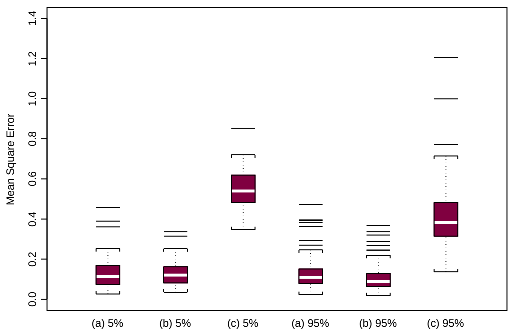
Exemple de résultats. Pour une présentation exhaustive des résultats obtenus, on pourra se reporter à [51]. La distribution empirique des erreurs $E_{n,\alpha}^{(\Theta)}$ pour $\Theta \in \{\mathbf{a},\mathbf{b},\mathbf{c}\}$ et $\alpha \in \{0.05,0.95\}$ est présentée figure 4.3. La première colonne présente les résultats obtenus sur le modèle **(M1)** avec $d \in \{3,9,13\}$. Il apparaît qu'il n'y pas de différence significative entre $E_{n,\alpha}^{(\mathbf{a})}$ et $E_{n,\alpha}^{(\mathbf{b})}$. L'estimation de la direction β par $\widehat{b}$ a peu de conséquences sur la qualité de l'estimation des courbes de référence, et ce quelle que soit la dimension. Par contre, les résultats obtenus par les estimateurs **(a)** et **(c)** sont très différents. L'estimateur **(a)** donne de meilleurs résultats que l'estimateur sans étape de réduction de dimension **(c)**. De plus la différence s'accentue avec le nombre p de covariables. Le fléau de la dimension constitue donc une limitation importante à l'estimateur **(c)** ce qui rend **(a)** particulièrement intéressant dans de telles situations. La seconde colonne présente les résultats obtenus sur le modèle **(M2)** avec $\theta \in \{0,0.6,1\}$. Les résultats obtenus avec l'estimateur **(a)** restent meilleurs ou comparables avec ceux obtenus par l'estimateur **(c)** lorsque la contamination ne dépasse pas 60%. Ce résultat est dû à la robustesse de la méthode SIR qui donne des estimations correctes des directions EDR dans cette plage de valeurs de θ.



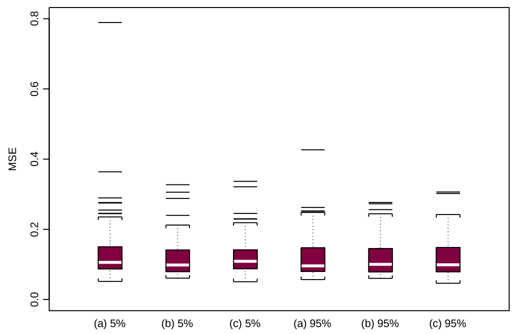
(a) Modèle **(M1)**, $d = 3$



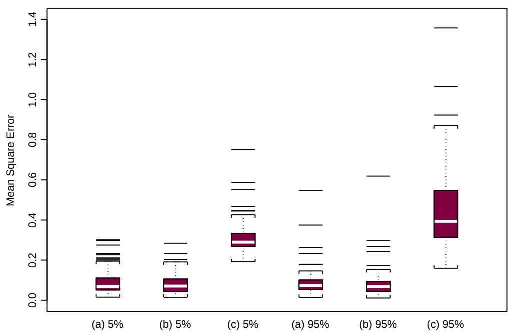
(d) Modèle **(M2)**, $\theta = 0$



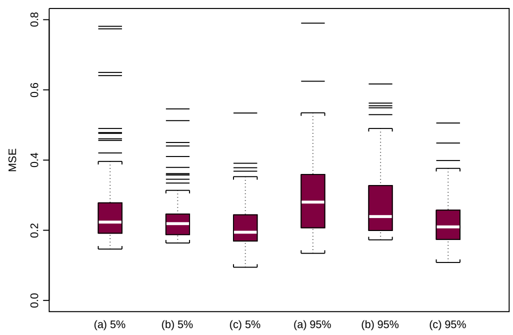
(b) Modèle **(M1)**, $d = 9$



(e) Modèle **(M2)**, $\theta = 0.6$



(c) Modèle **(M1)**, $d = 13$



(f) Modèle **(M2)**, $\theta = 1$

FIG. 4.3 – *Boîtes à moustaches pour les différentes erreurs quadratiques moyennes obtenues avec les estimateurs des quantiles conditionnels multidimensionnels.*

4.2.5 Application à des données réelles

Le contexte de cette étude est celui du paragraphe 4.1.4. L'étude complète est décrite dans [51]. Les covariables utilisées ici regroupent l'ensemble des variables décrites dans le paragraphe 4.1.4 ainsi que la température et l'humidité relative de l'environnement.

Procédure d'estimation. Notons $\mathcal{X}$ l'ensemble des p covariables et Y la variable d'intérêt.

– *Etape 1.* Le graphe représentant l'ébouli des valeurs propres obtenus par la méthode SIR permet de déterminer le nombre $\hat{d}$ de directions EDR à retenir. Ces directions sont notées $\hat{b}^1,\ldots,\hat{b}^{\hat{d}}$. On visualise alors la structure des données projetées $(y_i, {}^t\hat{b}^1\mathbf{x}_i,\ldots,{}^t\hat{b}^{\hat{d}}\mathbf{x}_i)$, $i=1,\ldots,n$.

– *Etape 2.* Le but de cette étape est de simplifier les indices ${}^t\hat{b}^k\mathbf{x}$, $k=1,\ldots,\hat{d}$ de façon à obtenir une interprétation plus aisée. Ainsi, pour chaque indice, une régression linéaire de ${}^t\hat{b}^k\mathbf{x}$ sur les covariables de $\mathcal{X}$ est effectuée avec une sélection *forward* des variables basée sur le critère AIC. On obtient ainsi un sous-ensemble $\mathcal{X}_1,\ldots,\mathcal{X}_{\hat{d}}$ de covariables sélectionnées parmi $\mathcal{X}$. Le sous-ensemble final de covariables retenues est alors $\widetilde{\mathcal{X}} = \cup_{k=1}^{\hat{d}}\mathcal{X}_k$. La procédure SIR est appliquée une nouvelle fois avec les covariables de $\widetilde{\mathcal{X}}$ et on obtient les $\hat{d}$ directions EDR estimées $\widetilde{b}^1,\ldots,\widetilde{b}^{\hat{d}}$. Enfin, on vérifie que les graphes $({}^t\hat{b}^k\mathbf{x}_i, {}^t\widetilde{b}^k\mathbf{x}_i)$, $i=1,\ldots,n$ ont une structure linéaire.

– *Etape 3.* Les courbes de références (qui sont en fait des hypersurfaces lorsque $\hat{d}>1$) sont alors calculées sur l'ensemble $(y_i, {}^t\widetilde{b}^1\mathbf{x}_i,\ldots,{}^t\widetilde{b}^{\hat{d}}\mathbf{x}_i)$, $i=1,\ldots,n$, par l'estimateur à noyau décrit paragraphe 4.2.2.

Résultats. A titre d'exemple nous avons choisi comme variable d'intérêt la conductance de la peau mesurée sur l'avant-bras. Les résultats de l'analyse montrent que six covariables entrent dans le modèle final : l'âge des volontaires, la température et l'humidité relative de l'environnement, le pH de la peau, sa capitance et sa perte insensible en eau. Les courbes de références à 90% sont présentées figure 4.4. Ces résultats ont été validés par les spécialistes du CERIES.

4.3 Perspectives

Afin de parachever l'étude de l'estimation des quantiles conditionnels dans le cas d'une covariable multidimensionnelle, il est nécessaire de compléter les résultats asymptotiques du paragraphe 4.2.3 en établissant la convergence en loi de $q_{n,\alpha}({}^t\hat{b}\mathbf{x})$. Cela doit permettre d'apprécier le gain de vitesse de convergence apporté par l'étape de réduction de dimension.

Du point de vue applicatif, il serait souhaitable de disposer d'un test permettant de comparer les espaces EDR obtenus sur deux échantillons indépendants. Ainsi, il serait possible de mettre en parallèle les courbes de références obtenues sur des populations différentes. Ce travail a été initié par Jérôme Saracco sous la forme d'une proposition de projet de DEA de Biostatistique à l'Université Montpellier 2. Dans le cadre des développements en cours sur le thème de l'estimation de courbes de référence, j'aimerais également citer les travaux de Ali Gannoun, Jérôme Saracco et Christiane Guinot [52] dans le cadre d'une variable d'intérêt Y multidimensionnelle. La réduction simultanée de la dimension de X et Y est rendue possible grâce à la méthode Alternating SIR [100] permettant de trouver les directions les plus prévisibles dans l'espace des valeurs de Y. Le calcul des quantiles conditionnels multivariés repose, quant à lui, sur le formalisme introduit dans [8].

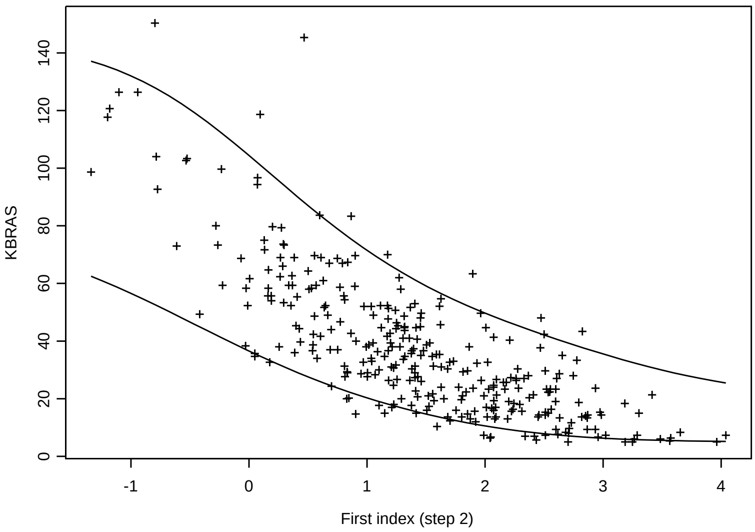


Fig. 4.4 – *Courbes de référence à 90% estimées pour la variable KBRAS (conductance mesurée sur l'avant-bras) à partir de l'indice calculé par la procédure décrite paragraphe 4.2.5.*

Chapitre 5

Perspectives

Les chapitres précédents reflètent l'état actuel de mes recherches sur les questions d'estimation de quantiles extrêmes, de frontière, et de courbes de référence ainsi que sur les problèmes de de réduction de dimension en analyse d'image. La présentation de chacun de ces quatre thèmes est conclue par mes travaux en cours et les perspectives soulevées dans le domaine de recherche correspondant. Ce dernier chapitre est dédié à la brève présentation d'une nouvelle direction de recherche, l'étude des copules, que je souhaite développer dans les années à venir en raison de ses applications probables dans mes thèmes de recherche actuels.

5.1 Construction et estimation de copules

J'ai débuté récemment mon travail sur les copules par l'introduction d'une nouvelle famille semi-paramétrique, l'étude de ses propriétés de dépendance et la mise en place d'une procédure d'estimation dans le cadre d'une collaboration avec Cécile Amblard (Université Grenoble 2).

Rappelons qu'une copule bivariée (pour simplifier les notations) est une fonction de répartition définie sur le carré unité et de marginales uniformes. Pour une présentation exhaustive de la théorie des copules, on pourra se reporter à [104]. Brièvement, l'étude des copules est motivée par le théorème de Sklar [114] établissant que toute fonction de répartition bivariée H de fonctions de répartition marginales F et G peut s'écrire $H(x,y) = C(F(x),G(y))$ où C est une copule. L'intérêt de cette décomposition réside dans le fait que les propriétés de dépendance entre les marginales sont entièrement déterminées par la copule C. Nous avons introduit [1] la famille de copules suivante :

$$C_\theta(u,v) = uv + \theta\phi(u)\phi(v), \;\; \theta \in [-1,1], \tag{5.1}$$

où ϕ est une fonction de [0,1] dans $\mathbb{R}$ vérifiant certaines conditions décrites dans [2], Théorème 1. La représentation (5.1) offre l'avantage d'englober différentes familles paramétriques existantes [109, 105] tout en les étendant. Les propriétés de dépendance modélisées par les copules de type (5.1) sont étudiées en détails dans [2] et les problèmes liés à l'estimation de la fonction ϕ sont abordés dans [3].

5.2 Domaines d'application

La théorie des copules est d'une grande utilité pour l'études des valeurs extrêmes multivariées. Par exemple, un couple de maxima de variables aléatoires peut être modélisé par une fonction de répartition bivariée de copule

$$C(u,v) = \exp(\log(uv)A(\log(u)/\log(uv))) \tag{5.2}$$

où A est une fonction univariée convexe, appelée fonction de dépendance [108]. L'estimation de A a été largement étudiée, en particulier dans [17]. Récemment, une famille de copules archimax incluant à la fois les copules de type (5.2) et les copules archimédiennes [62] a été introduite [18]. Plus généralement, l'étude des valeurs extrêmes multivariées est un domaine de recherche très actif actuellement et où de multiples questions intéressantes se posent. Ainsi, la définition même de la notion d'extrême multivarié est un problème délicat et peu de résultats existent encore sur ce thème en regard du cas univarié présenté au Chapitre 1 et des nombreuses applications potentielles. Je projette donc d'aborder les questions de modélisation et d'estimation posées en statistique des extrêmes multivariés en m'appuyant sur la théorie des copules.

J'aimerais également utiliser les copules pour étendre les travaux du Chapitre 2 à des supports de forme plus générale. L'estimation de régions de grande probabilité mise en œuvre [3] dans le cas des copules de type (5.1) semble être un bon point de départ pour cela.

Les méthodes de réduction de dimension et d'analyse d'image décrites Chapitre 3 peuvent également bénéficier de l'utilisation de la théorie des copules. Ainsi, il parait intéressant d'étudier la forme de la copule associée à une fonction auto-associative telle qu'elle est introduite Définition 3.1.1. Ce lien permettrait d'évaluer le degré de généralité des modèles auto-associatifs. De plus, l'extension des modèles actuels à des variétés paramétrables par des couples de directions pourrait également être basée sur l'utilisation de copules bivariées. Enfin, dans le cadre de l'application des méthodes de réduction de dimension à l'analyse d'image, la structure de dépendance spatiale inhérente aux données pourrait être modélisée elle aussi via l'utilisation de copules.

Bibliographie

[1] Amblard, C. et Girard, S. – A semiparametric family of symmetric bivariate copulas. *Comptes-Rendus de l'Académie des Sciences, Série I*, vol. 333, 2001, pp. 129–132.

[2] Amblard, C. et Girard, S. – Symmetry and dependence properties within a semiparametric family of bivariate copulas. *Nonparametric Statistics*, vol. 14, n° 6, 2002, pp. 715–727.

[3] Amblard, C. et Girard, S. – Estimation procedures for a semiparametric family of bivariate copulas. *Journal of Computational and Graphical Statistics*, 2004. – *A paraître.*

[4] Ballard, B. et Brown, C. – *Computer vision.* – Prentice-Hall, 1983.

[5] Beirlant, J., Broniatowski, M., Teugels, J.L. et Vynckier, P. – The mean residual life function at great age: Applications to tail estimation. *Journal of Statistical Planning and Inference*, vol. 45, n° 1-2, 1995, pp. 21–48.

[6] Beirlant, J., de Wet, T. et Goegebeur, Y. – *Nonparametric Estimation of Extreme Conditional Quantiles.* – Rapport technique n° TR 2002-07, KU Leuven - University Centre for Statistics, 2002.

[7] Beirlant, J., Dierckx, G., Goegebeur, Y. et Matthys, G. – Tail index estimation and an exponential regression model. *Extremes*, vol. 2, n° 2, 1999, pp. 177–200.

[8] Berlinet, A., Gannoun, A. et Matzner-Løber, E. – Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, vol. 35, 2001, pp. 139–169.

[9] Berred, M. – Record values and the estimation of the Weibull tail-coefficient. *Comptes-Rendus de l'Académie des Sciences, Série I*, vol. 312, 1991, pp. 943–946.

[10] Besse, P. et Ferraty, F. – A fixed effect curvilinear model. *Computational Statistics*, vol. 10, n° 4, 1995, pp. 339–351.

[11] Bingham, N.H., Goldie, C.M. et Teugels, J.L. – *Regular Variation.* – Cambridge University Press, 1987, *Encyclopedia of Mathematics and its application*, volume 27.

[12] Bouchard, G., Girard, S., Iouditski, A. et Nazin, A. – *Linear programming problems for frontier estimation.* – Rapport technique n° RT-0304, IAP network, 2003. `http://www.stat.ucl.ac.be/Iapdp/tr2003/TR0304.ps`.

[13] Bouchard, G., Girard, S., Iouditski, A. et Nazin, A. – Linear programming problems for frontier estimation. *Applied Stochastic Models in Business and Industry*, 2004. – *A paraître.*

[14] Bouchard, G., Girard, S., Iouditski, A. et Nazin, A. – Nonparametric frontier estimation by linear programming. *Automation and Remote Control*, vol. 65, n° 1, 2004, pp. 58–64.

[15] Breiman, L., Stone, C.J. et Kooperberg, C. – Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, vol. 37, 1990, pp. 127–149.

[16] Broniatowski, M. – On the estimation of the Weibull tail coefficient. *Journal of Statistical Planning and Inference*, vol. 35, n° 3, 1993, pp. 349–365.

[17] Capéraà, P., Fougères, A.-L. et Genest, C. – A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, vol. 84, n° 3, 1997, pp. 567–577.

[18] Capéraà, P., Fougères, A.-L. et Genest, C. – Bivariate distributions with given extreme value attractor. *Journal of Multivariate Analysis*, vol. 72, n° 1, 2000, pp. 30–49.

[19] Castillo, E. – *Extreme value theory in engineering.* – Academic Press, Inc., 1988, *Statistical Modeling and Decision Science.*

[20] Chalmond, B. – *Modeling and inverse problems in image analysis.* – Springer-Verlag, 2002, *Applied Mathematical Sciences*, volume 155.

[21] Chalmond, B. et Girard, S. – Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Pattern Analysis and Machine Intelligence*, vol. 21, n° 5, 1999, pp. 422–432.

[22] Cole, T.J. – Fitting smoothed centile curves to reference data. *Journal of Royal Statistical Society, Series A*, vol. 151, 1988, pp. 385–418.

[23] Cook, R.D. – *Regression graphics: Ideas for studying the regressions through graphics.* – New-York, Wiley, 1998.

[24] Cottrell, M. – *Analyse des données et réseaux de neurones.* – Rapport de recherche, SAMOS, Université Paris 1, 1994.

[25] Cowling, A. et Hall, P. – On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society B*, vol. 58, 1996, pp. 551–563.

[26] Cristianini, N. et Shawe-Taylor, J. – *An introduction to support vector machines.* – Cambridge University Press, 2000.

[27] Damsleth, E. – Conjugate classes for gamma distributions. *Scandinavian Journal of Statistics, Theory and Applications*, vol. 2, 1975, pp. 80–84.

[28] Davison, A. et Smith, R. – Models for exceedances over high thresholds. *Journal of Royal Statistical Society B*, vol. 52, n° 3, 1990, pp. 393–442.

[29] de Haan, L. et Rootzen, H. – On the estimation of high quantiles. *Journal of Statistical Planning and Inference*, vol. 35, n° 1, 1993, pp. 1–13.

[30] Demartines, P. – *Analyse de données par réseaux de neurones auto-organisés.* – Thèse de doctorat, PhD, Institut National Polytechnique de Grenoble, 1994.

[31] Deprins, D., Simar, L. et Tulkens, H. – Measuring labor efficiency in post offices. *In : The Performance of Public Enterprises: Concepts and Measurements*, éd. par M. Marchand, P. Pestieau et Tulkens, H. – Amsterdam, North Holland ed, 1984.

[32] Diebolt, J., Ecarnot, J., Garrido, M., Girard, S. et Lagrange, D. – Le logiciel Extremes, un outil pour l'étude des queues de distribution. *La revue de Modulad*, vol. 30, 2003, pp. 53–60.

[33] Diebolt, J., El-Aroui, M., Garrido, M. et Girard, S. – *Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling.* – Rapport technique n° RR-4803, INRIA, 2003. `http://www.inria.fr/rrrt/rr-4803.html`.

[34] Diebolt, J., Garrido, M. et Girard, S. – A goodness-of-fit test for the distribution tail. – 2002. Soumis pour publication.

[35] Diebolt, J., Garrido, M. et Girard, S. – Asymptotic normality of the ET method for extreme quantile estimation. Application to the ET test. *Comptes-Rendus de l'Académie des Sciences, Série I*, vol. 337, 2003, pp. 213–218.

[36] Diebolt, J. et Girard, S. – *Modélisation des queues de distributions et estimation de quantiles extrêmes.* – Rapport de contrat, INRIA–EDF, 1998.

[37] Diebolt, J. et Girard, S. – A Note on the asymptotic normality of the ET method for extreme quantile estimation. *Statistics and Probability Letters*, vol. 62, n° 4, 2003, pp. 397–406.

[38] Ditlevsen, O. – Distribution Arbitrariness in Structural Reliability. *In : Structural Safety and Reliability*, éd. par Schuller, Shinozuka et Yao, pp. 1241–1247. – Rotterdam, Balkema, 1994.

[39] Drees, H. et Kaufmann, E. – Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, vol. 75, n° 2, 1998, pp. 149–172.

[40] Durand, J-F. – Generalized principal component analysis with respect to instrumental variables via univariate spline transformations. *Computational Statistics and Data Analysis*, vol. 16, 1993, pp. 423–440.

[41] El-Aroui, M. A. et Diebolt, J. – On the use of the peaks over thresholds method for estimating out-of-sample quantiles. *Computational Statistics and Data Analysis*, vol. 39, n° 4, 2002, pp. 453–476.

[42] Embrechts, P., Klüppelberg, C. et Mikosh, T. – *Modelling Extremal Events.* – Springer-Verlag, 1997, *Applications of Mathematics*, volume 33.

[43] Farrel, M.J. – The measurement of productive efficency. *Journal of the Royal Statistical Society A*, vol. 120, 1957, pp. 253–281.

[44] Friedman, J.H. – Exploratory projection pursuit. *Journal of the American Statistical Association*, vol. 82, n° 397, 1987, pp. 249–266.

[45] Friedman, J.H. et Stuetzle, W. – Projection pursuit regression. *Journal of the American Statistical Association*, vol. 76, n° 376, 1981, pp. 817–823.

[46] Galambos, J. – *The Asymptotic Theory of Extreme Order Statistics.* – R.E. Krieger publishing compagny, 1987.

[47] Gannoun, A. – Estimation non paramétrique de la médiane conditionnelle : médianogramme et méthode du noyau. *Publications de l'Institut de Statistique de l'Université de Paris*, vol. XXXXV, 1990, pp. 11–22.

[48] Gannoun, A., Girard, S., Guinot, C. et Saracco, J. – Reference ranges based on nonparametric quantile regression. *Statistics in Medicine*, vol. 21, n° 20, 2002, pp. 3119–3135.

[49] Gannoun, A., Girard, S., Guinot, C. et Saracco, J. – Trois méthodes non paramétriques pour l'estimation de courbes de référence - application à l'analyse de propriétés biophysiques de la peau. *Revue de Statistique Appliquée*, vol. L, n° 1, 2002, pp. 65–89.

[50] Gannoun, A., Girard, S., Guinot, C. et Saracco, J. – Implémentation en C d'estimateurs non-paramétriques de quantiles conditionnels. Application au tracé de courbes de référence. *La revue de Modulad*, vol. 31, 2004, pp. 59–70.

[51] Gannoun, A., Girard, S., Guinot, C. et Saracco, J. – Sliced inverse regression in reference curves estimation. *Computational Statistics and Data Analysis*, vol. 46, n° 1, 2004, pp. 103–122.

[52] Gannoun, A., Guinot, C. et Saracco, J. – Reference curves estimation via alternating sliced inverse regression. *Environmetrics*, 2004. – *A paraitre.*

[53] Gardes, L. – Estimating the support of a Poisson process via the Faber-Shauder basis and extreme values. *Publications de l'Institut de Statistique de l'Université de Paris*, vol. XXXXVI, 2002, pp. 43–72.

[54] Gardes, L. – Double-thresholded estimator of extreme value index. *Comptes-Rendus de l'Académie des Sciences, Série I*, vol. 337, 2003, pp. 287–292.

[55] Gardes, L. – *Estimation d'une fonction quantile extrême.* – Thèse de doctorat, PhD, Université Montpellier 2, octobre 2003.

[56] Gardes, L. et Girard, S. – Asymptotic properties of a Pickands type estimator of the extreme value index. *In : Focus on probability theory*, éd. par Colombus, F. – New-York, Nova Science, 2004. *A paraître.*

[57] Gardes, L. et Girard, S. – Estimating extreme quantiles of Weibull tail-distributions. *Communication in Statistics - Theory and Methods*, 2004. – *A paraître.*

[58] Garrido, M. – *Modélisation des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution.* – Thèse de doctorat, PhD, Université Grenoble 1, juin 2002.

[59] Geffroy, J. – Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, vol. XIII, 1964, pp. 191–210.

[60] Geffroy, J. – *Sur un problème de loi-limite pour la distance L_1 entre une fonction et une approximation stochastique de celle-ci.* – Rapport technique n° 02-01, ENSAM-INRA-Université Montpellier II, 2002.

[61] Geffroy, J., Girard, S. et Jacob, P. – *Asymptotic normality of the L_1-error of a boundary estimate.* – Rapport technique n° 03-04, ENSAM-INRA-Université Montpellier II, 2003. `http://www.inrialpes.fr/is2/people/girard/RR0304.pdf`.

[62] Genest, C. et MacKay, R. – Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, vol. 14, 1986, pp. 145–159.

[63] Gijbels, I., Mammen, E., Park, B. U. et Simar, L. – On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association*, vol. 94, n° 445, 1999, pp. 220–228.

[64] Gijbels, I. et Peng, L. – Estimation of a support curve via order statistics. *Extremes*, vol. 3, 2000, pp. 251–277.

[65] Girard, S. – *Construction et apprentissage statistique de modèles auto-associatifs non-linéaires. Application à l'identification d'objets déformables en radiographie.* – Thèse de doctorat, PhD, Université de Cergy-Pontoise, octobre 1996.

[66] Girard, S. – A nonlinear PCA based on manifold approximation. *Computational Statistics*, vol. 15, n° 2, 2000, pp. 145–167.

[67] Girard, S. – A Hill type estimate of the Weibull tail-coefficient. *Communication in Statistics - Theory and Methods*, vol. 33, n° 2, 2004, pp. 205–234.

[68] Girard, S. – On the asymptotic normality of the L_1- error for Haar series estimates of Poisson point processes boundaries. *Statistics and Probability Letters*, vol. 66, 2004, pp. 81–90.

[69] Girard, S., Chalmond, B. et Dinten, J-M. – Designing non linear models for flexible curves. *Curves and Surfaces with Application in CAGD* , A. Le Méhauté, C. Rabut, and L.L. Schumaker (eds.), 1997, pp. 135–142.

[70] Girard, S., Chalmond, B. et Dinten, J-M. – Position of principal component analysis among auto-associative composite models. *Comptes-Rendus de l'Académie des Sciences, Série I*, vol. 326, 1998, pp. 763–768.

[71] Girard, S., Chalmond, B. et Dinten, J-M. – Une ACP non-linéaire basée sur l'approximation par variétés. *Revue de Statistique Appliquée*, vol. XLVI, n° 3, 1998, pp. 5–19.

[72] Girard, S. et Diebolt, J. – Consistency of the ET method and smooth variations. *Comptes-Rendus de l'Académie des Sciences, Série I*, vol. 329, 1999, pp. 821–826.

[73] Girard, S., Dinten, J-M. et Chalmond, B. – Building and training radiographic flexible prior models for object identification from incomplete data. *IEE proceedings on Vision, Image and Signal Processing*, vol. 143, n° 4, 1996, pp. 257–264.

[74] Girard, S. et Iovleff, S. – Auto-associative models and generalized principal component analysis. *Journal of Multivariate Analysis*, vol. 93, n° 1, 2005, pp. 21–39.

[75] Girard, S. et Jacob, P. – Extreme values and Haar series estimates of point process boundaries. *Scandinavian Journal of Statistics*, vol. 30, n° 2, 2003, pp. 369–384.

[76] Girard, S. et Jacob, P. – Projection estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, vol. 116, n° 1, 2003, pp. 1–15.

[77] Girard, S. et Jacob, P. – Asymptotic normality of the L_1-error for Geffroy's estimate of Poisson point process boundaries. *Publications de l'Institut de Statistique de l'Université de Paris*, 2004. – *A paraître.*

[78] Girard, S. et Jacob, P. – Extreme values and kernel estimates of point processes boundaries. *ESAIM: Probability and Statistics*, vol. 8, 2004, pp. 150–168.

[79] Girard, S. et Menneteau, L. – Central limit theorems for smoothed extreme value estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, 2004. – *A paraître.*

[80] Gomes, I. et Oliveira, O. – The bootstrap methodology in statistics of extremes. choice of the optimal sample fraction. *Extremes*, vol. 4, n° 4, 2001, pp. 331–358.

[81] Grimshaw, S.D. – Computing maximum likelihood estimates for the Generalized Pareto Distribution. *Technometrics*, vol. 35, n° 2, 1993, pp. 185–191.

[82] Hahn, G. et Meeker, W. – Pitfalls and practical considerations in product life analysis, part 1: Basic concepts and dangers of extrapolation. *Journal of Quality Technology*, vol. 14, 1982.

[83] Hall, P., Park, B. U. et Stern, S. E. – On polynomial estimators of frontiers and boundaries. *Journal of Multivariate Analysis*, vol. 66, n° 1, 1998, pp. 71–98.

[84] Hastie, T. et Stuetzle, W. – Principal curves. *Journal of the American Statistical Association*, vol. 84, n° 406, 1989, pp. 502–516.

[85] Häusler, E. et Teugels, J.L. – On asymptotic normality of Hill's estimator for the exponent of regular variation. *The Annals of Statistics*, vol. 13, 1985, pp. 743–756.

[86] Horn, P.S., Pesce, A.J. et Copeland, B.E. – A robust approach to reference interval estimation and evoluation. *Clinical Chemistry*, vol. 44, 1998, pp. 622–631.

[87] Hosking, J. et Wallis, J. – Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, vol. 29, n° 3, 1987, pp. 339–349.

[88] Huber, P.J. – Projection pursuit. *The Annals of Statistics*, vol. 13, n° 2, 1985, pp. 435–475.

[89] Imoussaten, A. – *Un test d'adéquation aux valeurs extrêmes.* – Mémoire de DEA, Université Grenoble 1, 2003.

[90] Jacob, P. et Suquet, P. – Estimating the edge of a Poisson process by orthogonal series. *Journal of Statistical Planning and Inference*, vol. 46, 1995, pp. 215–234.

[91] Jolliffe, I. – *Principal Component Analysis.* – New-York, Springer-Verlag, 1986.

[92] Jones, M.C. et Sibson, R. – What is projection pursuit? *Journal of the Royal Statistical Society A*, vol. 150, 1987, pp. 1–36.

[93] Karhunen, J. et Joutsensalo, J. – Generalizations of principal component analysis, optimization problems and neural networks. *Neural Networks*, vol. 8, 1995, pp. 549–562.

[94] Klüppeberg, C. et Villasenor, J.A. – Estimation of distribution tails - a semiparametric approach. *Bl. Dtsch. Ges. Versicherungsmath*, vol. 21, n° 2, 1993, pp. 213–235.

[95] Knight, K. – Limiting distributions of linear programming estimators. *Extremes*, vol. 4, n° 2, 2001, pp. 87–103.

[96] Korostelev, A.P. et Tsybakov, A.B. – *Minimax theory of image reconstruction.* – New-York, Springer-Verlag, 1993, *Lecture Notes in Statistics*, volume 82.

[97] Lebart, L. – Contiguity analysis and classification. *In: Data Analysis*, éd. par Gaul W., Opitz O. et M., Schader, pp. 233–244. – Berlin, Springer, 2000.

[98] Lepskij, O.V. – On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, vol. 35, n° 3, 1990, pp. 454–466.

[99] Li, K.C. – Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, vol. 86, 1991, pp. 316–342.

[100] Li, K.C., Aragon, Y., Shedden, K. et Thomas-Agnan, C. – Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, vol. 98, 2003, pp. 99–109.

[101] Mammen, E. et Tsybakov, A. B. – Asymptotical minimax recovery of set with smooth boundaries. *The Annals of Statistics*, vol. 23, n° 2, 1995, pp. 502–524.

[102] Matthys, G. et Beirlant, J. – Estimating the extreme value index and high quantiles with exponential regression models. *Statistica Sinica*, vol. 13, 2003, pp. 853–880.

[103] Moghaddam, B. et Pentland, A. – Probabilistic visual learning for object representation. *IEEE Pattern Analysis and Machine Intelligence*, vol. 19, n° 7, 1997, pp. 696–710.

[104] Nelsen, R. B. – *An introduction to copulas.* – New-York, Springer-Verlag, 1999, *Lecture Notes in Statistics*, volume 139.

[105] Nelsen, R. B., Quesada-Molina, J. J. et Rodríguez-Lallena, J. A. – Bivariate copulas with cubic sections. *Nonparametric Statistics*, vol. 7, 1997, pp. 205–220.

[106] Pan, J-X., Fung, W-K. et Fang, K-T. – Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference*, vol. 83, n° 1, 2000, pp. 153–167.

[107] Pickands, J. – Statistical inference using extreme order statistics. *The Annals of Statistics*, vol. 3, 1975, pp. 119–131.

[108] Pickands, J. – Multivariate extreme value distributions. *Bull. Int. Stat. Inst.*, vol. 49, n° 2, 1981, pp. 859–878.

[109] Quesada-Molina, J. J. et Rodríguez-Lallena, J. A. – Bivariate copulas with cubic sections. *Nonparametric Statistics*, vol. 5, 1995, pp. 323–337.

[110] Reiss, R. et Thomas, M. – *Statistical Analysis of Extreme Values.* – Birkhauser Verlag, 2001.

[111] Robert, C. – *Méthodes de Monte Carlo par chaînes de Markov.* – Paris, Economica, 1996, *Statistique mathématique et probabilité.*

[112] Royston, P. – Constructing time-specific reference ranges. *Statistics in Medicine*, vol. 10, 1991, pp. 675–690.

[113] Schölkopf, B. et Smola, A. – *Learning with kernels.* – Cambridge, MIT University Press, 2002.

[114] Sklar, A. – Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, vol. VIII, 1959, pp. 229–231.

[115] Smith, R.L. – Estimating tails of probability distributions. *The Annals of Statistics*, vol. 15, n° 3, 1987, pp. 1174–1207.

[116] Stone, C.J. – Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, vol. 10, 1982, pp. 1040–1053.

[117] Stone, C.J. – The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, vol. 14, 1986, pp. 590–606.

[118] Tipping, M.E. et Bishop, C.M. – Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, vol. 61, n° 3, 1999, pp. 611–622.

[119] Uenohara, M. et Kanade, T. – Use of Fourier and Karhunen-Loeve decomposition for fast pattern matching with a large set of templates. *IEEE on Pattern Analysis and Machine Intelligence*, vol. 19, n° 8, 1997, pp. 891–898.

[120] Yu, K. et Jones, M.C. – Local linear quantile regression. *Journal of the American Statistical Association*, vol. 93, 1998, pp. 228–237.