Étude des M-estimateurs et leurs versions pondérées pour des données clusterisées

Mohamed El Asri

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour l'obtention du grade de Docteur

SPÉCIALITÉ : Statistique

École Doctorale 356 «SAS»

Laboratoire de Mathématiques d'Avignon

Étude des M-estimateurs et leurs versions pondérées pour des données clusterisées

par **Mohammed EL ASRI**

Soutenue publiquement le jour mois année devant le jury composé de :

Mme	Michel BRONIATOWSKI	Professeur, LSTA, Paris	Rapporteur
M.	Anne RUIZ-GAZEN	Professeur, GREMAQ, Toulouse	Rapporteur
M.	Delphine BLANKE	Professeur, LMA, Avignon	Directrice de thèse
M.	Edith GABRIEL	Maître de Conférences, LMA, Avignon	Co-Directrice de thèse

Remerciments

Table des matières

Introduction **7**

1 Autour des M-estimateurs **13**

1.1 M-estimation dans le cas de variables aléatoires réelles 15

1.1.1 Définitions et notations 15

1.1.2 Exemples classiques de M-estimateurs 16

1.1.3 Unicité et convergence des M-estimateurs 18

1.1.4 Normalité asymptotique 20

1.1.5 Robustesse 24

1.2 Cas multidimensionnel 28

1.2.1 Définitions et notations 28

1.2.2 Comportement asymptotique 29

1.3 Le cas spécifique de la médiane 29

1.3.1 Les médianes multidimensionnelles 29

1.3.2 Le cas particulier de la médiane spatiale 30

2 M-estimateurs pour des données clusterisées **33**

2.1 Cadre d'étude 35

2.2 Convergence des M-estimateurs 36

2.3 Normalité asymptotique 40

2.4 Estimation de la variance 45

2.5 Cas particulier des Lp-estimateurs 48

2.6 Résultats numériques 49

3 M-estimateurs pondérés pour des données clusterisées **57**

3.1 Présentation des M-estimateurs pondérés 59

3.2 Propriétés asymptotiques 60

3.2.1 Convergence des estimateurs 60
3.2.2 Normalité asymptotique 63
3.3 Efficacité relative d'un M-estimateur pondéré 68
3.3.1 Estimation de la variance 68
3.3.2 Cadre des simulations 70
3.3.3 Résultats pour l'optimisation des poids 71
3.3.4 Résultats pour l'efficacité 71

4 Robustesse **79**
4.1 Points de rupture des M-estimateurs 81
4.2 Reformulation du point de rupture dans le cas pondéré 83
4.3 Comparaison des points de rupture entre versions pondérées et non pondérées 87
4.3.1 Remarques générales 87
4.3.2 Résultats numériques 88

Conclusions et perspectives **91**

Annexes **94**

A Outils probabilistes **95**
A.1 Notations et propriétés de o_p et O_p 97
A.2 Quelques propriétés de convergence stochastique 97
A.3 Théorèmes de convergence 98

B Programmes pour la M-estimation : calcul des estimateurs et des poids optimaux **99**
B.1 Calcul des M-estimateurs et M-estimateurs pondérés sous R 99
B.2 Optimisation des poids avec Matlab 106

Introduction

Ce travail de thèse s'inscrit dans le projet pluridisciplinaire ROLSES ("Robust and Optimal Locations for Sustainable Environment and Systems") de l'Agence Nationale de la Recherche (ANR) autour de la question de la localisation optimale des centres et leurs propriétés. Ce projet regroupe une équipe de géographes (UMR Espace) et des mathématiciens en optimisation et en statistique (Laboratoire de Mathématiques d'Avignon) de l'Université d'Avignon et des Pays de Vaucluse. Un centre est défini comme un point autour duquel se distribuent des phénomènes dans l'espace (centre de secours, hôpital, déchèterie,...) (Brunet et al., 1993). L'objectif du projet ROLSES est d'étudier les propriétés d'un centre suivant la méthode employée pour sa détermination. D'un point de vue statistique, le centre en question est un estimateur de localisation d'un ensemble de données. Le but est alors de présenter une sélection d'estimateurs adaptés aux données d'application ou aux données simulées pour une problématique.

La nature des données et le contexte de l'étude aident au choix de méthodes d'estimation d'un centre. Dans le contexte géographique, un centre, ou paramètre de centralité, répond souvent à un critère d'optimisation. En effet, selon Lévy and Lussault (2003, p.140), "la centralité d'un lieu ne prend véritablement sens que lorsqu'on associe à sa position dans l'espace physique la mesure du rayonnement des potentiels et des fonctions localisées dans ce même lieu et quand on considère les gradients et les "champs" qu'ils produisent et disposent dans l'espace". En géographie, les méthodes statistiques employées sont communément : le centre moyen, qui minimise un critère de distance au carré des données par rapport au centre et qui favorise les points extrêmes ; le centre médian, qui optimise équitablement les distances par rapport au centre ; ou encore les Lp-médianes, qui sont paramétrables selon un paramètre qui favorise plus ou moins un type de données pour le choix du centre dans l'optimisation (Lévy and Lussault, 2003).

En statistique, cela revient à étudier les estimateurs issus de l'optimisation d'un critère ou d'une fonction objective. Nous parlons alors de la M-estimation. Un exemple des M-estimateurs est celui du maximum de vraisemblance. Le "M" de "M-estimateur" provient de ce cas précis. Cette classe des M-estimateurs est introduite par Huber en 1964 pour l'étude de la robustesse des estimateurs (Huber, 1964). D'une manière générale, on définit un M-estimateur comme le résultat de la minimisation (ou maximisation) d'une fonction qui dépend des données et des paramètres du modèle ou de la loi utilisée (Hampel et al., 1986; Arcones, 1998; Van der Vaart, 2000).

Les données spatiales peuvent être issues d'un seul domaine géographique homo-

gène, ou de plusieurs domaines géographiques homogènes. Un domaine géographique peut être une zone de chalandise, un type d'occupation du sol (urbain, rural, industriel etc.), une zone météorologique, zone biologique etc. Nous pouvons alors regrouper les données par domaine. Pour généraliser, nous parlons ici de données en groupe ou des données clusterisées. Ces données clusterisées peuvent avoir une structure de dépendance ou de corrélation. Pour autant, les estimateurs classiques ne sont que trop peu étudiés dans ce cadre.

Pour approfondir la compréhension de cette large classe d'estimateurs dans le cas des variables aléatoires clusterisées, dans ce travail de thèse, je propose une étude générale des M-estimateurs dans ce cadre : convergence, normalité asymptotique et robustesse. J'étends l'analyse en étudiant l'influence de la pondération de ces estimateurs sur leurs propriétés asymptotiques, ainsi que l'apport de celle-ci sur la précision de l'estimation.

La thèse est composée de quatre chapitres. Dans le premier, je parcours quelques notations et définitions autours des M-estimateurs. Je présente ensuite quelques résultats issus de la littérature sur la convergence et la normalité asymptotique des M-estimateurs, en passant du cas réel au cas multivarié pour des données indépendantes et identiquement distribuées (i.i.d.). J'expose également un aperçu de la notion de robustesse des M-estimateurs.
Dans le deuxième chapitre, je propose une généralisation des résultats de convergence et de normalité asymptotique pour les M-estimateurs dans un cadre de variables aléatoires multivariées et clusterisées avec une structure de dépendance. Je donne quelques exemples et résultats de simulations pour illustrer ce comportement asymptotique.
Le troisième chapitre fournit les résultats de convergence et de normalité asymptotique pour la variante des M-estimateurs, la classe des M-estimateurs pondérés. J'annonce également la pertinence de l'introduction des poids dans la M-estimation, en étudiant leur efficacité relativement aux estimateurs non pondérés. Je conclus le chapitre par une étude de simulations.
Dans le dernier chapitre, je développe l'expression générale d'une mesure globale de la robustesse qui est le point de rupture pour cette classe, ainsi qu'une réécriture explicite de cette mesure. Après avoir établi l'influence des poids sur l'efficacité dans le chapitre précédent, je réponds dans ce chapitre à la question de l'évaluation de l'influence des poids sur la robustesse des M-estimateurs pondérés.
La thèse se conclut par des perspectives de recherche qu'il serait intéressant de dévelop-

per par la suite. Dans les annexes, je rappelle des résultats de probabilité utilisés dans la thèse (annotés par un astérisque dans le corps de texte), et donne les programmes sous R et Matlab ayant servi dans les simulations.

Chapitre 1

Autour des M-estimateurs

Sommaire

1.1 M-estimation dans le cas de variables aléatoires réelles 15

1.1.1 Définitions et notations 15

1.1.2 Exemples classiques de M-estimateurs 16

1.1.3 Unicité et convergence des M-estimateurs 18

1.1.4 Normalité asymptotique 20

1.1.5 Robustesse 24

1.2 Cas multidimensionnel 28

1.2.1 Définitions et notations 28

1.2.2 Comportement asymptotique 29

1.3 Le cas spécifique de la médiane 29

1.3.1 Les médianes multidimensionnelles 29

1.3.2 Le cas particulier de la médiane spatiale 30

Résumé

Ce chapitre propose un aperçu de la littérature autour de la M-estimation. Nous donnons donc quelques notations et résultats sur la convergence et la normalité asymptotique. Nous introduisons également la notion de robustesse pour un estimateur.

La classe des M-estimateurs engendre des estimateurs classiques, tels que l'estimateur du maximum de vraisemblance, la moyenne empirique et la médiane spatiale. Huber (1964) introduit les M-estimateurs dans le cadre de l'étude des estimateurs robustes. Parmi la littérature dédiée à ces estimateurs, on trouve en particulier les ouvrages Huber (1981) et Hampel et al. (1986) sur le comportement asymptotique et la robustesse via le point de rupture et la fonction d'influence. On peut se reporter à Ruiz-Gazen (2012) pour une synthèse détaillée autour de ces notions. Plus récemment, des résultats sur la convergence et la normalité asymptotique ont été établis par Van der Vaart (2000) dans le cadre multidimensionnel.

Dans ce chapitre, nous introduisons des notations et des définitions autour des M-estimateurs, nous donnons ensuite quelques propriétés de convergence et de normalité asymptotiques. Nous poursuivons par l'analyse de leur robustesse à travers le point de rupture et la fonction d'influence. Enfin, nous concluons avec une généralisation au cas multidimensionnel.

1.1 M-estimation dans le cas de variables aléatoires réelles

1.1.1 Définitions et notations

Dans le cas de variables aléatoires réelles, un M-estimateur est défini, d'une manière générale, comme étant une fonctionnelle définie en une loi de probabilité P, notée $\theta(P)$ et qui minimise la fonction objective en a, de la forme $\int \rho(x,a)dP(x)$, où l'intégrale est calculée sur $\mathbb{R}$. On note :

$$\theta(P) = \underset{a\in\Theta}{\operatorname{argmin}} \int \rho(x,a)dP(x), \tag{1.1}$$

où ρ est une fonction mesurable en x ; si cette fonction est suffisamment régulière, alors $\theta(P)$ est la solution de l'équation

$$\int \psi(x,\theta(P))dP(x) = 0 \tag{1.2}$$

avec $\psi(x,\theta(P)) = \frac{\partial\rho(x,a)}{\partial a}_{/a=\theta(P)}$.
L'estimateur paramétrique est un cas particulier où $P = P(.;\theta) := P_\theta$; et θ appartient à un ensemble de paramètres Θ. Dans ce cas, $\theta = \theta(P)$ pour tout $\theta \in \Theta$ et les expressions

(1.1) et (1.2) s'écrivent respectivement :

$$\theta = \underset{a\in\Theta}{\operatorname{argmin}} \int \rho(x,a)dP_\theta(x), \tag{1.3}$$

$$\int \psi(x,\theta)dP_\theta(x) = 0. \tag{1.4}$$

Soit $X_1,\dots,X_n$ un échantillon de variables aléatoires réelles de même loi P et soit P_n la mesure empirique associée à cet échantillon. On définit la version empirique d'un M-estimateur par $\theta(P_n)$. Les versions (1.1) et (1.2) s'écrivent dans ce cas :

$$\theta(P_n) = \underset{a\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(X_i,a), \tag{1.5}$$

$$\sum_{i=1}^{n} \psi(X_i,\theta(P_n)) = 0. \tag{1.6}$$

Dans le cas de l'estimation paramétrique, on peut étudier l'estimation d'un paramètre de localisation. La notion de localisation sous-entend qu'on a une loi concentrée autour d'un paramètre de centralité qui peut être sa moyenne ou sa médiane, ou encore son centre de symétrie. Formellement $\rho(x,a) = \rho(x-a)$, et le M-estimateur $\hat{\theta}_n$ vérifie l'une des deux équations :

$$\hat{\theta}_n = \underset{a\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(X_i - a), \tag{1.7}$$

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}_n) = 0. \tag{1.8}$$

Des exemples typiques de choix pour les fonctions ρ et ψ sont donnés dans le paragraphe suivant.

1.1.2 Exemples classiques de M-estimateurs

Maximum de vraisemblance

Soit $X_1,\dots,X_n$ une suite de variables aléatoires réelles i.i.d issues d'une loi de densité $f(.,\theta)$. L'estimateur du maximum de vraisemblance $\hat{\theta}_n$ maximise en a la vraisemblance $\prod_{i=1}^{n} f(X_i,a)$ ou la log-vraisemblance $\sum_{i=1}^{n} \ln f(X_i,a)$. Il est équivalent d'écrire que :

$$\hat{\theta}_n = \underset{a\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} -\ln f(X_i,a). \tag{1.9}$$

L'estimateur du maximum de vraisemblance est un M-estimateur. Si pour chaque x fixé, la densité $f(x,a)$ est différentiable sur un voisinage de θ et $f(x,a) > 0$, alors $\hat{\theta}_n$ résout également l'équation de type (1.2), avec $\psi(x,a) = \frac{\dot{f}(x,a)}{f(x,a)}$ et $\dot{f}(x,a) = \left.\frac{\partial f(x,t)}{\partial t}\right|_{t=a}$. Cette définition n'est pas vérifiée si on prend l'exemple de la loi uniforme où la fonction $\ln f(x,\theta) = \ln 1_{[0,\theta]}(x) - \ln\theta$ qui n'est pas continue sur tout voisinage de θ. La définition (1.1) est donc plus générale que celle donnée par (1.2), même si cette dernière est plus pratique pour les calculs.

Moyenne et médiane

Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires issues d'une loi paramétrique avec un paramètre de localisation θ. La moyenne empirique $\overline{X}_n$ et la médiane Med_n sont deux exemples classiques d'un M-estimateur de localisation. Nous les définissons respectivement avec les deux formulations :

$$\overline{X}_n = \underset{a\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} |X_i - a|^2 \,;\; \sum_{i=1}^{n} (X_i - \overline{X}_n) = 0 \tag{1.10}$$

$$Med_n = \underset{a\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} |X_i - a| \,;\; \sum_{i=1}^{n} \operatorname{sign}(X_i - Med_n) = 0 \tag{1.11}$$

où $|.|$ est la valeur absolue et

$$\operatorname{sign}(x) = \begin{cases} -1 & \text{si } x < 0, \\ 0 & \text{si } x = 0, \\ 1 & \text{si } x > 0. \end{cases}$$

Estimateur de Huber

Huber (1964) propose un compromis entre la moyenne empirique et la médiane qui s'écrit sous la forme (1.1) avec une fonction ρ définie par

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq k \\ k\,|x| - \frac{1}{2}k^2, & |x| > k \end{cases}$$

et sous la forme (1.2) avec la fonction ψ :

$$\psi(x) = \begin{cases} -k, & x < -k \\ x, & |x| \leq k \\ k, & x > k \end{cases}$$

Cet estimateur a le comportement de la médiane sur les grandes valeurs qui peuvent être des valeurs aberrantes ; ce qui est un avantage important sachant que la médiane garde de bonnes propriétés en présence de ce type de valeurs. Nous détaillons ces propriétés dans la suite de ce chapitre, en développant la notion de robustesse.

1.1.3 Unicité et convergence des M-estimateurs

Soit P_n la distribution empirique associée à l'échantillon $X_1, \ldots, X_n$ de variables aléatoires réelles de même loi P_θ. Nous considérons la définition (1.2) pour un M-estimateur, et nous posons $\lambda_P(a) = \int \psi(x,a) dP_\theta(x)$. On suppose que θ est une racine de l'équation $\lambda_P(a) = 0$, et que θ_n est une racine pour la version empirique :

$$\lambda_{P_n}(a) = \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, a) = 0.$$

Si, par exemple, θ_n est la moyenne empirique, alors sous les conditions de la loi faible des grands nombres, θ_n converge en probabilité vers θ. Il est alors naturel de chercher des conditions suffisantes assurant la convergence de θ_n vers θ. Dans cette section, nous utilisons la notation o_p (petit o en probabilité) dont la définition et les propriétés sont rappelées en annexe.

Lemme 1.1.1 *(Serfling, 1980, p.249) Soit θ une racine isolée de l'équation $\lambda_P(a) = 0$ et supposons que la fonction $a \mapsto \psi(x,a)$ soit monotone en a. Alors θ est unique et toute solution θ_n de l'équation empirique $\lambda_{P_n}(a) = 0$ converge presque sûrement vers θ. De plus, si $\psi(x,a)$ est continue en a sur un voisinage de θ alors θ_n existe.*

Lemme 1.1.2 *(Serfling, 1980, p.249) Soit θ une racine isolée de l'équation $\lambda_P(a) = 0$. Supposons que $a \mapsto \psi(x,a)$ est continue en a et bornée. Alors toute solution θ_n de l'équation empirique $\lambda_{P_n}(a) = 0$ converge presque sûrement vers θ.*

Ces résultats nous indiquent que, pour des lois adaptées, la continuité en a de la fonction ψ entraîne l'existence et la convergence de l'estimateur dès que ψ est soit monotone,

soit bornée en a.
Prenons maintenant le cas des M-estimateurs définis plus généralement par la relation (1.1). Nous définissons les deux fonctions objectives $M(a)$ et sa version empirique $M_n(a)$ par

$$M(a) = \int \rho(x, a) dP(x)$$

et

$$M_n(a) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, a)$$

ainsi que θ et $\hat{\theta}_n$ les valeurs qui les minimisent respectivement. La convergence asymptotique de $\hat{\theta}_n$ dépend fortement du comportement asymptotique de $M_n(a)$. Sous les conditions de la loi faible des grands nombres on a :

$$M_n(a) \xrightarrow{P} M(a), \quad \text{pour tout } a.$$

Sous des conditions suffisantes de régularité, on peut espérer que la valeur $\hat{\theta}_n$ minimisant M_n converge vers θ l'argument minimum de M. Une approche basée sur la convergence uniforme des fonctions objectives permet d'obtenir cette convergence.

Théorème 1.1.3 *(Van der Vaart, 2000, p.45)*
Soient M et sa version empirique M_n telles que pour tout $\epsilon > 0$

$$\sup_{a \in \Theta} |M_n(a) - M(a)| \xrightarrow{P} 0, \tag{1.12}$$

$$\sup_{a \,:\, |\theta - a| \geq \epsilon} M(a) > M(\theta). \tag{1.13}$$

Alors tout $\hat{\theta}_n$, minimisant $M_n(a)$ en a, converge en probabilité vers θ.

Prenons la définition (1.2) d'un M-estimateur ainsi que les notations que nous avons introduites dans cette section. Le théorème suivant est l'équivalent du théorème 1.1.3.

Théorème 1.1.4 *(Van der Vaart, 2000, p.46)*
Soit P une loi de probabilité, et P_n sa version empirique. Considérons les deux fonctions en a,

λ_P et sa version empirique λ_{P_n} telles que pour tout $\epsilon > 0$,

$$\sup_{a\in\Theta} |\lambda_{P_n}(a) - \lambda_P(a)| \xrightarrow{P} 0, \tag{1.14}$$

$$\inf_{a\,:\,|\theta-a|\geq\epsilon} |\lambda_P(a)| > 0 = \lambda_P(\theta). \tag{1.15}$$

Alors tout $\hat{\theta}_n$, vérifiant $\lambda_{P_n}(\hat{\theta}_n) = 0$, converge en probabilité vers θ.

Notons que si la convergence uniforme est vérifiée presque sûrement dans les deux théorèmes précédents alors il y a également convergence presque sûre de l'estimateur.

Dans le lemme qui suit, θ est considéré comme le point pour lequel la fonction $a \mapsto \lambda_P(a)$ change de signe.

Lemme 1.1.5 *(Serfling, 1980, p.249) On considère $\lambda_{P_n}(a)$ continue en a et admettant un unique zero $\hat{\theta}_n$, ou bien, croissante en a avec $\lambda_{P_n}(\hat{\theta}_n) = o_p(1)$. Soit θ le point vérifiant pour tout $\epsilon > 0$, $\lambda_P(\theta+\epsilon) < 0 < \lambda_P(\theta-\epsilon)$. Alors $\hat{\theta}_n \xrightarrow{p} \theta$.*

Exemple de la médiane : Soit Med_n, l'estimateur défini par la relation (1.11). On a $\lambda_P(a) = P_\theta(X < a) - P_\theta(X > a)$, pour tout a. Nous vérifions dans cet exemple que, pour des lois adaptées P_θ, Med_n converge vers la valeur θ telle que $P_\theta(X < \theta) = P_\theta(X > \theta)$ (la valeur médiane de la distribution). Nous pouvons appliquer le théorème 1.1.4 pour le montrer mais il ne sera pas aisé de montrer la convergence uniforme de λ_{P_n} vers λ_P (Van der Vaart, 2000). C'est pour cela que l'on applique le lemme (1.1.5) : la fonction $\lambda_{P_n}(a)$ est croissante en a. Si la valeur médiane θ vérifie $P_\theta(X < \theta-\epsilon) < \frac{1}{2} < P_\theta(X > \theta+\epsilon)$ (θ est unique) alors $\lambda_P(\theta-\epsilon) < 0 < \lambda_P(\theta+\epsilon)$. On a donc $Med_n \xrightarrow{P} \theta$.

1.1.4 Normalité asymptotique

Dans la section précédente, nous avons rappelé les propriétés de convergence presque sûre et en probabilité des M-estimateurs. Une autre question se pose alors, celle de la vitesse de convergence de ces estimateurs. Le théorème central limite nous donne une normalisation en $\sqrt{n}$. Dans cette partie, nous allons donner des théorèmes permettant d'établir cette normalité asymptotique. Les résultats de cette section ont été proposés principalement par Huber (1964) et repris par d'autres travaux comme Serfling (1980) et Van der Vaart (2000).
Supposons que θ est un zéro isolé de la fonction en a $\lambda_P(a)$, et qu'il existe une solution $\hat{\theta}_n$ pour l'équation empirique $\lambda_{P_n}(a)$. On a alors le théorème suivant :

Théorème 1.1.6 *(Serfling, 1980, p.251) Soit la fonction $\psi(x,a)$ monotone en a. Si $\lambda_P(a)$ est dérivable en θ telle que $\lambda'_P(\theta) \neq 0$ et si $\int \psi^2(x,a)dP(x)$ est continue en θ et finie sur un voisinage de θ, alors on a :*

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{Loi}{\to} \mathcal{N}\left(0, \frac{\int \psi^2(x,\theta)dP(x)}{(\lambda'_P(\theta))^2}\right). \tag{1.16}$$

Exemple : Soit l'estimateur de Huber dont la fonction ψ s'écrit :

$$\psi(x) = \begin{cases} -k, & x < -k \\ x, & |x| \leq k \\ k, & x > k. \end{cases}$$

Il est clair que $\psi(x-a)$ est décroissante en a. Sous les hypothèses du théorème, $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi vers une loi normale centrée, de variance :

$$\frac{\int_{\theta-k}^{\theta+k} x^2 dP(x) + k^2 \left(P\left(]-\infty, \theta-k[\right) + P\left(]-\theta+k, \infty[\right)\right)}{\left(P\left([\theta-k, \theta+k]\right)\right)^2}$$

Pour l'estimateur du maximum de vraisemblance, la fonction $\psi(x,a) = \frac{\dot{f}(x,a)}{f(x,a)}$ qui lui est associée n'est pas forcément monotone. Il faut donc d'autres méthodes pour établir cette normalité. Prenons par exemple le développement limité suivant :

$$\psi(x,\hat{\theta}_n) - \psi(x,\theta) = (\hat{\theta}_n - \theta)\frac{\partial\psi(x,a)}{\partial a}_{/a=\tilde{\theta}_n} \tag{1.17}$$

ce qui implique

$$\sum_{i=1}^n \psi(X_i,\hat{\theta}_n) - \sum_{i=1}^n \psi(X_i,\theta) = (\hat{\theta}_n - \theta)\sum_{i=1}^n \frac{\partial\psi(X_i,a)}{\partial a}_{/a=\tilde{\theta}_n} \tag{1.18}$$

où $\tilde{\theta}_n$ est un point intermédiaire entre θ et $\hat{\theta}_n$. Par définition, $\hat{\theta}_n$ est solution de l'équation empirique $\lambda_{P_n}(a) = 0$, donc $\sum_{i=1}^n \psi(X_i,\hat{\theta}_n) = 0$. L'équation (1.17) implique :

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{-\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i,\theta)}{\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(X_i,a)}{\partial a}_{/a=\tilde{\theta}_n}}. \tag{1.19}$$

Si $\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(X_i,a)}{\partial a}_{/a=\tilde{\theta}_n}$ converge vers une limite non nulle et finie, et si on vérifie les conditions du théorème central limite pour $\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i,\theta)$, alors, on assure la conver-

gence en loi de $\sqrt{n}(\hat{\theta}_n - \theta)$. Le théorème qui suit donne des conditions suffisantes pour avoir la normalité asymptotique de l'estimateur.

Théorème 1.1.7 *(Van der Vaart, 2000, p.52) Soit $\frac{\partial\psi(x,a)}{\partial a}$ continue en θ uniformément en x, telle que : $0 < \left|\int \frac{\partial\psi(x,a)}{\partial a}_{/a=\theta} dP(x)\right| < \infty$ ainsi que, $\int \psi^2(x,\theta)dP(x) < \infty$.*
Alors si $\hat{\theta}_n \xrightarrow{P} \theta$, on obtient :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{Loi} \mathcal{N}\left(0, \frac{\int \psi^2(x,\theta)dP(x)}{(\int \frac{\partial\psi(x,a)}{\partial a}_{/a=\theta} dP(x))^2}\right). \tag{1.20}$$

Continuons sur l'exemple de l'estimateur du maximum de vraisemblance où $\psi(x,a) = \frac{\dot{f}(x,a)}{f(x,a)}$. Sous les hypothèses du théorème, on a successivement :

$$\begin{aligned}
\frac{\partial\psi(x,a)}{\partial a} &= \frac{\ddot{f}(x,a)}{f(x,a)} - \left(\frac{\ddot{f}(x,a)}{f(x,a)}\right)^2, \\
\int \frac{\partial\psi(x,a)}{\partial a}_{/a=\theta} dP(x) &= \int \ddot{f}(x,\theta)dx - \int \left(\frac{\ddot{f}(x,\theta)}{f(x,\theta)}\right)^2 dx \\
&= -\int \left(\frac{\ddot{f}(x,a)}{f(x,a)}\right)^2 dx,
\end{aligned}$$

pour toutes les densités pour lesquelles on peut intervertir signe intégral et dérivation. Puisque l'intégrale d'une densité $f(.,a)$ est égal à 1, les dérivées première et seconde de cette intégrale en tout point a sont égales à 0. Par conséquent, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{Loi} \mathcal{N}\left(0, I_\theta^{-1}\right)$, où $I_\theta := \int \left(\frac{\dot{f}(x,\theta)}{f(x,\theta)}\right)^2 f(x,\theta)dx$ est l'information de Fisher.

La fonction objective de l'équation (1.1) n'est pas toujours dérivable. Prenons l'exemple de la fonction objective de la médiane. Elle n'est pas dérivable en 0, mais elle est presque partout dérivable. Le théorème suivant traite ce type d'estimateur dont la fonction objective est presque partout dérivable. Soit θ la valeur qui minimise $M(a) = \int \rho(x,a)dP(x)$ et supposons que $\hat{\theta}_n$ la valeur qui minimise la version empirique de $M(a)$, converge en probabilité vers θ. On suppose également que Θ, l'espace des paramètres, est un ouvert de $\mathbb{R}$.

Théorème 1.1.8 *(Van der Vaart, 2000, p.53) Soit $\rho(x,a)$ dérivable en θ x-presque partout, notons cette dérivée $\dot{\rho}(x,\theta)$, telle que, il existe une fonction $K(x)$ en x, avec $\int K^2(x)dP(x) < \infty$*

vérifiant : pour tout a_1 et a_2 dans un voisinage de θ,

$$|\rho(x,a_1) - \rho(x,a_2)| \leq K(x)\,|a_1 - a_2|\,. \tag{1.21}$$

Notons que si on a le développement de Taylor suivant pour la fonction $M(a)$, au voisinage de θ

$$M(a) = M(\theta) + \frac{1}{2}\ddot{M}(\theta)(a-\theta)^2 + o((a-\theta)^2),$$

avec une dérivée seconde $\ddot{M}(\theta)$ en θ non nulle, alors

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{Loi}{\to} \mathcal{N}\left(0, \frac{\int \dot{\rho}^2(x,\theta)dP(x)}{\ddot{M}^2(\theta)}\right). \tag{1.22}$$

Clairement, si ρ est dérivable alors $\psi(x,a) = \dot{\rho}(x,a)$, et $\ddot{M}(\theta) = \int \frac{\partial\psi(x,a)}{\partial a}_{/a=\theta} dP(x)$. On retrouve donc la normalité asymptotique établie au théorème 1.1.6.

Soit $f(x,\theta) := f(x)$ une densité de probabilité avec θ un paramètre de la médiane. On note $F_\theta := F$ la fonction de répartition associée. La médiane empirique Med_n minimise en a la fonction objective $\sum_{i=1}^n |X_i - a|$ et θ minimise $M(a) = \int \rho(x,a)dF(x)$, avec $\rho(x,a) = |x-a|$. On suppose que $\theta \geq 0$. Pour assurer l'existence de $M(a)$, on prend Θ borné et $\rho(x,a) = |x-a| - |x|$ (nous avons $|\rho(x,a)| \leq |a|$, et donc $M(a) < \infty$). Il est clair que ρ vérifie (1.21), avec $k(x) \equiv 1$ et elle est dérivable sauf en $a = \theta$.
Calculons $M(a)$ pour $a \geq 0$:

$$\begin{aligned}
M(a) &= \int_{-\infty}^{\infty} |x-a| - |x|\,dF(x) \\
&= \int_{-\infty}^{0} |x-a| - |x|\,dF(x) + \int_0^a |x-a| - |x|\,dF(x) + \int_a^{\infty} |x-a| - |x|\,dF(x) \\
&= aF(0) + \int_0^a a - 2x dF(x) - a\int_a^{\infty} dF(x) \\
&= aF(0) + [F(x)(a-2x)]_0^a + 2\int_0^a F(x)dx - a(1-F(a)) \\
&= 2\int_0^a F(x)dx - a.
\end{aligned}$$

La dérivabilité de F sur un voisinage de θ telle que $f(\theta) > 0$ implique que $M(a)$ est deux fois dérivable sur le même voisinage. On obtient ainsi le développement de Taylor : $M(a) = M(\theta) + f(\theta)(a-\theta)^2 + o((a-\theta)^2)$ et donc

$$\sqrt{n}(Med_n - \theta) \overset{Loi}{\to} \mathcal{N}\left(0, \frac{1}{4f^2(\theta)}\right).$$

1.1.5 Robustesse

Soit P une loi de probabilité sur $\mathbb{R}$ et soit $\theta(P)$ le M-estimateur calculé sous P. En prenant en compte l'aspect fonctionnel de notre estimateur, il est naturel d'essayer de comprendre le comportement de l'estimateur en fonction de P, par exemple, étudier sa continuité en P, évaluer ses limites, ou encore, ses propriétés après une perturbation de la loi de départ P. Cette étude de comportement entre dans le cadre de l'étude de robustesse de l'estimateur. Un estimateur est dit robuste s'il représente de bonnes propriétés après une petite perturbation du modèle de départ, et s'il ne prend pas des valeurs aberrantes si le modèle de départ est changé ou fortement perturbé. Dans la littérature, des auteurs comme Hampel et Huber ont introduit des grandeurs qui permettent de mesurer cette robustesse : la fonction d'influence et le point de rupture.

Fonction d'influence

Une première mesure de la robustesse est une mesure locale qui calcule l'influence d'un point aberrant sur l'estimateur. Cette mesure est la fonction d'influence (IF) introduite par Hampel et al. (1986); Hampel (1974) sous le nom de courbe d'influence. Nous gardons ici le nom le plus utilisé et qui est adopté dans le cadre vectoriel.
La IF est le comportement infinitésimal de la vraie valeur $\theta(P)$. On suppose que le domaine de définition de la fonction θ est un sous-ensemble convexe de l'ensemble des mesures signées. On prend la version où notre estimateur θ est dérivable au sens de Gâteaux. Elle est proposée par Huber (1977) et nous la trouvons également dans Hampel et al. (1986).

Définition 1.1.1
La fonction d'influence IF de θ en P est donnée par

$$IF(x,\theta,P) = \lim_{t\searrow 0} \frac{\theta((1-t)P + t\delta_x) - \theta(P)}{t}, \tag{1.23}$$

où x est un point dans $\mathbb{R}$ et δ_x la mesure de Dirac en ce point.

Huber (1977) montre que l'existence de cette fonction d'influence est une condition plus faible que la dérivabilité au sens de Gâteaux de l'estimateur.
Cette définition montre l'importance de la fonction d'influence. Elle décrit l'effet d'une contamination infinitésimale en un point x sur l'estimateur, et mesure le biais asymp-

totique après une contamination dans les observations.
On peut également faire le lien entre la fonction d'influence et la variance de l'estimateur. Considérons pour cela un échantillon $X_1, \ldots, X_n$ de loi P. La mesure empirique P_n converge vers P (théorème de Glivenko-Cantelli). Nous utilisons le développement de Von Mises de θ en P_n pour écrire (Huber, 1981, p.13-15) :

$$\theta(P_n) = \theta(P) + \int IF(x, \theta, P) dF_n(x) + R_n,$$

où R_n est un reste qui tend vers zéro (en probabilité ou presque sûrement). Nous développons l'expression sous le signe intégral :

$$\sqrt{n}\left(\hat{\theta}_n - \theta(P)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(X_i, \theta, P) + \sqrt{n} R_n$$

avec $\hat{\theta}_n = \theta(P_n)$. Nous utilisons un des théorèmes de normalité asymptotique (comme par exemple le théorème 1.1.7), pour établir que $\sqrt{n}\left(\hat{\theta}_n - \theta(P)\right)$ converge vers une loi normale $\mathcal{N}(0, V)$. Supposons que le terme du reste $\sqrt{n} R_n$ converge vers 0 quand n tend vers ∞ (en probabilité ou presque sûrement), alors la variance asymptotique est donnée par :

$$V = \int IF^2(x, \theta, P) dP(x). \tag{1.24}$$

Cette relation entre la variance de l'estimateur et sa fonction d'influence est très importante. Nous allons voir l'expression explicite de la fonction d'influence d'un M-estimateur et dans quelle mesure on peut exploiter cette relation.
Une mesure de la robustesse induite par la fonction d'influence est la sensibilité de l'erreur maximale (*gross-error sensitivity*) définie par

$$\gamma^* = \sup_x |IF(x, \theta, F)|. \tag{1.25}$$

Elle mesure la borne maximale de la fonction d'influence et donc le biais asymptotique maximal. Un estimateur a un bon comportement si la valeur de γ^* est finie. On parle dans ce cas de la B-robustesse (B pour biais) (Rousseeuw, 1981).
En prenant l'exemple d'un M-estimateur, défini par la relation (1.2), sous la loi $P_{t,x} = (1-t)P + t\delta_x$ et en dérivant par rapport à la variable t (sous les conditions suffisantes de continuité, dérivabilité et permutation de la dérivée avec l'intégration), nous pouvons

écrire :

$$0 = \int \frac{\partial \psi(y,a)}{\partial a} /_{\theta(P_{t,x})} \frac{\partial \theta(P_{t,x})}{\partial t} dP_{t,y} + \int \psi(y, \theta(P_{t,x})) d(\delta_x - P).$$

Avec le passage à la limite quand $t \to 0$ et la relation (1.23) nous obtenons :

$$0 = IF(x,\theta,P) \int \frac{\partial \psi(y,a)}{\partial a} /_{\theta(P)} dP(y) + \psi(x,\theta(P)) - \underbrace{\int \psi(y,\theta(P)) dP(y)}_{=0}$$

d'où

$$IF(x,\theta,P) = \frac{\psi(x,\theta(P))}{-\int \frac{\partial \psi(y,a)}{\partial a} /_{\theta(P)} dP(y)}, \tag{1.26}$$

sous l'hypothèse que $\int \frac{\partial \psi(y,a)}{\partial a} /_{\theta(P)} dP(y)$ diffère de 0. Sous les conditions précédentes, le M-estimateur est donc B-robuste si la fonction $x \mapsto \psi(x,\theta(P))$ est bornée. On en déduit que la variance de l'estimateur s'écrit :

$$V = \frac{\int \psi^2(x,\theta(P)) dP(x)}{[\int \frac{\partial \psi(y,a)}{\partial a} /_{\theta(P)} dP(y)]^2}. \tag{1.27}$$

La fonction d'influence ainsi introduite est une notion théorique ou asymptotique. Nous pouvons donc penser à utiliser sa version empirique pour des questions pratiques, en particulier dans des simulations. Si nous remplaçons la loi P par sa version empirique P_{n-1} issue d'un échantillon de taille $n-1$, et prenant $t = 1/n$ on obtient la version empirique suivante, souvent notée dans la littérature par SC_n (*sensitivity curve*) :

$$SC_n(x) = \frac{\left[\theta\left((1-\frac{1}{n})P_n + \frac{1}{n}\delta_x\right) - \theta(P_{n-1})\right]}{\frac{1}{n}}. \tag{1.28}$$

Cette version (1.28) illustre l'influence d'une observation en plus sur notre estimateur par rapport à l'échantillon de départ. Il existe une autre version de la fonction d'influence qui consiste à remplacer une observation de l'échantillon de départ par une valeur arbitraire x. La plus utilisée et la plus pratique est celle que nous avons définie

dans (1.28). L'étude de ces versions empiriques a été effectuée, en particulier, par Tukey and Wilk (1970); Andrews et al. (1972); Field and Hampel (1982).

Point de rupture

Le point de rupture d'un estimateur est une deuxième mesure de la robustesse bien plus utilisée que la fonction d'influence grâce à sa facilité d'interprétation. Il est introduit tout comme la fonction d'influence par Hampel (1968, 1971). Dans cette section, nous ne traitons que sa version empirique introduite par Donoho (1982) et Donoho and Huber (1983) pour un échantillon fini d'observations. D'autres auteurs ont élargi l'étude du point de rupture par exemple pour des estimateurs de localisation et d'échelle et pour des modèles de régression. Citons en particulier les travaux de Chen and Tyler (2004); Davies and Gather (1993); Davies et al. (2005); Huber (1997); Chao (1986).

Définition 1.1.2 *Soit X un ensemble de n variables aléatoires réelles $\{X_1, \ldots, X_n\}$ et soit Y_k un ensemble d'observations contenant $n-k$ observations en commun avec X et k valeurs arbitraires. Soient P_n et $Q_{n,k}$ deux distributions empiriques associées respectivement à X et Y_k. Ainsi un point de rupture de remplacement pour un estimateur $\hat{\theta}_n$ s'écrit :*

$$\epsilon_n^*(\hat{\theta}_n, X) = \inf_{k \in \{1,\ldots,n\}} \left\{ \frac{k}{n}, \textit{tel que } \sup_{Y_k} \left|\hat{\theta}_n(P_n) - \hat{\theta}_n(Q_{n,k})\right| = \infty \right\}. \tag{1.29}$$

Le point de rupture est par définition compris entre 0 et 1 et ses bornes sont atteintes en limite respectivement pour la moyenne empirique et pour un estimateur constant. Un estimateur est dit robuste si ϵ_n^* converge vers une limite strictement positive. Un exemple d'estimateur robuste est l'estimateur de la médiane qui a pour point de rupture $\frac{1}{n}\left[\frac{n-1}{2}\right]$. Cette valeur représente la borne maximale de ϵ_n^* que peut prendre un estimateur de localisation, pour lequel Maronna et al. (2006) montre que :

$$\epsilon_n^*(\hat{\theta}_n, X) \leq \frac{1}{n}\left[\frac{n-1}{2}\right]. \tag{1.30}$$

Un autre exemple est celui de l'estimateur de la moyenne tronquée à $\alpha \in]0, \frac{1}{2}[$, qui consiste à enlever aux deux extrémités de l'échantillon ordonné α% des données extrêmes. On montre que son point de rupture, quand n est assez grand, est de l'ordre de α.

Dans la littérature, nous trouvons également une définition alternative du point de rupture.

Définition 1.1.3 *Soit X un ensemble de n variables aléatoires réelles $\{X_1, \ldots, X_n\}$ et soit $Y_k = X \cup Y$, où Y un ensemble contenant k valeurs arbitraires, avec $k \in \{1, \ldots, n\}$. Soient P_n et $Q_{n,k}$ deux distributions empiriques associées respectivement à X et Y_k. Alors un point de rupture d'ajout pour un estimateur $\hat{\theta}$ s'écrit :*

$$\epsilon_n^{**}(\hat{\theta}_n, X) = \inf_{k \in \{1,\ldots,n\}} \left\{ \frac{k}{n+k}, \textit{tel que } \sup_{Y_k} \left| \hat{\theta}(P_n) - \hat{\theta}(Q_{n,k}) \right| = \infty \right\}. \tag{1.31}$$

A la limite, les deux définitions sont équivalentes (la première étant toutefois plus répandue), mais nous pouvons en privilégier une ou l'autre, pour en faciliter le calcul ou bien pour des raisons de simulation.

1.2 Cas multidimensionnel

1.2.1 Définitions et notations

Nous nous intéressons au cadre multidimensionnel où le paramètre à estimer est vectoriel. Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$ de loi paramétrique P_θ, $\theta \in \Theta$ où Θ est un ensemble de paramètres inclus dans $\mathbb{R}^d$. On reprend les définitions (1.1) et (1.2) du cas réel unidimensionnel, avec $a = (a_1, \ldots, a_d)$,

$$\psi = \left(\frac{\partial \rho}{\partial a_1}, \ldots, \frac{\partial \rho}{\partial a_d}\right)^T.$$

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires i.i.d dans $\mathbb{R}^d$ issue de la même loi P_θ de X. On suppose que pour tout $\theta \in \Theta$, $\theta = \underset{a \in \Theta}{\operatorname{argmin}} E_\theta(\rho(X, a)) = \underset{a \in \Theta}{\operatorname{argmin}} \int \rho(x, a) dP_\theta(x)$ où $\rho(x, .)$ est une fonction de $\mathbb{R}^d \to \mathbb{R}$ mesurable en x. Un M-estimateur est défini par :

$$\hat{\theta}_n = \underset{a \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, a),$$

ou par la valeur de a vérifiant vectoriellement

$$\sum_{i=1}^{n} \psi(X_i, a) = 0.$$

1.2.2 Comportement asymptotique

Quelques résultats asymptotiques vus précédemment peuvent être généralisés au cas multidimensionnel. Nous prenons les exemples du lemme 1.1.2 et du théorème 1.1.3 : les propriétés analytiques des fonctions utilisées dans ces résultats peuvent être étendues (la continuité et la bornitude ainsi que les propriétés des racines), en remplaçant la valeur absolue $|.|$ par la norme euclidienne $\|.\|$. En particulier la convergence en probabilité ou presque sûre des estimateurs sont obtenues avec les conditions de régularité adaptées à $\mathbb{R}^d$
Les résultats de la normalité asymptotique peuvent également être généralisés en remplaçant les produits dans $\mathbb{R}$ par des produits scalaire et matriciel. Dans ce cas, on montre que (voir par exemple Van der Vaart, 2000) :

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{Loi}{\to} \mathcal{N}\left(0, \ddot{M}(\theta)^{-1} \int \dot{\rho}(x,\theta)\dot{\rho}^T(x,\theta)dP(x)\ddot{M}(\theta)^{-1}\right)$$

où $\ddot{M}(\theta)$ est la matrice de dérivées partielles secondes de $\rho(x,\theta)$ et $\dot{\rho}(x,\theta)$ le vecteur de ces dérivées partielles premières, le symbole x^T étant la transposée du vecteur x.
Nous détaillerons les hypothèses qui nous permettrons d'établir ces résultats dans le cas multidimensionnel dans le chapitre 2 consacré aux données clusterisées.

1.3 Le cas spécifique de la médiane

1.3.1 Les médianes multidimensionnelles

La popularité de la médiane dans le cas unidimensionnel est due à ses propriétés fondamentales. La première caractéristique de cet estimateur est l'équivariance par des transformations affines des données, qui est une propriété géométrique très importante pour l'estimation d'un paramètre de localisation. La seconde propriété est sa robustesse, son point de rupture est maximal et est égal à 0.5. Elle possède des bonnes propriétés asymptotiques, comme notamment sa convergence presque sûre et la normalité asymptotique.
Dans la littérature, on trouve plusieurs propositions pour généraliser la médiane dans le cas multidimensionnel. Nous pouvons citer par exemple les travaux de Small (1990); Chaudhuri (1992) qui proposent une synthèse autour de la médiane multidimensionnelle. Dans cette partie, nous allons voir dans quelle mesure les propriétés de cet esti-

mateur dans le cas unidimensionnel s'étendent au cas multidimensionnel.
La version la plus populaire de la médiane est la L1-médiane que nous avons définie dans la relation (1.11) dans le cas particulier réel unidimensionnel. Cette médiane est aussi appelée médiane spatiale, Brown (1983), terminologie que nous adoptons dans ce travail.

1.3.2 Le cas particulier de la médiane spatiale

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires i.i.d dans $\mathbb{R}^d$ issue de la même loi P_θ. Nous pouvons définir la médiane spatiale par :

$$\hat{\theta}_n := \underset{a \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \|X_i - a\|, \tag{1.32}$$

$$\sum_{i=1}^{n} \frac{(X_i - \hat{\theta}_n)}{\|X_i - \hat{\theta}_n\|} = 0. \tag{1.33}$$

Dans un cas bidimensionnel, Brown (1983) a montré que $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi vers une loi normale bivariée, avec θ la médiane théorique recherchée. Plus récemment, dans le cas multidimensionnel nous pouvons citer les travaux de Chaudhuri (1992) et Van der Vaart (2000). De plus Lopuhaa and Rousseeuw (1991) ont montré que le point de rupture de la médiane spatiale est maximal et vaut 50%, i.e. $\epsilon_n^* = \frac{1}{2}$. Une des interprétations géométrique est que si sur un point θ on trouve superposées 50% des données, alors ce point est la médiane spatiale. La médiane spatiale est équivariante par toute transformation orthogonale. Si T est un opérateur orthogonal alors on a la propriété suivante :

$$\hat{\theta}_n(T(X_1), \ldots, T(X_n)) = T(\hat{\theta}_n(X_1, \ldots, X_n)), \tag{1.34}$$

Elle n'est cependant pas invariante pour n'importe quelle transformation affine. Cette notion d'équivariance est une propriété très appréciée pour un estimateur de localisation. Pour cette raison, Chakraborty and Chaudhuri (1996, 1998, 1999) ont proposé de rendre cette médiane spatiale affine équivariante en transformant les données de départ, suivie d'une transformation de la médiane à la fin du processus. Cette transformation et retransformation de la médiane spatiale est la seule version des médianes qui partage les propriétés désirables de la médiane du cas unidimensionnel.
Une autre version de la médiane multidimensionnelle est le vecteur de coordonnées

médianes, où chaque coordonnée de cet estimateur est la médiane unidimensionnelle de la même coordonnée des observations. Notons que cette version, dans un cas de données issues d'une loi symétrique, coïncide avec la médiane spatiale et sera également robuste avec un point de rupture maximal. Elle vérifie aussi les propriétés de convergence et de normalité, voir par exemple Bickel (1964). Par contre, cette version n'est pas équivariante pour une transformation affine équivariante.

Chapitre 2

M-estimateurs pour des données clusterisées

Sommaire

2.1 Cadre d'étude 35
2.2 Convergence des M-estimateurs 36
2.3 Normalité asymptotique 40
2.4 Estimation de la variance 45
2.5 Cas particulier des Lp-estimateurs 48
2.6 Résultats numériques 49

Résumé

Dans ce chapitre nous étudions la classe des M-estimateurs introduite par Huber en 1964. *Nous donnons des conditions suffisantes pour établir la convergence ainsi que la normalité asymptotique de cette famille d'estimateurs pour des vecteurs aléatoires clusterisés.*

2.1 Cadre d'étude

Dans ce chapitre nous proposons une étude des M-estimateurs dans un cas multidimensionnel avec une structure de dépendances pour les variables aléatoires qui généralise le cadre multidimensionnel i.i.d. L'existence et l'unicité d'un M-estimateur dépend, dans un premier temps, de la fonction objective qui le définit (dérivabilité, convexité, propriété de Lipschitz...), et d'un autre coté de la loi de probabilité des variables aléatoires.
Soit $X_1, \ldots, X_n$ une suite de n clusters indépendants, avec $X_i = (X_{i1}, \ldots, X_{im_i})$ où les variables aléatoires $X_{ij}, i = 1, 2, \ldots, n$ et $j = 1, 2, \ldots, m_i$, $m_i \geq 1$ sont à valeurs dans $\mathbb{R}^d$ et sont issues de la même loi paramétrique P_θ. Par la suite, nous supposons que le paramètre θ appartient à Θ, avec Θ un ouvert non vide, convexe et borné de $\mathbb{R}^d$. Ainsi, X_{ij} est associée au $j^{ème}$ élément du $i^{ème}$ cluster. Pour chaque cluster i, $i = 1, \ldots, n$, nous faisons l'hypothèse que $(X_{i1}, X_{i2}) \stackrel{d}{=} (X_{ik}, X_{ik'})$ pour tout $i = 1, \ldots, n$ et pour tout $k, k' = 1, .., m_i$ pour $k, k' = 1, .., m_i$, avec $k \neq k'$. Cette condition implique que la corrélation est la même pour chaque élément d'un cluster mais peut varier d'un cluster à l'autre. Nous notons $N_n := \sum_{i=1}^{n} m_i$ le nombre total de variables et nous supposons que $\lim_{n\to\infty} \frac{N_n}{n} = l$, $l \in]0, \infty[$.
Enfin, nous considérons la fonction $\rho(x, a) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ mesurable en x pour laquelle on a $\theta = \underset{a\in\Theta}{\text{argmin}} E_\theta(\rho(X_{11}, a)) = \underset{a\in\Theta}{\text{argmin}} \int \rho(x, a) dP_\theta(x)$ pour tout $\theta \in \Theta$.

Définition 2.1.1 *Le M-estimateur associé à la fonction ρ est défini par :*

$$\hat{\theta}_n = \underset{a\in\Theta}{argmin}\ M_n(a) \tag{2.1}$$

avec $M_n(a) = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \rho(X_{ij}, a)$. On notera également dans la suite, pour tout $a \in \Theta$

$$M(a) = E_\theta(\rho(X_{11}, a)).$$

Si $\rho(x, a)$ est différentiable pour tout $a = (a_1, \ldots, a_d)^T$ dans un voisinage de θ, alors le vecteur de ses dérivées partielles $\psi = (\frac{\partial\rho}{\partial a_1}, \ldots, \frac{\partial\rho}{\partial a_d})^T$ vérifie $E_\theta(\psi(X_{11}, \theta)) = 0$.

Définition 2.1.2 *Le M-estimateur $\hat{\theta}_n$ correspond à la valeur de a vérifiant les relations vectorielles :*

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} \psi(X_{ij}, a) = 0. \tag{2.2}$$

2.2 Convergence des M-estimateurs

Nous introduisons les hypothèses suivantes pour tout $\theta \in \Theta$.

Hypohèses 2.2.1

a) Pour tout $\epsilon > 0$: $\inf_{a \in \Theta : \|a-\theta\| > \epsilon} E_\theta \rho(X_{11}, a) > E_\theta \rho(X_{11}, \theta)$;

b) Pour tout $a \in \Theta$, $E_\theta(\rho^2(X_{11}, a)) < \infty$;

c) $\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} m_i^2 = 0$;

d) $\sum_{i \geq 1} \frac{m_i^2}{i^2} < \infty$.

L'hypothèse 2.2.1a), qui assure l'unicité du paramètre θ, est vérifiée si la fonction ρ est strictement convexe et si le support de la loi P_θ n'est pas concentré sur une ligne : cas de la médiane par exemple, Milasevic et al. (1987). La condition 2.2.1c) est utilisée pour la convergence en probabilité, alors que 2.2.1d) intervient pour la convergence presque sûre. De plus, cette dernière condition implique la précédente grâce au lemme de Kronecker. Ces deux conditions sont clairement vérifiées pour une suite (m_i) bornée, mais le cas non borné peut être considéré avec, par exemple, le choix $m_i = k \in \mathbb{N}^*$ pour $i = 2^k$ et $m_i = l$ sinon.

Nous commençons par un théorème général que nous utiliserons par la suite pour la convergence des estimateurs.

Théorème 2.2.1 *Pour tout* $\theta \in \Theta$ *tel que l'hypothèse 2.2.1a) est vérifiée et si*

$$\sup_{a \in \Theta} |M_n(a) - M(a)| \xrightarrow[n\to\infty]{P} 0, \tag{2.3}$$

alors

$$\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta. \tag{2.4}$$

Démonstration

On détermine les conditions suffisantes qui permettent de montrer que $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$. Par la définition de la convergence simple :

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(\|\hat{\theta}_n - \theta\| < \epsilon) = 1 \Leftrightarrow \hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta.$$

Rappelons que $M(a) := E_\theta(\rho(X_{11}, a))$. Sous l'hypothèse 2.2.1a), θ vérifie

$$\forall \epsilon > 0, \forall a : \|\theta - a\| > \epsilon,\ M(a) > M(\theta),$$

cela implique qu'il existe un $\eta > 0$, tel que si a vérifie $\|a - \theta\| > \epsilon$ alors $M(a) > M(\theta) + \eta$. L'événement $\left\{\|\hat{\theta}_n - \theta\| > \epsilon\right\}$ est donc inclus dans $\left\{M(\hat{\theta}_n) > M(\theta) + \eta\right\}$. Ainsi, par passage au complémentaire :

$$\left\{M(\theta) \leq M(\hat{\theta}_n) \leq M(\theta) + \eta\right\} \subset \left\{\|\hat{\theta}_n - \theta\| \leq \epsilon\right\}, \tag{2.5}$$

on obtient :

$$P\left(M(\hat{\theta}_n) - M(\theta) \leq \eta\right) \leq P\left(\|\hat{\theta}_n - \theta\| \leq \epsilon\right). \tag{2.6}$$

L'estimateur $\hat{\theta}_n$ est par définition la valeur qui minimise la fonction M_n. On a alors $M_n(\hat{\theta}_n) \leq M_n(\theta)$ et :

$$\begin{aligned} 0 \leq M(\hat{\theta}_n) - M(\theta) &= M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + M_n(\hat{\theta}_n) - M(\theta) \\ &\leq M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + M_n(\theta) - M(\theta) \\ &\leq 2 \sup_{a \in \Theta} |M_n(a) - M(a)|. \end{aligned}$$

Par hypothèse $\sup_{a\in\Theta} |M_n(a) - M(a)| \xrightarrow[n\to\infty]{P} 0$, donc $\left|M(\hat{\theta}_n) - M(\theta)\right| \xrightarrow[n\to\infty]{P} 0$ et par suite, nous obtenons la convergence en probabilité de notre M-estimateur. □

Remarquons que l'inégalité (2.6) provient de l'inclusion (2.5) à partir de laquelle nous pouvons déduire que la convergence de $\hat{\theta}_n$ est de la même nature que la convergence uniforme de la suite de fonctions M_n. Ainsi, si $\sup_{a\in\Theta} |M_n(a) - M(a)| \xrightarrow[n\to\infty]{p.s.} 0$ alors $\hat{\theta}_n \xrightarrow[n\to\infty]{p.s.} \theta$. Nous pouvons donc établir le théorème suivant :

Théorème 2.2.2 *Pour tout $\theta \in \Theta$, si pour tout x la fonction $a \mapsto \rho(x, a)$ est $k(x)$-Lipschitzienne, avec $E_\theta(k^2(X_{11})) < \infty$ alors :*

1. *Sous les hypothèses 2.2.1 a), b) et c) :*

$$\hat{\theta}_n \xrightarrow[n\to\infty]{P-L^2} \theta. \tag{2.7}$$

2. *Sous les hypothèses 2.2.1 a), b) et d) :*

$$\hat{\theta}_n \xrightarrow[n\to\infty]{p.s.} \theta. \tag{2.8}$$

Démonstration
D'après le théorème 2.2.1, il suffit d'avoir la convergence uniforme (en probabilité ou presque sûre) de $M_n(a)$ pour avoir la convergence (en probabilité ou presque sûre) de l'estimateur. Nous établissons cette convergence en deux étapes via les deux lemmes qui suivent.

Lemme 2.2.3 *Sous les hypothèses précédentes, $M_n(a)$ converge simplement vers $M(a)$.*

Preuve Les X_{ij} ont la même loi P_θ, donc d'après l'inégalité de Cauchy-Schwarz on a $E_\theta(\rho(X_{ij'},a)\rho(X_{ij},a)) \leq E_\theta(\rho(X_{ij},a)^2)$, et on obtient :

$$\begin{aligned} E_\theta\left(\sum_{j=1}^{m_i}\rho(X_{ij},a)\right)^2 &= \sum_{j=1}^{m_i} E_\theta\left(\rho^2(X_{ij},a)\right) + \sum_{j\neq j'}^{m_i} E_\theta\left(\rho(X_{ij},a)\rho(X_{ij'},a)\right) \\ &\leq m_i^2 E_\theta\left(\rho^2(X_{11},a)\right). \end{aligned}$$

On pose $Y_i = \sum_{j=1}^{m_i}\rho(X_{ij},a)$. Les $(Y_i)_{1\leq i}$ sont des variables aléatoires indépendantes d'espérance $\mu_i = m_i E_\theta(\rho(X_{11},a))$ dont la variance vérifie :

– Sous la condition 2.2.1c)

$$\frac{1}{n^2}\sum_{i=1}^{\infty} Var(Y_i) \leq \frac{1}{n^2}\sum_{i=1}^{\infty} m_i^2 E_\theta\left(\rho^2(X_{11},a)\right) \xrightarrow[n\to\infty]{} 0 \tag{2.9}$$

– Sous la condition 2.2.1d)

$$\sum_{i=1}^{\infty}\frac{Var(Y_i)}{i^2} \leq \sum_{i=1}^{\infty}\frac{m_i^2}{i^2}E_\theta\left(\rho^2(X_{11},a)\right) < \infty. \tag{2.10}$$

D'après le théorème de Kolmogorov$*$, pour tout $a \in \Theta$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho(X_{ij},a) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}M(a) \xrightarrow[n\to\infty]{p.s} 0.$$

De même, en utilisant le théorème de Chebyshev$*$, pour tout $a \in \Theta$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho(X_{ij},a) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}M(a) \xrightarrow[n\to\infty]{P-L^2} 0.$$

Par hypothèse, $\lim_{n\to\infty} \frac{N_n}{n} = l,\ l \in]0,\infty[$, on obtient donc $M_n(a) \xrightarrow[n\to\infty]{\text{p.s ou p}} M(a)$. □

Lemme 2.2.4 *Sous les conditions précédentes, $M_n(a)$ converge uniformément vers $M(a)$.*

Preuve L'ensemble $\overline{\Theta}$ est un compact, donc pour tout $\epsilon > 0$, il existe $h_\epsilon > 0$ et $a_1, \ldots, a_r$ dans Θ tels que $\Theta \subset \bigcup_{i=1}^{r} B(a_i, h_\epsilon)$, avec $B(a_i, h_\epsilon)$ la boule de centre a_i et de rayon h_ϵ donné.
Ainsi, pour tout $a \in \Theta$, il existe i tel que $a \in B(a_i, h_\varepsilon)$ et :

$$|M_n(a) - M(a)| \leq |M_n(a) - M_n(a_i)| + |M_n(a_i) - M(a_i)| + |M(a_i) - M(a)|.$$

La condition de Lipschitz faite sur la fonction ρ permet d'établir

$$|M_n(a) - M_n(a_i)| \leq \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} k(X_{ij}) \, \|a - a_i\|.$$

Comme $E_\theta(k(X_{ij})^2) < \infty$ par hypothèse et d'après le critère de Kolmogorov (ou de Chebyshev), on obtient la limite $k_n = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} k(X_{ij}) \xrightarrow[n\to\infty]{\text{p.s (ou p)}} E_\theta k(X_{ij})$. Nous continuons notre démonstration avec la convergence $p.s$. Le cas de la convergence en probabilité peut être établi en suivant le même procédé. Il existe ainsi un ensemble d'événements Ω_0 tel que $P(\Omega_0) = 1$ et pour tout $\omega \in \Omega_0$, il existe k' et $N_0(\omega)$ tels que $\forall n \geq N_0(\omega), k_n \leq k'$. De plus, la condition de Lipschitz sur $M(a)$ entraîne que :

$$|M(a) - M(a_i)| \leq E_\theta k(X_{11}) \, \|a - a_i\|.$$

Le choix de a nous permet d'écrire l'inégalité $||a - a_i|| < h_\varepsilon$. Enfin, la convergence simple de M_n vers M entraîne que pour $1 \leq i \leq r$ il existe Ω_1 tel que $P(\Omega_1) = 1$, pour tout $\omega \in \Omega_1$ et pour tout $\epsilon > 0$, il existe $N'(\omega)$ tel que :

$$\forall n > N'(\omega),\ |M_n(a) - M(a)| \leq k' h_\epsilon + \epsilon/3 + k' h_\epsilon.$$

La convergence uniforme presque sûre de $M_n(a)$ s'ensuit avec le choix h_ε tel que $h_\epsilon < \frac{\epsilon}{3k'}$. □

Remarque : Les hypothèses du théorème 2.2.2 ne sont pas très contraignantes. En effet, si pour tout x, la fonction $x \mapsto \psi(x, a)$ est continue et dominée par une fonction de carré intégrable indépendante de a, alors $\rho(x, a)$ est $\sup_{a\in\Theta} \|\psi(x, a)\|$-Lipschitzienne.

Dans le cas où $E_\theta \left\|X_{ij}\right\|^2 < \infty$, alors pour la moyenne empirique, la fonction ψ vérifie $\sup_{a\in\Theta} \|\psi(x,a)\| \leq 2\|x\| + 2\sup_{a\in\Theta} \|a\|$, donc $\sup_{a\in\Theta} \|\psi(X_{11},a)\|$ est de carré intégrable. La médiane spatiale définie par $\rho(x,a) = \|x-a\|$ et l'estimateur de Huber qui correspond à la fonction $\rho(x,a) = \frac{1}{2}\|x-a\|^2$ si $\|x-a\| \leq k$ et $\rho(x,a) = k\|x-a\| - \frac{1}{2}k^2$ si $\|x-a\| > k$ sont d'autres exemples de fonction ρ vérifiant les conditions du théorème.

2.3 Normalité asymptotique

Après une étude de la convergence des M-estimateurs, dans cette partie, nous allons établir une vitesse de convergence ainsi que la normalité asymptotique des estimateurs dans le cadre des données clusterisées. Dans un premier temps, nous donnons les hypothèses qui permettent d'annoncer ces résultats.

Hypohèses 2.3.1

Il existe un réel $\eta > 0$ tel que :

a) $E_\theta(\|\psi(X_{11},\theta)\|^{2+\eta}) < \infty$.

b) $\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} m_i^{2+\eta} < \infty$.

Hypohèse 2.3.1

Si C_i est la matrice définie par $C_i = E_\theta\psi(X_{ij},\theta)\psi^T(X_{ij'},\theta)$, pour tout $j \neq j'$, on suppose que

$$\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} m_i(m_i-1)C_i = \mathbb{C}_m.$$

Nous avons besoin des hypothèses 2.3.1 dans la démonstration de la normalité asymptotique des M-estimateurs. La matrice $C_i = E_\theta\psi(X_{ij},\theta)\psi^T(X_{ij'},\theta)$ ne dépend que du cluster i ; nous notons $C_i = C$ lorsque les corrélations intra-cluster sont les mêmes pour tous les clusters. L'hypothèse 2.3.1 est vérifiée si les corrélations intra-cluster sont les mêmes pour tous les clusters, et si $\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} m_i(m_i-1) = c_m < \infty$, alors $\mathbb{C}_m = c_m C$. Cette condition est également vérifiée pour des corrélations telles que $m_n C_n$ tend vers $\mathbb{C}_m$ quand n tend vers l'infini.

Théorème 2.3.1

Pour tout a dans un voisinage ouvert de θ et pour tout x, on suppose que la fonction $a \mapsto \psi(x,a)$ est deux fois différentiable, que sa dérivée seconde est continue, que les dérivées partielles secondes de $\psi(x,a)$ sont dominées par une fonction de x de carré intégrable, indépendante de

a, ainsi que les éléments de la matrice hessienne $\frac{\partial\psi(X_{ij},a)}{\partial a}/_{a=\theta}$ admettent des moments d'ordre 2 pour $i = 1,\dots,n$; $j = 1,\dots,m_i$ et que $E_\theta(\frac{\partial\psi(X_{ij},a)}{\partial a}/_{a=\theta})$ est inversible. Alors sous les hypothèses 2.2.1, 2.3.1 et 2.3.1 :

$$\sqrt{N_n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{L} N\left(0, V_\theta^{-1}\left(B + \mathbb{C}_m\right)V_\theta^{-1}\right),$$

avec $B := E_\theta\psi(X_{11},\theta)\psi^T(X_{11},\theta)$ et $V_\theta = E_\theta(\frac{\partial\psi(X_{11},a)}{\partial a}/_{a=\theta})$.

Démonstration

On définit la statistique $T_n(a)$ par :

$$T_n(a) = \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\psi(X_{ij},a).$$

On note également $\dot{T}_n$, $\ddot{T}_n$ les matrices de dérivées partielles premières et secondes de T_n. La démonstration de ce théorème se fait à l'aide des deux lemmes qui suivent :

Lemme 2.3.2 *Sous les hypothèses de régularité effectuées sur la fonction ψ, il existe dans un voisinage de θ, $\tilde{\theta}_n$ et une matrice $D(\hat{\theta}_n - \theta)$ tels que :*

$$-\sqrt{N_n}T_n(\theta) = \left\{\dot{T}_n^{\,T}(\theta) + \frac{1}{2}D(\hat{\theta}_n - \theta)\ddot{T}_n(\tilde{\theta}_n)\right\}\sqrt{N_n}(\hat{\theta}_n - \theta). \tag{2.11}$$

Preuve : Soit le vecteur de paramètre $a = (a_1,\dots,a_d)^T$ dans Θ, et soit $\psi(x,a) := (\frac{\partial\rho(x,a)}{\partial a_1},\dots,\frac{\partial\rho(x,a)}{\partial a_d})^T := (\psi_1(x,a),\dots,\psi_d(x,a))^T$. On écrit le développement de Taylor de $\psi_i(x,a)$, $i = 1,\dots,d$, pour tout a dans un voisinage convexe et ouvert de θ, alors il existe $\tilde{\theta}_i$ dans ce voisinage tel que :

$$\psi_i(x,a) = \psi_i(x,\theta) + \dot{\psi}_i(x,\theta)^T(a-\theta) + \frac{1}{2}(a-\theta)^T\ddot{\psi}_i(x,\tilde{\theta}_i)(a-\theta),$$

avec $\dot{\psi}_i(x,\theta) = (\frac{\partial\psi_i(x,\theta)}{\partial a_1},\dots,\frac{\partial\psi_i(x,\theta)}{\partial a_d})^T$ et

$$\ddot{\psi}_i(x,\tilde{\theta}_i) = \begin{pmatrix} \frac{\partial^2\psi_i(x,\tilde{\theta}_i)}{\partial {a_1}^2} & \cdots & \frac{\partial^2\psi_i(x,\tilde{\theta}_i)}{\partial a_1\partial a_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2\psi_i(x,\tilde{\theta}_i)}{\partial a_d\partial a_1} & \cdots & \frac{\partial^2\psi_i(x,\tilde{\theta}_i)}{\partial {a_d}^2} \end{pmatrix}.$$

En notant $\tilde{\theta} := (\tilde{\theta}_i, \ldots, \tilde{\theta}_d)$, on peut écrire la formule de Taylor vectorielle suivante :

$$\psi(x,a) = \psi(x,\theta) + \dot{\psi}(x,\theta)^T(a-\theta) + \frac{1}{2}D(a-\theta)\ddot{\psi}(x,\tilde{\theta})(a-\theta), \tag{2.12}$$

avec

$$D(a-\theta) = \begin{pmatrix} (a-\theta)^T & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & (a-\theta)^T \end{pmatrix} = (a-\theta)^T \otimes I$$

où I la matrice unité $d \times d$, et $\ddot{\psi}(x,\tilde{\theta}) = \begin{pmatrix} \ddot{\psi}_1(x,\tilde{\theta}_1) \\ \vdots \\ \ddot{\psi}_d(x,\tilde{\theta}_d) \end{pmatrix}$.

L'équation (2.12) implique qu'il existe un $\bar{\theta}$ tel que :

$$T_n(a) = T_n(\theta) + \dot{T}_n(\theta)^T(a-\theta) + \frac{1}{2}D(a-\theta)\ddot{T}_n(\bar{\theta})(a-\theta). \tag{2.13}$$

Le résultat est obtenu en appliquant le développement (2.13) pour $\hat{\theta}_n$, sachant que par définition $T_n(\hat{\theta}_n) = 0$. □

A partir du lemme 2.3.2 nous pouvons établir la normalité asymptotique de $\hat{\theta}_n$ en démontrant les lemmes suivants :

Lemme 2.3.3 *Sous les hypothèses du théorème 2.3.1,* $\sqrt{N_n}T_n(\theta) \xrightarrow[n\to\infty]{L} N(0, B + \mathbb{C}_m)$.

Lemme 2.3.4 *Sous les hypothèses du théorème 2.3.1,* $\dot{T}_n(\theta) \xrightarrow[n\to\infty]{p.s} V_\theta$ *et* $\ddot{T}_n(a) = O_p(1)$ *uniformément en a.*

Lemme 2.3.5 *Sous les hypothèses du théorème 2.3.1,* $\hat{\theta}_n \xrightarrow[n\to\infty]{p.s} \theta$.

Preuve du lemme 2.3.3

On utilise dans cette démonstration le théorème de Lindeberg$*$ qui permet d'obtenir cette convergence en loi. Nous introduisons les vecteurs $\xi_i := \sum_{j=1}^{m_i} \psi(X_{ij},\theta)$, $i = 1, \ldots, n$. Ces vecteurs sont indépendants, d'espérance $E_\theta(\xi_i) = 0$ et de matrices de covariance V_i, $i = 1, \ldots, n$. Pour chaque i, V_i est définie par : $V_i = \sum_{j=1}^{m_i} E_\theta\psi(X_{ij},\theta)\psi^T(X_{ij},\theta) + \sum_{j\neq j'}^{m_i} E_\theta\psi(X_{ij},\theta)\psi^T(X_{ij'},\theta)$. Comme tous les couples $(X_{ij'}, X_{ij})$ avec $j \neq j'$ ont la même corrélation dans le cluster i, on pose $C_i = E_\theta\psi(X_{ij},\theta)\psi^T(X_{ij'},\theta)$ et $B := E_\theta\psi(X_{ij},\theta)\psi^T(X_{ij},\theta)$.

En conséquent, sous l'hypothèse 2.3.1 on a la limite :

$$\frac{1}{N_n}\sum_{i=1}^{n} V_i \xrightarrow[n\to\infty]{} B + \mathbb{C}_m, \tag{2.14}$$

et comme $\lim_{n\to\infty} \frac{N_n}{n} = l$, on peut écrire :

$$\frac{1}{n}\sum_{i=1}^{n} V_i \xrightarrow[n\to\infty]{} l(B + C_m) := \Sigma.$$

On vérifie dans la suite que $E_n := \frac{1}{n}\sum_{i=1}^{n} E_\theta[\|\xi_i\|^2 1_{\|\xi_i\|>\epsilon\sqrt{n}}]$ tend vers 0, pour être dans les conditions du théorème de Lindeberg :

$$E_n = \frac{1}{n}\sum_{i=1}^{n} E_\theta[\|\xi_i\|^{-\eta}\,\|\xi_i\|^{2+\eta}\,1_{\|\xi_i\|>\epsilon\sqrt{n}}] \leq \sqrt{n}^{-\eta}\frac{1}{n}\sum_{i=1}^{n}(\epsilon)^{-\eta}E_\theta[\|\xi_i\|^{2+\eta}].$$

Or par l'inégalité de Minkowski :

$$E_\theta\,\|\xi_i\|^{2+\eta} \leq \left[\sum_{j=1}^{m_i}\left[E_\theta\,\|\psi(X_{ij},\theta)\|^{2+\eta}\right]^{1/(2+\eta)}\right]^{2+\eta} = E_\theta\,\|\psi(X_{ij},\theta)\|^{2+\eta}\,m_i^{2+\eta}.$$

Sous les hypothèses 2.3.1, et pour tout $\epsilon > 0$:

$$E_n \leq E_\theta[\|\psi(X_{ij},\theta)\|^{2+\eta}](\sqrt{n}\epsilon)^{-\eta}\frac{1}{n}\sum_{i=1}^{n} m_i^{2+\eta} \xrightarrow[n\to\infty]{} 0.$$

En conclusion $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i \xrightarrow{L} N\left(0, l(B+\mathbb{C}_m)\right)$, ou encore

$$\sqrt{N_n}T_n(\theta) \xrightarrow{L} N\left(0, B+\mathbb{C}_m\right).$$

□

Preuve du lemme 2.3.4

Dans la preuve du lemme 2.3.3, nous avons établi le critère de Kolmogorov$*$ pour la suite de variables aléatoires $\left(\sum_{j=1}^{m_i}\rho(X_{ij},a)\right)_{1\leq i\leq n}$. En remplaçant $\rho(X_{ij},a)$ par les éléments des matrices $\dot{\psi}(x,\theta)$ qui sont de carrés intégrables, on déduit le critère de Kolmogorov pour les éléments de la suite $\left(\sum_{j=1}^{m_i}\dot{\psi}(X_{ij},\theta)\right)_{1\leq i\leq n}$. Ainsi, $\dot{T}_n(\theta) \xrightarrow[n\to\infty]{p.s} V_\theta$.

Par hypothèse, il existe une fonction F de carré intégrable telle que $\|\ddot{\psi}(x,a)\| \leq F(x)$ pour tout a. Le critère de Kolmogorov étant vérifié pour $\sum_{j=1}^{m_i} F(X_{ij})$, on en déduit que $\ddot{T}_n(a)$ est un $O_p(1)$ uniformément en a. □

Remarque : nous utilisons le critère de Kolmogorov$*$, pour établir la convergence presque sûre qui nécessite l'hypothèse 2.2.1d), mais nous pouvons obtenir la convergence en probabilité par le critère de Chebyshev$*$ et l'hypothèse 2.2.1c).

Preuve du lemme 2.3.5

Pour montrer la convergence presque sûre de $\hat{\theta}_n$ vers θ, il suffit de vérifier que le Theorème 2.2.2 s'applique, c'est-à-dire que la fonction est $k(x)$ lipshitzienne avec k de carré intégrable. Nous avons d'après le théorème des accroissements finis, pour tous a_1 et a_2 :

$$|\rho(x,a_1) - \rho(x,a_2)| \leq \sup_{a\in\Theta} \|\psi(x,a)\| \, \|a_1 - a_2\| \, .$$

D'après l'équation (2.12), on a la majoration suivante :

$$\begin{aligned}\sup_{a\in\Theta} \|\psi(x,a)\| &\leq \|\psi(x,\theta)\| + \|\dot{\psi}(x,\theta)\| \sup_{a\in\Theta} \|a-\theta\| + \frac{1}{2} \sup_{a\in\Theta} \|\ddot{\psi}(x,\tilde{a})\| \, \|a-\theta\|^2 \\ &\leq \|\psi(x,\theta)\| + \alpha \, \|\dot{\psi}(x,\theta)\| + \frac{1}{2}\alpha^2 \, \|F(x)\| \, ,\end{aligned}$$

avec $\alpha = \sup_{a\in\Theta} \|a-\theta\| < \infty$. Par hypothèse, les fonctions $\|\psi(x,\theta)\|$, $\|\dot{\psi}(x,\theta)\|$ et $\|F(x)\|$ sont de carré intégrable, donc $\sup_{a\in\Theta} \|\psi(x,a)\|$ est aussi de carré intégrable. On est donc sous les hypothèses du théorème 2.2.2. On peut conclure que $\hat{\theta}_n \xrightarrow[n\to\infty]{p.s} \theta$. □

Les hypothèses du théorème 2.3.1 peuvent paraître contraignantes, mais elles sont satisfaites par des estimateurs classiques.

Exemple 2.3.1 L'estimateur du maximum de vraisemblance. Il entre dans le cadre du théorème si la fonction de densité $f_\theta(x) > 0$, avec $\psi(x,\theta) = \frac{\dot{f}_\theta(x)}{f_\theta(x)}$ et $\dot{\psi}(x,\theta) = \frac{\ddot{f}_\theta(x)}{f_\theta(x)} - \frac{\dot{f}_\theta(x)^T \dot{f}_\theta(x)}{f_\theta(x)^2}$. Si on suppose qu'on peut interchanger les signes de dérivée et d'intégrale alors on obtient $E_\theta \psi(X_{ij},\theta) = 0$ et $E_\theta \dot{\psi}(X_{ij},\theta) = -E_\theta \frac{\dot{f}_\theta(X_{ij})^T \dot{f}_\theta(X_{ij})}{f_\theta(X_{ij})^2} = -I_\theta$ et donc $\sqrt{N_n}(\hat{\theta}_n - \theta)$ converge vers une loi normale centrée de matrice de variance-covariance $I_\theta^{-1} + I_\theta^{-1} \mathbb{C}_m I_\theta^{-1}$, où I_θ est l'information de Fisher en θ. Sous les hypothèses 2.2.1, 2.3.1 et 2.3.1 les estimateurs L^p avec $p \geq 2$ intègrent parfaitement le théorème, avec en particulier, la moyenne empirique ($p = 2$).

Dans le cas spécifique des variables i.i.d. (où les variables intra-clusters sont indépendantes), les résultats précédents restent valables. On obtient ainsi la convergence

presque sûre et la normalité asymptotique des estimateurs avec la matrice de variance-covariance $V_\theta^{-1} B V_\theta^{-1}$.

2.4 Estimation de la variance

À présent, nous pouvons estimer les matrices de covariances apparaissant dans le résultat de normalité asymptotique. Nous considérons d'abord le cas particulier où les lois sont identiques pour les clusters de même taille et que les corrélations intra-cluster sont les mêmes pour tous les clusters. Les matrices de covariances $C_i \equiv C$ ne dépendent donc plus du cluster i, $i = 1, \ldots, n$. Nous avons le théorème suivant :

Théorème 2.4.1 *Sous l'hypothèse 2.3.1a) et si la suite (m_i) est à valeurs dans $\{1, \ldots, \mathcal{M}\}$, alors les estimateurs :*

$$\hat{B}_n = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \psi(X_{ij}, \theta) \psi^T(X_{ij}, \theta) \tag{2.15}$$

$$\hat{C}_n = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{j' \neq j}^{m_i} \psi(X_{ij}, \theta) \psi^T(X_{ij'}, \theta) \tag{2.16}$$

convergent respectivement presque sûrement vers B et $\mathbb{C}_m$ définies dans le théorème 2.3.1. L'estimateur $\hat{V}_n = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{\partial \psi(X_{ij}, a)}{\partial a} /_{a=\theta}$ converge également presque sûrement vers V_θ si les dérivées partielles de $\psi(X_{ij}, a)$ admettent un moment d'ordre 1.

Notons que la convergence en probabilité et presque sûre des matrices est équivalente respectivement, à la convergence en probabilité et presque sûre de chacun de ses éléments.

Démonstration

On peut écrire

$$\hat{B}_n = \frac{1}{N_n} \sum_{m=1}^{\mathcal{M}} \sum_{i=1}^{n} I_{\{m_i = m\}} \sum_{j=1}^{m} \psi(X_{ij}, \theta) \psi^T(X_{ij}, \theta).$$

On note $\alpha_n^m := \sum_{i=1}^{n} I_{\{m_i = m\}}$, le nombre de clusters contenant m éléments. Comme m varie dans $\{1, \ldots, \mathcal{M}\}$, nécessairement il existe au moins un m tel que $\alpha_n^m := \sum_{i=1}^{n} \mathbb{1}(m_i = m) \xrightarrow[n \to \infty]{} \infty$. En réordonnant les sommes précédentes par taille de clusters, on obtient

$$\hat{B}_n = \frac{1}{N_n} \sum_{m=1}^{\mathcal{M}} \sum_{i=1}^{\alpha_n^m} \sum_{j=1}^{m} \psi(\tilde{X}_{ij}, \theta) \psi^T(\tilde{X}_{ij}, \theta).$$

Soit E_n l'ensemble défini par $\{m \in \{1, \dots, \mathcal{M}\} : \alpha_n^m \xrightarrow[n\to\infty]{} \infty\}$, nous avons encore

$$\hat{B}_n = \frac{1}{N_n} \sum_{m=1, m\in E_n}^{\mathcal{M}} \alpha_n^m \Big\{ \frac{1}{\alpha_n^m} \sum_{i=1}^{\alpha_n^m} \sum_{j=1}^{m} \psi(\tilde{X}_{ij}, \theta)\psi^T(\tilde{X}_{ij}, \theta) \Big\}$$
$$+ \frac{1}{N_n} \sum_{m=1, m\notin E_n}^{\mathcal{M}} m\alpha_n^m \Big\{ \frac{1}{m\alpha_n^m} \sum_{i=1}^{\alpha_n^m} \sum_{j=1}^{m} \psi(\tilde{X}_{ij}, \theta)\psi^T(\tilde{X}_{ij}, \theta) \Big\}.$$

D'après la loi des grands nombres, le premier terme entre accolades converge vers $mE_\theta(\psi(X_{ij}, \theta)\psi^T(X_{ij}, \theta))$ presque sûrement. Par définition de E_n^c, il y a un nombre fini de termes pour le deuxième terme entre accolades. Comme de plus

$$N_n = \sum_{m=1, m\in E_n}^{\mathcal{M}} m\alpha_n^m + \sum_{m=1, m\notin E_n}^{\mathcal{M}} m\alpha_n^m,$$

on en déduit que $\frac{1}{N_n} \sum_{m=1, m\notin E_n}^{\mathcal{M}} m\alpha_n^m$ tend vers 0, alors que $\frac{1}{N_n} \sum_{m=1, m\in E_n}^{\mathcal{M}} m\alpha_n^m$ tend vers 1. Le résultat s'ensuit pour $\hat{B}_n$.
La démonstration est la même pour $\hat{C}_n$ et $\hat{V}_n$. Prenons l'exemple de la matrice $\hat{C}_n$, il suffit d'utiliser dans la démonstration les termes $\sum_{j=1}^{m} \sum_{j'\neq j}^{m} \psi(X_{ij}, \theta)\psi^T(X_{ij'}, \theta)$. □

Remarque : La convergence presque sûre de $\hat{V}_n$ vers V_θ, supposée inversible, entraîne que presque sûrement pour n assez grand, $\hat{V}_n$ est également inversible.

Nous pouvons considérer un cas plus général où les m_i ne sont plus supposées bornées et la corrélation C_i peut varier d'un cluster à un autre. Le critère de Kolmogorov permet alors, d'établir la convergence presque sûre des trois estimateurs en supposant l'existance des moments d'ordre 4 de la fonction ψ.

Théorème 2.4.2 *Nous obtenons les comportements asymptotiques suivants.*

1. *Si $\sum_{i\geq 1} \frac{m_i^2}{i^2} < \infty$ et si la fonction ψ admet un moment d'ordre 4, alors $\hat{B}_n$ converge presque sûrement vers B.*

2. *Si $\sum_{i\geq 1} \frac{m_i^4}{i^2} < \infty$ et si la fonction ψ admet un moment d'ordre 4, alors $\hat{C}_n$ converge presque sûrement vers $\mathbb{C}_m$.*

3. *Si $\sum_{i\geq 1} \frac{m_i^2}{i^2} < \infty$ et si la fonction la matrice $\frac{\partial\psi(X_{ij}, a)}{\partial a} /_{a=\theta}$ admet un moment d'ordre 2, alors $\hat{V}_n$ converge presque sûrement vers V_θ.*

Démonstration : Prenons l'élément $\hat{B}_n(k,l)$ de la matrice $\hat{B}_n$,

$$\hat{B}_n(k,l) = \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\psi_k(X_{ij},\theta)\psi_l^T(X_{ij},\theta).$$

Les variables $\sum_{j=1}^{m_i}\psi_k(X_{ij},\theta)\psi_l(X_{ij},\theta)$ sont indépendantes d'espérances

$$E_\theta\left(\sum_{j=1}^{m_i}\psi_k(X_{ij},\theta)\psi_l(X_{ij},\theta)\right) = m_i B(k,l)$$

et de variances notées $\tilde{V}_i$. Nous calculons, dans un premier temps, le moment d'ordre 2, par l'inégalité de Cauchy Schwarz, nous obtenons les majorations successives

$$\begin{aligned}
E_\theta\left(\sum_{j=1}^{m_i}\psi_k(X_{ij},\theta)\psi_l(X_{ij},\theta)\right)^2 &= \sum_{j=1}^{m_i} E_\theta\left(\psi_k^2(X_{ij},\theta)\psi_l^2(X_{ij},\theta)\right)\\
&+ \sum_{j\neq j'}^{m_i} E_\theta\left(\psi_k(X_{ij},\theta)\psi_l(X_{ij},\theta)\psi_k(X_{ij'},\theta)\psi_l(X_{ij'},\theta)\right)\\
&\leq \sum_{j=1}^{m_i} E_\theta\left(\psi_k^2(X_{ij},\theta)\psi_l^2(X_{ij},\theta)\right)\\
&+ \sum_{j\neq j'}^{m_i} \underbrace{E_\theta\left(\psi_k^2(X_{ij},\theta)\psi_l^2(X_{ij},\theta)\right)}_{\text{même corrélation intracluster}}\\
&\leq \sum_{j=1}^{m_i}\sqrt{E_\theta\left(\psi_k^4(X_{ij},\theta)\right)}\sqrt{E_\theta\left(\psi_l^4(X_{ij},\theta)\right)}\\
&+ \sum_{j\neq j'}^{m_i}\sqrt{E_\theta\left(\psi_k^4(X_{ij},\theta)\right)}\sqrt{E_\theta\left(\psi_l^4(X_{ij},\theta)\right)}\\
&\leq m_i^2\sqrt{E_\theta\left(\psi_k^4(X_{ij},\theta)\right)}\sqrt{E_\theta\left(\psi_l^4(X_{ij},\theta)\right)}
\end{aligned}$$

où la dernière assertion provient du fait que les X_{ij} ont la même loi. Nous vérifions dans la suite le critère de Kolmogorov :

$$\sum_{i\geq 1}\frac{\tilde{V}_i}{i^2} \leq \sqrt{E_\theta\left(\psi_k^4(X_{ij},\theta)\right)}\sqrt{E_\theta\left(\psi_l^4(X_{ij},\theta)\right)}\sum_{i\geq 1}\frac{m_i^2}{i^2}.$$

La série $\sum_{i\geq 1}\frac{\tilde{V}_i}{i^2}$ est donc convergente. Le théorème de Kolmogorov nous permet de conclure la convergence presque sûre de $\hat{B}_n(k,l)$ vers $B(k,l)$ (la notation $A(k,l)$ désigne l'élément issu de la $k$ème ligne et de la $l$ème colonne de la matrice A).

Avec le même procédé nous démontrons la convergence presque sûre des matrices $\hat{C}_n$ et $\hat{V}_n$. $\square$

2.5 Cas particulier des Lp-estimateurs

Nevalainen et al. (2007a) ont traité la médiane spatiale pour des données clusterisées. Cet estimateur entre bien dans le cadre du théorème 2.2.2 de convergence, mais il ne vérifie pas les conditions du théorème 2.3.1 de la normalité asymptotique. La médiane spatiale est un cas particulier de la famille des Lp-estimateurs que nous allons traiter comme exemple dans cette section.

Définition 2.5.1 *Un Lp-estimateur $\hat{\theta}_n$ dans ce cadre de données clusterisées est donné par la relation suivante : $\hat{\theta}_n = \underset{a\in\Theta}{argmin} M_n(a)$, où*

$$M_n(a) = \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|X_{ij}-a\right\|^p; \tag{2.17}$$

ou bien encore par

$$T_n(\hat{\theta}_n) = \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|X_{ij}-\hat{\theta}_n\right\|^{p-2}\left(X_{ij}-\hat{\theta}_n\right) = 0. \tag{2.18}$$

La fonction objective $\rho(x,a) = \|x-a\|^p$ est différentiable suivant les valeurs de p. Pour une valeur de p convenable nous avons :

$$\psi(x-a) = -p\,\|x-a\|^{p-2}(x-a); \tag{2.19}$$

$$\dot{\psi}(x-a) = p\,\|x-a\|^{p-2}\left(I_d + (p-2)\frac{(x-a)(x-a)^T}{\|x-a\|^2}\right). \tag{2.20}$$

Ainsi quand $p = 1$ (le cas de la médiane spatiale) la fonction ρ n'est pas différentiable en a. Pour des valeur de $p \geq 2$ les dérivées partielles de ψ secondes existent. L'estimateur peut donc vérifier les hypothèses du théorème 2.3.1 sous une loi paramétrique qui assurent l'existence des conditions sur les moments de la fonction ψ et de ses dérivées partielles.

Les estimateurs des matrices de variances-covariances, pour les Lp-estimateurs, sont

données sous les hypothèses du théorème 2.4.1 par :

$$\begin{aligned}
\hat{V}_n =& \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} p\left\|X_{ij}-\theta\right\|^{p-2}\left(I_d+(p-2)\frac{(X_{ij}-\theta)(X_{ij}-\theta)^T}{\left\|X_{ij}-\theta\right\|^2}\right);\\
\hat{B}_n =& \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} p^2\left\|X_{ij}-\theta\right\|^{2p-4}(X_{ij}-\theta)(X_{ij}-\theta)^T;\\
\hat{C}_n =& \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'=1,j'\neq j}^{m_i} p^2\left\|X_{ij}-\theta\right\|^{p-2}\left\|x_{ij'}-\theta\right\|^{p-2}(X_{ij}-\theta)(x_{ij'}-\theta)^T.
\end{aligned}$$

Ces estimateurs convergent presque sûrement.

2.6 Résultats numériques

Dans cette partie, nous étudions les propriétés asymptotiques d'un M-estimateur dans un cas clusterisé. L'objectif ici est d'illustrer à partir de simulations, la convergence asymptotique et d'observer l'influence de la corrélation ainsi que la taille des clusters sur la qualité de l'estimation effectuée.

Nous choisissons $X_1,\dots,X_n$ n clusters indépendants issus de P_θ, où P_θ est une loi normale bivariée d'espérance nulle ($\theta=0$) et de matrice de covariance $Var(X_{i1},X_{i2}) = \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$, avec $0<\rho<1$. Nous supposons donc que chaque cluster présente la même corrélation interne choisie égale à ρ. Nous prenons comme M-estimateurs, les Lp-estimateurs définis par : $\hat{\theta}_n = \underset{a\in\Theta}{\operatorname{argmin}}\frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|X_{ij}-a\right\|^p$ que nous allons étudier pour différentes valeurs de p.
Nous considérons dans cette partie un nombre identique d'individus par clusters, soit $m_i\equiv m=10$, pour tout $i=1,\dots,n$, pour différentes valeurs de m et de ρ. La figure 2.1 illustre un échantillon de variables aléatoires clusterisées de loi normale P_0 centrée et de matrice de covariance $\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$, $\rho=0.2,0.8$, avec un ellipsoïde de confiance à 98%. On remarque que la possibilité d'avoir des clusters très éloignés du centre augmente avec ρ. On s'attend donc à une estimation d'autant plus difficile pour de grandes valeurs de ρ.

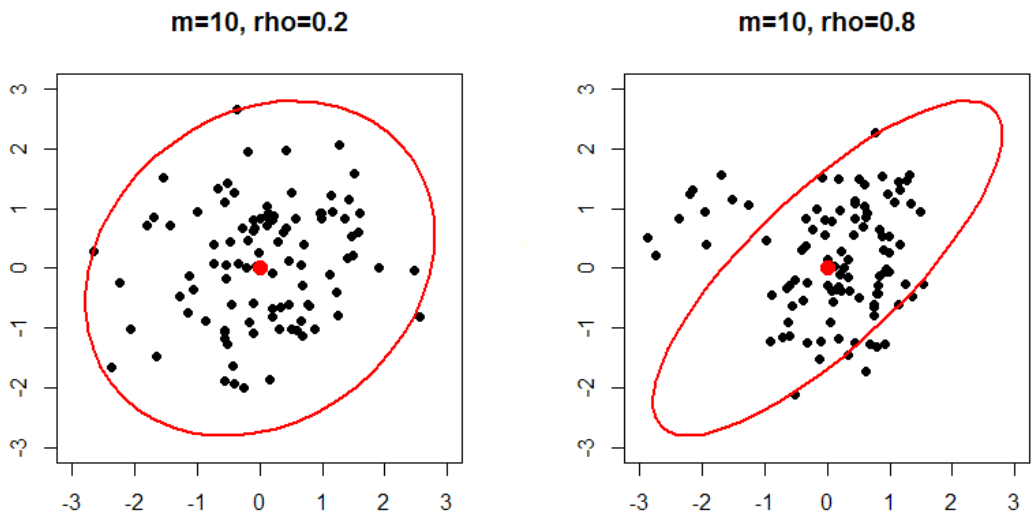

FIGURE 2.1 – *Ellipsoïdes de confiance pour des données clusterisées issues d'un loi normale bivariée centré*

Soit θ_n le Lp-estimateur. La moyenne de l'erreur quadratique (MSE) s'écrit :

$$\underbrace{E_\theta\left((\theta_n-\theta)(\theta_n-\theta)^T\right)}_{MSE(\theta_n)} = \underbrace{E_\theta(\theta_n\theta_n^T)-E_\theta(\theta_n)E_\theta^T(\theta_n)}_{Var(\theta_n)} + \underbrace{(\theta-E_\theta(\theta_n))(\theta-E_\theta(\theta_n))^T}_{SQB(\theta_n)}$$

où $Var(\theta_n)$ est la matrice de variance-covariance de θ_n et $SQB(\theta_n)$ est la matrice du biais quadratique.
Nous effectuons une étude de Monte Carlo avec $r = 1000$ échantillons et nous notons θ_n^i le Lp-estimateur calculé pour le $i^{\text{ème}}$ échantillon, $i = 1,\ldots,r$. L'erreur quadratique empirique s'écrit alors :

$$\hat{MSE}(\theta_n) = \frac{1}{r}\sum_{i=1}^{r}(\theta_n^i-\bar{\theta}_n)(\theta_n^i-\bar{\theta}_n)^T + \bar{\theta}_n\bar{\theta}_n^T,$$

avec $\bar{\theta}_n = \frac{1}{r}\sum_{i=1}^{r}\theta_n^i$.
Nous prenons comme critère de mesure $\sqrt{det(\hat{MSE}(\theta_n))}$, où $det(.)$ est le déterminant de la matrice. Sous les hypothèses du théorème 2.3.1, $\sqrt{det(\hat{MSE}(\theta_n))}$ converge vers $\sqrt{det(V_\theta^{-1}\,(B+\mathbb{C}_m)\,V_\theta^{-1})}$.

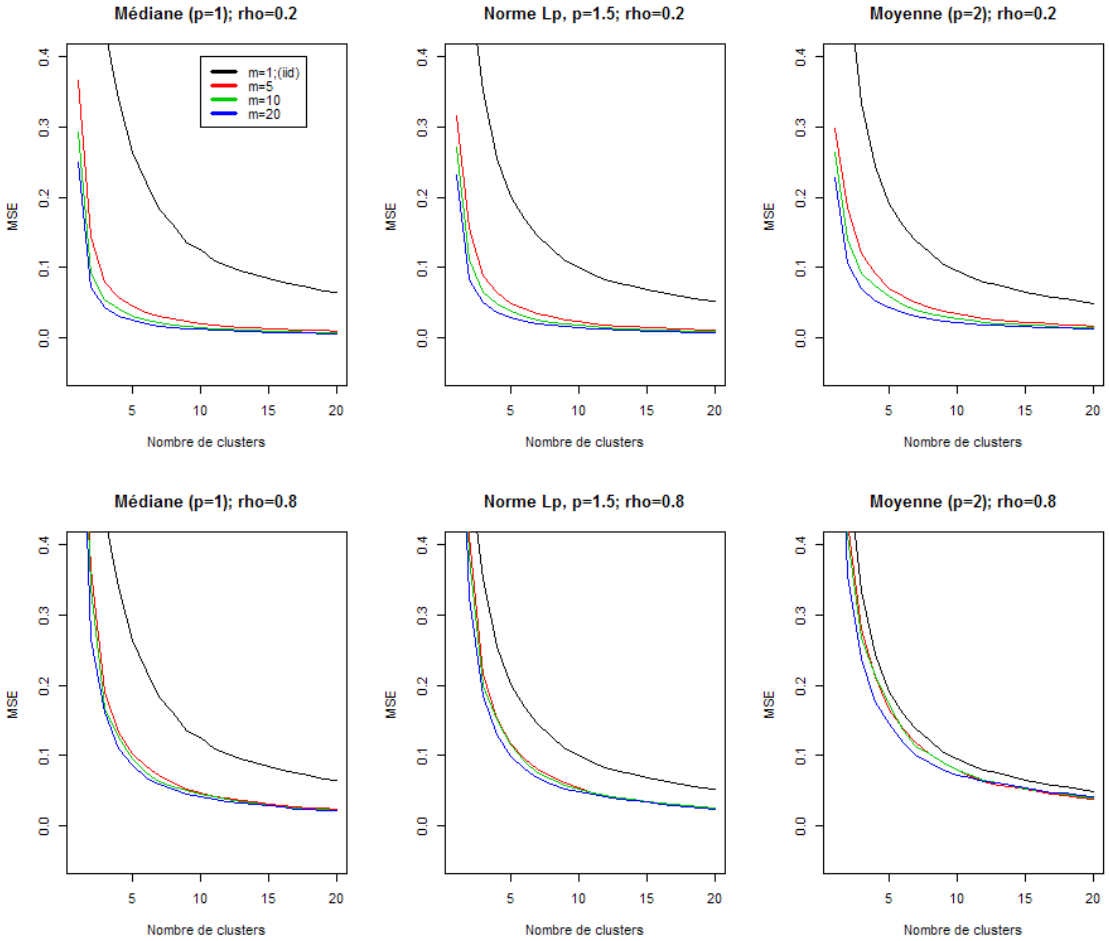

FIGURE 2.2 – *Influence de la taille des clusters et de la corrélation sur la convergence*

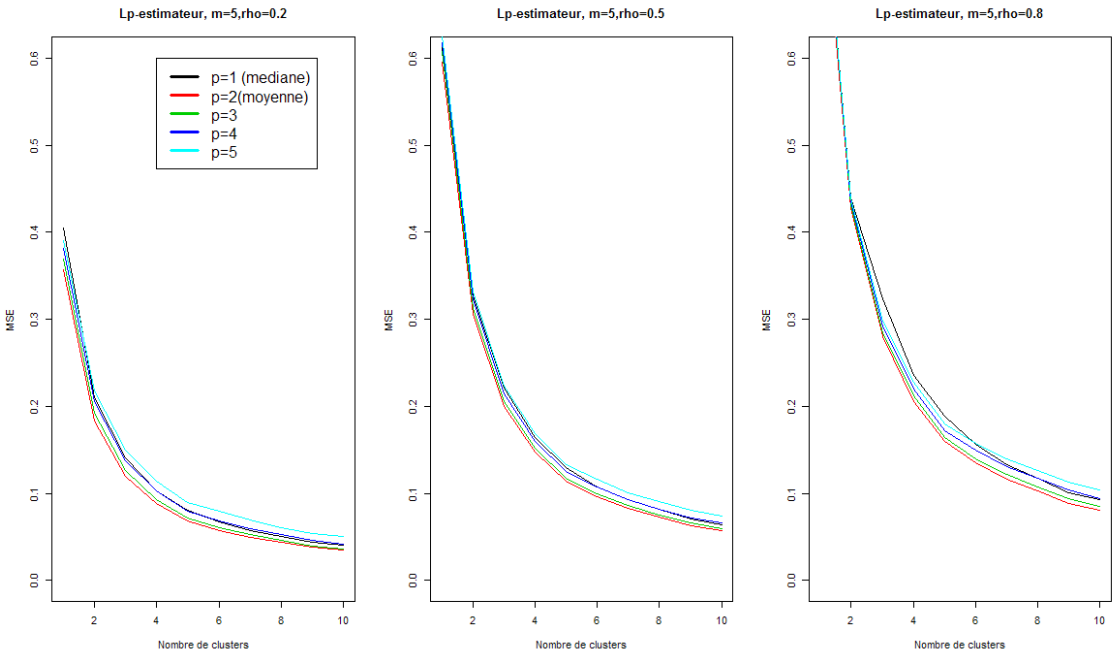

FIGURE 2.3 – *Convergence des Lp-estimateurs $p = 1 - 5$ dans le cas d'une loi normale centrée avec de gauche à droite $\rho = 0.2, 0.5, 0.8$*

On estime $\theta = 0$ la moyenne de la loi normale bivariée. L'erreur quadratique converge vers 0 et cette convergence est plus au moins rapide en fonction des paramètres et les estimateurs choisis (figure 2.2). Nous prenons comme référence le cas $m = 1$ (où le nombre de valeurs est simplement égal au nombre de clusters), ce qui permet de voir le comportement de l'estimateur dans un cas i.i.d. On voit que pour un nombre de clusters suffisamment grand, la taille des clusters a peu d'influence sur la convergence, notamment avec une valeur de corrélation importante. Cependant, pour un plus petit nombre de clusters avec une faible corrélation la convergence est d'autant meilleure que la taille m est importante. On note également qu'avec l'augmentation de la valeur de p, le comportement de l'erreur quadratique moyenne des estimateurs se dégrade en présence d'une forte corrélation et se rapproche de l'erreur quadratique moyenne du cas de référence (comportant moins de valeurs pour l'estimation).

Dans le cas d'une loi normale bivariée les Lp-estimateurs ($p = 1 - 5$) convergent (figure 2.3) et la convergence est plus rapide en diminuant p ($p \geq 2$) quelque soit la corrélation ρ choisie (figure 2.4).

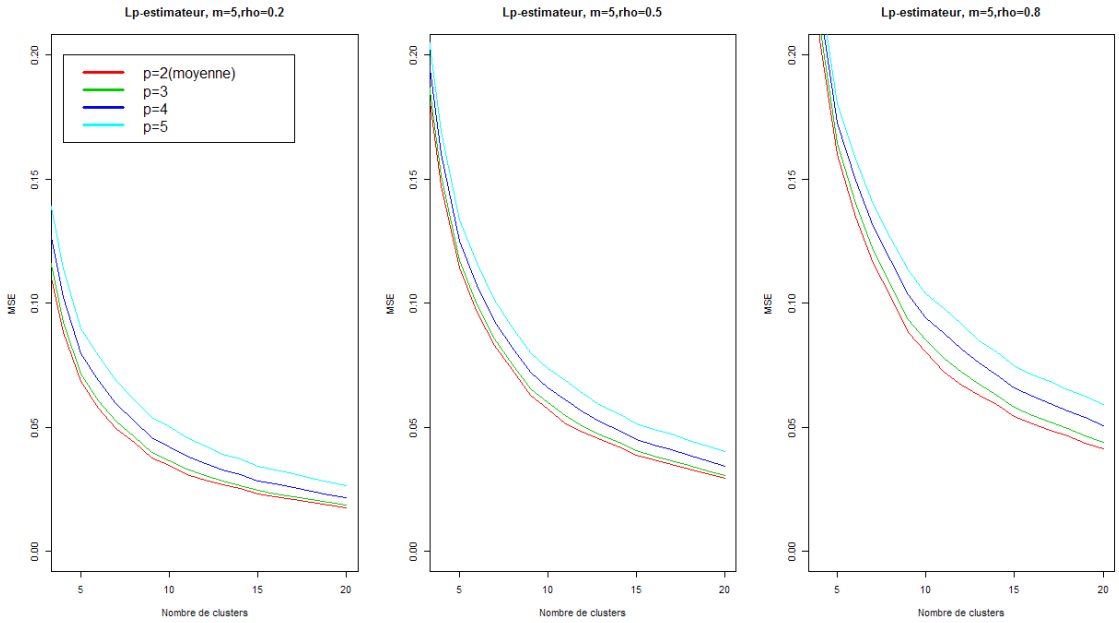

FIGURE 2.4 – *Agrandissement pour la convergence des Lp-estimateurs $p = 2 - 5$ dans le cas d'une loi normale centrée avec $\rho = 0.2, 0.5, 0.8$*

La fonction ψ pour les Lp-estimateurs est $\psi(x, \theta) = -p \|x - \theta\|^{p-2} (x - \theta)$. Pour la normalité asymptotique on a besoin que l'hypothèse 2.3.1a) soit vérifiée, c'est-à-dire $E_\theta(\|x - \theta\|^{2p-2+\eta}) < \infty$. Dans le cas de la loi normale, cette hypothèse est vérifiée avec l'ensemble des hypothèses de convergence (figures 2.2, 2.3 et 2.4) ainsi que dans le cas de la médiane où fonction ψ est bornée. Par contre, dans le cas d'une loi de Student bivariée de ν degrés de liberté, l'hypothèse 2.3.1a) ou l'hypothèse 2.2.1b) voire même l'hypothèse 2.2.1a) n'est pas automatiquement vérifiable. En effet, les moments qui existent dans ce cas sont ceux d'ordre inférieur strictement à ν. Pour les Lp-estimateurs avec $p > 1$, il faut que $2p - 2 < \nu$ pour que l'hypothèse 2.3.1a) soit vérifiée, $p < \nu$ pour que l'hypothèse 2.2.1a) et $2p < \nu$ pour l'hypothèse 2.2.1b). Prenons l'exemple de $\nu = 3$, dans ce cadre les Lp-estimateurs avec p=3, 4, 5 ne semblent plus converger (figure 2.5).

Si nous prenons une loi de Student avec $\nu = 9$, les hypothèses sont vérifiées pour les Lp-estimateurs avec $p \leq 4$, et donc ils sont convergent d'après nos théorèmes de convergence comme le montre la figure 2.6 pour $p = 3, 4, 5$. On remarque également qu'on a une convergence du Lp-estimateur $p = 5$ même si l'hypothèse 2.2.1b) n'est pas vérifiée ($2p > 9$), cela montre bien que cette hypothèse et suffisante mais pas nécessaire.

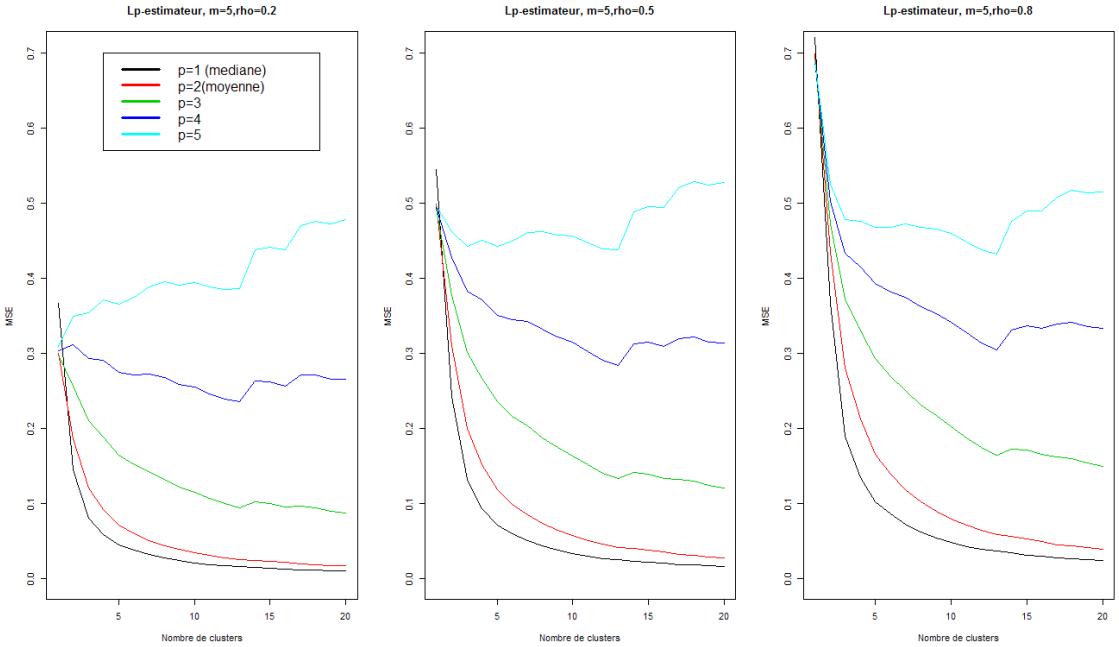

FIGURE 2.5 – *Convergence des Lp-estimateurs* $p = 1 - 5$ *dans le cas d'une loi Student* $\nu = 3$ *et de correlation intra-cluster* $\rho = 0.2, 0.5, 0.8$

On peut formuler la même remarque pour $p = 2$ dans le cas d'une loi de Student avec $\nu = 3$.

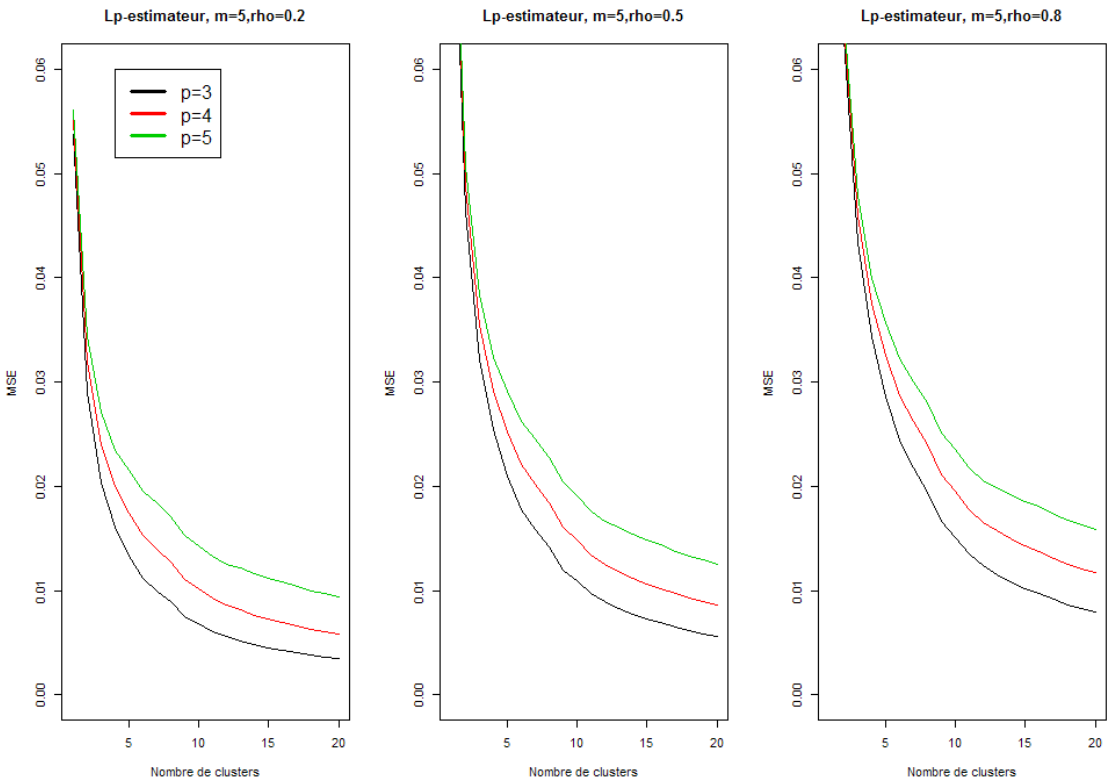

FIGURE 2.6 – *La convergence des Lp-estimateurs* $p = 3 - 5$ *dans le cas d'une loi Student* $\nu = 9$ *et de corrélation intra-cluster* $\rho = 0.2, 0.5, 0.8$

Chapitre 3

M-estimateurs pondérés pour des données clusterisées

Sommaire

3.1 Présentation des M-estimateurs pondérés **59**

3.2 Propriétés asymptotiques **60**

3.2.1 Convergence des estimateurs 60

3.2.2 Normalité asymptotique 63

3.3 Efficacité relative d'un M-estimateur pondéré **68**

3.3.1 Estimation de la variance 68

3.3.2 Cadre des simulations 70

3.3.3 Résultats pour l'optimisation des poids 71

3.3.4 Résultats pour l'efficacité 71

Résumé

Nous étudions la classe des M-estimateurs pondérés, qui est une variante des M-estimateurs vus dans le chapitre 2. Nous donnons également des conditions suffisantes pour établir la convergence ainsi que la normalité asymptotique et nous réalisons une étude comparative par rapport au M-estimateur non pondérés par le moyen de l'efficacité des estimateurs.

3.1 Présentation des M-estimateurs pondérés

La classe des M-estimateurs pondérés consiste à attribuer des poids dans la fonction objective des M-estimateurs. Cela peut, par exemple, permettre d'adapter un M-estimateur au cas de données clusterisées, avec un choix de poids favorisant certains clusters pour une raison de taille ou de valeurs prises. Cela peut permettre également d'améliorer l'estimateur, par exemple, en réduisant sa variance. Un des exemples de M-estimateurs est la médiane spatiale pondérée. Elle est traitée dans Nevalainen et al. (2007b) où les poids sont calculés numériquement pour réduire la variance de l'estimateur non pondéré.

Nous considérons de nouveau n clusters indépendants de loi P_θ, avec les mêmes hypothèses portant sur $\theta \in \Theta$. Nous utilisons les mêmes notations : ainsi X_{ij} est le j-ième élément du cluster i (comprenant m_i vecteurs aléatoires).

Pour chaque cluster i, la corrélation intra-cluster est encore supposée la même, l'hypothèse générale étant que $(X_{i1}, X_{i2}) \stackrel{d}{=} (X_{ik}, X_{ik'})$ pour tout $i = 1, \ldots, n$ et pour tout $k, k' = 1, .., m_i$, avec $k \neq k'$. On note $N_n := \sum_{i=1}^{n} m_i$ le nombre total d'éléments vérifiant $\lim_{n\to\infty} \frac{N_n}{n} = l,\ l \in]0, \infty[$. Enfin, on note encore $M(a) = E_\theta(\rho(X_{11}, a))$ et on suppose que θ vérifie

$$\underset{a\in\Theta}{\operatorname{argmin}} M(a) = \underset{a\in\Theta}{\operatorname{argmin}} \int \rho(x, a) dP_\theta,$$

où pour tout $a \in \Theta$, où la fonction ρ remplit les mêmes conditions générales qu'au chapitre 2.

Un M-estimateur pondéré associé à la fonction ρ est défini par :

$$\hat{\theta}_n^w = \underset{a\in\Theta}{\operatorname{argmin}}\ M_n^w(a)$$

avec $M_n^w(a) = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} \rho(X_{ij}, a)$ où les w_{ij} sont des poids strictement positifs choisis par le statisticien. Si $\rho(x, a)$ est différentiable dans un voisinage de θ, alors, pour $a = (a_1, \ldots, a_d)^T$ et $\psi = (\frac{\partial\rho}{\partial a_1}, \ldots, \frac{\partial\rho}{\partial a_d})^T$ le vecteur de dérivées partielles de ρ, on a $E_\theta(\psi(X_{11}, \theta)) = 0$.

L'estimateur $\hat{\theta}_n^w$ est alors la valeur de a vérifiant, également, les relations vectorielles :

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} \psi(X_{ij}, a) = 0.$$

3.2 Propriétés asymptotiques

3.2.1 Convergence des estimateurs

Les hypothèses qui permettent d'établir la convergence sont les suivantes :

Hypohèses 3.2.1

a) $\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} = 1$;

b) $\sum_{i\geq 1} \frac{w_{i.}^2}{i^2} < \infty$, *avec* $w_{i.} = \sum_{j=1}^{m_i} w_{ij}$.

Selon les cas d'étude, on peut prendre des valeurs de pondération qui leurs sont adaptées, par exemple, le même poids pour tous les éléments d'un cluster donné et un poids spécifique pour chaque cluster ; on pose dans ce cas $w_{ij} = w_i$. Si $\lim_{n\to\infty} w_n = 1$, alors d'après le théorème de Cesàro généralisé $\lim_{n\to\infty} \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} m_i w_i = \lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} m_i w_i = 1$, les hypothèses 3.2.1 sont vérifiées. De même, le choix $w_i = \frac{l}{m_i}$ entraîne clairement ces conditions.

Le premier des résultats pour les M-estimateurs pondérés, qu'on annoncera dans ce chapitre, est l'équivalent de la convergence presque sûre établie au théorème 2.2.2 pour les M-estimateurs non pondérés. Nous reprenons les hypothèses 2.2.1 données en page 36.

Théorème 3.2.1

Sous les hypothèses 2.2.1a)-b) et 3.2.1, si pour tout x la fonction $a \mapsto \rho(x,a)$ est $k(x)$-Lipschitzienne, avec $E_\theta(k^2(X_{11})) < \infty$, alors $\hat{\theta}_n^w \xrightarrow[n\to\infty]{p.s} \theta$.

Démonstration

La démonstration est basée sur l'équivalence :

$$\forall \epsilon > 0,\ \lim_{n\to\infty} P(\|\hat{\theta}_m^w - \theta\| < \epsilon, \forall m > n) = 1 \Leftrightarrow \hat{\theta}_n^w \xrightarrow[n\to\infty]{p.s} \theta.$$

De manière similaire à l'inclusion (2.5), pour tout $\epsilon > 0$ il existe $\eta > 0$ tel que

$$\left\{M(\theta) \leq M(\hat{\theta}_n^w) \leq M(\theta) + \eta\right\} \subset \left\{\|\hat{\theta}_n^w - \theta\| \leq \epsilon\right\}.$$

Comme dans la démonstration du théorème 2.2.1 on établit la majoration :

$$0 \leq M(\hat{\theta}_n^w) - M(\theta) \leq 2 \sup_{a \in \Theta} |M_n^w(a) - M(a)|.$$

Ainsi la convergence uniforme presque sûre de $M_n^w(a)$ implique la convergence presque sûre de l'estimateur. Comme dans le théorème 2.2.2, on détermine cela en deux lemmes prenant en compte la pondération.

Lemme 3.2.2 *Sous les condition du théorème 3.2.1, $M_n^w(a)$ converge simplement vers $M(a)$.*

Preuve On commence par vérifier les conditions de Kolmogorov$*$. Les X_{ij} ont la même loi P_θ, donc d'après l'inégalité de Cauchy-Schwarz on a

$$E_\theta(\rho(X_{ij'}, a)\rho(X_{ij}, a)) \leq E_\theta(\rho(X_{ij}, a)^2),$$

et on obtient :

$$\begin{aligned} E_\theta\left(\sum_{j=1}^{m_i} w_{ij}\rho(X_{ij}, a)\right)^2 &= \sum_{j=1}^{m_i} w_{ij}^2 E_\theta\left(\rho^2(X_{ij}, a)\right) + \sum_{j \neq j'}^{m_i} w_{ij} w_{ij'} E_\theta\left(\rho(X_{ij}, a)\rho(X_{ij'}, a)\right) \\ &\leq \left(\sum_{j=1}^{m_i} w_{ij}\right)^2 E_\theta\left(\rho^2(X_{ij}, a)\right). \end{aligned}$$

On pose $Y_i = \sum_{j=1}^{m_i} w_{ij}\rho(X_{ij}, a)$, les (Y_i) sont des variables aléatoires indépendantes d'espérance $\mu_i = w_{i.}E_\theta(\rho(X_{11}, a))$ où $w_{i.} = \sum_{j=1}^{m_i} w_{ij}$, et :

$$\sum_{i=1}^{\infty} \frac{Var(Y_i)}{i^2} \leq \sum_{i=1}^{\infty} \frac{w_{i.}^2}{i^2} E_\theta\left(\rho^2(X_{ij}, a)\right)$$

qui est finie grâce à l'hypothèse 3.2.1b). D'après le théorème de Kolmogorov$*$, on obtient pour tout $a \in \Theta$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\rho(X_{ij}, a) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}M(a) \xrightarrow[n\to\infty]{p.s} 0.$$

On conclut la démonstration grâce à la condition $\lim_{n\to\infty} \frac{N_n}{n} = l,\ l \in]0, \infty[$ et $\lim_{n\to\infty} \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} = 1$, et la condition 3.2.1a).

Lemme 3.2.3 *Sous les conditions du théorème 3.2.1, $M_n^w(a)$ converge uniformément vers $M(a)$.*

Preuve Nous exploitons de nouveau le fait que l'ensemble $\overline{\Theta}$ soit un compact et donc pour tout $\epsilon > 0$, il existe $h_\epsilon > 0$ et $a_1, \ldots, a_r$ dans Θ tels que $\Theta \subset \bigcup_{i=1}^{r} B(a_i, h_\epsilon)$, avec $B(a_i, h_\epsilon)$ la boule de centre a_i et de rayon h_ϵ. Ainsi nous avons, pour tout $a \in \Theta$:

$$|M_n^w(a) - M(a)| \leq |M_n^w(a) - M_n^w(a_i)| + |M_n^w(a_i) - M(a_i)| + |M(a_i) - M(a)|.$$

La fonction ρ est par hypothèse lipschitzienne, nous en déduisons la majoration :

$$|M_n^w(a) - M_n^w(a_i)| \leq \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} k(X_{ij}) \, \|a - a_i\|.$$

Comme $E_\theta(k(X_{ij})^2) < \infty$, on obtient d'après le critère de Kolmogorov$*$:

$$k_n = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} k(X_{ij}) \xrightarrow[n \to \infty]{p.s} E_\theta k(X_{ij}).$$

Il existe ainsi un Ω_1 vérifiant $P(\Omega_1) = 1$ et pour tout $\omega \in \Omega_1$, il existe k' et $N_0(\omega)$ tels que $\forall n \geq N_0(\omega)$, $k_n \leq k'$. De plus, la condition de Lipschitz sur $M(a)$ entraîne :

$$|M(a) - M(a_i)| \leq E_\theta k(X_{11}) \, \|a - a_i\|$$

avec a_i choisi tel que $\|a - a_i\| \leq h_\epsilon$. Enfin la convergence simple de M_n^w vers M entraîne que pour $1 \leq i \leq r$, il existe Ω_2 tel que $P(\Omega_2) = 1$, pour tout $\omega \in \Omega_2$ et pour tout $\epsilon > 0$, il existe $N'(\omega)$ tel que $\forall n > N'(\omega)$ on a :

$$|M_n^w(a) - M(a)| \leq k' h_\epsilon + \epsilon/3 + k' h_\epsilon$$

On choisit h_ϵ tel que $h_\epsilon < \frac{\epsilon}{3k'}$, et on déduit ainsi la convergence uniforme de $M_n^w(a)$. $\square$

De même que dans le cas non pondéré (théorème 2.2.2), le théorème 3.2.1 intègre les estimateurs avec une fonction lipschitzienne tels que la médiane spatiale pondérée et l'estimateur de Huber pondéré. L'espace des paramètres est tel que $\overline{\Theta}$ est compact, ainsi cette propriété intègre d'autres estimateurs comme la moyenne empirique.

3.2.2 Normalité asymptotique

Nous étudions dans cette section la normalité asymptotique des M-estimateurs pondérés. Nous introduisons les hypothèses nous permettant d'établir ce résultat.

Hypohèses 3.2.2

a) $\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}^2 = c_w$, *avec* $c_w < \infty$;

b) $\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{j'\neq j} w_{ij} w_{ij'} C_i = \mathbb{C}_w$, *avec* $C_i = E_\theta \psi(X_{ij},\theta)\psi^T(X_{ij'},\theta)$;

c) Il existe un réel $\eta > 0$ *tel que :*

$$\lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} w_{i.}^{2+\eta} < \infty.$$

Si on reprend le cas $w_{ij} = w_i$, avec $\lim_{n\to\infty} w_n = w = 1$, alors l'hypothèse 3.2.2a) est également vérifiée. Pour ce même type de pondération, si les corrélations intra-cluster sont les mêmes pour tous les clusters, et si $\lim_{n\to\infty} m_n = m' > 1$ alors $\mathbb{C}_w = (m'-1)C$ et la condition 3.2.2b) est vérifiée.

Pour établir la normalité, nous utilisons les hypothèses 2.2.1a)-b) et 2.3.1a) données respectivement en pages 36 et 40.

Théorème 3.2.4

Pour tout a dans un voisinage ouvert de θ*, nous reprenons les hypothèses du théorème 2.3.1 sur la fonction* $a \mapsto \psi(x,a)$*, ses dérivées et leurs moments. Sous les hypothèses 2.2.1a)-b), 2.3.1a), 3.2.1 et 3.2.2, nous obtenons :*

$$\sqrt{N_n}(\hat{\theta}_n^w - \theta) \xrightarrow[n\to\infty]{L} N\left(0, V_\theta^{-1}\left(c_w B + \mathbb{C}_w\right) V_\theta^{-1}\right),$$

avec $B := E_\theta \psi(X_{ij},\theta)\psi^T(X_{ij},\theta)$ *et* $V_\theta = E_\theta(\frac{\partial \psi(X_{ij},a)}{\partial a}/_{a=\theta})$.

Démonstration

Comme pour le théorème 2.2.2, la normalité asymptotique de $\hat{\theta}_n^w$ découle d'un développement de Taylor à l'ordre 2 de la fonction ψ. La démonstration peut se décomposer en plusieurs parties.

Lemme 3.2.5 *Sous les hypothèses du théorème 3.2.4, il existe dans un voisinage de* θ*,* $\tilde{\theta}_n^w$ *et une*

matrice $D(\hat{\theta}_n^w - \theta)$ *tels que :*

$$-\sqrt{N_n}T_n^w(\theta) = \left\{ \dot{T}_n^w(\theta)^T + \frac{1}{2}D(\hat{\theta}_n^w - \theta)\ddot{T}_n^w(\tilde{\theta}_n^w) \right\} \sqrt{N_n}(\hat{\theta}_n^w - \theta) \tag{3.1}$$

où on a noté $T_n^w(a) = \dfrac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\psi(X_{ij}, a)$.

Preuve : Nous reprenons la démonstration du lemme 2.3.2 ainsi que ses notations pour écrire le développement vectoriel suivant à partir de l'équation (2.12) : il existe un $\bar{\theta}$ tel que :

$$T_n^w(a) = T_n^w(\theta) + \dot{T}_n^w(\theta)^T(a - \theta) + \frac{1}{2}D(a - \theta)\ddot{T}_n^w(\bar{\theta})(a - \theta). \tag{3.2}$$

Nous appliquons (3.2) pour le M-estimateur pondéré $\hat{\theta}_n^w$ (par définition $T_n^w(\hat{\theta}_n^w) = 0$), alors il existe $\tilde{\theta}_n^w$ tel que :

$$-\sqrt{N_n}T_n^w(\theta) = \left\{ \dot{T}_n^w(\theta)^T + \frac{1}{2}D(\hat{\theta}_n^w - \theta)\ddot{T}_n^w(\tilde{\theta}_n^w) \right\} \sqrt{N_n}(\hat{\theta}_n^w - \theta). \tag{3.3}$$

□

Le lemme 3.2.5 nous conduit à démontrer les lemmes suivants :

Lemme 3.2.6 *Sous les hypothèses du théorème 3.2.4,* $\sqrt{N_n}T_n^w(\theta) \xrightarrow[n\to\infty]{L} N(0, c_w B + \mathbb{C}_w)$.

Lemme 3.2.7 *Sous les hypothèses du théorème 3.2.4,* $\dot{T}_n^w(\theta) \xrightarrow[n\to\infty]{p.s} V_\theta$ *et* $\ddot{T}_n^w(a) = O_p(1)$ *uniformément en a.*

Lemme 3.2.8 *Sous les hypothèses du théorème 3.2.4,* $\hat{\theta}_n^w \xrightarrow[n\to\infty]{p.s} \theta$.

Preuve du lemme 3.2.6 : On utilise dans cette démonstration le théorème Lindeberg$*$ qui permet d'obtenir cette convergence en loi. Les vecteurs $\xi_i^w := \sum_{j=1}^{m_i} w_{ij}\psi(X_{ij}, \theta)$ avec $i = 1, \ldots, n$, sont indépendantes d'espérances $E_\theta(\xi_i^w) = 0$ et de variances $V_i, i = 1, \ldots, n$. Pour chaque i, V_i est définie par :

$$V_i = \sum_{j=1}^{m_i} w_{ij}^2 E_\theta\psi(X_{ij}, \theta)\psi^T(X_{ij}, \theta) + \sum_{j\neq j'}^{m_i} w_{ij}w_{ij'}E_\theta\psi(X_{ij}, \theta)\psi^T(X_{ij'}, \theta).$$

Comme tous les couples $(X_{ij'}, X_{ij})$ avec $j \neq j'$ ont la même corrélation dans le cluster i, on pose $C_i = E_\theta\psi(X_{ij}, \theta)\psi^T(X_{ij'}, \theta)$ et $B := E_\theta\psi(X_{ij}, \theta)\psi^T(X_{ij}, \theta)$. Alors sous les

hypothèses 3.2.2a)-b) on a :

$$\frac{1}{N_n}\sum_{i=1}^{n} V_i \xrightarrow[n\to\infty]{} c_w B + \mathbb{C}_w, \tag{3.4}$$

comme $\lim_{n\to\infty} \frac{N_n}{n} = l$ alors on peut écrire :

$$\frac{1}{n}\sum_{i=1}^{n} V_i \xrightarrow[n\to\infty]{} l(c_w B + \mathbb{C}_w) := \Sigma.$$

On vérifie dans la suite que $E_n := \frac{1}{n}\sum_{i=1}^{n} E_\theta[\|\xi_i^w\|^2 1_{\|\xi_i^w\|>\epsilon\sqrt{n}}]$ tend vers 0, pour vérifier les conditions de Lindeberg$*$. Nous avons :

$$E_n = \frac{1}{n}\sum_{i=1}^{n} E_\theta[\|\xi_i^w\|^{-\eta}\,\|\xi_i^w\|^{2+\eta}\,1_{\|\xi_i^w\|>\epsilon\sqrt{n}}] \leq \sqrt{n}^{-\eta}\frac{1}{n}\sum_{i=1}^{n}(\epsilon)^{-\eta}E_\theta[\|\xi_i^w\|^{2+\eta}],$$

or par l'inégalité de Minkowski :

$$E_\theta\,\|\xi_i^w\|^{2+\eta} \leq \left[\sum_{j=1}^{m_i} w_{ij}\left[E_\theta\,\|\psi(X_{ij},\theta)\|^{2+\eta}\right]^{1/(2+\eta)}\right]^{2+\eta} = E_\theta\,\|\psi(X_{ij},\theta)\|^{2+\eta}\left[\sum_{j=1}^{m_i} w_{ij}\right]^{2+\eta}.$$

Sous les conditions 2.3.1a), 3.2.2c), et pour tout $\epsilon > 0$:

$$E_n \leq E_\theta[\|\psi(X_{ij},\theta)\|^{2+\eta}](\sqrt{n}\epsilon)^{-\eta}\frac{1}{n}\sum_{i=1}^{n} w_{i.}^{2+\eta} \xrightarrow[n\to\infty]{} 0.$$

Donc par conclusion $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i^w \xrightarrow{L} N\left(0, l(c_w B + \mathbb{C}_w)\right)$, ce qui entraîne la normalité asymptotique de $\sqrt{N_n}T_n^w(\theta)$ avec une matrice de covariance asymptotique donnée par $c_w B + \mathbb{C}_w$.

Preuve du lemme 3.2.7 : Nous avons :

$$\dot{T}_n^w(\theta) = \frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\dot{\psi}(X_{ij},\theta) \tag{3.5}$$

et

$$\ddot{T}_n^w(a) = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} \ddot{\psi}(X_{ij}, a). \tag{3.6}$$

Par hypothèse, les éléments de la matrice $\dot{\psi}(X_{ij}, \theta)$ sont de carré intégrable et de même pour les éléments de $\ddot{\psi}(X_{ij}, a)$. Nous vérifions le critère de Kolmogorov$*$ en suivant la procédure que nous avons utilisée pour montrer la convergence simple de $M_n^w(a) = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} \rho(X_{ij}, a)$, ainsi nous avons les convergences simples presque sûrement de l'équation (3.5). Pour tout a, les éléments de $\ddot{\psi}(X_{ij}, a)$ sont dominés par une fonction $F(X_{ij})$ de carré intégrable et indépendante du paramètre a, impliquant la vérification du critère de Kolmogorov pour la suite $\sum_{j=1}^{m_i} w_{ij} F(X_{ij})$ et par suite, nous avons $\ddot{T}_n^w(a) = O_p(1)$ uniformement en a. □

Preuve du lemme 3.2.8 : Nous montrons que la fonction $a \mapsto \rho(x, a)$ est $\sup_{a \in \Theta} \|\psi(x, a)\|$-lipschitzienne, et que $\sup_{a \in \Theta} \|\psi(x, a)\|$ est de carré intégrable en reproduisant les étapes de la démonstration du lemme 2.3.5. On vérifie au final les hypothèses du théorème 3.2.1, ce qui nous mène au résultat annoncé. □

En prenant la version pondérée de l'exemple 2.3.1, nous obtenons la convergence en loi de $\sqrt{N_n}(\hat{\theta}_n^w - \theta)$ vers une loi normale $N(0, c_w I_\theta^{-1} + I_\theta^{-1} \mathbb{C}_w I_\theta^{-1})$, où I_θ est l'information de Fisher. En travaillant sur un voisinage de θ, l'estimateur de Huber pondéré a le comportement de la moyenne empirique pondérée, qui vérifie les conditions du théorème, ainsi que plus généralement les Lp-estimateurs pondérés avec $p \geq 2$.

La régularité supposée pour la fonction ψ dans le théorème 3.2.4 exclut des estimateurs de fonction objective. La médiane spatiale pondérée, qui a une fonction objectif non différentiable, est abordée par Nevalainen et al. (2007a). Le théorème suivant propose un résultat de convergence pour un M-estimateur pondéré ne supposant pas l'existence de $\ddot{\psi}$, mais impose une condition plus forte sur $\dot{\psi}$.

Théorème 3.2.9 *Soit θ un paramètre de localisation d'une loi continue par rapport à la mesure de Lebesgue. On suppose que θ est l'unique zéro de la fonction $a \mapsto E_\theta(\psi(X_{11} - a))$. Nous supposons également que la fonction $a \mapsto \psi(x - a)$ est k - Hölderienne (uniformément en x). De plus $E_\theta(\frac{\partial \psi(X_{11} - a)}{\partial a}\Big|_{a=\theta})$ existe et est supposée inversible. Alors, sous les hypothèses 2.3.1a), 3.2.1 et 3.2.2 :*

1. $\hat{\theta}_n^w \xrightarrow[n \to \infty]{p.s} \theta$;

2. $\sqrt{N_n}(\hat{\theta}_n^w - \theta) \xrightarrow[n\to\infty]{L} N\left(0, V_\theta^{-1}\left(c_w B + \mathbb{C}_w\right) V_\theta^{-1}\right)$

Démonstration

Nous avons, à partir de l'unicité de θ, que pour tout $\epsilon > 0$

$$\inf_{\|a-\theta\|>\epsilon} \|E_\theta \psi(X_{11} - a)\| > \|E_\theta \psi(X_{11} - \theta)\| = 0. \tag{3.7}$$

Le paramètre θ peut être également défini par $\theta = \underset{a\in\Theta}{\operatorname{argmin}} \|E_\theta \psi(X_{11} - a)\|$. On reprend la démonstration du théorème 3.2.1 en remplaçant l'équation (3.7) par l'hypothèse 2.2.1a), $\|E_\theta \psi(X_{11} - a)\|$ par $M(a)$, et $\left\|\frac{1}{N_n}\sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}\psi(X_{ij}, a)\right\|$ par $M_n^w(a)$. On se retrouve dans les conditions du théorème 3.2.1(la condition de Lipschitz peut être assouplie par une régularité Hölderienne), ainsi $\hat{\theta}_n^w \xrightarrow[n\to\infty]{p.s} \theta$.

D'après un développement de Taylor avec reste intégral, nous obtenons la majoration : pour tout a dans un voisinage de θ

$$\psi(x-a) = \psi(x-\theta) + \dot{\psi}(x-\theta)^T(a-\theta) + R(x,\theta)(a-\theta),$$

avec $|R(x,\theta)| = o(1)$ uniformément en x. Prenons $a = \hat{\theta}_n^w \xrightarrow[n\to\infty]{p.s} \theta$ et $x = X_{ij}$:

$$\psi(X_{ij} - \hat{\theta}_n^w) = \psi(X_{ij} - \theta) + \dot{\psi}(X_{ij} - \theta)(\hat{\theta}_n^w - \theta) + (\hat{\theta}_n^w - \theta)o_p(1) \tag{3.8}$$

Introduisons la somme empirique :

$$\frac{1}{N_n}\sum_{i=1}^n\sum_{j=1}^{m_i} w_{ij}\psi(X_{ij} - \hat{\theta}_n^w) = \frac{1}{N_n}\sum_{i=1}^n\sum_{j=1}^{m_i} w_{ij}\psi(X_{ij} - \theta) + \frac{1}{N_n}\sum_{i=1}^n\sum_{j=1}^{m_i} w_{ij}\dot{\psi}(X_{ij} - \theta)(\hat{\theta}_n^w - \theta) + (\hat{\theta}_n^w - \theta)o_p(1) \tag{3.9}$$

Nous conservons les notations utilisées dans la démonstration du théorème 3.2.4, et nous obtenons le développement :

$$-\sqrt{N_n}T_n^w(\theta) = \left(\dot{T}_n^w(\theta) + o_p(1)\right)\sqrt{N_n}(\hat{\theta}_n^w - \theta). \tag{3.10}$$

Sous les hypothèses 3.2.2 on applique le lemme 3.2.6 pour avoir

$$\sqrt{N_n}T_n^w(\theta) \xrightarrow[n\to\infty]{L} N\left(0, c_w B + \mathbb{C}_w\right).$$

La fonction $\psi(x,a)$ est $k(x)$ - Lipschitzienne, les éléments de la matrice $\dot{\psi}(X_{ij}-\theta)$ sont donc dominés par $k(x)$ qui est de carré intégrable par hypothèse. Dans ce cas, on se retrouve sous les conditions du lemme 3.2.7, nous avons, par conclusion,
$\dot{T}_n^w(\theta) \xrightarrow[n\to\infty]{p.s} V_\theta$ $\square$.

3.3 Efficacité relative d'un M-estimateur pondéré

Dans ce paragraphe, nous étudions l'efficacité des M-estimateurs pondérés par rapport à leurs versions non pondérées. Notre objectif est de montrer l'influence des poids sur la matrice de variance-covariance des estimateurs et donc sur leurs précisions. La variance $V_\theta^{-1}\left(c_w B + \mathbb{C}_w\right) V_\theta^{-1}$ peut être comparée avec celle trouvée dans le cas non pondéré $V_\theta^{-1}\left(B + c_m C\right) V_\theta^{-1}$ à travers la mesure d'efficacité relative suivante :

$$E_f = \left[\frac{det(V_\theta^{-1}\left(c_w B + \mathbb{C}_w\right) V_\theta^{-1})}{det(V_\theta^{-1}\left(B + c_m C\right) V_\theta^{-1})}\right]^{\frac{1}{d}} \tag{3.11}$$

Dans les simulations, nous utilisons les estimateurs de la variance. Dans la section qui suit, nous proposons des estimateurs pondérés des matrices de la variance.

3.3.1 Estimation de la variance

Nous nous plaçons ici dans les conditions des théorèmes de la normalité asymptotique afin d'établir la convergence asymptotique des estimateurs de la matrice de variance-covariance d'un M-estimateur pondéré. On suppose également que la même importance est donnée aux variables d'un cluster, cela revient à distribuer les poids par cluster (non par individu) et les poids sont donc identiques au sein d'un même cluster ($w_{ij} \equiv w_i$). Dans ce cadre d'étude, la matrice $\mathbb{C}_w$ s'écrit :

$$\mathbb{C}_w = \lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} w_i^2 m_i (m_i - 1) C_i. \tag{3.12}$$

Si on fait de plus l'hypothèse que la corrélation intra-cluster est la même pour tous les clusters, $C_i \equiv C$, la variance asymptotique V_w de $\sqrt{N_n}(\hat{\theta}_n^w - \theta)$ s'écrit

$$V_w = V_\theta^{-1}\left(c_w B + c_{w_2} C\right) V_\theta^{-1} \text{ avec } c_{w_2} = \lim_{n\to\infty} \frac{1}{N_n} \sum_{i=1}^{n} w_i^2 m_i (m_i - 1).$$

Cette variance dépend des poids uniquement par les termes c_w et c_{w_2}. Les matrices V_θ, B et C n'étant pas fonction des poids, nous pouvons utiliser les estimateurs proposés dans le théorème 2.4.1 pour lesquels la convergence presque sûre est établie (sous l'hypothèse des poids bornés et d'une loi intra-cluster identique pour tous les clusters de même taille).

Une méthode alternative est d'estimer directement la matrice de variance-covariance à partir des versions empiriques des matrices qui la composent et d'établir l'analogue du théorème 2.4.2 pour ce cadre pondéré. On n'effectue que l'hypothèse $w_{ij} \equiv w_i$ et pour l'estimation de $c_w B$, on note $\hat{B}_w$ la matrice définie par :

$$\hat{B}_w = \frac{1}{N_n} \sum_{i=1}^{n} w_i^2 \sum_{j=1}^{m_i} \psi(X_{ij}, \theta)\psi^T(X_{ij}, \theta).$$

De la même façon, nous définissons les estimateurs de C_w et V_θ respectivement par :

$$\hat{C}_w = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i^2 \sum_{j' \neq j}^{m_i} \psi^T(X_{ij'}, \theta)\psi(X_{ij}, \theta)$$

et

$$\hat{V}_\theta^{\,w} = \frac{1}{N_n} \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} \frac{\partial \psi^T(X_{ij}, a)}{\partial a}\Big|_{a=\theta}.$$

Par une démonstration totalement analogue à celle effectuée pour le théorème 2.4.2, nous obtenons ainsi le résultat suivant :

Lemme 3.3.1 *Dans les conditions du théorème 3.2.4 (ou du théorème 3.2.9), et si de plus*

1) $\sum_{i \geq 1} \frac{w_i^4 m_i^2}{i^2} < \infty$ *et* ψ *admet un moment d'ordre 4, alors* $\hat{B}_w$ *converge presque sûrement vers* $c_w B$.

2) $\sum_{i \geq 1} \frac{w_i^4 m_i^4}{i^2} < \infty$ *et* ψ *admet un moment d'ordre 4, alors* $\hat{C}_w$ *converge presque sûrement vers* $\mathbb{C}_w$.

3) $\sum_{i \geq 1} \frac{w_i^2 m_i^2}{i^2} < \infty$ *et la matrice* $\frac{\partial \psi(X_{ij}, a)}{\partial a}\Big|_{a=\theta}$ *admet un moment d'ordre 2, alors* $\hat{V}_\theta^w$ *converge presque sûrement vers* V_θ.

Remarquons qu'en renforçant les hypothèses sur les lois intra-clusters, notamment en considérant que ces lois ne sont définies que par l'espérance et la matrice de covariance, nous pouvons assouplir les conditions précédentes sur les moments en établissant un

résultat analogue à celui du théorème 2.4.1 sous l'hypothèse d'une corrélation identique pour chaque cluster et de l'existence d'un moment d'ordre 2 pour ψ. Cela sera notamment le cas dans l'étude de simulation, où nous utiliserons les deux versions pondérées et non pondérées des estimateurs de la matrice de variance-covariance pour estimer l'efficacité relative définie en (3.11).

3.3.2 Cadre des simulations

L'efficacité relative d'un M-estimateur pondéré relativement à sa version non pondéré est donnée par l'équation (3.11). Dans nos simulations, nous utilisons sa version empirique $\hat{E}_f$ qui est fonction des estimateurs des variances. Nous considérons la fonction $f(w) = det(\hat{V}_w)$, où $\hat{V}_w$ est un estimateur de V_w. Notre premier objectif sera donc de trouver les poids qui minimisent cette fonction.
Les simulations seront réalisées pour quatre configurations de 10 clusters, et pour chacune de ces configurations nous aurons 100 réalisations de vecteurs aléatoires (v.a.) :

1. C1 : 9 clusters de 4 vecteurs aléatoires et un cluster contenant 64 v.a..
2. C2 : 5 clusters de 4 v.a. et 5 clusters de 16 v.a..
3. C3 : 2 clusters de 4 v.a., un cluster de 8 v.a. et 7 clusters de 12 v.a..
4. C4 : 10 clusters comprenant respectivement $\{5, 6, 7, 8, 9, 11, 12, 13, 14, 15\}$ v.a..

Nous utilisons des modèles bivariés, avec deux composantes indépendantes, présentant la même corrélation intra-cluster pour tous les clusters, et de loi $\mathcal{L}$ définie par son espérance et sa matrice de covariance, avec donc

$$Cov(X_{ij}, X_{ij'}) = \rho \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ pour tout } j \neq j' \text{ et } i = 1, \ldots, 10$$

et

$$Cov(X_{ij}, X_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ pour tout } j \text{ et } i = 1, \ldots, 10.$$

avec $\rho \in]0, 1[$ ($\rho = 0.2$ ou 0.8 pour les résultats présentés).
Nous simulons des v.a. de loi $\mathcal{L}$ normale ou de Student avec ν degrés de liberté où ν est choisi parmi les valeurs $\{1, 3, 9\}$ (la valeur $\nu = 1$ correspondant à une loi de Cauchy). Les simulations et les calculs des estimateurs sont réalisés sous le logiciel R

et l'optimisation de la fonction $f(w)$ avec le logiciel *Matlab*. Les M-estimateurs choisis pour l'étude sont : la médiane, la moyenne empirique, l'estimateur de Huber et la L_p-médiane avec $p = 3, 4, 5, 6$.

3.3.3 Résultats pour l'optimisation des poids

Nous rappelons que par hypothèse les poids des variables aléatoires d'un même cluster sont identiques ($w_{ij} = w_i$). À l'issue de l'optimisation, on obtient des poids de même ordre de grandeur entre les clusters de même taille quelque soit l'estimateur et la corrélation considérés (tableaux 3.1 à 3.4). De plus, on observe que le poids des variables d'un cluster est chaque fois décroissant en fonction de la taille de ce dernier. Enfin, à corrélation égale, les poids choisis pour l'estimateur de Huber et l'estimateur L_3 semblent avoir le même ordre de grandeur.

Tableau 3.1 *– Poids optimaux pour l'estimateur de Huber pondéré dans le cas d'une loi normale bivariée centrée avec $\rho = 0.2$*

C1		C2		C3		C4	
m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$
4	2,2682	4	1,8753	4	1,7597	5	1,5668
4	2,3430	4	1,9372	4	1,8160	6	1,4386
4	2,2749	4	1,8810	8	1,1985	7	1,3007
4	2,2158	4	1,8319	12	0,9021	8	1,1710
4	2,2587	4	1,8677	12	0,8943	9	1,0894
4	2,3377	16	0,7971	12	0,9215	11	0,9779
4	2,2894	16	0,7758	12	0,9095	12	0,9024
4	2,2421	16	0,7525	12	0,8790	13	0,8293
4	2,3518	16	0,8097	12	0,9397	14	0,8329
64	0,2761	16	0,7667	12	0,8964	15	0,7487

3.3.4 Résultats pour l'efficacité

En analysant les résultats obtenus dans le cas d'une loi normale bivariée centrée avec $\rho = 0.2$ (tableau 3.5), nous remarquons que, quelque soit la configuration des clusters, l'efficacité relative de tous les M-estimateurs pondérés de manière optimale, est améliorée par rapport à leur version non pondérée. On remarque néanmoins une variation de cette amélioration selon les configurations des clusters. L'efficacité des estimateurs de la première configuration ($C1$) est meilleure que les autres, suivie (dans l'ordre décroissant des résultats d'efficacité) de $C2$, $C4$ et $C3$. En effet, on obtient, par

Tableau 3.2 *– Poids optimaux pour l'estimateur de Huber pondéré dans le cas d'une loi normale bivariée centrée avec* $\rho = 0.8$

C1		C2		C3		C4	
m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$
4	2,5196	4	2,4701	4	2,4466	6	1,6592
4	2,4675	4	2,4191	8	1,2394	7	1,4105
4	2,4505	4	2,4024	12	0,8303	8	1,2274
4	2,4638	4	2,4154	12	0,8354	9	1,1043
4	2,5220	16	0,6503	12	0,8545	11	0,9279
4	2,5063	16	0,6470	12	0,8497	12	0,8483
4	2,4408	16	0,6314	12	0,8289	13	0,7645
4	2,5627	16	0,6618	12	0,8694	14	0,7461
64	0,1626	16	0,6287	12	0,8257	15	0,6625

Tableau 3.3 *– Poids optimaux pour le Lp estimateur (p=3) pondéré dans le cas d'une loi normale bivariée centrée avec* $\rho = 0.2$

C1		C2		C3		C4	
m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$
4	2,3144	4	1,8974	4	1,7875	6	1,4783
4	2,3307	4	1,9108	8	1,2087	7	1,2922
4	2,3110	4	1,8939	12	0,8998	8	1,1918
4	2,1451	4	1,7586	12	0,8606	9	1,0360
4	2,2406	16	0,7766	12	0,9232	11	0,9636
4	2,2707	16	0,7859	12	0,9108	12	0,9018
4	2,0954	16	0,7344	12	0,8542	13	0,7884
4	2,5505	16	0,8623	12	1,0099	14	0,8889
64	0,2875	16	0,7867	12	0,9218	15	0,7791

Tableau 3.4 *– Poids optimaux pour le Lp estimateur (p=3) pondéré dans le cas d'une loi normale bivariée centrée avec* $\rho = 0.8$

C1		C2		C3		C4	
m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$	m_i	$w_i * 100$
4	2,6321	4	2,5772	4	2,5538	6	1,7484
4	2,5378	4	2,4848	8	1,2736	7	1,4423
4	2,5125	4	2,4600	12	0,8487	8	1,2563
4	2,2735	4	2,2261	12	0,7720	9	1,0176
4	2,4241	16	0,6290	12	0,8267	11	0,8970
4	2,4567	16	0,6346	12	0,8330	12	0,8310
4	2,2610	16	0,5866	12	0,7700	13	0,7108
4	2,8079	16	0,7235	12	0,9502	14	0,8142
64	0,1703	16	0,6591	12	0,8660	15	0,6943

exemple, pour l'estimateur de Huber pondéré une efficacité de 0.406 pour $C1$, 0.881 pour $C2$, 0.953 pour $C4$ et 0.964 pour $C3$. En mettant ces résultats en regard avec les caractéristiques des configurations des clusters, on note l'impact de la taille des clusters sur l'amélioration de la variance des estimateurs : il apparaît que plus l'écart entre les tailles des clusters au sein d'une configuration est grand, plus la pondération optimale des estimateurs permet de diminuer leur variance. De plus, le choix des estimateurs semble peu influer sur les valeurs trouvées pour E_f.

Tableau 3.5 *– Estimations et efficacité relative pour une loi normale centrée avec $\rho = 0.2$ et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+C}_w)\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
	Configuration C1		
Médiane	0,0019086	0,01003492	0,43611448
Moyenne	0,0014	0,0088	0,39886202
Huber	0,001503759	0,009105587	0,4063826
Lp-médiane ; p=3	0,001571769	0,008940633	0,41928581
Lp-médiane ; p=4	0,002209924	0,009740658	0,4763153
Lp-médiane ; p=5	0,003646194	0,011300737	0,56802375
	Configuration C2		
Médiane	0,0011	0,0014	0,88640526
Moyenne	0,000166115	0,001235924	0,36661301
Huber	0,001003217	0,001291184	0,8814616
Lp-médiane ; p=3	0,001062472	0,001347751	0,88787932
Lp-médiane ; p=4	0,001403585	0,001682487	0,91336335
Lp-médiane ; p=5	0,002158981	0,00239793	0,94886879
	Configuration C3		
Médiane	0,001044156	0,00111218	0,96893631
Moyenne	0,000844332	0,000913767	0,96125576
Huber	0,000892671	0,00096124	0,9636733
Lp-médiane ; p=3	0,000946128	0,001011128	0,96732374
Lp-médiane ; p=4	0,001238507	0,001301117	0,97564307
Lp-médiane ; p=5	0,001863664	0,001913617	0,98686178
	Configuration C4		
Médiane	0,001010169	0,001098677	0,95887501
Moyenne	0,000826841	0,000914741	0,95074022
Huber	0,000873324	0,000961087	0,95324882
Lp-médiane ; p=3	0,000926597	0,001010915	0,95738824
Lp-médiane ; p=4	0,001209292	0,001294527	0,96651799
Lp-médiane ; p=5	0,001804459	0,001883408	0,97881648

En augmentant la corrélation intra-cluster avec $\rho = 0.8$ dans le cas d'une loi normale bivariée centrée comme précédemment (tableau 3.6), l'efficacité des M-estimateurs pondérés est améliorée : lorsque la corrélation intra-cluster $\rho = 0.8$ pour l'estimateur d'Huber pondéré et dans le cas de la configuration $C1$ $E_f = 0.249$, alors que, pour $\rho = 0.2$, $E_f = 0.406$. Ce constat est le même pour les autres M-estimateurs pondérés, quelque soit la configuration. On remarque le même classement des configurations en termes d'efficacité que dans le cas où la corrélation intra-cluster avec $\rho = 0.2$.

Tableau 3.6 *– Estimations et efficacité relative pour une loi normale centrée avec $\rho = 0.8$ et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+\mathbf{C}}_w)\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
	Configuration C1		
Médiane	0,01122996	0,17281001	0,25492044
Moyenne	0,007272	0,11843232	0,24779451
Huber	0,00830953	0,13421632	0,24882018
Lp-médiane ; p=3	0,00821798	0,12911198	0,25228957
Lp-médiane ; p=4	0,01181483	0,17499006	0,2598406
Lp-médiane ; p=5	0,01929972	0,26215187	0,27133074
	Configuration C2		
Médiane	0,00921496	0,01596246	0,75979567
Moyenne	0,00689569	0,01223565	0,75071567
Huber	0,00756989	0,01338423	0,75205227
Lp-médiane ; p=3	0,00838464	0,01466955	0,75602115
Lp-médiane ; p=4	0,01351215	0,02329204	0,7616552
Lp-médiane ; p=5	0,02719672	0,04615815	0,76759843
	Configuration C3		
Médiane	0,00875783	0,01043083	0,91630239
Moyenne	0,00686948	0,00823936	0,91309324
Huber	0,00754503	0,00904854	0,91314786
Lp-médiane ; p=3	0,00831216	0,00988204	0,91713604
Lp-médiane ; p=4	0,01317475	0,01557201	0,91981147
Lp-médiane ; p=5	0,02562071	0,0300795	0,92291205
	Configuration C4		
Médiane	0,00866745	0,01048374	0,90925888
Moyenne	0,00679455	0,00826239	0,90683353
Huber	0,00746373	0,00907923	0,90667901
Lp-médiane ; p=3	0,00814464	0,00982062	0,91068133
Lp-médiane ; p=4	0,01265439	0,01518394	0,91291094
Lp-médiane ; p=5	0,02376532	0,02834946	0,91558658

Que ce soit avec la loi bivariée de Cauchy (tableaux 3.7 et 3.8) ou de Student (ta-

bleaux 3.9 et 3.10) le schéma des résultats est similaire à celui observé avec la loi normale bivariée. L'exemple de l'estimateur L_3 apparaît dans les tableaux 3.11 et 3.12. La compa-

Tableau 3.7 *– Estimations et efficacité relative pour une loi de Cauchy avec $\rho = 0.2$ et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+\mathbf{C}_w})\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
	Configuration C1		
Médiane	0,010814731	0,054554955	0,44523649
Huber	0,01458914	0,078990703	0,42976086
	Configuration C2		
Médiane	0,006367514	0,007796083	0,90374672
Huber	0,008586175	0,010677137	0,89675221
	Configuration C3		
Médiane	0,005624282	0,005966512	0,97089722
Huber	0,007473544	0,007971481	0,968264
	Configuration C4		
Médiane	0,006011515	0,006500481	0,96165487
Huber	0,007665358	0,008338495	0,95878755

Tableau 3.8 *– Estimations et efficacité relative pour une Loi de Cauchy avec $\rho = 0.8$ et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+\mathbf{C}_w})\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
	Configuration C1		
Médiane	0,0590576	0,883735189	0,25850969
Huber	0,082399488	1,271491642	0,25456899
	Configuration C2		
Médiane	0,048451229	0,082970412	0,76417139
Huber	0,066249132	0,114869971	0,75942841
	Configuration C3		
Médiane	0,048364424	0,057388705	0,91801505
Huber	0,064154169	0,076456175	0,91602253
	Configuration C4		
Médiane	0,048093136	0,057983915	0,9107261
Huber	0,065496014	0,079243014	0,909132

raison des différentes lois selon les deux niveaux de corrélation intra-cluster ($\rho = 0.2$ et $\rho = 0.8$), nous permet de confirmer les observations faites sur les résultats d'optimisation sous la loi normale et d'établir que plus la corrélation intra-cluster est importante, meilleure est l'efficacité des estimateurs, et que plus la différence de taille des clusters est importante dans une configuration, plus la pondération du M-estimateur permet de diminuer sa variance.

Tableau 3.9 *– Estimations et efficacité relative pour une loi de Student à 3 degrés de liberté avec* $\rho = 0.2$ *et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+C}_w)\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
	Configuration C1		
Médiane	0,000383829	0,001982887	0,43996682
Moyenne	0,001454416	0,009223789	0,39709064
Huber	0,000484392	0,002975047	0,40350725
	Configuration C2		
Médiane	0,000217085	0,000268096	0,89984873
Moyenne	0,000937076	0,001189441	0,887597
Huber	0,000287908	0,000369686	0,88249147
	Configuration C3		
Médiane	0,000196438	0,000208621	0,97036409
Moyenne	0,000833429	0,000848679	0,99097453
Huber	0,000251313	0,000270225	0,96437323
	Configuration C4		
Médiane	0,000204325	0,000221723	0,95996576
Moyenne	0,000883147	0,000917008	0,98136317
Huber	0,000263744	0,000290224	0,95328901

Tableau 3.10 *– Estimations et efficacité relative pour une loi de Student à 3 degrés de liberté avec* $\rho = 0.8$ *et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+C}_w)\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
	Configuration C1		
Médiane	0,002064043	0,031299353	0,25679801
Moyenne	0,008121454	0,124576557	0,25532816
Huber	0,002548411	0,041058835	0,24913309
	Configuration C2		
Médiane	0,001614083	0,002782014	0,76169875
Moyenne	0,006846101	0,011699045	0,76497361
Huber	0,002142422	0,003780795	0,7527677
	Configuration C3		
Médiane	0,001726909	0,002052645	0,91722899
Moyenne	0,006792496	0,007619202	0,94419126
Huber	0,002097062	0,002513413	0,9134268
	Configuration C4		
Médiane	0,001649014	0,00199159	0,90993881
Moyenne	0,007268894	0,008306544	0,93545735
Huber	0,002273213	0,002764808	0,9067499

Tableau 3.11 *– Estimations et efficacité relative pour l'estimateur L_3 dans le cas d'une Loi de Student à 9 degrés de liberté avec $\rho = 0.2$ et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+C}_w)\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
C1	6,9747E-05	0,000405752	0,41460306
C2	3,95409E-05	4,9491E-05	0,89384044
C3	3,47965E-05	3,68854E-05	0,97127198
C4	3,8191E-05	4,08775E-05	0,96658075

Tableau 3.12 *– Estimations et efficacité relative pour l'estimateur L_3 dans le cas d'une Loi de Student à 9 degrés de liberté avec $\rho = 0.8$ et les configurations de cluster C1-C4*

	$det(V_\theta^{-1}\,(c_w\widehat{B+C}_w)\,V_\theta^{-1})$	$det(V_\theta^{-1}\,(\widehat{B+C})\,V_\theta^{-1})$	E_f
C1	0,000423426	0,00679401	0,24964663
C2	0,00031543	0,000544111	0,76139092
C3	0,000287671	0,000337713	0,92294202
C4	0,000333174	0,000385364	0,92982258

Chapitre 4

Robustesse

Sommaire

4.1 Points de rupture des M-estimateurs **81**

4.2 Reformulation du point de rupture dans le cas pondéré **83**

4.3 Comparaison des points de rupture entre versions pondérées et non pondérées **87**

4.3.1 Remarques générales 87

4.3.2 Résultats numériques 88

Résumé

Nous traitons ici de la robustesse globale d'un M-estimateur pondéré à partir du point de rupture. Nous donnons sa formulation ainsi qu'une étude du gain apporté par les poids pour la robustesse.

4.1 Points de rupture des M-estimateurs

Dans le chapitre précédent, nous avons présenté des M-estimateurs pondérés avec des poids optimaux qui améliorent leur efficacité. Nous développons maintenant la question de leur robustesse. En effet, lors de l'analyse d'un cas développé précédemment dans lequel on considère un cluster largement plus grand que les autres, on obtient pour celui-ci un poids plus faible que ceux obtenus pour les clusters de plus petite taille. Dans cette situation, si le cluster contient des valeurs aberrantes ou extrêmes, du fait de la pondération considérée, on peut espérer qu'elles perturbent de façon moindre notre estimateur. Pondérer un estimateur pourrait donc améliorer sa robustesse. Le graphique qui suit (figure 4.1) illustre le faible effet d'une perturbation d'un grand cluster sur l'estimation du centre.
Nous allons désormais tester l'hypothèse selon laquelle un choix pertinent de poids pourrait améliorer la robustesse d'un M-estimateur pondéré par rapport à sa version non-pondérée en utilisant le point de rupture comme mesure globale de la robustesse.

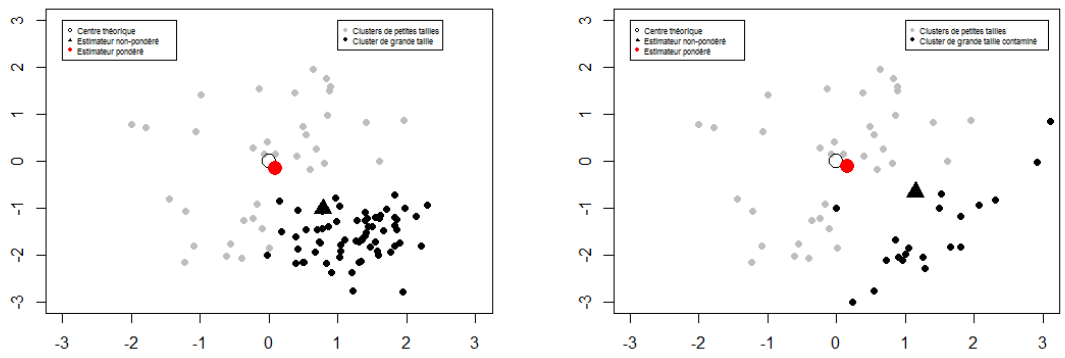

FIGURE 4.1 – *Comparaison de deux estimateurs, pondéré et non-pondéré, pour des variables clusterisées en présence de perturbations $\delta \approx N(1,5)$ sur le grand cluster avec $X_{ij} + \delta$*

Nous commençons par rappeler l'expression du point de rupture de remplacement pour un estimateur donné.

Définition 4.1.1 *(Donoho and Huber, 1983)*
Le point de rupture de remplacement pour un estimateur $\hat{\theta}(X)$ fondé sur n observations est

défini par :

$$\epsilon_n^* = \min_{1\leq k\leq n} \left\{ \frac{k}{n} : \sup_{Y_k} \left\| \hat{\theta}(Y_k) - \hat{\theta}(X) \right\| = \infty \right\} \tag{4.1}$$

où Y_k est obtenu en remplaçant k observations de X par des valeurs arbitraires.

Notons $\hat{\theta}^w(X)$ le M-estimateur pondéré associé à l'ensemble des données X et l'estimateur $\hat{\theta}(X)$ sa version non pondérée. Nous rappelons que ces deux estimateurs sont obtenus par :

$$\hat{\theta}^w(X) = \underset{a\in\Theta}{\operatorname{argmin}} \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\rho(X_{ij}, a) \tag{4.2}$$

$$\hat{\theta}(X) = \underset{a\in\Theta}{\operatorname{argmin}} \frac{1}{N_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \rho(X_{ij}, a). \tag{4.3}$$

Nous réorganisons l'indexation de l'ensemble des observations ainsi que celle des poids en $X = \{X_1, \ldots, X_{N_n}\}$ et les poids qui leurs sont attribués $W = \{w_1, \ldots, w_{N_n}\}$. Nous pouvons donc réécrire les estimateurs définis dans (4.2) et (4.3) comme suit :

$$\hat{\theta}^w(X) = \underset{a\in\Theta}{\operatorname{argmin}} \frac{1}{N_n} \sum_{i=1}^{N_n} w_i\rho(X_i, a) \tag{4.4}$$

$$\hat{\theta}^w(X) = \underset{a\in\Theta}{\operatorname{argmin}} \frac{1}{N_n} \sum_{i=1}^{N_n} \rho(X_i, a). \tag{4.5}$$

Nous gardons les mêmes notations des estimateurs avec les nouvelles indexations. Cette réécriture des estimateurs a pour avantage de faire disparaître l'effet des clusters, ce qui nous facilitera le développement du point de rupture (dont la définition est indépendante de la structure même des variables). Les points de rupture de ces estimateurs sont donc donnés respectivement par

$$\epsilon_{N_n}^{w*} = \min_{1\leq k\leq N_n} \left\{ \frac{k}{N_n} : \sup_{Y_k} \left\| \hat{\theta}^w(Y_k) - \hat{\theta}^w(X) \right\| = \infty \right\} \tag{4.6}$$

et

$$\epsilon_{N_n}^{*} = \min_{1\leq k\leq N_n} \left\{ \frac{k}{N_n} : \sup_{Y_k} \left\| \hat{\theta}(Y_k) - \hat{\theta}(X) \right\| = \infty \right\}, \tag{4.7}$$

où Y_k est toujours obtenu en remplaçant k observations de X par des valeurs arbitraires.

4.2 Reformulation du point de rupture dans le cas pondéré

Nous nous intéressons à l'aspect empirique non asymptotique du point de rupture et nous considérons pour simplifier que $\frac{1}{N_n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} = 1$, ce qui se réécrit avec la nouvelle indexation $\sum_{i=1}^{N_n} w_i = N_n$. Le théorème suivant donne l'expression du point de rupture $\epsilon_{N_n}^{w*}$ en fonction des poids.

Théorème 4.2.1 *Supposons que la fonction objective $a \mapsto \rho(x,a)$ est convexe et que le point de rupture de $\hat{\theta}(X)$ vérifie la condition $\epsilon_{N_n}^* \leq \epsilon_{LN_n}^* \leq \epsilon_0^*$ pour tout entier $L \geq 1$. Alors la version pondérée $\hat{\theta}^w(X)$ de $\hat{\theta}(X)$ a un point de rupture $\frac{k_w^*}{N_n}$ tel que $k_w^* \in [[k_1^*, k_0^*]]$ où k_0^* et k_1^* sont respectivement définis par :*

$$k_0^* = \min_{1\leq k\leq N_n} \left\{k : \textit{il existe } i_1,\ldots,i_k \in \{1,\ldots,N_n\} \textit{ tels que } w_{i_1}+\ldots+w_{i_k} \geq \epsilon_0^* N_n\right\},$$
$$k_1^* = \min_{1\leq k\leq N_n} \left\{k : \textit{il existe } i_1,\ldots,i_k \in \{1,\ldots,N_n\} \textit{ tels que } w_{i_1}+\ldots+w_{i_k} \geq \epsilon_{N_n}^* N_n\right\}.$$

Remarques : L'hypothèse de convexité sur la fonction ρ n'est utilisée que dans la généralisation au cas de poids réels, elle n'est donc pas utile si les poids sont choisis sous la forme de rationnels. La condition $\epsilon_{N_n}^* \leq \epsilon_{LN_n}^* \leq \epsilon_0^*$ est vérifiée par la médiane spatiale dont le point de rupture est donné par $\frac{\lfloor \frac{N_n-1}{2} \rfloor}{N_n}$ (Croux and Rousseeuw, 1992). Nous avons alors $\epsilon_0^* = \frac{1}{2}$ et la condition $\epsilon_{N_n}^* \leq \epsilon_{LN_n}^*$ est vraie pour tout $L \geq 1$.

Démonstration

Soit k le nombre de valeurs remplacées dans l'ensemble des données initiales X, ainsi $k = \#\{Y_k \setminus (Y_k \cap X)\}$, où $\#A$ représente le cardinal de l'ensemble A. À partir de la définition du point de rupture de $\hat{\theta}(X)$ donnée en (4.7), remarquons tout d'abord que nous avons l'équivalence :

$$\sup_{Y_k} \left\|\hat{\theta}(Y_k) - \hat{\theta}(X)\right\| = \infty \Longleftrightarrow k \geq \epsilon_{N_n}^* N_n. \tag{4.8}$$

En terme de cardinaux d'ensembles, (4.8) peut se récrire de manière équivalente sous la forme :

$$\sup_{Y_k} \left\|\hat{\theta}(Y_k) - \hat{\theta}(X)\right\| = \infty \Longleftrightarrow \#\{Y_k \setminus (Y_k \cap X)\} \geq \epsilon^*_{N_n} \#\{X\}. \tag{4.9}$$

Nous nous plaçons, dans un premier temps, dans le cas où les poids sont sous forme rationnelle et donnés par $w_i = \frac{\ell_i}{L}$, pour tout $i = 1, \ldots, N_n$ et $\ell_i, L \in \mathbb{N}^*$. Le M-estimateur pondéré (4.4) s'écrit donc :

$$\begin{aligned}\hat{\theta}^w(X) &= \underset{a\in\Theta}{\operatorname{argmin}} \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{\ell_i}{L} \rho(X_i, a) \\ &= \underset{a\in\Theta}{\operatorname{argmin}} \frac{1}{N_n L} \sum_{i=1}^{N_n} \ell_i \rho(X_i, a).\end{aligned}$$

Ainsi $\hat{\theta}^w(X)$, associé à l'ensemble de données X, est également le M-estimateur non pondéré $\hat{\theta}(\tilde{X})$ calculé pour le nouvel ensemble de données $\tilde{X}$ construit en répétant ℓ_i fois chaque élément X_i de l'échantillon X, et de même on construit $\tilde{Y}_k$ à partir de Y_k. Suite à ces transformations, nous réécrivons le point de rupture défini en (4.6) par :

$$\epsilon^{w*}_{N_n} = \min_{1\leq k\leq N_n} \left\{ \frac{k}{N_n} : \sup_{\tilde{Y}_k} \left\|\hat{\theta}(\tilde{Y}_k) - \hat{\theta}(\tilde{X})\right\| = \infty \right\} \tag{4.10}$$

On a de plus $\sum_{i=1}^{N_n} \ell_i = LN_n$. En réécrivant l'équation (4.10) avec l'aide de l'équivalence (4.9), on obtient alors :

$$\epsilon^{w*}_{N_n} = \min_{1\leq k\leq N_n} \left\{ \frac{k}{N_n} : \#\{\tilde{Y}_k - (\tilde{Y}_k \cap \tilde{X})\} \geq \epsilon^*_{LN_n} \#\{\tilde{X}\} \right\}. \tag{4.11}$$

Si $X_{i_1}, \ldots, X_{i_k}$ représentent les k variables aléatoires de l'ensemble X remplacées par des valeurs arbitraires, alors il faut remplacer $l_{i_1} + \ldots + l_{i_k}$ variables dans $\tilde{X}$ pour obtenir $\tilde{Y}_k$. Comme de plus $\#\{\tilde{X}\} = \sum_{i=1}^{N_n} \ell_i$, la formule (4.11) devient ainsi :

$$\epsilon^{w*}_{N_n} = \min_{1\leq k\leq N_n} \left\{ \frac{k}{N_n} : \text{il existe } i_1, \ldots, i_k \in \{1, \ldots, N_n\} \text{ tels que } l_{i_1} + \ldots + l_{i_k} \geq \epsilon^*_{LN_n} \sum_{i=1}^{N_n} \ell_i \right\} \tag{4.12}$$

$$= \min_{1\leq k\leq N_n} \left\{ \frac{k}{N_n} : \text{il existe } i_1, \ldots, i_k \in \{1, \ldots, N_n\} \text{ tels que } w_{i_1} + \ldots + w_{i_k} \geq \epsilon^*_{LN_n} N_n \right\}. \tag{4.13}$$

Notons k_w^* le point défini par (4.13). Comme k_0^* est défini par

$$\min_{1\leq k\leq N_n}\{k : \text{il existe } i_1,\ldots,i_k \in \{1,\ldots,N_n\} \text{ tels que } w_{i_1}+\ldots+w_{i_k} \geq \epsilon_0^* N_n\}$$

et que par hypothèse, $\epsilon_0^* \geq \epsilon_{LN_n}^*$, nous en déduisons que nécessairement $k_w^* \leq k_0^*$ car

$$w_{i_1}+\cdots+w_{i_{k_0^*}} \geq \epsilon_0^* N_n \Longrightarrow w_{i_1}+\cdots+w_{i_{k_0^*}} \geq \epsilon_{LN_n}^* N_n.$$

De la même façon, la définition de k_1^* et la condition $\epsilon_{N_n}^* \leq \epsilon_{LN_n}^*$ entraînent que $k_w^* \geq k_1^*$. Ceci achève la démonstration du théorème 4.2.1 dans le cas des poids $w_i = \frac{\ell_i}{L}$, $i = 1,\ldots,N_n$.

Nous traitons maintenant le cas des pondérations réelles où nous allons montrer que le point de rupture peut alors s'approcher par le point de rupture précédent. Rappelons que $\hat{\theta}^w = \underset{a\in\Theta}{\operatorname{argmin}} M_n(a)$, avec $M_n(a) = \frac{1}{N_n}\sum_{i=1}^{N_n} w_i\rho(X_i,a)$. On peut approcher $M_n(a)$ par $M_n^*(a) = \frac{1}{N_n}\sum_{i=1}^{N_n} w_i^*\rho(X_i,a)$, où les poids w_i^* sont donnés par $w_i^* = \frac{\ell_i}{L}$, $\ell_i, L \in \mathbb{N}^*$, $i = 1,\ldots,N_n$ et seront choisis ultérieurement. Soient $\epsilon > 0$ et $B_\epsilon(\hat{\theta}^w)$ la boule ouverte de rayon ϵ et de centre $\hat{\theta}^w$. Le M-estimateur $\hat{\theta}^w$ minimise M_n, ce qui implique que pour tout $a \notin B_\epsilon(\hat{\theta}^w)$ il existe $\delta > 0$ tel que :

$$M_n(a) \geq M_n(\hat{\theta}^w) + \delta. \tag{4.14}$$

Puisque $M_n(a)$ et $M_n^*(a)$ sont des combinaisons linéaires finies des w_i et w_i^*, nous utilisons la densité de l'espace des nombres rationnels dans l'espace des nombres réels. Soit $\alpha > \epsilon$, nous choisissons les w_i^* proches des w_i tels que pour tout $a \in B_\alpha(\hat{\theta}^w)$:

$$|M_n(a) - M_n^*(a)| \leq \frac{\delta}{3}. \tag{4.15}$$

Nous avons donc en particulier

$$M_n^*(\hat{\theta}^w) \leq M_n(\hat{\theta}^w) + \frac{\delta}{3} \tag{4.16}$$

et

$$M_n^*(a) \geq M_n(a) - \frac{\delta}{3}. \tag{4.17}$$

Prenons $a \in B_\alpha(\hat{\theta}^w) \setminus B_\epsilon(\hat{\theta}^w)$, alors nous avons les implications suivantes :

$$(4.14), (4.17) \Rightarrow M_n^*(a) \geq M_n(\hat{\theta}^w) + 2\frac{\delta}{3} \tag{4.18}$$

$$(4.14), (4.16) \Rightarrow M_n(a) \geq M_n^*(\hat{\theta}^w) + 2\frac{\delta}{3}. \tag{4.19}$$

Les minorations (4.16) et (4.18) entraînent à leur tour que pour tout a dans $B_\alpha(\hat{\theta}^w) \setminus B_\epsilon(\hat{\theta}^w)$:

$$M_n^*(a) \geq M_n^*(\hat{\theta}^w) + \frac{\delta}{3}. \tag{4.20}$$

Notons $\hat{\theta}^{w*}$ le M-estimateur pondéré défini pour les poids rationnels, c'est-à-dire :

$$\hat{\theta}^{w*} = \underset{a \in \Theta}{\operatorname{argmin}} M_n^*(a). \tag{4.21}$$

Nous déduisons de l'équation (4.20) que l'estimateur $\hat{\theta}^{w*} \notin B_\alpha(\hat{\theta}^w) \setminus B_\epsilon(\hat{\theta}^w)$ et donc $\hat{\theta}^{w*} \in \partial B_\alpha(\hat{\theta}^w) \cup B_\epsilon(\hat{\theta}^w)$, où le signe ∂A désigne la frontière du domaine A. Nous allons, avec un raisonnement par l'absurde, commencer par montrer que nécessairement $\hat{\theta}^{w*} \in B_\epsilon(\hat{\theta}^w)$. Supposons donc que $\hat{\theta}^{w*} \in \partial B_\alpha(\hat{\theta}^w)$ et choisissons $\tilde{a} = (1-t)\hat{\theta}^{w*} + t\hat{\theta}^w$ un point dans le segment d'extrémités $\hat{\theta}^{w*}$ et $\hat{\theta}^w$ tel que $\tilde{a} \in B_\alpha(\hat{\theta}^w) \setminus B_\epsilon(\hat{\theta}^w)$ et $t \in]0,1[$. La convexité de la fonction $\rho(.,a)$ entraîne celle de $M_n^*(a)$ (combinaison linéaire finie de fonctions convexes) :

$$\begin{aligned} M_n^*(\tilde{a}) &\leq (1-t)M_n^*(\hat{\theta}^{w*}) + tM_n^*(\hat{\theta}^w) \\ &\leq (1-t)M_n^*(\hat{\theta}^w) + tM_n^*(\hat{\theta}^w) = M_n^*(\hat{\theta}^w). \end{aligned} \tag{4.22}$$

L'équation (4.16) implique alors que

$$M_n^*(\hat{\theta}^w) \leq M_n(\hat{\theta}^w) + \frac{\delta}{3}. \tag{4.23}$$

Les inégalités (4.18) et (4.22) nous mènent à des équations absurdes :

$$M_n(\hat{\theta}^w) + 2\frac{\delta}{3} \leq M_n^*(\tilde{a}) \leq M_n(\hat{\theta}^w) + \frac{\delta}{3}. \tag{4.24}$$

Nous concluons donc que l'alternative $\hat{\theta}^{w*} \in \partial B_\alpha(\hat{\theta}^w)$ n'est pas possible quelque soit $\alpha > \epsilon$. Le seul choix possible est donc $\hat{\theta}^{w*} \in B_\epsilon(\hat{\theta}^w)$.
Pour des poids rationnels w_k^* qui convergent vers w_k, les M-estimateurs $\hat{\theta}^{w*}$ et $\hat{\theta}^w$ sont de la même nature : si $\hat{\theta}^{w*}$ est borné alors $\hat{\theta}^w$ l'est aussi et inversement. Le point de

rupture de $\hat{\theta}^w$ s'écrit donc comme une limite du point de rupture de $\hat{\theta}^{w*}$ par densité des rationnels dans le corps des réels, ce qui achève la démonstration. □

4.3 Comparaison des points de rupture entre versions pondérées et non pondérées

4.3.1 Remarques générales

Par sa définition, le point de rupture d'un M-estimateur pondéré dépend plus des poids attribués aux valeurs qu'à leur nature potentiellement aberrantes. L'examen de la démonstration du théorème 4.2.1, montre également que pour des poids w_i rationnels écrits sous la forme $w_i = \frac{l_i}{L}$, l'expression exacte du point de rupture de l'estimateur pondéré est donnée par l'équation (4.13) :

$$\epsilon_{N_n}^{w*} = \min_{1 \leq k \leq N_n} \left\{ \frac{k}{N_n} : \text{il existe } i_1, \ldots, i_k \in \{1, \ldots, N_n\} \text{ tels que } w_{i_1} + \ldots + w_{i_k} \geq \epsilon_{LN_n}^* N_n \right\}.$$

La première remarque que l'on peut formuler est que si $\epsilon_{LN_n}^*$ est très proche de ϵ_0^*, cette expression s'écrit

$$\epsilon_{N_n}^{w*} = \min_{1 \leq k \leq N_n} \left\{ \frac{k}{N_n} : \text{il existe } i_1, \ldots, i_k \in \{1, \ldots, N_n\} \text{ tels que } w_{i_1} + \ldots + w_{i_k} \geq \epsilon_0^* N_n \right\}$$

et généralise aux M-estimateurs la définition donnée par Nevalainen et al. (2007b) pour la médiane spatiale pondérée (car leur démonstration est basée sur un choix du point de rupture égal à 0.5 pour la médiane spatiale non pondérée).

Supposons maintenant que l'estimateur pondéré atteigne son point de rupture maximal $\epsilon_{N_n}^{w*} \frac{k_0^*}{N_n}$ avec k_0^* donné par

$$k_0^* = \min_{1 \leq k \leq N_n} \left\{ k : \text{il existe } i_1, \ldots, i_k \in \{1, \ldots, N_n\} \text{ tels que } w_{i_1} + \ldots + w_{i_k} \geq \epsilon_0^* N_n \right\}.$$

Ce résultat montre que le point de rupture maximal sera obtenu en ordonnant les w_i par ordre croissant et en choisissant de remplacer les observations correspondant aux $k = k_0^*$ poids les plus grands (afin de garantir la minimalité du choix effectué pour k). En notant $w_{(i)}$ les poids ordonnés par ordre croissant, on obtient donc que k_0^* a pour

expression :

$$\min_{1\leq k\leq N_n}\left\{k : \text{il existe } i_1,\ldots,i_k \in \{1,\ldots,N_n\} \text{ tels que } w_{(N_n)}+\ldots+w_{(N_n-k_0^*+1)} \geq \epsilon_0^* N_n\right\}. \tag{4.25}$$

En considérant que l'estimateur non pondéré a un point de rupture égal à une proportion α_0^* ($0 \leq \alpha_0^* \leq \frac{1}{2}$), on remarque que pour améliorer ce point de rupture initial, il faut sur-pondérer $\alpha_0^* N_n$ observations. On voit ainsi que malheureusement, on ne peut pas espérer augmenter simultanément l'efficacité de l'estimateur (qui implique d'imposer des poids faibles sur les plus grands clusters, soit sur le plus grand nombre d'observations) et son point de rupture.

4.3.2 Résultats numériques

Le théorème 4.2.1 indique que le point de rupture maximal de l'estimateur pondéré est défini par :

$$\epsilon_n^{w*} = \frac{k_0^*}{N_n} \tag{4.26}$$

avec k_0^* donné par l'équation (4.25). Nous prenons deux estimateurs avec un point de rupture maximal $\epsilon_0^* = 0.5$, la médiane spatiale et l'estimateur de Huber. Le cadre choisi pour les simulations est le même que celui utilisé dans le chapitre 3 où nous avons obtenu les poids optimaux permettant d'obtenir la meilleure efficacité relative. Nous considérons donc huit cas de figure pour chaque estimateur avec 4 cas de configurations pour les effectifs de clusters :

1. C1 : 9 clusters de 4 v.a. et un cluster de 64 v.a..
2. C2 : 5 clusters de 4 v.a. et 5 clusters de 16 v.a..
3. C3 : 2 clusters de 4 v.a., un cluster de 8 v.a. et 7 clusters de 12 v.a..
4. C4 : Des clusters avec les nombres de v.a. suivants $5, 6, 7, 8, 9, 11, 12, 13, 14, 15$.

De plus, les variables aléatoires sont issues d'une loi normale centrée et nous prenons deux valeurs de corrélation intra-cluster $\rho = 0.2$ et $\rho = 0.8$. Nous rappelons que dans les exemples de simulations, les variables aléatoires d'un cluster i ont le même poids w_i, avec un total de 10 clusters ($i = 1, \ldots, 10$). Dans les tableaux 4.1 et 4.2, nous donnons les poids optimaux calculés pour les deux M-estimateurs pondérés.

Comme déjà mis en avant dans le chapitre 3, plus les clusters sont grands, plus leurs poids optimaux sont petits. Les tableaux montrent que le point de rupture obtenu

Tableau 4.1 *– Optimisation des poids pour la médiane spatiale pondérée avec $\rho = 0.2$ et $\rho = 0.8$ dans le cas d'une Loi Normale*

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
					$\rho = 0.2$					
C1	2,251	2,324	2,232	2,203	2,252	2,291	2,227	2,192	2,281	0,297
C2	1,818	1,877	1,803	1,780	1,819	0,812	0,789	0,769	0,821	0,784
C3	1,712	1,764	1,180	0,915	0,911	0,926	0,907	0,899	0,929	0,900
C4	1,519	1,420	1,262	1,171	1,093	0,989	0,904	0,841	0,840	0,763
					$\rho = 0.8$					
C1	2,487	2,520	2,464	2,460	2,473	2,486	2,482	2,454	2,498	0,167
C2	2,412	2,444	2,389	2,386	2,398	0,654	0,656	0,638	0,658	0,636
C3	2,382	2,414	1,243	0,837	0,846	0,851	0,855	0,831	0,859	0,827
C4	1,939	1,645	1,407	1,241	1,114	0,920	0,852	0,767	0,738	0,665

Tableau 4.2 *– Optimisation des poids pour l'estimateur de Huber pondéré avec $\rho = 0.2$ et $\rho = 0.8$ dans le cas d'une Loi Normale*

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
					$\rho = 0.2$					
C1	2,268	2,343	2,275	2,216	2,259	2,338	2,289	2,242	2,352	0,276
C2	1,875	1,937	1,881	1,832	1,868	0,797	0,776	0,752	0,810	0,767
C3	1,760	1,816	1,198	0,902	0,894	0,922	0,909	0,879	0,940	0,896
C4	1,567	1,439	1,301	1,171	1,089	0,978	0,902	0,829	0,833	0,749
					$\rho = 0.8$					
C1	2,464	2,520	2,468	2,451	2,464	2,522	2,506	2,441	2,563	0,163
C2	2,416	2,470	2,419	2,402	2,415	0,650	0,647	0,631	0,662	0,629
C3	2,393	2,447	1,239	0,830	0,835	0,854	0,850	0,829	0,869	0,826
C4	1,941	1,659	1,410	1,227	1,104	0,928	0,848	0,765	0,746	0,663

Tableau 4.3 *– Point de rupture ϵ_n^{w*} pour l'estimateur de Huber pondéré dans les différentes cas des simulations*

	C1	C2	C3	C4
			$\rho = 0.2$	
$w_{(N_n)} + \ldots + w_{(N_n-k-1)}$	50,118	50,326	50,055	50,611
$\epsilon_{N_n}^{w*}$	22%	36%	45%	41%
			$\rho = 0.8$	
$w_{(N_n)} + \ldots + w_{(N_n-k-1)}$	51,985	50,443	50,117	50,219
$\epsilon_{N_n}^{w*}$	21%	23%	41%	36%

pour chaque cas de figure pour les deux estimateurs est inférieur à 50%. Il varie selon l'estimateur, le nombre de variables par clusters et la corrélation. Les résultats que nous obtenons dans les tableaux 4.3 et 4.4 permettent de constater que la forte corrélation dégrade le point de rupture par rapport à une faible corrélation, et cette dégradation

Tableau 4.4 *– Point de rupture $\epsilon_{N_n}^{w*}$ pour la médiane spatiale pondérée dans les différents cas des simulations*

	C1	C2	C3	C4
$\rho = 0.2$				
$w_{(N_n)} + \ldots + w_{(N_n-k-1)}$	51,921	50,173	50,817	50,081
$\epsilon_{N_n}^{w*}$	23%	37%	46%	41%
$\rho = 0.8$				
$w_{(N_n)} + \ldots + w_{(N_n-k-1)}$	52,099	50,079	50,177	50,293
$\epsilon_{N_n}^{w*}$	21%	23%	41%	36%

est accentuée en présence d'une grande hétérogénéité des tailles de cluster, comme par exemple pour la configuration $C1$ dans les tableaux 4.3 et 4.4. De manière cohérente en ce qui concerne le point de rupture, la configuration $C1$ (qui contient un cluster comportant plus de 50% des observations) et la configuration $C2$ (où 5 clusters de taille identique regroupent 80% des observations) ont le point de rupture le plus dégradé. En conclusion, travailler avec des poids optimaux améliore de manière significative l'efficacité de l'estimateur par rapport à sa version pondérée mais ne peut que dégrader son point de rupture initial.

Conclusions et perspectives

Dans ce travail, nous avons établi les propriétés asymptotiques des M-estimateurs : la convergence en probabilité et presque-sûre ainsi que la normalité asymptotique. Nous avons illustré ces résultats au travers de simulations qui ont mis en évidence l'influence de la structure (la corrélation intra-cluster et leur taille) des variables aléatoires clusterisées. Nous avons obtenu des résultats similaires pour la classe des M-estimateurs pondérés. De plus, une étude d'optimisation de la variance en fonction des poids nous a permis d'identifier qu'un choix pertinent de la pondération permet d'améliorer l'efficacité des M-estimateurs pondérés par rapport aux non pondérés. Pour finir, nous avons donné l'expression explicite du point de rupture maximal pour des M-estimateurs pondérés, en fonction de celui associé à leur version non pondérée, ce qui nous a menés à la conclusion que la pondération ne pouvait pas permettre d'améliorer à la fois l'efficacité et la robustesse d'un M-estimateur.

Il serait opportun d'appliquer une sélection de M-estimateurs pondérés ou non à des données clusterisées relatives à des cas réels. En effet, cette étude permettrait de valider nos estimateurs, jusqu'à présent étudiés dans un cadre théorique, et d'évaluer la pertinence de la localisation de centres existants, ou encore de proposer de nouveaux centres plus pertinents.

Prenons les poids optimaux que nous avons trouvés dans le chapitre 3, en particulier, l'exemple du tableau 3.4. Pour les clusters de même taille, les valeurs des poids obtenues dans la première configuration C_1 sont très proches et choisis d'autant plus petits que la taille du cluster est importante. Il serait intéressant d'établir, par exemple en spécifiant la loi des clusters, un résultat théorique sur les valeurs optimales des poids permettant d'obtenir la meilleure efficacité. En premier lieu, on pourrait étudier, pour une corrélation intra-cluster identique, les M-estimateurs pondérés avec des poids choisis en fonction des tailles des clusters $w_{ij} = w_{m_i}$. Un autre cas intéressant sera d'appliquer des poids qui dépendent de la corrélation intra-cluster $w_{ij} = w_{\rho_i}$, avec ρ_i qui représente la corrélation qui caractérise les variables aléatoires du cluster i.

Jusqu'à présent, les poids w_{ij} utilisés sont déterministes. Si nous supposons que les poids sont aléatoires, indépendants, et indépendants des observations X_i, les résultats du 3 peuvent se généraliser moyennant quelques adaptations simples, par exemple en supposant $E(w_{ij}) = w_{ij}$ pour tout i et j. Nous pouvons envisager par la suite un cadre moins restrictif, en prenant par exemple des poids choisis en fonction du cluster i. Cela pourrait s'envisager notamment pour des données provenant de processus ponctuels agrégés où la taille des clusters M_i devient aléatoire. Nous pourrions également prendre

en compte la dépendance des poids vis à vis des variables aléatoires $w_{ij} = f(X_{ij})$, avec en particulier un choix du type $w_{ij} = I_{X_{ij} \in D}$, où D est un domaine donné. Enfin, une application aux données réelles serait particulièrement intéressante pour comparer les différentes versions des estimateurs proposés dans cette thèse.

Annexe A

Outils probabilistes

A.1 Notations et propriétés de o_p et O_p

Les notations o_p et O_p sont fréquemment utilisées dans la littérature. Nous reprenons ici les propriétés données dans le livre de Van der Vaart (2000, pages 12-13).

Définition A.1.1 *La notation $o_p(1)$ désigne une suite de variables aléatoires qui converge en probabilité vers 0.*
La notation $O_p(1)$ désigne une suite de variables aléatoires bornée en probabilité.
Prenons R_n une suite de variables aléatoires :

- *Si une suite $X_n = o_p(R_n)$ alors $X_n = R_n o_p(1)$.*
- *Si une suite $X_n = O_p(R_n)$ alors $X_n = R_n O_p(1)$.*

Si X_n et R_n sont deux suites non aléatoires, alors les notions de o_p et de O_p coïncident respectivement avec les notations usuelles, o et O, utilisées en analyse. Quelques propriétés autour de o_p et de O_p sont données ci-dessous.

Proposition A.1.1

$$\begin{aligned} o_p(1) + o_p(1) &= o_p(1); \\ o_p(1) + O_p(1) &= O_p(1); \\ o_p(1)O_p(1) &= o_p(1); \\ (1 + o_p(1))^{-1} &= O_p(1); \\ o_p(O_p(1)) &= o_p(1). \end{aligned}$$

A.2 Quelques propriétés de convergence stochastique

Nous commençons par rappeler le théorème de Slutsky pour la continuité de la convergence en probabilité. Des extensions de ce résultat sont possibles (voir par exemple Serfling, 1980, page 24). Nous avons donc en particulier le théorème suivant pour trois types de convergences stochastiques.

Théorème A.2.1 *Soit $g : \mathbb{R}^k \mapsto \mathbb{R}^k$ une fonction continue en tout point $x \in D$ avec $P(X \in D) = 1$.*

1. *Si $X_n \overset{Loi}{\to} X$, alors $g(X_n) \overset{Loi}{\to} g(X)$.*
2. *Si $X_n \overset{P}{\to} X$, alors $g(X_n) \overset{P}{\to} g(X)$.*
3. *Si $X_n \overset{p.s.}{\to} X$, alors $g(X_n) \overset{p.s.}{\to} g(X)$.*

A.3 Théorèmes de convergence

Nous annonçons ici deux théorèmes de convergence qui sont des variantes de la loi des grands nombres, et qui n'imposent pas la condition d'équidistribution sur les variables aléatoires, voir par exemple Serfling (1980).

Théorème A.3.1 *(Chebyshev)*
Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes de moyennes $\mu_1, \mu_2, \ldots$ et de variances $\sigma_1^2, \sigma_2^2, \ldots$. Si $\sum_{i\geq 1} \sigma_i^2 = o(n^2)$ quand $n \to \infty$, alors

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \overset{P,L^2}{\to} 0. \tag{A.1}$$

Théorème A.3.2 *(Kolmogorov)*
Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes de moyennes $\mu_1, \mu_2, \ldots$ et de variances $\sigma_1^2, \sigma_2^2, \ldots$. Si la série $\sum_{i\geq 1} \frac{\sigma_i^2}{i^2}$ converge, alors

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \overset{p.s.}{\to} 0. \tag{A.2}$$

De la même façon nous avons l'extension suivante du théorème central limite.

Théorème A.3.3 *(Lindeberg) Soit $X_1, \ldots, X_n$ une suite des v.a. indépendantes issues des distributions $F_1, \ldots, F_n$, de moyennes $\mu_1, \ldots, \mu_n$ et de matrices de variance-covariances $\Sigma_1, \ldots, \Sigma_n$ respectivement. On suppose que :*

$$\lim_{n\to\infty} \frac{\Sigma_1 + \ldots + \Sigma_n}{n} = \Sigma,$$

ainsi que pour tout $\epsilon > 0$:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \int_{\|x-\mu_i\|\geq\epsilon\sqrt{n}} \|x-\mu_i\|^2 \, dF_i(x) = 0.$$

Alors quand $n \to \infty$:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i\right) \overset{Loi}{\to} N(0, \Sigma).$$

Annexe B

Programmes pour la M-estimation : calcul des estimateurs et des poids optimaux

B.1 Calcul des M-estimateurs et M-estimateurs pondérés sous R

1. Paramètres et notations
   (a) Nombre total de clusters : $np = 100$
   (b) Vecteur des tailles des clusters. Dans cet exemple tous les cluster ont la même taille : $ne = rep(1, np)$.
   (c) Vecteur de correlations intraclusters, qui sont identiques dans l'exemple : $rho = rep(0.2, np)$.
   (d) Nombre de simulations : $N = 1000$.
   (e) Le vecteur poids qui est le même pour toutes les simulations : $set.seed(1); wt < -rexp(max(ne) * np, 1)$.
   (f) Vecteurs vides dont lesquels on enregistre les coordonnées des variables aléatoires pour toutes les simulation : $x << -array(NA, c(max(ne), np, N))$;$y << -x$.
   (g) Ce vecteur est la réorganisation en deux dimensions du vecteur wt, afin d'avoir pour chaque couple (x_{ij}, y_{ij}) une valeur de piods W_{ij} : $W << -x$.

2. Simulation des lois de probabilités multidimensionnelles

   (a) Loi normale :
   On peut dans ce cas utiliser deux fonctions de R "rmvnorm" du package **mvtnorm** ou "mvrnorm" du package **MASS**.

   (b) Loi de Student (t-distribution) :
   On utilise la fonction "rmvt" du package **mvtnorm** avec différents degré de liberté df (1 pour une loi de couchy et > 1 pour avoir des moments ≥ 1).

   (c) Ici, on simule N échantillons de variables aléatoires (x,y) de loi de Student de degré de liberté 3 et on attribut les poids associés :

```
for(j in 1:np) {
matcov<-diag(1,nrow=ne[j],ncol=ne[j])+rho[j]*
(matrix(1,nrow=ne[j],
ncol=ne[j])-diag(1,nrow=ne[j],ncol=ne[j]))
set.seed(j)
xx1<- rmvt(n=N,sigma=matcov/3,df=3)
set.seed(np+j)
xx2<- rmvt(n=N,sigma=matcov/3,df=3)
for(i in 1:ne[j]) {W[i,j,]=wt[j*i]; x[i,j,]<-xx1[,i];
y[i,j,]<-xx2[,i]}
  }
```

3. Variables pour enregistrer le calcul des fonctions objectives pour $M_n(mu1, mu2)$

```
Moyenne=array(0,c(2,np,N));
WMoyenne=array(0,c(2,np,N))
Mediane=array(0,c(2,np,N));
WMediane=array(0,c(2,np,N))
Huber=array(0,c(2,np,N));
WHuber=array(0,c(2,np,N));
```

4. Variables pour enregistrer les risques empiriques

```
Rmoyenne<<-matrix(0,2,2]
MSEMoyenne<-rep(0,np);
WRmoyenne<<-matrix(0,2,2)
WMSEMoyenne<-rep(0,np);
Rmediane<<-matrix(0,2,2)
```

```
MSEMediane<-rep(0,np);
WRmediane<<-matrix(0,2,2)
WMSEMediane<-rep(0,np);
RHuber<<-matrix(0,2,2)
MSEHuber<-rep(0,np);
WRHuber<<-matrix(0,2,2)
WMSEHuber<-rep(0,np);
```

5. La boucle de calcul : pour chaque nombre de cluster considéré "s" (qui varie de 1 à np) on réalise le calcul suivant :

```
for(s in 1:np) {
  Rmoyenne<<-matrix(0,2,2); WRmoyenne<<-matrix(0,2,2);
  Rmediane<<-matrix(0,2,2); WRmediane<<-matrix(0,2,2);
  RHuber<<-matrix(0,2,2); WRHuber<<-matrix(0,2,2)
}
```

   (a) Pour chaque simulation i (de 1 à N), on définit la fonction objective à optimiser qui est spécifique pour chaque estimateur :
   $for(i\ in\ 1:N)\{$
   - Pour la moyenne empirique

     ```
     Moyennec<-function(mu) {
     mu1<-mu[1]; mu2<-mu[2];
     sum(((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2),na.rm=T)
     }
     ```

   - Pour la moyenne empirique pondérée

     ```
     WMoyennec<-function(mu) {
     mu1<-mu[1]; mu2<-mu[2];
     sum(W[,1:s,i]*((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2),na.rm=T)
     }
     ```

   - Pour la Médiane spatiale

     ```
     Medianec<-function(mu) {
     mu1<-mu[1]; mu2<-mu[2]
     sum(sqrt((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2),na.rm=T)
     ```

```
}
```

– Pour la Médiane spatiale pondérée

```
WMedianec<-function(mu) {
mu1<-mu[1]; mu2<-mu[2]
sum(W[,1:s,i]*sqrt((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2),na.rm=T)
}
```

– Pour l'estimateur de Huber

```
Huberc<-function(mu) {
mu1<-mu[1]; mu2<-mu[2]; kh=1.9
dif<-sqrt((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2)
f1<-0.5*dif^2; f2<-kh*dif-0.5*kh^2
Huberc=sum(f1[dif<=kh],na.rm=T)+sum(f2[dif>kh],na.rm=T)
}
```

– Pour l'estimateur de Huber pondéré

```
WHuberc<-function(mu) {
mu1<-mu[1]; mu2<-mu[2]; kh=1.9;
dif<-sqrt((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2);
f1<-W[,1:s,i]*(0.5*dif^2); f2<-W[,1:s,i]*(kh*dif-0.5*kh^2);
WHuberc=sum(f1[dif<=kh],na.rm=T)+sum(f2[dif>kh],na.rm=T);
}
```

– D'autres fonctions sont implémentées pour la norme Lp pour différentes valeurs de p. Exemple (p=0.5) :

```
NormLp0.5c<-function(mu) {
      mu1<-mu[1];mu2<-mu[2]
      p=0.5
      dif<-sqrt((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2)
      NormLp0.5c=sum(dif^p,na.rm=T)
      }

WNormLp0.5c<-function(mu) {
      mu1<-mu[1];mu2<-mu[2]
```

```
p=0.5
dif<-sqrt((x[,1:s,i]-mu1)^2+(y[,1:s,i]-mu2)^2)
WNormLp0.5c=sum(W[,1:s,i]*dif^p,na.rm=T)
}
```

(b) Calcul des estimateurs par optimisation des fonctions objectives définies précédemment avec l'aide de la fonction "optim". Avant de donner le code de cette optimisation, on a effectué un test pour évaluer "optim" dans le cas où on peut donner la vraie valeur optimale (par exemple la moyenne empirique) :
soit "a" un échantillon de variables aléatoires i.i.d simulé à partir d'une loi gaussienne bidimensionnelle d'espérance $(0,0)$ et de matrice de variance la matrice identité de dimension de 2×2
$(a = rmvnorm(1000, c(0,0), diag(1, nrow = 2, ncol = 2)))$. On note "Meanf" la fonction objective à optimiser et "GrMeanf" son gradient :

```
Meanf<-function(mu) {
mu1<-mu[1];  mu2<-mu[2]
sum(((a[,1]-mu1)^2+(a[,2]-mu2)^2),na.rm=T)
    }

GrMeanf<-function(mu) {
mu1<-mu[1];  mu2<-mu[2]
-2*c(sum(a[,1]-mu1,na.rm=T),sum(a[,2]-mu2,na.rm=T))
  }
```

L'expression de l'estimateur de la moyenne est
$Lamoyenne = c(mean(a[,1]), mean(a[,2]))$. On compare cette valeur avec les résultats de la fonction "optim".
L'évaluation basique de "optim" qui permet d'optimiser la fonction objective. Cette méthode est adaptée pour les fonctions non différentiables : $Mean1 = optim(c(-0.1, 0.1), Meanf)\$par$.
Si la fonction objective est différentiable (le cas ici), alors on peut utiliser sa fonction gradient pour améliorer ou accélérer le calcul. Ce gradient peut être donné par l'utilisateur (ici GrMeanf) ou par une approximation numérique (on précise 'NULL' comme paramètre de la fonction) .

Plusieurs méthodes d'optimisation sont implémentées dans "optim". Nous n'entrerons pas dans le détail de ces méthodes qu'on peut trouver facilement dans l'aide sous R. Nous donnons seulement les valeurs d'optimisation avec 3 méthodes (l'argument "method" dans "optim") pour comparer avec la valeur réelle :

```
Mean1=optim(c(-0.1,0.1),Meanf) $par
Mean2=optim(c(-0.1,0.1),Meanf,GrMeanf,method="BFGS") $par
Mean3=optim(c(-0.1,0.1),Meanf,NULL,method="BFGS")$par
Mean4=optim(c(-0.1,0.1),Meanf,GrMeanf,method="CG")$par
Mean5=optim(c(-0.1,0.1),Meanf,NULL,method="CG")$par
Mean6=optim(c(-0.1,0.1),Meanf,GrMeanf,method="L-BFGS-B")$par
Mean7=optim(c(-0.1,0.1),Meanf,NULL,method="L-BFGS-B")$par
```

Pour chaqu'une des méthodes on calcule les valeurs Mean1 à Mean7 et leurs écarts à la valeur théorique
$Lamoyenne = (-0.01164814, -0.01626191)$
$Basique = (-0.01123831, -0.01629605)$
$Ecart = (-4.098325e-04, 3.413573e-05)$

Methode	Avec gradient	Avec approximation du gradient
BFGS	(-0.01164814, -0.01626191)	(-0.01164815, -0.01626192)
l'écart	(-3.338252e-11, 2.395806e-11)	(9.826777e-09, 7.470087e-09)
L-BFGS-B	(-0.01164814, -0.01626191)	(-0.01164814, -0.01626191)
l'écart	(-1.734723e-17, -6.938894e-18)	(4.148071e-14, 3.547510e-14)
CG	(-0.01164814, -0.01626191)	(-0.01164814, -0.01626191)
l'écart	(3.610442e-09, -4.750970e-09)	(3.609334e-09, -4.749535e-09)

- calcul de la moyenne (optimisation)

```
Moyenne[,s,i]=optim(c(-0.1,0.1),Moyennec)$par
WMoyenne[,s,i]=optim(c(-0.1,0.1),WMoyennec)$par
```

- calcul de Médiane

```
Mediane[,s,i]=optim(c(-0.1,0.1),Medianec)$par
```

```
WMediane[,s,i]=optim(c(-0.1,0.1),WMedianec)$par
```

– calcul de l'estimateur de Huber

```
Huber[,s,i]=optim(c(-0.1,0.1),Huberc)$par
WHuber[,s,i]=optim(c(-0.1,0.1),WHuberc)$par
```

(c) Calcul des risques quadratiques pour chaque estimateur

– risque quadratique de la moyenne

```
    Rmoyenne=Rmoyenne+Moyenne[,s,i]\%*\%t
(Moyenne[,s,i])
    WRmoyenne=WRmoyenne+WMoyenne[,s,i]\%*\%t
(WMoyenne[,s,i])
```

– risque quadratique de la médiane

```
    Rmediane=Rmediane+Mediane[,s,i]\%*\%t
(Mediane[,s,i])
    WRmediane=WRmediane+WMediane[,s,i]\%*\%t
(WMediane[,s,i])
```

– risque quadratique de l'estimateur de Huber

```
    RHuber=RHuber+Huber[,s,i]\%*\%t(Huber[,s,i])
    WRHuber=WRHuber+WHuber[,s,i]\%*\%t(WHuber[,s,i])
```

(d) fin de calcul pour la boucle pour toutes les simulations : }.

6. On calcul les MSE et on ferme la boucle sur le nombre de cluster considéré ("s") :

– MSE Moyenne

```
MSEMoyenne[s]=det(Rmoyenne/N)
WMSEMoyenne[s]=det(WRmoyenne/N)
```

– MSE Mediane

```
MSEMediane[s]=det(Rmediane/N)
WMSEMediane[s]=det(WRmediane/N)
```

– MSE de l'estimateur de Huber

```
MSEHuber[s]=det(RHuber/N)
WMSEHuber[s]=det(WRHuber/N)
```

7. Enregistrement dans un tableau des MSE des estimateurs :

```
simut_m1r2=data.frame(MSEMoyenne,WMSEMoyenne,
MSEMediane,WMSEMediane,MSEHuber,WMSEHuber)$
```

B.2 Optimisation des poids avec Matlab

Sous Matlab, nous réalisons principalement l'optimisation sous contrainte de la matrice de variance-covariance des M-estimateurs pondérés, dans un cadre de données clusterisées bivariées. Les variables aléatoires que nous utilisons sont simulées avec le logiciel R. Nous les importons ensuite dans Matlab.
Dans un premier temps, nous calculons les estimateurs pour les matrices de variance-covariance, et nous appliquons les deux types de simulations, avec ou sans poids (voir l'estimation de la variance chapitre 3 page-68).

– La fonction suivante permet de calculer ces estimateurs pour un Lp-estimateur non pondéré :

```
function [Bchap,Bchap2,Cchap,Cchap2,Vchap,Vchap2]
=bchap_cchap_Lp(x,y,ne,p)
np=10;N=1000;
cm=0;Nn=0;
Bchap=[0 0;0 0];
Bchap2=[0 0;0 0];
Cchap=[0 0;0 0];
Cchap2=[0 0;0 0];
Vchap=[0 0;0 0];
Vchap2=[0 0;0 0];
I=[1 0;0 1];

for i = 1:N
  Bchap1(:,:)=[0 0;0 0] ;
  Bchap12(:,:)=[0 0;0 0] ;
```

```
  Cchap1(:,:)=[0 0;0 0] ;
  Cchap12(:,:)=[0 0;0 0] ;
  Vchap1(:,:)=[0 0;0 0];
  Vchap12(:,:)=[0 0;0 0];
  for s = 1:np
    for j = 1:ne(1,s)
     normj=sqrt(x(j,s,i).^2+y(j,s,i).^2);
     for k = 1:ne(1,s)
       normk=sqrt(x(k,s,i).^2+y(k,s,i).^2);
        if k==j
            Bchap1(:,:)=Bchap1(:,:)+(([x(j,s,i)y(j,s,i)]'*
(p.*normj.^(p-2)))*([x(j,s,i)y(j,s,i)]*(p.*normj.^(p-2))));
            Vchap1(:,:)=Vchap1(:,:)+I*p*normj^(p-2)+
([x(j,s,i)y(j,s,i)]'*[x(j,s,i)y(j,s,i)])*
(p.*(p-2).*normj.^(p-4));
        end
          if k~=j
            Cchap1(:,:)=Cchap1(:,:)+([x(j,s,i) y(j,s,i)]'*
(p.*normj.^(p-2)))*([x(k,s,i)y(k,s,i)]*(p.*normk.^(p-2)));
          end
        end
    end
    norm1=sqrt(x(1,s,i).^2+y(1,s,i).^2);
norm2=sqrt(x(2,s,i).^2+y(2,s,i).^2);
    Bchap12(:,:)=Bchap12(:,:)+(([x(1,s,i) y(1,s,i)]'*
(p.*norm1.^(p-2)))*([x(1,s,i) y(1,s,i)]*(p.*norm1.^(p-2))));
     Cchap12(:,:)=Cchap12(:,:)+([x(1,s,i) y(1,s,i)]'*
(p.*norm1.^(p-2)))*([x(2,s,i)y(2,s,i)]*(p.*norm2.^(p-2)));
     Vchap12(:,:)=Vchap12(:,:)+I*p*norm1^(p-2)+
([x(1,s,i) y(1,s,i)]'*[x(1,s,i) y(1,s,i)])*
(p.*(p-2).*norm1.^(p-4));
    end
  Bchap=Bchap+Bchap1(:,:)/sum(ne)/N;
  Bchap2=Bchap2+Bchap12(:,:)/np/N;
  Cchap=Cchap+Cchap1(:,:)/sum(ne)/N/
```

```
sum((ne.*(ne-1))/sum(ne));
  Cchap2=Cchap2+Cchap12(:,:)/np/N;
  Vchap=Vchap+Vchap1(:,:)/sum(ne)/N;
  Vchap2=Vchap2+Vchap12(:,:)/np/N;
  i %compteur d'iteration
end
end
```

– Nous exécutons la fonction pour avoir les matrices estimées :

```
importfile('xcas1');importfile('ycas1');
%les varaibles aléaoires simulées
importfile('wcas1');importfile('wcas1');
%importation des variables aléatoires
[Bchap1,Bchap21,Cchap1,Cchap21,Vchap1,Vchap21]=
bchap_cchap_Lp(x1,y1,ne(1,:),p)
%p est la valeur de la norme Lp et nous calculons les estimateurs
```

– Nous donnons la fonction à optimiser en fonction des poids :

```
function Wvarmedianef2=Mohammedcas2(w,x,y,ne,p)
np=10;N=1000;
cm=0;Nn=0;
Bchap=[0 0;0 0];
Cchap=[0 0;0 0];
for i = 1:N
  Bchap1(:,:)=[0 0;0 0] ;
  Cchap1(:,:)=[0 0;0 0] ;
  for s = 1:np
    for j = 1:ne(1,s)
     normj=sqrt(x(j,s,i).^2+y(j,s,i).^2);
     psi1=p*(normj.^(p-2))*[x(j,s,i) y(j,s,i)];
     for k = 1:ne(1,s)
       normk=sqrt(x(k,s,i).^2+y(k,s,i).^2);
       psi2=p*(normk.^(p-2))*[x(k,s,i) y(k,s,i)];
```

```
            if k==j
              Bchap1(:,:)=Bchap1(:,:)+(w(1,s).^2)*(psi1' * psi1);
            end
            if k~=j
              Cchap1(:,:)=Cchap1(:,:)+(w(1,s).^2)*(psi1' * psi2);
            end
          end
      end
     end
  Bchap=Bchap+Bchap1(:,:);
  Cchap=Cchap+Cchap1(:,:);
  i
end
Wvarmedianef2=det((Bchap/sum(ne)/N)+(Cchap/sum(ne)/N))
end
```

avec les paramètres d'entrée : w la variable poids, les données x et y, les cas de configuration des clusters ne et le paramètre p de la norme Lp.

– On optimise cette fonction des poids :

```
[wcas1,val1]=fmincon(@Mohammedcas2,x0,[],[],ne(1,:),
sum(ne(1,:)),lb,[],[],[],x1,y1,ne(1,:),p);
save('wcas1.mat','wcas1','val1');
```

avec $wcas1$ le vecteur des poids optimaux et $val1$ la valeur optimale de la fonction.

– Nous calculons le point de rupture avec la fonction suivante :

```
function[sw,cont]=prupture(w,k,N,e0)
ne=[ones(1,9)*4 64;ones(1,5)*4 ones(1,5)*16;
ones(1,2)*4 8 ones(1,7)*12;5 6 7 8 9 11 12 13 14 15];
cas=ne(k,:);sw=0;cont=0;

    for i=1:10
    for j=1:cas(i)
```

```
        if sw<(e0*N)
            sw=sw+w(i);
            cont=cont+1;
        end
    end
    end
end
```

avec $e0$ le point de rupture pour l'estimateur non pondéré, w vecteur des poids optimaux, k le cas de configuration des clusters (nous avons 4 configurations qui sont regroupées dans la matrice ne) et N le nombre de variables.

– On exécute la fonction, pour obtenir le point de rupture (ici $k = 1$ et le point de rupture est cont1%) :

```
[sw1,cont1]=prupture(wcasR081,1,100,0.5)
[sw2,cont2]=prupture(wcasR082,2,100,0.5)
[sw3,cont3]=prupture(wcasR083,3,100,0.5)
[sw4,cont4]=prupture(wcasR084,4,100,0.5)
res08=[sw1 sw2 sw3 sw4;cont1 cont2 cont3 cont4]
save('rup08.mat','res08');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[sw1,cont1]=prupture(wcas1,1,100,0.5)
[sw2,cont2]=prupture(wcas2,2,100,0.5)
[sw3,cont3]=prupture(wcas3,3,100,0.5)
[sw4,cont4]=prupture(wcas4,4,100,0.5)
res02=[sw1 sw2 sw3 sw4;cont1 cont2 cont3 cont4]
save('rup02.mat','res02');
```

Bibliographie

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust estimates of location : survey and advances*, volume 173. Princeton University Press Princeton, NJ. 27

Arcones, M. A. (1998). Asymptotic theory for m-estimators over a convex kernel. *Econometric Theory*, 14(4) :387–422. 9

Bickel, P. (1964). On some robust estimates of location. In *Annals of Mathematical Statistics*, volume 35, page 1403. Inst Mathematical Statistics. 31

Brown, B. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 25–30. 30

Brunet, R., Ferras, R., and Théry, H. (1993). *Les mots de la géographie : dictionnaire critique*. Reclus, Paris. 9

Chakraborty, B. and Chaudhuri, P. (1996). On a transformation and re-transformation technique for constructing an affine equivariant multivariate median. *Proceedings of the American Mathematical Society*, 124(8) :2539–2547. 30

Chakraborty, B. and Chaudhuri, P. (1998). On an adaptive transformation–retransformation estimate of multivariate location. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 60(1) :145–157. 30

Chakraborty, B. and Chaudhuri, P. (1999). A note on the robustness of multivariate medians. *Statistics & Probability Letters*, 45(3) :269–276. 30

Chao, M.-T. (1986). On m and p estimators that have breakdown point equal to 12. *Statistics & probability letters*, 4(3) :127–131. 27

Chaudhuri, P. (1992). Multivariate location estimation using extension of r-estimates through u-statistics type approach. *The Annals of Statistics*, pages 897–916. 29, 30

Chen, Z. and Tyler, D. E. (2004). On the finite sample breakdown points of redescending M-estimates of location. *Statist. Probab. Lett.*, 69(3) :233–242. 27

Croux, C. and Rousseeuw, P. J. (1992). A class of high-breakdown scale estimators based on subranges. *Communications in statistics-theory and methods*, 21(7) :1935–1951. 83

Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423) :782–792. 27

Davies, P. L., Gather, U., et al. (2005). Breakdown and groups. *The Annals of Statistics*, 33(3) :977–1035. 27

Donoho, D. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA. 27, 81

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http ://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf. 27

Field, C. A. and Hampel, F. R. (1982). Small-sample asymptotic distributions of M-estimators of location. *Biometrika*, 69(1) :29–46. 27

Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), University of California, Berkeley. 27

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896. 27

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346) :383–393. 24

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York. The approach based on influence functions. 9, 15, 24

Huber, P. (1981). *Robust statistics*. Wiley, New York. 15, 25

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1) :73–101. 9, 15, 17, 20

Huber, P. J. (1977). Robust methods of estimation of regression coefficients 1. *Statistics : A Journal of Theoretical and Applied Statistics*, 8(1) :41–53. 24

Huber, P. J. (1997). Robustness : Where are we now ? *Lecture Notes-Monograph Series*, pages 487–498. 27

Lévy, J. and Lussault, M. (2003). *Dictionnaire de la géographie et de l'espace des sociétés*. 9

Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, pages 229–248. 30

Maronna, R., Martin, R., and Yohai, V. (2006). Robust statistics : Theory and methods. 27

Milasevic, P., Ducharme, G., et al. (1987). Uniqueness of the spatial median. *The Annals of Statistics*, 15(3) :1332–1333. 36

Nevalainen, J., Larocque, D., and Oja, H. (2007a). On the multivariate spatial median for clustered data. *Canadian Journal of Statistics*, 35(2) :215–231. 48, 66

Nevalainen, J., Larocque, D., and Oja, H. (2007b). A weighted spatial median for clustered data. *Statistical Methods and Applications*, 15(3) :355–379. 59, 87

Rousseeuw, P. J. (1981). A new infinitesimal approach to robust estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56(1) :127–132. 25

Ruiz-Gazen, A. (2012). Robust statistics : a functional approach. *Ann. Inst. Statist. Univ. Paris*, 56(2-3) :49–64. 15

Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York. 18, 20, 21, 97, 98

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277. 29

Tukey, J. W. and Wilk, M. B. (1970). Data analysis and statistics : techniques and approaches. *The quantitative analysis of social problems*, pages 370–390. 27

Van der Vaart, A. (2000). *Asymptotic statistics*. Cambridge University Press, Cambridge. 9, 15, 19, 20, 22, 29, 30, 97