République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

N° Attribué par la bibliothèque

Année univ.: 2017/2018

Estimation robuste de la régression

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université Dr Moulay Tahar - Saïda

Discipline : MATHEMATIQUES

Spécialité : Analyse stochastique, statistique des processus et applications (ASSPA)

par

Samira Souci

Sous la direction de

Dr. Mme. F. Benziadi

Soutenu le 21/06/2018 devant le jury composé de

Dr. L. Bousmaha	Université Dr Tahar Moulay - Saïda	Présidente
Dr. F. Benziadi	Université Dr Tahar Moulay - Saïda	Encadreure
Dr. F. Mokhtari	Université Dr Tahar Moulay - Saïda	Examinatrice
Pr. N. Hachemi	Université Dr Tahar Moulay - Saïda	Examinatrice

Remerciements

En préambule de ce mémoire, je tiens tout d'abord à remercier mon dieu, le tout puissant et miséricordieux, qui a bien voulu me donner la force et la patience pour effectuer le présent travail.

Mes remerciements vont en second lieu à m'encadreure Mme Fatima Benziadi pour tous ses précieux conseils, sa confiance, ses idées, ses encouragements, ses corrections et ses remarques durant toute la période de la réalisation de ce travail.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.

Je tiens aussi à remercier très chaleureusement mes chers parents et mon ami pour ses encouragements et ses aides précieuses.

Je remercie sincèrement Prof A. Kandouci qui m'a ouvert les portes du laboratoire LMSSA et je n'oublie pas, bien évidemment, tous mes enseignants durant les années des études.

Enfin, je tiens à remercier tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Dédicace

Je dédie ce travail à mes chers parents et mon ami.

Toutes mes sœurs et mon frère.

tous les fils de ma sœur Ishek, Youssef et Sajdda.

Tous mes enseignants de département de mathématiques.

Toutes mes camarades de promotion 2017 /2018.

Merci à tous et à toutes.

Table des matières

Introduction

Un problème courant en statistiques est d'essayer d'expliquer comment une variable d'intérêt Y est reliée à une variable explicative X. En statistique, la régression est l'outil principal pour répondre à cette question. Cependant, cette estimation vu comme la moyenne conditionnelle de Y sachant X peut être inadaptée dans certaines situations. Par exemple, la présence de données aberrantes peut amener à des résultats non pertinents.

La régression robuste a été introduite pour résoudre ce genre de problèmes. Depuis les premiers résultats obtenus dans les années soixante, notamment par Huber $\big((1964), [24]\big)$, dont il a obtenu la consistance et la normalité asymptotique d'une classe d'estimateurs pour cette fonction. De nombreux auteurs ont développé ce domaine : Robinson $\big((1984), [26]\big)$, Härdle $\big((1984), [19]\big)$ et Härdle et Tsybakov $\big((1989), [20]\big)$ ont établi sous des conditions de mélange la normalité asymptotique d'une famille d'estimateurs issue de la méthode du noyau pour la fonction de régression. Parallèlement, Boente et Fraiman $\big((1989), [8]\big)$, $\big((1990), [9]\big)$ ont utilisé l'estimateur de Robinson $\big((1984), [26]\big)$ pour étudier simultanément les deux paramètres de position et d'échelle. La consistance des estimateurs construits est obtenue sous des conditions générales et dans les deux cas indépendants et fortement mélangeants. La convergence uniforme de l'estimateur robuste de la fonction de régression a été obtenu par Collomb et Härdle $\big((1986), [12]\big)$ en considérant des observations φ [1]- mélangeantes. Laïb

1. On dit que la famille $\{\Delta_i, i \in \mathbb{Z}\}$ est φ-mélangeante si la suite

$$\varphi(n) = \sup_{k\in\mathbb{Z}} \sup_{\{A\in\sigma_{-\infty}^{k},\, B\in\sigma_{n+k}^{\infty}\}} | \mathbb{P}(B/A) - \mathbb{P}(B) |$$

tend vers 0 quand n tend vers à l'infinie.

et Ould-Saïd ((2000), [25]) ont adapté l'estimateur de Collomb et Härdle ((1986), [12]) pour le modèle d'auto-régression d'un processus stationnaire ergodique. Ils ont obtenu la convergence uniforme de cet estimateur même lorsque la fonction objective est non bornée. Cai et Ould-Saïd ((2003), [11]) ont utilisé une version robuste de l'estimation par la méthode des polynômes locaux pour la fonction de la régression, ils ont démontré sous des conditions standards et lorsque les observations sont α-mélangeantes, la normalité asymptotique et la convergence presque sûr de ces estimateurs.Récemment, Azzedine et al. ((2008), [3], [4]) ont étudié la convergence presque complète d'estimateurs robustes à noyau. Dans le même cadre, Attouch et al. ((2007), [5]) ont étudié la normalité asymptotique de ces estimateurs.

Le but de ce travail est d'étudier l'estimation robuste de la fonction de régression.

Dans le 1^{ier} chapitre, nous discuterons sur les moyens de détection et de controle les outliers dans l'ensemble de données avant ou pendant le processus d'estimation et nous introduirons la notion de robustesse au début de ce chapitre.

Dans le 2^{ieme} chapitre, nous nous concentrerons sur les methodes robustes disponibles de régression linéaire dans la littérature.

Le chapitre 3 sera consacré à l'étude d'estimation à noyau de la fonction de régression robuste dans le cas où l'estimation est non paramétrique,en considérant les deux cas : le cas où les observations sont indépendantes identiquement distribuées et le cas des observations fortement mélangeantes.

Bien entendu, ce mémoire s'achèvera par une conclusion générale sur notre travil.

Chapitre 1

Introduction à la notion de valeur aberrante et robustesse

1.1 Robustesse

D'une façon générale, la robustesse est définie comme la capacité d'un système à maintenir ses performances malgré des changements dans les conditions d'utilisation ou la présence d'incertitudes liées à ses paramètres ou à ses composants. On ne cherchera pas à supprimer les causes d'une variabilité mais plutôt à en minimiser les effets. De fait, la robustesse implique une insensibilité aux écarts dûs à une non-conformité des hypothèses sous-jacentes au problème traité. Les méthodes robustes garantissent de bons résultats pour une grande collection d'hypothèses sans pour autant avoir les meilleurs résultats pour une en particulier.

1.2 Points aberrants

1.2.1 Définition

Le terme "point aberrant" est l'équivalent du terme anglais "outlier", mot qui est également employé dans la littérature francophone. Nous utiliserons donc indifféremment l'un ou l'autre de ces termes dans la suite de ce manuscrit.

En 1978, Barnett et Lewis [6] ont défini un point aberrant comme étant : une observation, ou un sous-ensemble d'observations, qui paraît être incohérent avec le reste de l'ensemble des données.

Cette définition, très générale, induit que la décision de traiter une observation comme étant un outlier est le fruit du jugement subjectif des chercheurs. Par la suite, Judge et al. $\big((1988),[17]\big)$ ont utilisé le terme d'outlier pour désigner les observations induisant une importante valeur résiduelle. Puis Krasker et al. (1983), Hampel et al. (1986), ainsi que Rousseeuw et Leroy (1987), ont présenté une classification des outliers en différentes catégories sans pour autant en donner une définition plus formelle.

En 1993, Davies et Gather ont présenté une définition considérée plus précise. Soit F la distribution cible. Pour une simplicite de présentation, F sera choisie comme une distribution normale univariable avec sa moyenne notée μ (pouvant être inconnue) et sa variance notée σ^2 (pouvant être inconnue aussi). Pour α strictement compris entre 0 et 1, Davies et Gather introduisent le concept de région d'outliers comme :

$$out(\alpha, \mu, \sigma^2) = \big\{x : \quad \mid x - \mu \mid > z_{1-\alpha/2}\sigma\big\} \tag{1.1}$$

avec z_q le q-quantile de la distribution normale standard. Une observation x_i sera appelée un α-outlier selon F si $x_i \in out\big(\alpha, \mu, \sigma^2\big)$. Par conséquent, une "bonne" observation, c'est à dire appartenant a la distribution cible F, pourra être classée comme un α-outlier. Il est à noter qu'avec cette définition, x_i peut être n'importe quel nombre réel, il ne doit pas coïncider nécessairement avec l'une des observations.

Le point important de cette définition est qu'elle se base sur les paramètres réels μ et σ^2. Elle ne dépend donc pas de la méthode de détection des outliers choisie. Selon la méthode utilisée pour estimer les paramètres μ et σ^2, les observations labelisées comme étant des outliers seront différentes. La définition de Davies et Gather permet de faire une distinction entre d'un côté les "vrais" outliers et de l'autre les observations labelisées comme outliers a partir d'une méthode d'identification.

À partir de ces différentes définitions, nous allons maintenant présenter une classification des points aberrants utilisée dans la littérature.

1.2.2 Classification

Les outliers peuvent être de différents types selon leurs positions dans l'ensemble de données. Une classification possible des outliers dans le cas d'un modèle de régression linéaire[1], c'est à dire $Y = aX + b$, est illustrée sur la figure 1.1 qui suit.

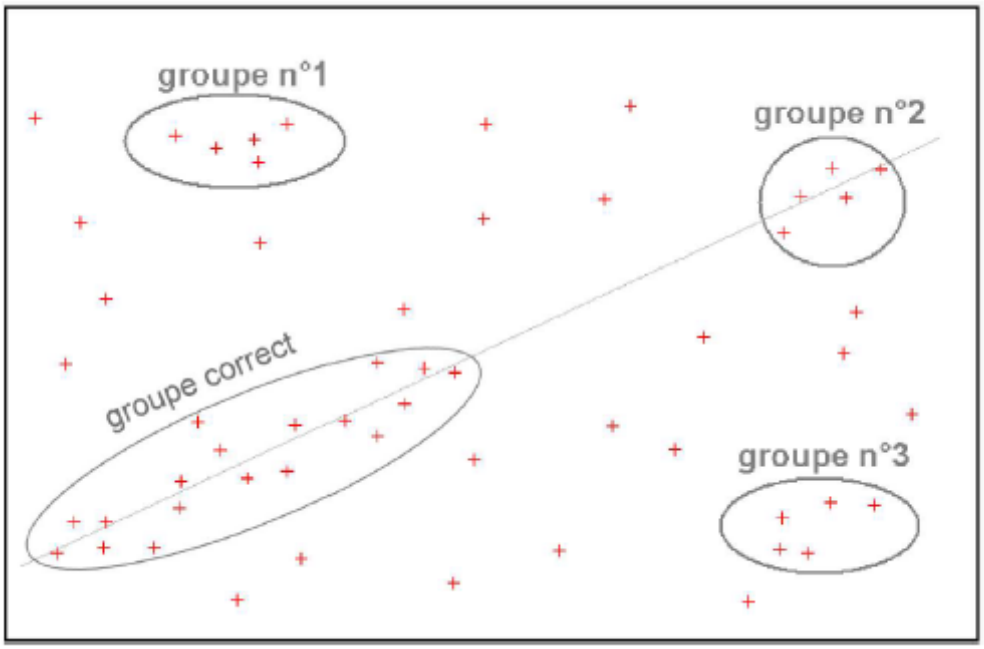

FIGURE 1.1 – Classification des outliers dans une régression linéaire simple

La figure nous présente un ensemble de données composé d'un certain nombre de points que nous classerons en différentes catégories. Le regroupement en bas à gauche est supposé représenter les données correctes, la structure principale dont nous souhaitons obtenir la relation. Les groupes d'outliers numérotés 1, 2 et 3, ainsi

1. Lorsqu'une variable réponse, notée Y , semble liée à une variable explicative, notée X, par une relation de type affine, c'est à dire que $Y = aX + b$ dans un problème à 2 dimensions, on nomme régression linéaire simple l'estimation des deux paramètres a et b.

que les outliers uniformément distribués dans l'espace, sont les données qui posent problème. Ces données peuvent être structurées, c'est le cas des groupes 1, 2 et 3, mais elles peuvent aussi être totalement aléatoires et ne correspondre à aucune structure. Nous identifions trois classes d'outliers :

- **les outliers verticaux** (suivant l'axe des y) : il s'agit des points appartenant à la plage des abscisses de la majorité des données, mais se trouvant éloignées de l'estimation linéaire des données. Elles possèdent une valeur résiduelle plus importante que la plupart des autres données. Ces outliers verticaux sont représentés par le groupe n°1 sur la figure 1.1. Ils ont une influence relativement faible sur la solution de la régression, il perturbe généralement le biais b plutôt que la pente a de la droite solution.
- **les outliers horizontaux**, dits à effet levier (suivant l'espace des x) : il s'agit des points n'appartenant pas à la plage des abscisses de la majorité des données, c'est à dire qu'elles provoquent une discontinuité dans les données suivant l'axe des abscisses. On peut séparer ces observations en 2 groupes :
  - **les outliers à effet levier favorable** : le groupe des outliers à effet favorable se trouve proche de l'estimation linéaire des données, c'est à dire que ces outliers possèdent une valeur résiduelle du même ordre que celle de la majorité des données, ils sont réprésentés par le groupe n°2.
  - **les outliers à effet levier défavorable** : le groupe des outliers à effet défavorable est placé relativement loin de l'estimation linéaire, ces outliers obtiennent une valeur résiduelle plus importante que la majorité des données, ils sont représentés par le groupe n°3.

Maintenant que nous avons défini ce que sont les outliers, et que nous en avons proposé une classification, une question se pose quant à la résolution de notre problème de régression. Comment pouvons-nous détecter, avant ou pendant le processus d'estimation, les outliers dans l'ensemble de données ? L'objectif de cette détection sera de minimiser l'impact de ces erreurs sur la solution, ou tout simplement les ignorer dans le calcul de la solution.

1.3 Techniques de détection et de contrôle des outliers

Les approches existantes pour la détection d'outliers peuvent être regroupées en 2 catégories : les mesures basées sur le K-voisinage et les mesures basées sur le voisinage géométrique :

1.3.1 Les approches basées sur le K-voisinage

ces approches reposent sur l'idée proposée par Breunig et al. [10] qui commencent par définir La k- distance, notée $Dist_k$ et relative au facteur local d'aberration [2], noté LOF, de chaque point-donnée, qui dépend de la densité locale de son voisinage.
$Dist_k(x)$ est égale à $d(x, y)$ pour une observation y de sorte que pour au moins k observations $y^{'}$ de l'ensemble de données noté D :

$$d(x, y') \leq d(x, y).$$

$V_k(x)$, le voisinage de k-distance de x est l'ensemble des observations vérifiant :

$$V_k(x) = \{x \in D \quad / \quad d(x, y) \leq Dist_k(x)\}.$$

Ce voisinage est défini par la distance entre le point-observation et son k-ième plus proche voisin.
La distance d'atteignabilité (reachability distance) entre deux observations est définie par :

$$RDist_k(x, y) = \max\{Dist_k(y), d(x, y)\}.$$

La densité locale d'atteignabilité (local reachability density) en x est ensuite définie par l'inverse de la moyenne de la distance d'atteignabilité dans le k- voisinage de x :

$$LRDens(x) = \left(\frac{\sum_{y \in V_k(x)} RDist_k(x, y)}{card(V_k(x))} \right)^{-1}$$

2. ≪ *local outlier factor* ≫ en anglais

Finalement :

$$LOF_k(x) = \frac{\sum_{y \in V_k(x)} LRDens(y)\big(LRDens(x)\big)^{-1}}{card(V_k(x))}$$

K est une variable définissant un nombre minimal de points dans le voisinage, le choix de sa valeur est important. Les points possédant un LOF élevé sont étiquetés comme outliers. Un exemple est représenté sur la figure 1.3.1.

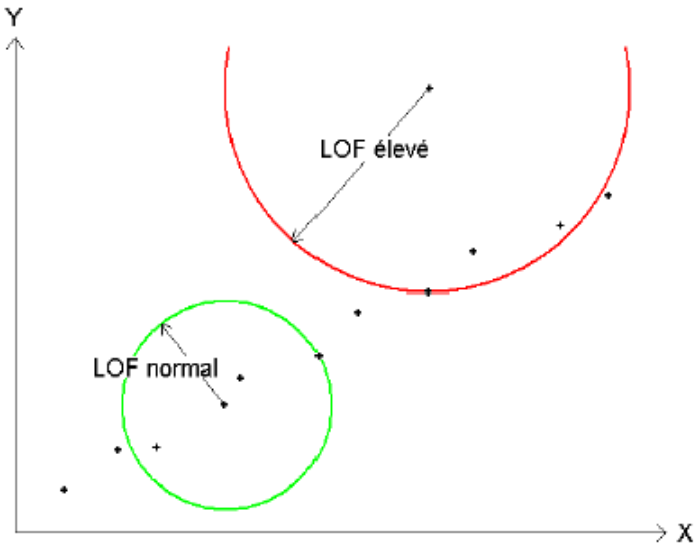

Fig.1.3.1- Facteur local d'aberration LOF (K=3).

Le point aberrant en haut à droite de la figure possède un LOF élevé, la distance entre ce point et son 3ème plus proche voisin est importante. En revanche, la "bonne" donnée en bas à gauche possède un LOF "normal", c'est à dire que son LOF n'est pas significativement plus important que le LOF moyen des données.

1.3.2 Les approches basées sur le voisinage géométrique

Elles exploitent l'idée proposée originellement par Knorr et Ng en 1997. Une observation dans l'ensemble de données noté x_P est un outlier selon la distance si au moins une fraction β des observations de x_P se trouvent éloignés d'au moins la distance r de

celui-ci. Cette définition d'un outlier est basée sur un critère simple et global déterminé par les paramètres r et β . On peut utiliser cette méthode pour les problèmes où l'ensemble de données possède dans le même temps des régions denses et des régions éparses. Un exemple est présenté sur la figure 1.3.2.

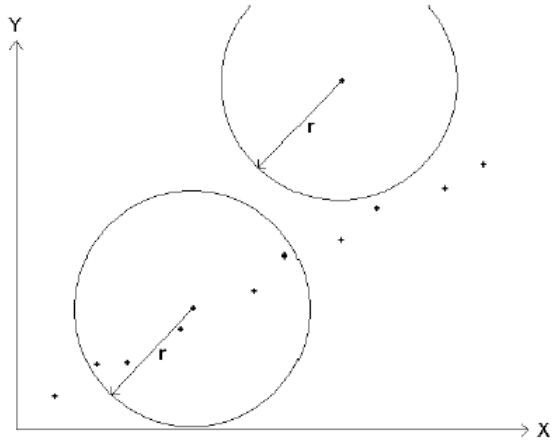

Fig.1.3.2 Voisinage géométrique.

Le point aberrant en haut à droite est éloigné des 11 autres données d'une distance plus grande que r, alors que le "bon" point en bas à gauche possède 5 points dans son voisinage géométrique.

1.4 La diagnostic de la régression

Hormis ces deux types d'approches de détection d'outliers, il est également possible d'utiliser les outils de diagnostic de la régression [14][7] [18] afin de contrôler, voire supprimer les outliers ayant une forte influence sur le modèle de régression. Nous allons présenter parmi ces différents outils qui mesurent l'influence des observations : la matrice "chapeau" notée H, les résidus, ainsi que la distance de Cook.

1.4.1 La matrice chapeau

Soit X la matrice d'observation, ou matrice des prédicteurs, on construit la matrice chapeau[3] H telle que :

$$H = X\left(X^T X\right)^{-1} X^T \tag{1.2}$$

Nous avons donc une expression des prédictions $\widehat{y}_i$ comme une combinaison linéaire des réponses y_i :

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \tag{1.3}$$

Les éléments qui nous intéresse dans cette matrice sont ceux appartenant à la diagonale :

$$h_{ij} = \frac{x_i x_i^T}{\sum_{j=1}^{n} x_j x_j^T} \tag{1.4}$$

Les éléments h_{ii} représentent le poids de la ième réponse y_i sur la ième valeur prédite $\widehat{y}_i$. Ils permettent donc de mesurer l'importance de la contribution de y_i dans la prédiction de $\widehat{y}_i$. Ce critère est également relié à l'éloignement de x_i par rapport à la moyenne $\overline{x}$. La valeur moyenne des h_{ii} est $\overline{h}_{ii} = p/n$, où p est le nombre de coefficients $\theta = (\theta_i)_{i=1,\,\cdots,\,p}$ du modèle $Y = X\theta$ et n le nombre d'observations. D'après la littérature [21], il convient de porter une attention particulière aux observations dont la valeur h_{ii} serait supérieure à 2 ou 3 fois la valeur moyenne. En effet, un élément sur la diagonale de la matrice chapeau H qui possède une valeur importante, va attirer l'hyperplan de régression vers l'observation correspondante.

1.4.2 Les résidus

Les résidus mesurent l'écart entre la ième réponse y_i et la ième valeur prédite $\widehat{y}_i$ par l'estimation de la régression.

$$\epsilon_i = y_i - \widehat{y}_i \tag{1.5}$$

3. ≪ *hat matrix* ≫ en anglais est une solution par la méthode de moindre carré.

Même si les erreurs sont indépendantes et de même variance, les résidus issus de l'estimation n'ont pas la même variance : $\mathbb{E}[\epsilon_i] = 0$ et $Var(\epsilon_i) = \sigma^2(1 - h_{ii})$. On va donc calculer des versions standardisées afin de les rendre comparables.

Les résidus standardisés

Les résidus peuvent être divisés par une estimation de leur propre variance afin d'obtenir des résidus dits standardisés ou studentisés.

$$\epsilon_{i_{stand}} = \frac{\epsilon_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}} \tag{1.6}$$

Avec :

$$\frac{1}{n} \leq H_{ii} < 1 \tag{1.7}$$

Et :

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \frac{\epsilon_i^2}{n - p} \tag{1.8}$$

Le problème avec ces résidus standardisés est que le numérateur et le dénominateur ne sont pas indépendants, on utilise ϵ_i pour estimer la variance σ^2. Pour ne pas souffrir de ce biais, les résidus Jackknife sont préférés [18].

Les résidus Jackknife

Les résidus Jackknife, notés aussi résidus R-Student, sont les quotients des résidus sur leurs propres variances en estimant la variance σ^2 avec toutes les observations exceptée l'observation i.

$$\epsilon_{i_{R-Student}} = \frac{\epsilon_i}{\widehat{\sigma}_{-i}\sqrt{1 - h_{ii}}} \tag{1.9}$$

Avec :

$$\widehat{\sigma}^2_{-i} = \left(\sum \frac{\epsilon_i^2}{n - p}\right) - \frac{\epsilon_i^2}{n - p} \tag{1.10}$$

Il est admis, dans la littérature, qu'il convient d'examiner les observations pour lesquelles la valeur absolue du résidu Jackknife est supérieure à 2.

1.4.3 La distance de Cook

La distance qui a été introduit par Cook [13] est l'un des critères les plus utilisés pour juger de l'influence d'une observation x_i. Elle consiste à comparer les paramètres estimés du modèle utilisant toutes les observations avec ceux estimés sans la ième observation. On peut l'écrire sous cette forme :

$$D_{i_{cook}} = \frac{\left(\widehat{\theta} - \widehat{\theta}_{-i}\right)^T \left(X^T X\right) \left(\widehat{\theta} - \widehat{\theta}_{-i}\right)}{p\widehat{\sigma}^2} \tag{1.11}$$

Où $\widehat{\theta}$ et $\widehat{\theta}_{(-i)}$ représentent respectivement les paramètres estimés avec toutes les observations et sans l'observation x_i. On peut aussi écrire la distance de Cook sous une forme plus simple qui est :

$$D_{i_{cook}} = \frac{\epsilon^2_{i_{stand}}}{p} \frac{h_{ii}}{1 - h_{ii}} \tag{1.12}$$

Ce critère mesure donc l'influence d'une observation sur l'ensemble des prévisions en prenant en compte l'effet levier et l'importance des résidus. En pratique, la stratégie de détection des valeurs atypiques consiste à rechercher les distances de Cook avec une valeur supérieure à 1.

Chapitre 2

Quelques estimateurs robustes de la régression linéaire

Les méthodes de régression que nous nommons déterministes utilisent la totalité des données pour estimer les paramètres du modèle recherché. Elles emploient différentes fonctions pertes [1] qui leur garantissent une certaine robustesse. On regroupe sous le signe X-estimateurs une famille d'estimateurs comprenant : les M-estimateurs, les W-estimateurs, les S-estimateurs, les R-estimateurs.

1. Posons $(X, Y, \widehat{Y})$ le triplet composé de : la matrice d'observations X appartenant à l'ensemble des matrices d'observations, noté $\mathcal{X}$, du vecteur réponse Y ainsi que de la prédiction $\widehat{Y}$ appartenant tous deux à l'ensemble des vecteurs réponses, noté $\mathcal{Y}$. La fonction c définie par :

$$c \,:\, \mathcal{X} \,\times\, \mathcal{Y} \,\times \mathcal{Y} \longrightarrow\, [0,\, \infty)$$

telle que :

$$c\big(X,\, Y,\, \widehat{Y} = Y\big) = 0$$

sera nommée une fonction de perte.

2.1 M-estimateurs

La méthode des moindres carrés [2] n'est pas robuste, aussi bien pour les outliers suivant y que suivant x. Cette non-robustesse est due à la fonction objectif de la forme : $\sum_{i=1}^{n} \mid \epsilon_i \mid$, qui augmente très rapidement avec le résidu ϵ_i. Huber, en 1973, a introduit la notion de M-estimateur, dit estimateur du maximum de vraisemblance, afin de limiter l'influence des données érronées sur l'estimation. Estimer θ par la méthode du maximum de vraisemblance, c'est proposer comme valeur de θ celle qui rend maximale la vraisemblance, à savoir la probabilité d'observer les données comme réalisation d'un échantillon suivant une certaine loi de probabilité. Pour calculer le maximum de vraisemblance, il faut déterminer les valeurs pour lesquelles la dérivée de la vraisemblance s'annule.

Il s'agit de la méthode la plus simple tant au niveau calculatoire que théorique. Elle est encore très utilisée dans le domaine de l'analyse de données où la contamination est principalement située dans le vecteur réponse y. Au lieu d'utiliser la fonction perte quadratique, comme dans les moindres carrés, à laquelle est associée une loi de probabilité gaussienne, les estimateurs d'Huber du type M- estimateur minimisent une somme de valeurs résiduelles calculées via une fonction perte c qui croît moins rapidement que la quadratique :

$$\widehat{\theta}_M = \min_{\theta} Q_M(\theta) \tag{2.1}$$

2. La méthode des moindres carrés, notée LS pour Least Squares en anglais, est la méthode d'estimation de la régression la plus communément rencontrée. On peut aussi la rencontrer sous le nom de L2-régression.L'estimation θ_{LS} de θ de l'équation $Y = X\theta$ est définie comme le p-vecteur qui minimise la somme avec toutes les valeurs résiduelles au carré, c'est à dire :

$$\theta_{LS} = \min_{\widehat{\theta}} \sum_{i=1}^{n} \epsilon_i^2$$

où

$$\epsilon_i = y_i - x_i\widehat{\theta}.$$

et

$$Q_M(\theta) = \sum_{i=1}^{n} c(\epsilon_i) \tag{2.2}$$

L'estimation optimale est déterminée en annulant les dérivées de la somme par rapport aux p coefficients de θ, soit :

$$\frac{\partial c(\epsilon_i)}{\partial \theta_j} = \psi(\epsilon_i) x_{ij} \tag{2.3}$$

et

$$\sum_{i=1}^{n} \psi(\epsilon_i) x_{ij} = 0, \quad \forall j = 1, \cdots, p \tag{2.4}$$

où la fonction ψ est la dérivée de la fonction perte c. La M- estimation est obtenue en résolvant ce système de p équations nonlinéaires. Cependant, la solution n'est pas équivariante par rapport à l'echelle. En effet, si l'on multiplie les résidus par une valeur quelconque, c'est à dire que l'on modifie l'echelle, la solution obtenue sera différente. On doit donc standardiser les résidus à l'aide d'une estimation de l'écart type $\widehat{\sigma}$. Ainsi la solution peut s'écrire :

$$\sum_{i=1}^{n} \psi(\epsilon_i / \widehat{\sigma}) x_i = 0 \tag{2.5}$$

Où l'écart type $\widehat{\sigma}$ doit être estimé simultanément. Une des possibilités souvent employée pour son estimation est d'utiliser un multiple de la médiane de la déviation absolue $\big($notée MAD pour median absolute deviation$\big)$. Cette utilisation suppose implicitement que le taux de contamination dû au bruit soit de 50% . La médiane de la déviation absolue est définie ainsi :

$$MAD(X_i) = med\big\{ \mid X_j - med_j(X_j) \mid \big\} \tag{2.6}$$

Et donc l'estimateur de l'écart type s'écrit :

$$\widehat{\sigma} = \beta.MAD \tag{2.7}$$

Où β est le facteur multiplicatif. La valeur utilisée communément pour β est 1.483, cette valeur ajuste l'échelle pour une efficacité maximale lorsque les données proviennent d'une distribution gaussienne.

Le tableau 2.1 présente les Différents M-estimateurs utilisant différentes fonctions pertes.

méthode	fonction d'influence
L1-régression	$\psi(\epsilon) = sgn(\epsilon)$
L2-régression	$\psi(\epsilon) = \epsilon$
Huber minimax	$\psi(\epsilon) = \begin{cases} \epsilon, & si \mid \epsilon \mid < B \\ B.sgn(\epsilon), & si \mid \epsilon \mid \geq B. \end{cases}$
Minimax descendant	$\psi(\epsilon) = \begin{cases} \epsilon, & si \mid \epsilon \mid < A \\ B.sgn(\epsilon).tanh\left[\frac{1}{2}B\big(C- \mid \epsilon \mid \big)\right], & si\ A \leq \mid \epsilon \mid < C \\ 0, & autrement \end{cases}$
Hampel	$\psi(r) = \begin{cases} r, & si \mid \epsilon \mid < A \\ A.sgn(\epsilon), & si\ A \leq \mid \epsilon \mid < B \\ \frac{C-\mid\epsilon\mid}{C-B}.A.sgn(\epsilon), & si\ B \leq \mid \epsilon \mid < C \\ 0, & autrement \end{cases}$
Andrew	$\psi(\epsilon) = \begin{cases} sin(\epsilon), & si\ -\pi \leq \mid \epsilon \mid < \pi \\ 0, & autrement \end{cases}$
Tukey	$\psi(\epsilon) = \begin{cases} \epsilon\big(1-(\frac{\epsilon}{C})^2\big)^2, & si \mid \epsilon \mid < C \\ 0, & autrement \end{cases}$

Tab. 2.1- Exemples de fonctions d'influence utilisées dans les M-estimateurs.

Dans le but de réduire l'influence des données contaminées, la fonction perte c doit être choisie en accord avec la densité de probabilité qui définie la loi des erreurs. La fonction c doit respecter certaines conditions. Elle doit être symétrique, positive avec un minimum unique en zéro et avec une croissance moins rapide que la fonction quadratique.

Bien que les M-estimateurs soient plus robustes que la $L1$-régression ou les moindres carrés en ce qui concerne les points aberrants verticaux (notion explicitée dans le paragraphe 1.2.2). Ils restent très vulnérables face à des points aberrants horizontaux. Afin d'améliorer leurs robustesses contre ces erreurs, les W-estimateurs ont été introduits.

2.2 W-estimateurs

Les W-estimateurs, ou Generalized M-estimators, sont des M-estimateurs pondérés. Chaque W-estimateur possède une fonction poids caractéristique, notée $\omega(.)$, qui représente l'importance de chaque observation dans l'estimation de θ. L'estimation optimale est déterminée en résolvant le système de p équations non-linéaires suivant :

$$\sum_{i=1}^{n} \omega(X_i)\psi(\epsilon_i/\widehat{\sigma})X_{ij} = 0, \quad \forall j = 1, \cdots, p \tag{2.8}$$

2.3 R-estimateurs

Dans cette approche, chaque résidu est pondéré par une fonction score basée sur lerang du résidu. Soit R_i le rang du résidu $\epsilon_i = Y_i - X_i\widehat{\theta}$, l'objectif est alors de minimiser suivant $\widehat{\theta}$ la somme des résidus pondérés par cette fonction score :

$$\widehat{\theta}_R = \min_{\theta} Q_R(\theta) \tag{2.9}$$

$$Q_R(\theta) = \sum_{i=1}^{n} a_n(R_i)\epsilon_i \tag{2.10}$$

où la fonction score $a_n(R_i)$ est monotone et satisfait :

$$\sum_{i=1}^{n} a_n(R_i) = 0 \tag{2.11}$$

Différentes fonctions scores ont été proposées, parmi lesquelles :
– la fonction score de Wilcoxon :

$$a_n(R_i) = R_i - \frac{n+1}{2} \tag{2.12}$$

– la fonction score de Van der waerden :

$$a_n(R_i) = \Phi^{-1}\left(\frac{R_i}{n+1}\right) \tag{2.13}$$

où Φ^{-1} est la fonction cumulative inverse d'une distribution normale.

– la fonction score médiane :

$$a_n(R_i) = sgn\left(R_i - \frac{n+1}{2}\right) \tag{2.14}$$

Un inconvénient majeur des R-estimateurs est qu'ils ne sont pas aisément optimisables. De plus, la définition de la fonction score nécessite implicitement une connaissance a priori du taux de contamination des données. Enfin, les R-estimateurs ne sont pas robustes contre les outliers dans la direction x. En revanche, un avantage important des R-estimateurs comparés aux M-estimateurs est qu'ils sont automatiquement équivariants par rapport à l'echelle, aucune estimation de l'écart type n'est nécessaire.

2.4 S-estimateurs

Les S-estimateurs sont une autre classe de M-estimateurs automatiquement équivariants. Ils sont définis par la minimisation de la dispersion des résidus :

$$\widehat{\theta}_s = \min_{\theta} Q_s(\theta) \tag{2.15}$$

et

$$Q_s(\theta) = s\big(\epsilon_1, \cdots, \epsilon_n\big) \tag{2.16}$$

où la dispersion $s(\epsilon_1, \cdots, \epsilon_n)$ est définie comme la solution de :

$$\frac{1}{n}\sum_{i=1}^{n} c\big(\frac{\epsilon_i}{s}\big) = K \tag{2.17}$$

Avec c la fonction perte et K souvent posé égal à $E_\Phi[c]$, Φ étant la distribution normale standard.

Chapitre 3

Estimation robuste de la régression non paramétrique

3.1 Modèle et estimation

Soit $\big(X, Y\big)$ un couple de variables aléatoires à valeurs dans $\mathcal{F}\times\mathbb{R}$, où $\mathcal{F}$ est un espace semimétrique. On note d la semi métrique sur $\mathcal{F}$. Pour $x \in \mathcal{F}$, on considère le modèle de régression non paramétrique suivant

$$Y = \theta_x(X) + \epsilon.$$

On considère une fonction réelle, mesurable, notée ψ_x. Le paramètre fonctionnel étudié dans ce travail, noté θ_x, est la solution (supposée unique) de l'équation en t définie par :

$$\Psi(t, x) = \mathbb{E}\Big[\psi_x(Y - t)\big/X = x\Big] = 0 \tag{3.1}$$

En général, la fonction ψ_x est fixée par le statisticien en fonction de la situation à laquelle il est confronté. Des exemples classiques de ψ_x conduisent à l'estimation de la moyenne conditionnelle $\big($si $\psi_x(t) = t\big)$ ou de quantiles conditionnels $\big($si $\psi_x(t) = \mathbb{1}_{[0,\infty[}(t) - (1-\alpha)$, $\alpha \in [0,1]\big)$, voir Ferraty et Vieu $\big($(2006), [16]$\big)$ et Attouch et al. $\big($(2007), [2]$\big)$.

Etant donné $\big(X_1, Y_1\big), \cdots, \big(X_n, Y_n\big)$ une suite des observations de même loi que le couple $\big(X, Y\big)$, l'estimateur à noyau de $\Psi(t, x)$ est donnée par :

$$\widehat{\Psi}(t,x) = \frac{\displaystyle\sum_{i=1}^{n} K\left(h_K^{-1} d\big(x, X_i\big)\right) \psi_x\big(Y_i - t\big)}{\displaystyle\sum_{i=1}^{n} K\left(h_K^{-1} d\big(x, X_i\big)\right)}, \quad \forall t \in \mathbb{R},$$

où K est un noyau et $h_K = h$ est une suite de nombres réels positifs tend vers 0 quand n tend vers à l'infini.
L'estimateur naturel de θ_x, noté $\widehat{\theta}_x$, est tel que

$$\widehat{\Psi}\big(\widehat{\theta}_x, x\big) = 0 \tag{3.2}$$

Quand ψ_x est égale à l'identité, $\widehat{\theta}_x$ est identifié à l'estimateur de Ferraty et Vieu (2006) [16] pour la régression fonctionnelle.

3.2 Outils

Soient $\big(\Omega,\ \mathcal{A},\ \mathbb{P}\big)$ un espace de probabilité et $\{\Delta_i\}_{i\in\mathbb{Z}}$ une famille des variables aléatoires définie sur $\big(\Omega,\ \mathcal{A},\ \mathbb{P}\big)$ à valeurs dans un espace probabilisable $\big(E, \xi\big)$.

Le lemme suivant donne l'inégalité de Hoeffding.

Lemme 3.2.1. *[22] Soit $(\Delta_n)_{n\in\mathbb{N}}$ une suite des variables aléatoires réelles centrées, indépendantes et identiquement distribuées, telles qu'il existe deux réels positifs d et δ vérifiant :*

$$\mid \Delta_1 \mid \leq d \quad et \quad \mathbb{E}\big[\Delta_1^2\big] \leq \delta^2,$$

alors, pour tout $\epsilon \in]0, \frac{\delta^2}{d}[$, on a

$$\mathbb{P}\left(n^{-1} \mid \sum_{i=1}^{n} \Delta_i \mid > \epsilon\right) \leq 2e^{-\frac{n\epsilon^2}{4\delta^2}}.$$

Le lemme suivant donne l'inégalité de Fuk-Nageav.

Lemme 3.2.2. *[22] Soit* $\{\Delta_i,\ i \in \mathbb{N}\}$ *une suite des variables aléatoires réelles* α*-mélangeante*[1]*, de coefficient de mélange* $\alpha(n)$ *vérifiant :* $\exists c \in \mathbb{R}^{*+}$, $a \in \mathbb{R}^{*+}$

$$\alpha(n) \leq cn^{-a}$$

et si $\forall i,\ \| \Delta_i \|_\infty < \infty$[2]*, alors, pour tout* $\epsilon > 0$ *et* $r > 0$*, on a*

$$\mathbb{P}\Big(\mid \sum_{k=1}^{n} \Delta_k \mid > 4\epsilon \Big) \leq \left(1 + \frac{\epsilon^2}{rS_n^2}\right)^{\frac{-r}{2}} + 2ncr^{-1}\left(\frac{2r}{\epsilon}\right)^{a+1} \tag{3.3}$$

où

$$S_n^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \mid cov(\Delta_i, \Delta_j) \mid .$$

L'inégalité du lemme suivant s'appelle inégalité de covariance et elle est très utile pour le calcul de S_n^2, définie dans le lemme 3.2.2.

Lemme 3.2.3. *[16] Soit* $\{\Delta_i,\ i \in \mathbb{N}\}$ *une suite des variables aléatoires réelles* α*-mélangeante, de coefficient de mélange* $\alpha(n)$*, telle que* $\| \Delta \|_\infty < \infty$, $\forall i$*. On a, pour tout* $i \neq j$ *:*

$$\mid cov(\Delta_i,\ \Delta_j) \mid \leq 4 \| \Delta_i \|_\infty \| \Delta_j \|_\infty \alpha\big(\mid i - j \mid \big).$$

1. Soient $\big(\Omega,\ \mathcal{A},\ \mathbb{P}\big)$ un espace de probabilité et $\{\Delta_i\}_{i\in\mathbb{Z}}$ une famille des variables aléatoires définie sur $\big(\Omega,\ \mathcal{A},\ \mathbb{P}\big)$ à valeurs dans un espace probabilisable $\big(E, \xi\big)$. On note $\big(\sigma_i^j\big)_{i\neq j\in\mathbb{Z}\cup\{-\infty,+\infty\}}$, la tribu engendrée par $\big\{\Delta_k, i < k < j\big\}$ et par $L_2(\sigma_i^j)$ l'espace des variables aléatoires σ_i^j- mesurables et de carrée sommable.

On dit que la famille $\{\Delta_i, i \in \mathbb{Z}\}$ est α- mélangeante si la suite

$$\alpha(n) = \sup_{\left\{k\in\mathbb{Z},\ A\in\sigma_{-\infty}^{k},\ B\in\sigma_{n+k}^{+\infty}\right\}} \mid \mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B) \mid$$

tend vers 0 quand n tend vers à l'infinie. La suite α_n est appelée coefficient de mélange forte.

2. Pour $\Delta = (\Delta_1, \cdots, \Delta_n) \in \mathbb{R}^n$ on note

$$\| \Delta \|_\infty = \max_{1\leq i\leq n} \mid \Delta_i \mid$$

Lemme 3.2.4. ***Inégalité*** *Cr [23]*

$$\mathbb{E}\big[\mid X+Y\mid^r\big] \leq c_r\mathbb{E}\big[\mid X\mid^r\big] + c_r\mathbb{E}\big[\mid Y_r\mid\big]$$

où $c_r = 1$ *(resp.* 2^{r-1}*) selon que* $r \leq 1$ *(resp.* $r \geq 1$*).*

3.3 Résultats asymptotiques : Cas i.i.d

Le but principal de cette section est d'étudier l'estimateur à noyau de la fonction de régression robuste, dans le cas où les observations sont indépendantes et identiquement distribuées. Cette section est divisée en deux parties : dans la première partie, on étudie la convergence presque complète [3]. On traite la normalité asymptotique de notre estimateur dans la deuxième partie.

3.
- On dit que la suite $(X_n)_{n\in\mathbb{N}}$ converge presque complètement vers X si $\forall\epsilon > 0$,

$$\sum_{n=0}^{\infty}\mathbb{P}\Big(\mid X_n - X\mid > \epsilon\Big) < \infty$$

et on note $X_n \longrightarrow X$ en p.co.

- On dit que la suite $X_n = O\big(Y_n\big)$ en p.co. s'il existe un $\epsilon > 0$ vérifiant :

$$\sum_{n=0}^{\infty}\mathbb{P}\Big(X_n > \epsilon Y_n\Big) < \infty.$$

- Soient X_n, Y_n deux suites des variables aléatoires. La suite $\left(\frac{X_n}{Y_n}\right)_{n\in\mathbb{N}} \to 0$ en p.co si

$$X_n \to 0 \quad \textit{en p.co.}$$

et

$$\exists\delta > 0, \quad \sum_{n=0}^{\infty}\mathbb{P}\Big(\mid Y_n\mid < \delta\Big) < \infty.$$

3.3.1 Convergence presque complète

Hypothèses

On fixe un point $x \in \mathcal{F}$, on note $B(x,h) = \left\{ x' \in \mathcal{F} / \; d(x,x') < h \right\}$ la boule de centre x et de rayon h et on introduit les hypothèses suivantes :

(H1) $\mathbb{P}\big(X \in B(x,h)\big) = \phi_x(h) > 0$, $\forall h > 0$.

(H2) Il existe $C_1 > 0$ et $b > 0$ tel que $\forall x_1, x_2 \in \mathcal{N}_x$, $\forall t \in \mathbb{R}$

$$\mid \Psi(t,x_1) - \Psi(t,x_2) \mid \leq C_1 d^b(x_1,x_2).$$

(H3) La fonction ψ_x est continument différentiable, strictement monotone par rapport à la deuxième composante et sa dérivée est telle que $\mid \psi_x^{'}(t) \mid > C_2 > 0$, $\forall t \in \mathbb{R}$.

(H4) K est une fonction continue de support [0,1] et tel que $0 < C < K(t) < C^{'} < \infty$.

(H5) $\lim\limits_{n\to\infty} h_K = 0$ et $\lim\limits_{n\to\infty} \dfrac{\log n}{n\phi_x(h_K)} = 0$.

Commemtaire

1. L'hypothèse (H1) sur la loi marginale de la variable explicative X est n'est pas très restrictive. En effet, est une condition sur la fonction de répartition et non pas sur la densité. Dans le cas fini dimensionnel, l'hypothèse (H1) est même moins restrictive que la positivité stricte de la densité de la variable explicative (une condition indispensable dans le cas fini dimensionnel) car on peut avoir une boule centrée en un point et tel que la densité est nulle bien que la probabilité de cette boule soit strictement positive. Prenons l'exemple de la loi du χ_2 avec un degré de liberté supérieur à deux.
2. L'hypothèse (H2) est une hypothèse de régularité qui caractérise l'espace fonctionnel de notre modèle, ce qui justifie l'emploi des méthodes non paramétriques pour le problème considéré.
3. La condition (H3) contrôle la robustesse de notre modèle.
4. Les hypothèses (H4)- (H5) sont des hypothèses techniques nécessaires à l'obtention de nos résultats.

Résultat

On établit le résultat suivant

Théorème 3.3.1. *Sous les hypothèses (H1)-(H5), $\widehat{\theta}_x$ existe et est unique pour n assez grand, et on a*

$$\widehat{\theta}_x - \theta_x = O(h^b) + O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), \quad p.co. \tag{3.4}$$

Démonstration du théorème 3.3.1 Dans ce qui suit, on note par C une constante strictement positive et $K_i = K\left(\frac{d(x,X_i)}{h}\right)$.

Sous l'hypothèse (H3), nous avons

$$\widehat{\Psi}\big(\widehat{\theta}_x, x\big) = \widehat{\Psi}\big(\theta_x, x\big) + \big(\widehat{\theta}_x - \theta_x\big)\widehat{\Psi}'\big(\xi_{x,n}, x\big)$$

où $\xi_{x,n}$ est un point entre $\widehat{\theta}_x$ et θ_x.
La dérivabilité de ψ_x dans (H3), nous permet d'écrire, $\exists C_2 > 0$, $\forall \epsilon_0 > 0$,

$$\mathbb{P}\left(\mid \widehat{\theta}_x - \theta_x \mid \geq \epsilon_0\big(h^b + \sqrt{\frac{\log n}{n\phi_x(h)}}\big)\right) \leq \mathbb{P}\left(\mid \widehat{\Psi}(\theta_x, x) - \Psi(\theta_x, x) \mid \geq C_2\epsilon_0\big(h^b + \sqrt{\frac{\log n}{n\phi_x(h)}}\big)\right).$$

Alors, il suffit de montrer que :

$$\widehat{\Psi}(\theta_x, x) - \widehat{\Psi}(\theta_x, x) = O\left(h^b + \sqrt{\frac{\log n}{n\phi_x(h)}}\right), \quad p.co. \tag{3.5}$$

La preuve de l'équation 3.5 repose sur la décomposition suivante :

$$\begin{aligned}\forall t \in \mathbb{R}, \quad \widehat{\Psi}(t,x) - \Psi(t,x) = \; & \tfrac{1}{\widehat{\Psi}_D(x)}\left[\left(\widehat{\Psi}_N(t,x) - \mathbb{E}\big[\widehat{\Psi}_N(t,x)\big]\right) - \left(\Psi(t,x) - \mathbb{E}\big[\widehat{\Psi}_N(t,x)\big]\right)\right] \\ & - \tfrac{\Psi(t,x)}{\widehat{\Psi}_D(x)}\left[\widehat{\Psi}_D(x) - \mathbb{E}\big[\widehat{\Psi}_D(x)\big]\right]\end{aligned} \tag{3.6}$$

où

$$\widehat{\Psi}_D(x) = \frac{1}{n\mathbb{E}[K_1]} \sum_{i=1}^{n} K_i,$$

et

$$\widehat{\Psi}_N(t,x) = \frac{1}{n\mathbb{E}[K_1]} \sum_{i=1}^{n} K_i \psi_x(Y_i - t).$$

Finalement, la preuve du théorème 3.3.1 est achevée à partir des lemmes suivants.

Lemme 3.3.1. *Sous les hypothèses (H1) et (H4)-(H5), on a*

$$\widehat{\Psi}_D(x) - \mathbb{E}[\widehat{\Psi}_D(x)] = O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), \quad p.co.$$

Corollaire 3.3.1. *Sous les hypothèses de lemme 3.3.1, on a*

$$\sum_{n\geq 1} \mathbb{P}\left(\mid \widehat{\Psi}_D(x) \mid \leq \frac{1}{2} \right) \leq \sum_{n\geq 1} \mathbb{P}\left(\mid \widehat{\Psi}_D(x) - 1 \mid > \frac{1}{2} \right) < \infty.$$

Lemme 3.3.2. *Sous les hypothèses (H1)-(H2) et (H4)-(H5), on a pour tout $t \in \mathbb{R}$,*

$$\Psi(t,x) - \mathbb{E}\big[\widehat{\Psi}_N(t,x)\big] = O(h^b).$$

Lemme 3.3.3. *Sous les hypothèses (H1)-(H3) et (H5), on a pour tout $t \in \mathbb{R}$,*

$$\widehat{\Psi}_N(t,x) - \mathbb{E}\big[\widehat{\Psi}_N(t,x)\big] = O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), \quad p.co.$$

Lemme 3.3.4. *Sous les hypothèses du théorème 3.3.1, $\widehat{\theta}_x$ existe et est unique presque sûrement pour n assez grand.*

Preuve du lemme 3.3.1

Nous avons,

$$\widehat{\Psi}_D(x) - \mathbb{E}\big[\widehat{\Psi}_D(x)\big] = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{K_i}{\mathbb{E}[K_1]} - 1 \right).$$

On pose

$$\Delta_i = \frac{K_i}{\mathbb{E}[K_1]}.$$

D'aprés l'hypothèse (H4), on a

$$\mathbb{E}\big[C\mathbb{1}_{B(x,h)}(X_i)\big] < \mathbb{E}\big[K_1\big] < \mathbb{E}\big[C'\mathbb{1}_{B(x,h)}(X_i)\big],$$

alors,

$$C\phi_x(h) < \mathbb{E}\big[K_1\big] < C'\phi_x(h)$$

et puisque la fonction K est bornée, on peut majorer directement $\mid \Delta_i \mid$ par $\frac{C}{\phi_x(h)}$ et on montre que $\mathbb{E}[\Delta_i^2] \leq \frac{C'}{\phi_x(h)}$.

On applique maintenant l'inégalité de Hoeffding du lemme 3.2.1 aux variable $\mid \Delta_i \mid$,

$$\begin{aligned}\mathbb{P}\Big(\mid \widehat{\Psi}_D(x) - \mathbb{E}\big[\widehat{\Psi}_D(x)\big] \mid > \epsilon \Big) &= \mathbb{P}\Big(\tfrac{1}{n} \mid \sum \Delta_i \mid > \epsilon \Big) \\ &\leq 2\exp\big(\tfrac{-n\epsilon^2 C}{4\phi_D(x)}\big).\end{aligned}$$

On prend

$$\epsilon = \eta\sqrt{\frac{\log n}{n\phi_x(h)}}.$$

On arrive à

$$\mathbb{P}\Big(\mid \widehat{\Psi}_D(x) - \mathbb{E}\big[\widehat{\Psi}_D(x)\big] \mid > \eta\sqrt{\frac{\log n}{n\phi_x(h)}} \Big) \leq 2n^{-C\epsilon_0^2}, \tag{3.7}$$

donc,

$$\sum_n \mathbb{P}\Big(\mid \widehat{\Psi}_D(x) - \mathbb{E}\big[\widehat{\Psi}_D(x)\big] \mid > \eta\sqrt{\frac{\log n}{n\phi_x(h)}} \Big) \leq 2n^{-C\eta^2}.$$

Il suffit de choisir $\eta > \frac{1}{\sqrt{C}}$ pour que la série converge.■

Preuve du lemme 3.3.2

L'équidistribution des couples (X_i, Y_i) et la condition (H4) impliquent

$$\Psi(t,x)-\mathbb{E}\big[\widehat{\Psi}_N(t,x)\big] = \frac{1}{\mathbb{E}[K_1]}\mathbb{E}\Bigg[\Big(K_1\mathbb{1}_{B(x,h)}(X_1)\Big)\Big(\Psi(t,x)-\mathbb{E}\big[\psi_x(Y_1-t)/X=X_1\big]\Big)\Bigg]$$

où $\mathbb{1}$ est la fonction indicatrice.

L'hypothèse (H2) nous permet d'écrire,

$$K_1\mathbb{1}_{B(x,h)}(X_1) \mid \Psi(t,X_1)-\Psi(t,x) \mid \leq C_1 h^b,$$

alors,

$$\mid \Psi(t,x)-\mathbb{E}\big[\widehat{\Psi}_N(t,x)\big] \mid \leq C_1 h^b.\blacksquare$$

Preuve du lemme 3.3.3

La preuve de ce résultat est similaire que la preuve du lemme 3.3.1. On prend

$$\Lambda_i = \frac{K_i\psi_x\big(Y_i-t\big)-\mathbb{E}\big[K_1\psi_x\big(Y_1-t\big)\big]}{\mathbb{E}[K_1]}$$

Puisque la fonction ψ_x est bornée, alors $\mid \Lambda_i \mid \leq C/\phi_x(h)$ et $\mathbb{E}[\Lambda_i^2] \leq C'/\phi_x(h)$, pour tout $i \leq n$.

Comme le lemme 3.3.1, il suffit d'appliquer l'inégalité de Hoeffding pour obtenir le résultat.$\blacksquare$

Preuve du lemme 3.3.4

Pour tout $\epsilon > 0$, la monotonie stricte de ψ_x implique

$$\Psi\big(\theta_x-\epsilon,x\big) < \Psi\big(\theta_x,x\big) < \Psi\big(\theta_x+\epsilon,x\big).$$

Les lemmes 3.3.1, 3.3.3, 3.3.4 et le corollaire 3.3.1, montrent que

$$\widehat{\Psi}\big(\theta_x,x\big)-\Psi\big(\theta_x,x\big) = O\Bigg(h^b+\sqrt{\frac{\log n}{n\phi_x(h)}}\Bigg), \quad p.co$$

pour tout t un réel fixé, alors, pour n suffisament grand,

$$\widehat{\Psi}\big(\theta_x-\epsilon,x\big) \leq 0 \leq \widehat{\Psi}\big(\theta_x+\epsilon,x\big), \quad p.co$$

Puisque ψ_x et K sont des fonctions continues, donc $\widehat{\Psi}(t,x)$ est continue, alors il existe $t_0 = \widehat{\theta}_x \in \big[\theta_x - \epsilon, \theta_x + \epsilon\big]$ tel que $\widehat{\Psi}(\widehat{\theta}_x, x) = 0$.

Finalement, l'unicité de $\widehat{\theta}_x$ est une conséquence directe de la monotonie stricte de ψ_x et la positivité de la fonction K.■

Preuve du corollaire 3.3.1

Nous avons,

$$\mathbb{P}\big(\mid \widehat{\Psi}_D(x) \mid \leq \frac{1}{2}\big) \leq \mathbb{P}\big(\mid \widehat{\Psi}_D(x) - 1 \mid > \frac{1}{2}\big).$$

Notons que $\widehat{\Psi}_D(x) - 1 = \widehat{\Psi}_D(x) - \mathbb{E}\big[\widehat{\Psi}_D(x)\big]$, en appliquant le lemme précédent, on peut écrire $\sum_{i=1}^{n} \mathbb{P}\big(\widehat{\Psi}_D(x) < \frac{1}{2}\big) < \infty$.■

Remarque 3.3.1. *La démonstration du théorème 3.3.1 peut être diviser en deux parties : une partie biais et une partie dispersion. Dans la partie biais, on utilise le faite que les observations sont identiquement distribuées. Par contre, la démonstration de la partie dispersion repose sur le faite que les observations sont indépendantes.*

3.3.2 Normalité asymptotique

La propriété asymptotique abordé dans cette section est la normalité asymptotique, il s'agit d'un sujet très important en statistique. En effet, la normalité asymptotique nous permet de construire les intervalles de confiance et de faire les tests.

Hypothèses et résultat

On suppose que

$$\lambda_\gamma\big(u,t\big) = \mathbb{E}\Big[\psi^\gamma\big(Y-t\big)/X = u\Big] \text{ et } \Gamma_\gamma\big(u,t\big) = \mathbb{E}\Big[(\psi')^\gamma\big(Y-t\big)/X = u\Big]$$

pour $\gamma \in \{1,2\}$, on garde les mêmes notations que la section précédente, la même hypothèse (H3) et on remplace (H1), (H2), (H4) et (H5) par :

(H1') Il existe une fonction positive différentiable ϕ et une fonction positive g telle

que :

$$\mathbb{P}\big(X \in B(x,r)\big) = \phi(r)g(x) + o\big(\phi(r)\big).$$

(H2')
(i) La fonction $\lambda_\gamma(.,.)$ satisfait la condition de lipschitz par rapport à la première composante, c'est-à-dire : il existe une constante b_γ strictement positive tel que :

$$\forall (u_1,u_2) \in N_x \times N_x, \quad \forall t \in \mathbb{R}, \quad \mid \lambda_\gamma(u_1,t) - \lambda_\gamma(u_2,t) \mid \leq C_1 d(u_1,u_2)^{b_\gamma}.$$

(ii) La fonction $\Gamma_\gamma(.,.)$ satisfait la condition de lipschitz par rapport à la première composante, c'est-à-dire : il existe une constante d_γ strictement positive tel que :

$$\forall (u_1,u_2) \in N_x \times N_x, \quad \forall t \in \mathbb{R}, \quad \mid \Gamma_\gamma(u_1,t) - \Gamma_\gamma(u_2,t) \mid \leq C_2 d(u_1,u_2)^{d_\gamma}.$$

(H4') Le noyau K est une fonction différentiable sur le support $[0,\ 1]$, sa dérivée K' existe et telle que $-\infty < C_3 < K'(t) < C_4 < 0$.
(H5') Le paramètre de lissage h satisfait :

$$h \downarrow 0, \quad \forall t \in [0,1], \quad \lim \frac{\phi(th)}{\phi(h)} = \beta(t) \quad et \quad n\phi(h) \to 0 \quad quand \quad n \to \infty$$

Notre résultat principal est donné par le théorème suivant.

Théorème 3.3.2. *Sous les hypothèses $(H1')-(H2')$, $(H3')$ et $(H4')-(H5')$, $\widehat{\theta}_x$ existe et est unique avec une probabilité tend vers 1, et pour tout $x \in \mathcal{A}$, on a*

$$\left(\frac{n\phi(h)}{\sigma^2\big(x,\theta_x\big)}\right)^{\frac{1}{2}} \Big(\widehat{\theta}_x - \theta_x - B_n(x)\Big) \xrightarrow{L} \mathcal{N}(0,1) \quad quand\ n \to \infty \tag{3.8}$$

où

$$B_n(x) = \frac{h}{\phi(h)\alpha_1\Gamma_1(x,\theta_x)} \int_0^1 K(t)\varphi_x(th)\phi'(th)dt + o(1).$$

avec $\varphi_x(s) = \mathbb{E}\Big[\psi\big(Y,\theta_x\big)/d\big(X,x\big) = s\Big]$,

$$\sigma^2(x,\theta_x) = \frac{\alpha_2\lambda_2(x,\theta_x)}{\alpha_1^2 g(x)(\Gamma_1(x,\theta_x\big)^2)}$$

avec $\alpha_j = \int_0^1 (K^j)'(s)\beta(s)ds$, *pour* $j = 1, 2$,

$$\mathcal{A} = \left\{ x \in \mathcal{F},\ g(x)\lambda_1(x,\theta_x)\Gamma_1(x,\theta_x) \neq 0 \right\}$$

Corollaire 3.3.2. *Sous les hypothèses du théorème 3.3.2 et si le paramètre de lissage* h *satisfait* $nh^{2b_1}\phi(h) \to 0$ *quand* $n \to \infty$, *on a*

$$\left(\frac{n\phi(h)}{\sigma^2(x,\theta_x)}\right)^{\frac{1}{2}} (\widehat{\theta}_x - \theta_x) \xrightarrow{L} \mathcal{N}(0,1) \quad \textit{quand} \quad n \to \infty. \tag{3.9}$$

Démonstation du théorème 3.3.2

Pour simplifier les notations, on pose, pour $i = 1, \cdots, n$, $K_i = K\left(\frac{d\left(x,X_i\right)}{h}\right)$,

$\psi_i(t) = \psi\left(Y_i, t\right)$ et

$$\widehat{\Psi}_N(x,t) = \frac{1}{n\mathbb{E}[K_1]} \sum_{i=1}^{n} K_i\psi_i(t).$$

On utilise le développement de Taylor d'ordre 1 autour de θ_x, on obtient

$$\widehat{\Psi}_N\left(x,\widehat{\theta}_x\right) = \widehat{\Psi}_N\left(x,\theta_x\right) + \left(\widehat{\theta}_x - \theta_x\right)\widehat{\Psi}'_N\left(x,\xi_n\right)$$

avec $\xi_n \in (\widehat{\theta}_x, \theta_x)$.

Par définition de $\widehat{\theta}_x$, on a

$$\widehat{\theta}_x - \theta_x = -\frac{\widehat{\Psi}_N\left(x,\theta_x\right)}{\widehat{\Psi}'_N\left(x,\xi_n\right)}.$$

Finalement, on utilise la décomposition suivante

$$\sqrt{n\phi(h)}\left(\widehat{\theta}_x - \theta_x\right) = \frac{-\sqrt{n\phi(h)}\left(\widehat{\Psi}_N(x,\theta_x) - \mathbb{E}\left[\widehat{\Psi}_N\left(x,\theta_x\right)\right]\right)}{\widehat{\Psi}'_N(x,\xi_n)} - \frac{\sqrt{n\phi(h)}\mathbb{E}\left[\widehat{\Psi}_N(x,\theta_x)\right]}{\widehat{\Psi}'_N\left(x,\xi_n\right)}$$

La preuve du théorème 3.3.2 est basée sur les lemmes suivants.

Lemme 3.3.5. *Sous les hypothèses (H1')-(H2'), (H3') et (H4')-(H5'), on a pour tout* $x \in \mathcal{A}$

$$\left(\frac{n\phi(h)\alpha_1^2 g(x)}{\alpha_2 \lambda_2(x,\theta_x)}\right)^{1/2}\left(\widehat{\Psi}_N\big(x,\theta_x\big) - \mathbb{E}\big[\widehat{\Psi}_N\big(x,\theta_x\big)\big]\right) \xrightarrow{L} \mathcal{N}(0,1) \quad \textit{quand } n \to \infty.$$

Lemme 3.3.6. *Sous les hypothèses (H1') et (H5'), on a*

$$\mathbb{E}\big[\widehat{\Psi}_N\big(x,\theta_x\big)\big] = \frac{h}{\phi(h)\alpha_1}\int_0^1 K(t)\varphi_x(th)\phi'(th)dt + o(1).$$

Lemme 3.3.7. *Sous les hypothèses (H1')-(H2'), (H3') et (H4')-(H5'), on a pour tout* $t \in \mathbb{R}$

$$\mathbb{E}\big[\widehat{\Psi}_N(x,t)\big] = \lambda_1\big(x,\theta_x\big) + O\big(h^{b_1}\big).$$

Lemme 3.3.8. *Sous les hypothèses du théorème 3.3.2, on a*

$$\widehat{\Psi}'_N\big(x,\xi_n\big) \xrightarrow{\mathbb{P}} \Gamma_1\big(x,\theta_x\big) \quad \textit{quand } n \to \infty.$$

Preuve du lemme 3.3.5

La démonstration de ce lemme est basée sur le théorème central limite qui exige comme condition suffisante l'existence d'une constante $\delta > 0$ vérifiant :

$$\frac{\sum_{i=1}^{n}\mathbb{E}\Big[\mid L_i(x) - \mathbb{E}[L_i(x)]\mid^{2+\delta}\Big]}{\Big(Var\Big[\sum_{i=1}^{n} L_i(x)\Big]\Big)^{(2+\delta)/2}} \to 0$$

où

$$L_i(x) = \frac{1}{n\mathbb{E}[K_1]}\Big[K_i\psi_i(\theta_x)\Big].$$

D'où la nécessité de calculer $n\phi(h)Var\big[\widehat{\Psi}_N(x,t)\big]$, par définition de $\widehat{\Psi}_N(x,t)$, on a

$$\begin{aligned} Var\left[\widehat{\Psi}_N(x,t)\right] &= \tfrac{1}{(n\mathbb{E}[K_1])^2}\sum_{i=1}^{n} Var\left[K_i\psi_i(t)\right] \\ &= \tfrac{1}{n\mathbb{E}^2[K_1]} Var\left[K_1\psi_1(t)\right] \\ &= \tfrac{1}{n\mathbb{E}^2[K_1]}\mathbb{E}\left[K_1^2\psi^2(Y,t)\right] - \tfrac{1}{n\mathbb{E}^2[K_1]}\left(\mathbb{E}\left[K_1\psi(Y,t)\right]\right)^2 \\ &= \tfrac{\mathbb{E}[K_1^2]}{n\mathbb{E}^2[K_1]}\mathbb{E}\left[\tfrac{K_1^2\psi^2(Y,t)}{\mathbb{E}[K_1^2]}\right] - \tfrac{1}{n}\left(\mathbb{E}\left[\tfrac{K_1\psi(Y,t)}{\mathbb{E}[K_1]}\right]\right)^2. \end{aligned}$$

Donc

$$n\phi(h)Var\big[\Psi_N(x,t)\big] = \frac{\phi(h)\mathbb{E}[K_1^2]}{n\mathbb{E}^2[K_1]}\left(\mathbb{E}\left[\frac{K_1^2\psi^2(Y,t)}{\mathbb{E}[K_1^2]}\right]\right) - \phi(h)\left(\mathbb{E}\left[\frac{K_1\psi(Y,t)}{\mathbb{E}[K_1]}\right]\right)^2. \tag{3.10}$$

D'une part, nous évaluons la limite de seconde terme de la côté à droite de l'équation 3.10.

Puisque $\phi(h) \to 0$, alors, il suffit de montrer que

$$\forall t \in \mathbb{R}, \quad \mathbb{E}\left[\frac{K_1\psi(Y,t)}{\mathbb{E}[K_1]}\right] \to \lambda_1(x,t). \tag{3.11}$$

En effet,

$$\mid \mathbb{E}\left[\frac{K_1\psi(Y,t)}{\mathbb{E}[K_1]}\right] - \lambda_1(x,t) \mid = \frac{1}{\mathbb{E}[K_1]} \mid \mathbb{E}[K_1]\mathbb{1}_{B(x,h)}(X_1) \times \Big(\lambda_1(X_1,t) - \lambda_1(x,t)\Big) \mid .$$

On utilise l'hypothèse (H3') (avec $\gamma = 1$), on obtient

$$\mathbb{1}_{B(x,h)}(X_1) \times \mid \lambda_1(X_1,t) - \lambda_1(x,t) \mid \leq ch^{b_1},$$

Alors, $\mid \mathbb{E}\left[\frac{K_1\psi(Y,t)}{\mathbb{E}[K_1]}\right] - \lambda_1(x,t) \mid \leq ch^{b_1} \to 0$. D'une autre part, pour le premier terme

de la partie à gauche de l'équation 3.10, on va montrer que :

$$\forall t \in \mathbb{R}, \quad \mathbb{E}\left[\frac{K_1^2 \psi_x^2(Y,t)}{\mathbb{E}[K_1^2]}\right] \to \lambda_2(x,t). \tag{3.12}$$

En effet, l'hypothèse (H2), nous permet d'obtenir

$$\mid \mathbb{E}\left[\frac{K_1^2 \psi_x^2(Y,t)}{\mathbb{E}[K_1^2]}\right] - \lambda_2(x,t) \mid = \frac{1}{\mathbb{E}[K_1^2]} \mid \mathbb{E}[K_1]\mathbb{1}_{B(x,h)}(X_1) \times \Big(\lambda_2(X_1,t) - \lambda_2(x,t)\Big) \mid .$$

Finalement, l'hypothèse (H3') (avec $\gamma = 2$) donne,

$$\frac{1}{\mathbb{E}[K_1^2]} \mid \mathbb{E}[K_1^2]\mathbb{1}_{B(x,h)}(X_1) \times \Big(\lambda_2(X_1,t) - \lambda_2(x,t)\Big) \mid \leq Ch^{b_1},$$

On va calculer maintenant les deux termes $\mathbb{E}[K_1]$ et $\mathbb{E}[K_1^2]$. En effet, on a

$$\mathbb{E}\left[K\left(\frac{d(x,X_1)}{h}\right)\right] = \int_0^1 K(u) d\mathbb{P}^{\frac{d(x,X_1)}{h}}(u).$$

Il est claire que $\int_0^u K'(t)dt = K(u) - K(0)$, donc

$$\begin{aligned}
\mathbb{E}\left[K\left(\tfrac{d(x,X_1)}{h}\right)\right] &= \textstyle\int_0^1 K(0) d\mathbb{P}^{\frac{d(x,X_1)}{h}}(u) + \int_0^1 \int_0^u K'(t)dt d\mathbb{P}^{\frac{d(x,X_1)}{h}}(u) \\
&= K(0)\phi_x(h) + \textstyle\int_0^1 \int_0^1 K'(t)\mathbb{1}_{[0,u]}(t)dt d\mathbb{P}^{\frac{d(x,X_1)}{h}}(u) \\
&= K(0)\phi_x(h) + \textstyle\int_0^1 \int_0^1 K'(t)\mathbb{1}_{[t,1]}(u)dt d\mathbb{P}^{\frac{d(x,X_1)}{h}}(u) \\
&= K(0)\phi_x(h) + \textstyle\int_0^1 K'(u)\mathbb{P}\Big(t < \frac{d(x,X_1)}{h} < 1\Big)dt \\
&= K(0)\phi_x(h) + \phi_x(h)K(1) - K(0)\phi_x(h) - g(x)\textstyle\int_0^1 K^{(1)}(t)\phi_x(th)dt + o(\phi(h)) \\
&= -g(x)\textstyle\int_0^1 K^{(1)}(t)\phi_x(th)dt + o(\phi(h)) \qquad (car \quad K(1) = 0).
\end{aligned}$$

De même, pour le deuxième terme,

$$\mathbb{E}[K_1^2] = -g(x)\int_0^1 (K^2)'(t)\phi_x(th)dt + o(\phi(h)).$$

Alors,

$$\frac{\phi(h)\mathbb{E}[K_1^2]}{\mathbb{E}^2[K_1]} \to \frac{\alpha_2}{\alpha_1 g(x)} \quad n \to \infty. \tag{3.13}$$

Les équations 3.11, 3.12 et 3.13, nous permet d'obtenir,

$$n\phi(h)Var\left[\Psi_N(x,t)\right] \to \frac{\alpha_2\lambda_2(x,t)}{\alpha_1 g(x)}. \tag{3.14}$$

Il reste maintenant à évaluer la limite de nomérateur, pour laquelle on utilise l'inégalité Cr du lemme 3.2.4, ce qui donne

$$\begin{aligned}\big(n\phi(h)\big)^{1+\delta/2}\sum_{i=1}^{n}\mathbb{E}\Big[\mid L_i(x)-\mathbb{E}\big[L_i(x)\big]\mid^{2+\delta}\Big] &\leq 2^{1+\delta}\big(n\phi(h)\big)^{(2+\delta)/2}\sum_{i=1}^{n}\mathbb{E}\Big[\mid L_i(x)\mid^{2+\delta}\Big] \\ &\quad +2^{1+\delta}\big(n\phi(h)\big)^{(2+\delta)/2}\sum_{i=1}^{n}\mid\mathbb{E}\big[L_i(x)\big]\mid^{2+\delta}.\end{aligned} \tag{3.15}$$

On remarque que $\forall j>0, \quad \mathbb{E}\big[K_1^j\big] = O\big(\phi(h)\big)$, alors, la positivité de la fonction ψ et l'hypothèse (H2), nous permet d'obtenir

$$\begin{aligned}2^{1+\delta}\big(n\phi(h)\big)^{1+\delta/2}\sum_{i=1}^{n}\mathbb{E}\Big[\mid L_i(x)\mid^{2+\delta}\Big] &= 2^{1+\delta}n^{-\delta/2}\big(\phi(h)\big)^{(-1-\delta)/2}\left(\frac{\phi(h)}{\mathbb{E}[K_1]}\right)^{2+\delta}\mathbb{E}\Big[K_1^{2+\delta}\mid\psi_1(\theta_x)\mid^{2+\delta}\Big] \\ &\leq C\big(n\phi(h)\big)^{-\delta/2}\left(\frac{\phi(h)}{\mathbb{E}[K_1]}\right)^{2+\delta}\mathbb{E}\Big[K_1^{2+\delta}/\phi(h)\Big]\to 0.\end{aligned}$$

En ce qui concerne le deuxième terme de l'équation 3.15, on a

$$\begin{aligned}2^{1+\delta}\big(n\phi(h)\big)^{1+\delta/2}\sum_{i=1}^{n}\mid\mathbb{E}\big[L_i(x)\big]\mid^{2+\delta} &\leq 2^{1+\delta}n^{-\delta/2}\big(\phi(h)\big)^{(1+\delta)/2}\mathbb{E}[K_1]^{2+\delta}\mid\mathbb{E}\Big[K_1\psi_1(\theta_x)\Big]\mid^{2+\delta} \\ &\leq Cn^{-\delta/2}\big(\phi(h)\big)^{1+\delta/2}\to 0.\blacksquare\end{aligned}$$

Preuve du lemme 3.3.6

On a

$$\mathbb{E}\Big[\widehat{\Psi}_N(x,\theta_x)\Big] = \frac{\mathbb{E}\Big[K_1\psi_1(\theta_x)\Big]}{\mathbb{E}\big[K_1\big]}.$$

Par conditionnement à $d(x,X_1)$, on obtient

$$\mathbb{E}\Big[\widehat{\Psi}_N(x,\theta_x)\Big] = \frac{\mathbb{E}\Big[K_1\mathbb{E}\big[\psi_1(\theta_x)/d(x,X1)\big]\Big]}{\mathbb{E}\big[K_1\big]}.$$

L'intégration par rapport à la première composante montre que

$$\mathbb{E}\Big[\widehat{\Psi}_N(x,\theta_x)\Big] = \frac{hg(x)\int_0^h K\big(h^{-1}t\big)\varphi_x(t)\phi'(t)dt + O\big(\phi(h)\big)}{-g(x)\int_0^h K'(t)\phi(th)dt + O\big(\phi(h)\big)}.$$

On prend le changement de variable $h^{-1}t = s$

$$\mathbb{E}\Big[\widehat{\Psi}_N(x,\theta_x)\Big] = \frac{hg(x)\int_0^1 K(s)\varphi_x(hs)\phi'(hs)ds + O\big(\phi(h)\big)}{-g(x)\int_0^h K'(t)\phi(th)dt + O(\phi(h))}.$$

Clairement d'après la définition de α_1, le dénominateur normalisé par $g(x)\phi(h)$ converge vers α_1 ce qui implique

$$\mathbb{E}\Big[\widehat{\Psi}_N(x,\theta_x)\Big] = \frac{hg(x)}{\phi(h)g(x)\alpha_1 + O\big(\phi(h)\big)}\left(\int_0^1 K(s)\varphi_x(hs)\phi'(hs)ds + O\big(\phi(h)\big)\right).\blacksquare$$

Preuve du lemme 3.3.7

Nous avons,

$$\mid \mathbb{E}\big[\widehat{\Psi}_N(x,t)\big] - \lambda_1(x,t) \mid \leq \mid \mathbb{E}\left[\frac{1}{\mathbb{E}[K_1]}K_1\mathbb{1}_{B(x,h)}(X_1)\Big(\lambda_1(X_1,t) - \lambda_1(X,t)\Big)\right] \mid .$$

Sous (H2'), on obtient

$$\mid \mathbb{E}\left[\frac{1}{\mathbb{E}[K_1]}K_1\mathbb{1}_{B(x,h)}(X_1)\Big(\lambda_1(X_1,t) - \lambda_1(X,t)\Big)\right] \mid \leq Ch^{b_1}.\blacksquare$$

Preuve du lemme 3.3.8

On utilise la décomposition suivante

$$| \Psi_N^{'}(x,\xi_n) - \Gamma_1(x,\theta_x) | \leq | \Psi_N^{'}(x,\xi_n) - \Psi_N^{'}(x,\theta_x) | + | \Psi_N^{'}(x,\theta_x) - \Gamma_1(x,\theta_x) | . \quad (3.16)$$

Concernant le premier terme, on remarque que

$$| \Psi_N^{'}(x,\xi_n) - \Psi_N^{'}(x,\theta_x) | \leq \sup_{y\in\mathbb{R}} | \frac{\partial\psi(y,\xi_n)}{\partial t} - \frac{\partial\psi(y,\theta_x)}{\partial t} | \widehat{\Psi}_D(x),$$

et puisque $\frac{\partial\psi(y,t)}{\partial t}$ est continue au θ_x uniformement à y. La proposition 3.3.1 et la convergence en probabilité de $\widehat{\Psi}_D(x)$ vers 1 montre que le premier terme de 3.16 converge en probabilité vers 0. Cependant, la limite de seconde terme est obtenue par évaluer séparément le biais et la variance de $\Psi_N^{'}(x,\theta_x)$ Clairement, des arguments similaires de la preuve du lemme 3.3.3 peuvent être utilisés pour obtenir

$$\mathbb{E}\left[\Psi_N^{'}(x,\theta_x)\right] \to \Gamma_1(x,\theta_x).$$

De plus, on utilise des arguments analogues que 3.14, on peut montrer,

$$n\phi(h)Var\left[\Psi_N^{'}(x,\theta_x)\right] \to \frac{\alpha_2\Gamma_2\big(x,\theta_x\big)}{\alpha_1^2 g(x)}.$$

Finalement, sous les hypothèses (H1') et (H4') la preuve est achevée.■

Corollaire 3.3.3. *Sous les hypothèses du lemme 3.3.7 et si le paramètre de lissage h satisfait $nh^{2b_1}\phi(h) \to 0$, quand $n \to \infty$, on a*

$$\sqrt{n\phi(h)}\mathbb{E}\left[\widehat{\Psi}_N(x,\theta_x)\right] \to 0, \quad quand \quad n \to \infty.$$

Preuve du corollaire 3.3.3

Le lemme 3.3.7 permet d'obtenir facilement

$$\sqrt{n\phi(h)}\mathbb{E}\left[\widehat{\Psi}_N(x,\theta_x)\right] = \sqrt{n\phi(h)h^{2b_1}}.$$

Proposition 3.3.1. *(voir Azzedine et al. (2008) [5])*
Supposons que (H1')-(H2'), (H3') et (H4')-(H5') sont satisfaites, donc $\widehat{\theta}_x$ existe et il est unique avec une probabilité tend vers 1, et on a

$$\widehat{\theta}_x - \theta_x \xrightarrow{\mathbb{P}} 0 \quad quand\ n \to \infty. \blacksquare \tag{3.17}$$

3.4 Résultats asymptotiques : Cas dépendant

Le but de cette section est la généralisation du résultat donné dans la section précédente à des observations mélangantes. On établit la convergence presque complète de l'estimateur à noyau de la fonction de régression robuste.

3.4.1 Propriétes asymptotiques

Afin d'établir la convergence presque complète de l'estimateur à noyau défini dans la premiere section, on garde les mêmes hypothèses, ainsi les mêmes notations du section précédente et on ajoute les hypothèses suivantes.
(H6) $(X_i, Y_i)_{i\geq 1}$ est une suite α-mélangeante avec ces coefficients satisfaient

$$\alpha(n) = O(n^{-a}) \quad pour\ a > 0$$

(H7) $0 < \sup_{i\neq j} \mathbb{P}\Big(\big(X_i, Y_j\big) \in B(x,h) \times B(x,h) \Big) = O\Big(\frac{\big(\phi_x(h)\big)^{(a+1)/a}}{n^{1/a}} \Big)$.
(H8) Il existe $\eta > 0$, tel que

$$Cn^{\frac{3-a}{a+1}+\eta} \leq \phi_x(h) \leq C'n^{\frac{1}{1-a}+\eta} \text{ avec } a > \tfrac{5+\sqrt{17}}{2}.$$

Remarque 3.4.1. *Les hypothèses (H6)-(H8) sont ajoutées pour éviter l'expression de covariance dans la vitesse de convergence. Autrement dit, on peut démontrer la convergence presque complète sans ces hypothèses. Cependant, la vitesse de convergence sera donné en fonction de covariance des observations et elle sera lente par rapport à la vitesse du cas indépendant. Ainsi, nous établissons la convergence presque complète avec la même précision, mais, sous des conditions un plus fort que le cas i.i.d.*

Théorème 3.4.1. *Sous les hypothèses (H1)-(H8) et si* $\Gamma(x,\theta_x) \neq 0$, *alors* $\widehat{\theta}_x$ *existe et il est unique p.s pour n assez grand, et nous avons*

$$\widehat{\theta}_x - \theta_x = O(h^{b_1}) + O\left(\sqrt{\frac{\log(n)}{n\phi_x(h)}}\right), \ p.co \quad quand\ n \to \infty.$$

3.4.2 Démonstation des résultats techniques

La démonstration est essentiellement basée sur les mêmes arguments analytiques utilisés dans la démonstration du théorème 3.3.1 à savoir le développement du Taylor de l'estimateur et la décomposition 3.6 et en vertu de la remarque 3.3.1, on peut dire que la propriété de l'indépendance des observations n'aucune influence sur la partie biais de la vitesse. Autrement dit, la vitesse de convergence de la partie biais de théorème 3.3.1 sera la même dans le cas de mélange. Cependant, la partie dispersion est basée sur les deux lemmes suivants.

Lemme 3.4.1. *Sous les hypothèses (H1), (H3)-(H8), on a*

$$\widehat{\Psi}_D - \mathbb{E}\big[\widehat{\Psi}_D\big] = O\left(\sqrt{\frac{\log(n)}{n\phi_x(h)}}\right), \ p.co \quad quand\ n \to \infty.$$

Preuve du lemme 3.4.1

On pose $\Delta_i = K_i(x) - \mathbb{E}[K_i(x)]$, alors,

$$\widehat{\Psi}_D - \mathbb{E}[\widehat{\Psi}_D] = \frac{1}{n\mathbb{E}[K_1(x)]}\sum_{i=1}^{n}\Delta_i(x)$$

On applique l'inégalité de Fuck-Nagaev du lemme 3.2.2, on obtient pour tout $r > 0$ et $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\ |\ \mathbb{E}\big[\widehat{\Psi}_D(x)\big] - \widehat{\Psi}_D(x)\ |> \epsilon\right) &\leq \mathbb{P}\left(\ |\ \sum_{i=1}^{n}\Delta_i(x)\ |> \epsilon n\mathbb{E}\big[K_1(x)\big]\right) \\ &\leq C\Big(A_1(x) + A_2(x)\Big),\end{aligned} \tag{3.18}$$

où

$$A_1(x) = \left(1 + \frac{\epsilon^2 n^2 \big(\mathbb{E}[K_1(x)]\big)^2}{r S_n^2}\right)^{-r/2}, \quad A_2(x) = nr^{-1}\left(\frac{r}{\epsilon n \mathbb{E}[K_1(x)]}\right)^{a+1}$$

et

$$\begin{aligned} S_n^2(x) &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) \\ &= S_n^{2*}(x) + nVar\Big[\Delta_1(x)\Big] \end{aligned} \tag{3.19}$$

avec

$$S_n^{2*}(x) = \sum_{i\neq j} Cov\Big(\Delta_i(x), \Delta_j(x)\Big). \tag{3.20}$$

On utilise les techniques de Masry (1986) et on partage cette somme sur les deux ensembles

$$E_1 = \Big\{(i,j)\ tels\ que\ 1 \leq \mid i-j \mid \leq m_n\Big\}$$

et

$$E_2 = \Big\{(i,j)\ tels\ que\ m_n + 1 \leq \mid i-j \mid \leq n-1\Big\}$$

où $m_n \to \infty$, quand $n \to \infty$ où m_n est une suite des entiers tend vers l'infinie quand n tend vers à l'infinie. Alors,

$$S_n^{2*}(x) = \sum_{i\neq j} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) = \sum_{E_1} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) + \sum_{E_2} Cov\Big(\Delta_i(x), \Delta_j(x)\Big)$$

En ce qui concerne la première partie, on a

$$J_{1,n} = \sum_{E_1} \mid Cov\Big(\Delta_i(x), \Delta_j(x)\Big) \mid \leq \sum_{E_1} \mid \mathbb{E}\Big[K_i(x)K_j(x)\Big] - \mathbb{E}^2\Big[K_1(x)\Big] \mid$$

On utilise les hypothèses (H1) et (H7)-(H8), on obtient

$$J_{1,n} = Cnm_n\phi_x(h)\left(\left(\frac{\phi_x(h)}{n}\right)^{1/a} + \phi_x(h)\right).$$

Pour E_2, on a d'après l'inégalité de covariance

$$\mid Cov\Big(K_i(x), K_j(x)\Big) \mid \leq C\alpha\Big(\mid i-j \mid\Big).$$

Par conséquent,

$$J_{2,n} = \sum_{E_2} \mid Cov\Big(K_i(x), K_j(x)\Big) \mid \leq n^2 m_n^{-a}.$$

Choisissons $m_n = \left(\frac{\phi_x(h)}{n}\right)^{1/a}$ permet d'obtenir, sous (H8)

$$S_n^{2*}(x) = J_{1,n} + J_{2,n} = O\Big(n\phi_x(h)\Big). \tag{3.21}$$

Concernant la variance, l'hypothèse (H1) nous permet d'obtenir

$$Var\Big[\Delta_1(x)\Big] \leq C\Big(\phi_x(h) + (\phi_x(h))^2\Big). \tag{3.22}$$

Finalement, par les équations 3.27, 3.20, 3.28 et 3.29, on obtient

$$S_n^2(x) = O\big(n\phi_x(h)\big). \tag{3.23}$$

Maintenant, on applique 3.18 avec

$$\epsilon = \lambda\frac{\sqrt{n\log n\phi_x(h)}}{n\mathbb{E}\big[K_1(x)\big]} \; et \; r = C\big(\log n\big)^2. \tag{3.24}$$

Sous (H8), nous avons,

$$A_2(x) \leq Cn^{-1-\eta(a+1)/2}(\log n)^{(3a-1)/2}.$$

Alors, il existe $\nu > 0$ tel que

$$A_2(x) \leq Cn^{-1-\nu}. \tag{3.25}$$

On utilise 3.24 et 3.23, on obtient

$$A_1 \leq C \exp\left(-\lambda^2 \frac{\log n}{2} \right) = Cn^{-\lambda^2/2}$$

Donc, pour λ assez grand

$$\exists \nu' > 0, \quad A_1(x) \leq Cn^{-\lambda^2/2} \leq Cn^{-1-\nu'}. \tag{3.26}$$

Finalement, le résultat est déduit facilement par 3.31, 3.20 et 3.26.■

Lemme 3.4.2. *Sous les hypothèses (H1), (H3)-(H8), on a*

$$\widehat{\Psi}_N(x) - \mathbb{E}\big[\widehat{\Psi}_N(x)\big] = O\left(\sqrt{\frac{\log(n)}{n\phi_x(h)}}\right), \ p.co \ quand \ n \to \infty.$$

Preuve du lemme 3.4.2 On va calculer

$$S_n^2(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) = \sum_{ij}^{n} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) + nVar\Big[\Delta_1(x)\Big] \tag{3.27}$$

avec $\Delta_i = K_i(x)\psi\Big(Y_i - t\Big) - \mathbb{E}\Big[K_i(x)\psi(Y_i - t)\Big]$. On partage cette somme sur les deux ensembles

$$S_1 = \Big\{(i,j) \ tels \ que \ 1 \leq \mid i-j \mid \leq u_n\Big\}$$

et

$$S_2 = \Big\{(i,j) \ tels \ que \ u_n + 1 \leq \mid i-j \mid \leq n-1\Big\}$$

Alors,

$$J'_{1,n} = \sum_{S_1} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) \leq \sum_{S_1} \mid \mathbb{E}\Big[K_i(x)K_j(x)\Big] - \mathbb{E}^2\Big[K_1(x)\Big] \mid$$

On utilise les hypothèses (H1), (H7) et (H8), on obtient

$$J_{1,n}^{'} = Cnu_n\phi_x(h)\left(\left(\frac{\phi_x(h)}{n}\right)^{1/a} + \phi_x(h)\right)$$

et on a,

$$J_{2,n}^{'} = \sum_{S_2} \mid Cov\Big(K_i(x), K_j(x)\Big) \mid \leq n^2 u_n^{-a}.$$

On prend $u_n = \left(\frac{\phi_x(h)}{n}\right)^{1/a}$, sous (H8), on obtient,

$$\sum_{i=1}^{n}\sum_{ij}^{n} Cov\Big(\Delta_i(x), \Delta_j(x)\Big) = O\Big(n\phi_x(h)\Big). \tag{3.28}$$

Concernant la variance, l'hypothèse (H1) nous permet d'obtenir

$$Var\Big[\Delta_1(x)\Big] \leq C\Big(\phi_x(h) + \big(\phi_x(h)\big)^2\Big). \tag{3.29}$$

On applique l'inégalité de Fuk-Nagaev du lemme 3.2.2, on obtient, $\forall r > 0$ et $\forall \epsilon > 0$,

$$\begin{aligned}\mathbb{P}\Big(\mid \mathbb{E}\big[\widehat{\Psi}_N(x)\big] - \widehat{\Psi}_N(x) \mid > \epsilon\Big) &\leq \mathbb{P}\Big(\mid \sum_{i=1}^{n}\Delta_i(x) \mid > \epsilon n\mathbb{E}\big[K_1(x)\big]\Big)\\ &\leq C\Big(A_1'(x) + A_2'(x)\Big)\end{aligned}$$

où

$$A_1'(x) = \left(1 + \frac{\epsilon^2 n^2\Big(\mathbb{E}\big[K_1(x)\big]\Big)^2}{rS_n^2}\right)^{-r/2}, \quad A_2'(x) = nr^{-1}\left(\frac{r}{\epsilon n\mathbb{E}\big[K_1(x)\big]}\right)^{a+1}$$

On prend

$$\epsilon = \lambda\frac{\sqrt{n\log n\phi_x(h)}}{n\mathbb{E}[K_1(x)]} \; et \; r = C(\log n)^2. \tag{3.30}$$

on obtient,

$$A_2'(x) \leq Cn^{-1-\eta(a+1)/2}(\log n)^{(3a-1)/2}.$$

Alors, il existe $\nu' > 0$ tel que

$$A_2'(x) \leq Cn^{-1-\nu'}. \tag{3.31}$$

Finalement, les équations 3.18, 3.31, 3.26 nous permet de conclure le résultat.■

Chapitre 4

Application sur des données réelles

Cette application est présentée par Ould Saïd. Les données sont disponibles sur le site web http ://lib.stat.cmu.edu/datasets/tecator.

4.1 Les données

On se donne un échantillon de 215 morceaux de viande, le but est d'estimer le taux de matière grasse. Pour cela, on observe le spectre pour 100 longueurs d'onde réparties entre 250 et 1050 nanomètres.

On observe alors pour chaque morceau de viande i, la variable fonctionnelle $X_i(t)$, $t \in [850, 1050]$ qui est la courbe spectromètrique du morceau de viande i, la répartition graphique des 215 spectres montre l'aspect fonctionnel des données.

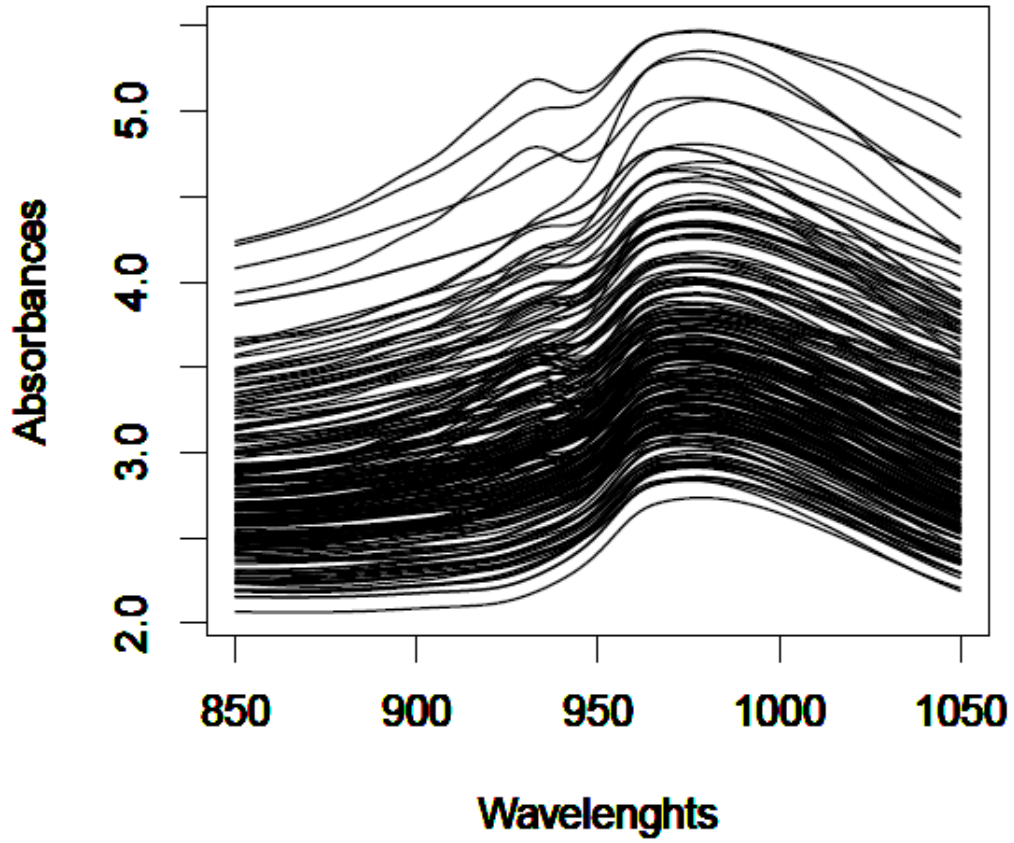

FIGURE 4.1 – 215 courbes spectromètriques, $X_i(t),\ t \in [850, 1050],\ i = 1, \cdots, 215$.

4.2 Le modèle

On se propose de comparer deux méthodes a savoir : la régression classique $r(x) = \mathbb{E}\big[Y/X = x\big]$ dont un estimateur est donné par (i.e $\psi(t) = t$)

$$\widehat{r}(x) = \frac{\sum_{i=1}^{n} K\big(h_K^{-1} d(x, X_i)\big)\psi_x\big(Y_i - t\big)}{\sum_{i=1}^{n} K\big(h_K^{-1} d(x, X_i)\big)}, \quad \forall t \in \mathbb{R}$$

et la régression robuste θ_x associée a $\psi(t) = \frac{1}{\sqrt{1+t^2/2}}$ en présence de points abérrants.

4.3 L'Algorithme

On utilise alors l'algorithme suivant :

- utiliser le noyau $K(x) = \frac{3}{2}(1 - x^2)\mathbb{1}_{[0,1]}$.
- Choisir le paramètre de lissage par la méthode de la L^1-validation croisée[1] sur le nombre de voisins les plus proches.
- **Etape 1**. On divise nos données en deux ensembles :
- $\big(X_j, Y_j\big)_{j=1,\cdots,170}$ échantillon d'apprentissage,
- $(X_i, Y_i)_{i=171,\cdots,215}$ échantillon test.
- **Etape 2**. On calcule alors l'estimateur $\widehat{\theta}_{X_j}$, pour tout $j = 1, \cdots, 170$. en utilisant l'échantillon d'apprentissage.
- **Etape 3**. Pour chaque X_i dans l'échantillon test, on calcule : $i_* = Argmin_{j=1,\cdots,170} d\big(X_i, X_j\big)$.
- **Etape 4**. Pour tout $i = 171, \cdots, 215$ on pose $\widehat{Y}_i = \widehat{X}_{i_*}$.
- **Etape 5**. L'erreur utilisée pour la comparaison est la moyenne des erreurs absolues

1. $\ll Cross - validation \gg$ en anglais.

(MAE) :

$$\frac{1}{45}\sum_{171}^{215} \mid \widehat{Y}_i - \widehat{T}_{X_i} \mid,$$

où $\widehat{T}$ désigne l'estimateur utilisé : Classique ou régression robuste.

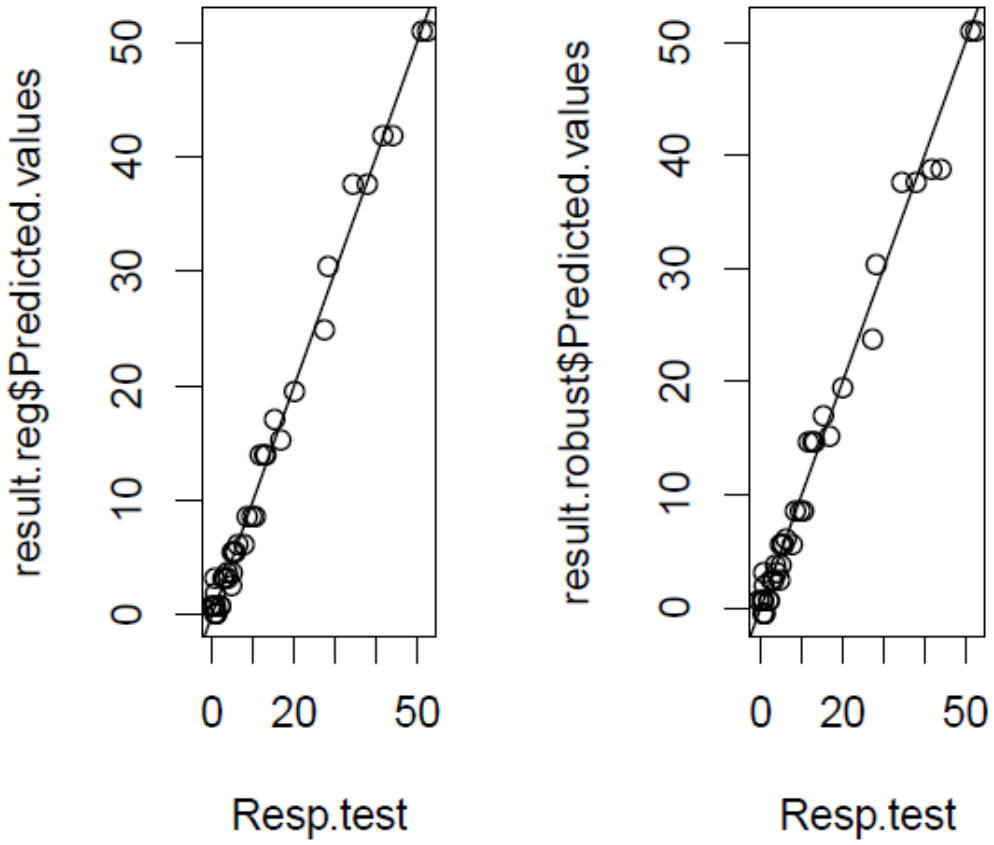

FIGURE 4.2 – Comparaison entre les deux méthodes en absence de données aberrantes

4.4 Résultat

La deuxième illustration est donnée la table suivante où on observe des valeurs abérrantes, la régression robuste donne de meilleurs résultats que la méthode

classique, Dans le sens que, même si la Moyenne des Erreurs Absolues (MEA), augmente avec le nombre de points perturbés, mais elle reste moins élevée pour la méthode robuste.

Nbre de pts perturbés C.M.	MEA. Méth. Robuste	MEA. Méth. Classique
une obs	2.388	4.955
12 obs perturbées	32.504	105.183
24 obs perturbées	49.538	143.794

Table 1. Comparaison des deux méthodes en présence de valeurs abérrantes.

Conclusion

Le travail présenté concerne un problème de régression, qui devient difficile à résoudre lorsque les données sont bruitées. Le cas de la méthode des moindres carrés en est le meilleur exemple : une seule donnée érronée suffit à biaiser l'estimation des paramètres.

L'utilisation de méthodes de régression robustes s'avère donc être une nécessité afin d'obtenir une estimation fiable des paramètres. Ces méthodes de régression tirent leur robustesse de différents outils qui aident à détecter les points aberrants, à diminuer leurs influences. Nous venons de parcourir les familles de méthodes de régression robustes, nous avons présenté les plus connues dans deux cas : lineaire et non paramétrique.

Bibliographie

[1] R. Andersen. (2008). Modern Methods For Robust Regression. Thousand Oaks : SAGE Publications.

[2] M. Attouch, A. Laksaci, and E. Ould-Saïd. (2007). Asymptotic distribution of robust estimator for functional nonparametric models. Prèpublication, LMPA No 314.

[3] M. Attouch, A. Laksaci, E. Ould-Saïd. (2008a). Asymptotic distribution of robust estimator for functional nonparametric models. Comm. Statist. Theory and Methods, in press.

[4] M. Attouch, A. Laksaci, E. Ould-Saïd. (2008b). Asymptotic distribution of robust estimator for functional dependent data. Preprint, LMPA No. 378, Janvier 2008. Univ. du Littoral cite d'Opale. submitted.

[5] N. Azzeddine, A. Laksaci, E. Ould-Saïd. (2008). On the robust nonparametric regression estimation for functional regressor. Statist. Probab. Lett., in press.

[6] V. Barnett et T. Lewis. Outliers in Statistical Data. Wiley, 1978.

[7] D. A. Belsley, E. Kuh, et R. E. Welsch. Regression Diagnostics. Wiley, 1980.

[8] G. Boente, R. Fraiman. (1989). Nonparametric regression estimation. J. Multivariate Anal., 29, 180-198.

[9] G. Boente, R. Fraiman. (1990). Asymptotic distribution of robust estimators for nonparametric models from mixing processes. Ann. Statist., 18, 891-906.

[10] M.M. Breuning, H.-P. Kriegel, R.T. Ng, et J. Sander. LOF : Identifying density-based local outliers. Dans ACM International Conference on Management of Data SIGMOD, pages 93-104, 2000.

[11] Z. Cai, E. Ould Saïd. (2003). Local M-estimator for nonparametric time series. Statist. and Probab. Lett., 65, 433-449.

[12] G. Collomb, W. Härdle. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction : Kernel regression estimation from dependent observations. Stoch. Proc. Appl., 23, 77-89.

[13] R.D. Cook. Detection of Influential Observations in Linear Regression. Technometrics, 19 : 15-18, 1977.

[14] R.D. Cook et S. Weisberg. Residuals and Influence in Regression. Chapman & Hall, 1982.

[15] N. R, Draper and K. Smith. (1998). Applied Regression Analysis. Third edition. New York : Wiley.

[16] F. Ferraty, Ph. Vieu, Modèle de régression pour variables aléatoires uni, multi et ∞-dimensionnées (2004). Cours de DEA.

[17] G.G. Judge, R.C. Hill, W.E. Griffiths, H. Lütkepohl, et T.-C. Lee. Introduction to the Theory and Practice of Econometrics. Wiley, 1988.

[18] D.G. Kleinbaum, L.L. Kupper, K.E. Muller, et A. Nizam. Applied regression analysis and other multivariable methods. Duxbury press, 1998.

[19] W. Härdle. (1984). Robust and nonlinear time series analysis. Lecture Notes in Statistics, Springer- Verlag, New York, 26.

[20] W. Härdle, A. B. Tsybakov. (1988). Robust nonparametric regression with simultaneous scale curve estimation. Ann. Statist., 16, No.1, 120-135.

[21] D. Hoaglin et R. Welsch. The hat matrix in regression and anova. The American Statistician, 1978.

[22] W. Hoeffding (1963). Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc., 58, 15-30.

[23] M. Loève. (1963). Probability Theory. Third Edition. Van Nostrand Princeton.

[24] P.J. Huber. (1964). Robust estimation of a location parameter, Ann. Math. Statist., 35, 73-101.

[25] N. Laïb, E. Ould-Saïd. (2000). A robust nonparametric estimation of the autoregression function under ergodic hypothesis. Canad. J. Statist., 28, 817-828.

[26] R. Robinson. (1984). Robust nonparametric autoregression. Lecture Notes in Statistics, 26, 247-255. Springer-Verlag, New York.

[27] P. J, Rousseeuw and A. M. Leroy (1987). Robust Regression and Outlier Detection. Hoboken : Wiley.