MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU

FACULTE DES SCIENCES

DEPARTEMENT DE MATHEMATIQUES

THESE DE DOCTORAT

SPECIALITE: MATHEMATIQUES

OPTION: PROBABILITES ET STATISTIQUE

Présentée par:

M^{elle} **CHEIKH Malika**

Sujet:

Contribution à l'étude de l'influence en analyse des données

Devant le jury d'examen composé de :

Mr Morsli Mohamed;	Professeur;	U.M.M.T.O;	Président.
Mme Bedouhene Fazia;	Professeur;	U.M.M.T.O;	Rapporteur.
Mr Fellag Hocine;	Professeur;	U.M.M.T.O;	Examinateur.
Mr Necir Abdelhakim;	Professeur;	U. Biskra;	Examinateur.
Mr Yousfate Abderrahmane;	Professeur ;	U.S.Belabbes;	Examinateur.
Mme Djeddour Khedidja;	Maître de conférence A;	U.S.T.H.B;	Examinatrice.

Soutenue: le 15 /05 / 2014

Remerciements

Je tiens à exprimer toute ma reconnaissance à Monsieur Mohamed-Hamou IBAZIZEN mon premier directeur de thèse, qui a proposé le thème de ce travail, et qui m'a fait découvrir ce domaine si riche de l'analyse des données et de la statistique robuste.

Je tiens à exprimer toute ma gratitude à Mme Fazia BEDOUHENE ma directrice actuelle de thèse, Professeur à l'université Mouloud MAMMERI de Tizi-Ouzou, qui a accepté de m'encadrer et de diriger ce travail. Ses remarques, ses conseils m'ont été indispensable dans la rédaction de ce document. Je la remercie de m'avoir guidé jusqu'au bout pour la réalisation de ce manuscrit.

Il m'est très agréable de remercier le Professeur Mohamed MORSLI de l'université Mouloud MAMMERI de Tizi-Ouzou, qui m'a fait l'honneur de présider le jury de cette thèse.

Je tiens à remercier le Professeur Hocine FELLAG de l'université Mouloud MAMMERI de Tizi-Ouzou, d'avoir accepté de faire partie de mon jury.

Mes remerciements vont particulièrement vers le Professeur Abdelhakim NECIR de l'université de Biskra, d'avoir accepté mon invitation et être parmi les membres de jury.

J'ai l'honneur de remercier le Professeur Abderrahmane YOUSFATE de l'université de Sidi Bel-Abbes, de faire partie de mon jury et d'avoir accepté d'examiner ce travail.

Mes remerciements chaleureux s'adressent également à Mme Khedidja DJEDDOUR de l'université de Bab Ez-zouar, d'avoir accepté d'examiner mon travail et de faire partie de mon jury d'examination.

Table des matières

Introduction générale

L'analyse des données recouvre un ensemble de techniques ayant pour objectif la description statistique des grands tableaux. Ces techniques permettent de rechercher les structures cachées dans les données, et d'obtenir une description de nature statistique pour un certain phénomène qui a donné lieu au recueil de mesures ou observations trop nombreuses et dépendantes les unes des autres pour être interprétables en première lecture.

Les domaines d'utilisation de l'analyse des données sont nombreux et diversifiés : biométrie, psychométrie, économétrie, medecine, etc .... Le besoin d'outils permettant d'obtenir des variables synthétiques résumant l'information disponible se fait alors ressentir; c'est précisément l'objectif de l'analyse des données et des méthodes d'analyse multivariée. Elles se caractérisent par leur objectif exploratoire et par l'abandon d'hypothèses probabilistes (contrairement à la statistique inférentielle) au profit de la géométrie euclidienne.

Il existe une multitude de méthodes d'analyse multivariée permettant de traiter différentes structures de données, notamment, l'analyse en composantes principales (A.C.P.) pour un tableau de variables quantitatives, l'analyse factorielle des correspondances pour les tables de contingence, l'analyse factorielle multiple pour les variables qualitatives, l'analyse discriminante pour la prise en compte d'une partition des individus en groupe et les méthodes de couplage pour l'analyse d'une paire de tableaux.

L'origine de ces méthodes remonte au moins à K. Pearson [44], mais leurs pratique n'est devenue courante que depuis l'ère informatique et ont été surtout développées en France dans les années 60 et 70, en particulier par Jean-Paul Benzécri ([3], [4], [5], [6]) qui a beaucoup exploité les aspects géométriques et les représentations graphiques.

Ainsi, à partir d'un tableau rectangulaire de données comportant les observations de p variables quantitatives sur n individus, on peut obtenir des représentations géométriques de ces individus qui sont dans $\mathbb{R}^p$ dans un sous espace E_q de faible dimension $(q < p)$, grâce à l'analyse en composantes principales. Cette méthode consiste à chercher ce sous espace tel que l'inertie du nuage projeté sur E_q soit maximale. En projetant les individus sur E_q, on obtiendra de nouvelles variables appelées *composantes principales*.

Le vecteur moyen et la matrice de covariance classiques sont des instruments

très utiles en A.C.P. et dans la plupart des méthodes statistiques multivariées (Analyse discriminante, Régression multiple,...). Ces derniers sont malheureusement très sensibles à des observations aberrantes qui ne sont parfois pas liées au phénomène étudié (appelées en anglais "outliers").

Dans la littérature, il existe plusieurs travaux de chercheurs qui se proposent de diminuer l'influence de ces outliers, la première approche consiste à remplacer le vecteur moyen et la matrice de covariance empiriques par leurs versions robustes tel que le M-estimateur (Maronna, [42]). Cette approche est étroitement liée à la distance de Mahalanobis, définie par:

$$d^2(x) = (x-\mu)'\Sigma^{-1}(x-\mu),$$

où x est un vecteur aléatoire dans $\mathbb{R}^p$ de moyenne μ et de matrice de covariance Σ. Cette distance joue un rôle fondamental dans la M-estimation multidimensionnelle de μ et Σ.

L'idée est de construire des "poids" w_i, inversement proportionnels à la distance $d(x_i)$. Les points extrêmes voient ainsi leurs poids diminuer et les M-estimateurs de μ et Σ sont définis par :

$$\hat{\mu}_M = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad \text{et} \quad \hat{\Sigma}_M = \frac{\sum_{i=1}^n w_i (x_i-\hat{\mu})(x_i-\hat{\mu})'}{\sum_{i=1}^n w_i}.$$

D'autres variantes de cette approche ont été construites, entre autres, les estimateurs "minimum covariance determinant estimator" (MCD) et leurs versions repondérées (Rousseeuw, [50]; Croux & Haesbroeck, [15]), le S-estimateur (Davies, [20]), et ceux proposés par Ma & Genton [41] et Kamiya [34].

Dans le même contexte, une autre approche primordiale est basée sur la notion de "fonction d'influence", introduite par Hampel [26], qui est un "outil" mesurant directement l'influence (relative) d'une observation sur un estimateur T.

La fonction d'influence a été d'abord employée comme outil-diagnostic, accompagnant les méthodes statistiques classiques unidimensionnelles. On peut citer dans ce cadre les travaux de Critchley [13], Jolliffe et Morgan [33], Croux & Joossens [17], et Croux & Gazen [18].

Par ailleurs, elle a été utilisée pour établir la robustesse d'un estimateur T, au sens ou la robustesse est synonyme de fonction d'influence bornée.

La fonction d'influence joue un rôle fondamental dans la détection des observations influentes en ACP, en particulier, sur le sous espace principal E_q $(q < p)$ engendré par les q premières composantes principales. Dans cette optique, de nouvelles mesures

d'influence ont été aussi construites. Plusieurs auteurs ont étudié l'influence sur le sous espace E_q, notamment:

- L'approche de Tanaka [57] est basée sur le calcul des fonctions d'influence du projecteur orthogonal P sur E_q et de Q définis par:

$$P = \sum_{j=1}^{q} v_j v_j',$$

$$Q = \sum_{j=1}^{q} \lambda_j v_j v_j',$$

où λ_j désigne la $j^{ème}$ valeur propre de la matrice de covariance ou de corrélation, v_j le vecteur propre correspondant et v_j' représente la transposé de v_j.

- Bénasséni [2] a calculé la fonction d'influence de coefficient de sensibilité ρ défini comme suit:

$$\rho = 1 - (\frac{1}{q}\sum_{j=1}^{q} \|v_j - \tilde{P}v_j\|),$$

où v_j désigne le $j^{ème}$ vecteur propre de la matrice de covariance et $\tilde{P} = \tilde{V}\tilde{V}'$, $\tilde{V} = (\tilde{v}_1, \cdots, \tilde{v}_q)$ est la matrice des q premiers vecteurs propres de la matrice de covariance associée à la loi contaminée $\tilde{F}$, où $\tilde{F} = (1-\varepsilon)F + \varepsilon\delta_x$, et δ_x est la mesure de Dirac au point x de $\mathbb{R}^p$.

Ce coefficient mesure la proximité entre le sous espace E_q et le sous espace $\tilde{E}_q$ résultat d'une perturbation $\tilde{F}$ de la loi des observations F.

- L'approche de Prendergast [46] basée sur la mesure suivante:

$$\tilde{\rho}_S(C_0,F_n,x) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}} \frac{y_j^2 y_r^2}{(\hat{\lambda_j} - \hat{\lambda_r})^2},$$

avec $y_j = \hat{v_j'}(x - \hat{\mu})$, où $\hat{v_j}$ et $\hat{\lambda_j}$, $j = 1, \cdots, p$ sont les éléments propres de la matrice de covariance empirique, et $\hat{\mu}$ représente l'estimateur classique du vecteur moyen.

- Récemment, Prendergast & Li Wai Suen [47] propose une nouvelle mesure (SCI) (squared canonical influence), qui est basée sur la moyenne des carrées des corrélations canoniques, elle est donnée comme suit:

$$SCI = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}} \frac{\hat{\lambda_r}}{\hat{\lambda_j}} \frac{y_j^2 y_r^2}{(\hat{\lambda_j} - \hat{\lambda_r})^2},$$

avec $\hat{\lambda_j}$, $j = 1, \cdots, p$ sont les valeurs propres de la matrice de covariance empirique.

Toutes ces mesures sont basées sur la matrice de covariance ou de corrélation classiques. Dans cette thèse, nous nous intéressons au coefficient de Bénasséni en ACP robuste. Ainsi dans notre première contribution (voir Cheikh & Ibazizen [11]), nous reprenons la démarche de Bénasséni, mais appliquée cette fois-ci à une A.C.P. basée sur un estimateur robuste de la matrice de covariance, en l'occurrence le MCD[1]:

- Premièrement, nous avons généralisé la fonction d'influence de ρ ($IF(\rho,F)$) dans la cas d'une A.C.P basée sur un estimateur quelconque C_n de Σ.

- Deuxièmement, nous avons caractérisé $IF(\rho,F)$ lorsque C_n est le MCD[1] estimateur.

- Cette étude est achevée par une étude comparative entre les fonctions d'influence de ρ dans le cas classique et robuste à travers deux exemples numériques.

Dans notre seconde contribution (voir Cheikh [12]), nous avons étendu les résultats obtenus dans Cheikh & Ibazizen [11] aux cas des estimateurs MCD et le S-estimateurs. Plus précisément, nous avons déduit la formule caractérisant la fonction d'influence correspondante à ces deux estimateurs.

- La représentation graphique de la fonction d'influence de ρ nous a permis d'établir une comparaison entre l'estimateur classique, le MCD, le MCD[1] et le S-estimateur. Cette comparaison est faite pour $p = 2$. Lorsque $p > 2$, nous avons pris en considération un autre critère de comparaison qui est la sensibilité aux grosses erreurs $\text{GES}(\rho_C,F) = \sup_x |IF(x,\rho,F)|$, cette quantité mesure la plus mauvaise influence possible qu'une petite fraction de contamination peut provoquer sur la valeur d'un estimateur .

- Dans la troisième partie, nous avons effectué une étude empirique de l'erreur quadratique moyenne de $\hat{\rho}$ (MSE ($\hat{\rho}$)), où $\hat{\rho}$ est un estimateur de ρ. Pour la loi de l'échantillon, nous avons pris deux lois: la loi multinormale sans contamination et la loi multinormale avec contamination. Grâce à cette étude, nous avons pu conclure que le S-estimateur est l'estimateur le plus résistant aux valeurs aberrantes.

Notre manuscrit est constitué d'une introduction générale, de 5 chapitres et d'une conclusion générale.

Le premier chapitre de notre travail est centré sur L'ACP, Nous rappellerons l'essentielle de la méthode qui consiste à chercher un sous espace E_q de dimension faible ($q < p$) tel que l'inertie du nuage projeté sur E_q soit maximale.

Le deuxième chapitre sera consacré aux notions de base de la théorie de la robustesse, à savoir: les valeurs aberrantes, maximum du biais et la fonction d'influence. Nous donnons aussi les définitions ainsi que les algorithmes de calcul des estimateurs suivants: Le MCD, le MCD^1 et le S-estimateur.

Le troisième chapitre est dédié aux différentes mesures d'influence utilisées pour étudier l'influence des observations influentes en Analyse en composantes principales (A.C.P.). Nous citerons le coefficient de Bénasséni [2], la mesure de Prendergast [46] et la mesure de Prendergast & Li Wai Suen [47].

Le quatrième chapitre constitue notre première contribution: Coefficient de sensibilité en A.C.P robuste. D'une part, nous avons caractérisé la fonction d'influence du coefficient de sensibilité de Bénasséni lorsqu'on utilise le MCD^1 estimateur, d'une autre part, nous avons effectué une étude comparative des fonctions d'influences de ρ dans la cas classique et robuste.

La seconde contribution est dans le cinquième chapitre, où nous utiliserons deux estimateurs, le MCD et le S-estimateur. Une étude comparative de ces estimateurs est effectuée en calculant les fonctions d'influence du coefficient de sensibilité ρ correspondantes à ces estimateurs, et l'erreur quadratique moyenne (MSE) des estimateurs de ρ.

Notre manuscrit se termine par une conclusion générale et quelques perspectives de recherche.

Chapitre 1

Analyse en composantes principales

1.1 Introduction

L'analyse en composantes principales (A.C.P.) est une méthode de statistique exploratoire permettant de décrire un grand tableau de données de type individus / variables. Lorsque les individus sont décrits par un nombre important de variables, aucune représentation graphique simple ne permet de visualiser le nuage de points formé par les données.

L'ACP propose une représentation dans un espace de dimension réduite, permettant ainsi de mettre en évidence d'éventuelles structures au sein des données. Pour cela, nous recherchons les sous-espaces dans lesquels la projection du nuage déforme le moins possible le nuage initial.

Ce chapitre est consacré à l'analyse en composantes principales (A.C.P), il comporte deux parties. Après avoir rappelé les définitions et les notations nécessaires pour ce chapitre: la matrice des poids, le vecteur moyen, la matrice de covariance, espace des individus et de variables dans la première partie, nous passerons à la partie essentielle de ce chapitre qui est l'analyse, qui consiste à trouver un sous espace E_q de dimension faible $(q < p)$ tel que l'inertie du nuage projeté sur E_q soit maximale.

1.2 Notations et définitions

1.2.1 Tableau des données

Les observations de p variables sur n individus sont rassemblés dans un tableau rectangulaire X à n lignes et p colonnes:

$$X = \begin{pmatrix} X_1^1 & X_1^2 & . & . & . & . & X_1^p \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ X_i^1 & X_i^2 & . & . & . & . & X_i^p \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ X_n^1 & X_n^2 & . & . & . & . & X_n^p \end{pmatrix}. \tag{1.1}$$

On identifie la $j^{ième}$ variable à la $j^{ième}$ colonne de X par:

$$X^j = \begin{pmatrix} X_1^j \\ X_2^j \\ . \\ . \\ . \\ X_n^j \end{pmatrix}, \tag{1.2}$$

ce sont les valeurs prises par X^j $(j = 1, \cdots, p)$ sur les n individus.

On identifie le $i^{ième}$ individu à la $i^{ième}$ ligne de X noté x_i $(i = 1,2,\ldots,n)$, avec :

$$x_i = (X_i^1, X_i^2, ..., X_i^p)'.$$

1.2.2 Matrice des poids

Si les données ont été recueillies à la suite d'un tirage aléatoire à probabilités égales, les n individus ont tous les mêmes importances, $\frac{1}{n}$, dans le calcul des caractéristiques de l'échantillon. Il n'en est pas toujours ainsi et il est utile pour certaines applications de travailler avec des poids p_i, $(i = 1, \cdots, p)$ éventuellement différents d'un individu à l'autre (échantillon redressés; données regroupées, $\cdots$). Soit $p_1, p_2, p_3, \cdots, p_n$ les poids respectifs des individus, avec $p_i \geq 0$, $(\forall i = 1, \cdots, n)$, et $\sum_{i=1}^{n} p_i = 1$.
Ces poids sont regroupés dans une matrice diagonale D de taille n:

$$D = \begin{pmatrix} p_1 & 0 & . & . & . & . & 0 \\ 0 & p_2 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & 0 & . & . \\ . & . & . & . & . & . & 0 \\ 0 & 0 & . & . & . & 0 & p_n \end{pmatrix} \tag{1.3}$$

D est appelé *matrice diagonale des poids.*

Si les poids sont uniformes i.e. $p = \dfrac{1}{n}$, on a:

$$D = \frac{1}{n} I_n,$$

I_n est la matrice identité d'ordre n.

1.2.3 Vecteur Moyen ou centre de gravité

Le vecteur moyen empirique ou centre de gravité du nuage est donné par :

$$g = \begin{pmatrix} \bar{X^1} \\ \bar{X^2} \\ . \\ . \\ . \\ \bar{X^p} \end{pmatrix}, \tag{1.4}$$

où $\quad \bar{X^j} = \sum_{i=1}^{n} p_i X_i^j \ (i,j = 1, \cdots, p)$.

Remarquons que g peut s'écrire comme suit:

$$g = X^{'} D 1_n,$$

où $\quad 1_n = (1,1,...,1)^{'}$, et D est donnée par (1.3).

Au tableau X, on peut associer le tableau centré suivant:

$$Y = \begin{pmatrix} X_1^1 - \bar{X^1} & X_1^2 - \bar{X^2} & . & . & . & . & X_1^p - \bar{X^p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ X_i^1 - \bar{X^1} & X_i^2 - \bar{X^2} & . & . & . & . & X_i^p - \bar{X^p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ X_n^1 - \bar{X^1} & X_n^2 - \bar{X^2} & . & . & . & . & X_n^p - \bar{X^p} \end{pmatrix}. \tag{1.5}$$

Les nouvelles variables $X^j - \bar{X^j}, \quad (j = 1, \cdots, p)$ sont centrées, i.e. de moyenne nulle, dans ce cas le centre de gravité vaut $g = 0_{\mathbb{R}^p}$.

La formule (1.5) s'écrit sous la forme matricielle suivante:

$$Y = X - \mathbf{1}_n g',$$

où X est donnée par (1.1) et g par (1.4).

1.2.4 Matrice de variance-covariance

On appelle matrice de variance-covariance empirique notée, Σ_e, associé à p variables aléatoires X^1,X^2,$\cdots$,X^p, mesurées sur un ensemble de n individus, la matrice carrée d'ordre p contenant sur sa diagonale principale les variances empiriques des p variables, et ailleurs, les covariances empiriques de ces variables deux à deux, autrement dit :

$$\Sigma_e = \begin{pmatrix} var(X^1) & cov(X^1,X^2) & . & . & cov(X^1,X^p) \\ cov(X^2,X^1) & var(X^2) & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ cov(X^p,X^1) & . & . & . & var(X^p) \end{pmatrix} = \begin{pmatrix} \delta_1^2 & \delta_{12} & \delta_{13} & . & . & \delta_{1p} \\ \delta_{12} & \delta_2^2 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \delta_{p1} & . & . & . & . & \delta_P^2 \end{pmatrix}, \tag{1.6}$$

où

$$\delta_{jk} = cov(X^j,X^k) = \sum_{i=1}^{n} p_i(X_i^j - \bar{X}^j)(X_i^k - \bar{X}^k), \ \ avec \ \ (j,k = 1,\cdots,p),$$

et

$$\delta_j = \Big(var(X^j)\Big)^{1/2} = \Big(\sum_{i=1}^{n} p_i(X^j - \bar{X}^j)^2\Big)^{1/2}, \ \ pour(j = 1,\cdots,p).$$

Σ_e peut s'écrire comme suit :

$$\Sigma_e = X^{'}DX - gg^{'} = Y^{'}DY, \tag{1.7}$$

où X est donnée par (1.1), D par (1.3), g par (1.4) et Y par (1.5).

Cas particulier: Si $g = 0_{\mathbb{R}^p}$, alors:

$$\Sigma_e = X^{'}DX.$$

1.2.5 Matrice de corrélation

La matrice de corrélation empirique R_e associé aux variables $(X^1,X^2,\cdots,X^p)$ est donnée par:

$$R_e = \begin{pmatrix} 1 & r_{12} & r_{13} & . & . & r_{1p} \\ r_{21} & 1 & . & . & . & . \\ r_{31} & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ r_{p1} & . & . & . & . & 1 \end{pmatrix}, \tag{1.8}$$

où

$$r_{jk} = cor(X^j, X^k) = \frac{\delta_{jk}}{\delta_j \delta_k}, \qquad et \qquad (j,k = 1, \cdots, p).$$

Si l'on note $D_{\frac{1}{\delta}}$ la matrice des inverses des écarts-types, celle-ci est donnée comme suit:

$$D_{\frac{1}{\delta}} = \begin{pmatrix} \frac{1}{\delta_1} & 0 & . & . & 0 \\ 0 & \frac{1}{\delta_2} & . & . & 0 \\ . & 0 & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \frac{1}{\delta_p} \end{pmatrix}. \tag{1.9}$$

Le tableau des données centrées et réduites, Z, s'écrit sous la forme matricielle suivante:

$$Z = Y D_{\frac{1}{\delta}},$$

où Y et donnée par (1.5).

Dans ce cas, R_e peut s'écrire comme suit:

$$R_e = D_{\frac{1}{\delta}} \Sigma_e D_{\frac{1}{\delta}} = D_{\frac{1}{\delta}} (Y'DY) D_{\frac{1}{\delta}} = Z'DZ, \tag{1.10}$$

où D est définie par (1.3) et Σ_e par (1.6).

Dans le cas centré réduit, on aura :

$$R_e = I_p \Sigma_e I_p = \Sigma_e.$$

On peut dire que R_e est la matrice de variance-covariance des données centrées et réduites, et résume la structure de dépendance linéaire entre deux variables .

1.3 Espace des individus

Chaque individu est représenté par $x_i = (X_i^1, X_i^2, \cdots, X_i^p)'$ de $\mathbb{R}^p$. L'ensemble $\mathbb{R}^p$ est considéré comme l'espace des individus. L'ensemble des individus forme un nuage $\mathcal{N}$ dans $\mathbb{R}^p$.

1.3.1 Le Rôle de la métrique

Comment mesurer la distance entre deux individus de $\mathbb{R}^p$? En physique, la distance entre deux points de l'espace se calcule facilement avec la formule de Pythagore: le carrée de la distance est la somme des carrés des différences des coordonnées, car les dimensions sont de même nature. Il n'en est pas de même en statistique, où chaque dimension correspond à un caractère qui s'exprime avec son unité particulière: comment alors calculer la distance entre deux individus décrits par les trois variables (âge, salaire, nombre d'enfants)? il y a manifestement un problème d'unité de mesure. Illustrons cela par l'exemple suivant:

Soit les trois observations.

	Âge	Salaire	Nombre d'enfants
x_1	30	30000	2
x_2	31	31000	3
x_3	60	30000	10

L'utilisation de la formule de Pythagore, entraîne :
$d^2(x_1,x_2) = \|x_1 - x_2\|^2 = (31-30)^2 + (30000-31000)^2 + (3-2)^2 = 2.10^6$.
$d^2(x_1,x_3) = \|x_1 - x_3\|^2 = (60-30)^2 + (30000-30000)^2 + (10-2)^2 = 964$.

D'après ces calculs, x_1 est proche de x_3, loin de x_2. Mais, logiquement d'après le tableau, on remarque que x_1 et x_2 sont beaucoup plus proche c'est-à-dire (même catégorie) contrairement à x_3 qui parait loin de x_1. Il faudra penser à introduire une distance qui donne une importance relative à chaque variable, pourquoi ne pas prendre une formule de type:

$$d^2(x_i,x_k) = b_1(X_i^1 - X_k^1)^2 + b_2(X_i^2 - X_k^2)^2 + \cdots + b_p(X_i^p - X_k^p)^2.$$

Ce qui revient à multiplier chaque variable X^j par $\sqrt{b_j}$ (on prendra bien sûr des b_j positifs).

On utilisera donc la formulation générale suivante: la distance entre deux individus x_i et x_k est donnée par:

$$d_M(x_i,x_k) = \|x_i - x_k\|_M = (\prec x_i - x_k, x_i - x_k \succ_M)^{\frac{1}{2}} = \sqrt{(x_i - x_k)' M (x_i - x_k)},$$

où M est une matrice symétrique définie positive de taille p. L'espace des individus est donc muni du *produit scalaire:*

$$\prec x_i, x_k \succ_M = x_i' M x_k.$$

La matrice M qui définit le produit scalaire et donc des distances entre individus et appelée *métrique.*

En A.C.P, on utilise deux métriques:

$$M = I_p = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{1.11}$$

ou bien

$$M = D_{\frac{1}{\delta^2}} = \begin{pmatrix} \frac{1}{\delta_1^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\delta_2^2} & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\delta_p^2} \end{pmatrix}. \tag{1.12}$$

Remarques 1.1.

1. Le choix de $M = D_{\frac{1}{\delta^2}}$, revient à diviser chaque variable sur son écart type δ, ce qui est très utile lorsque les variables ne s'expriment pas avec les mêmes unités. Dans ce cas, la distance entre deux individus ne dépend pas des unités de mesures car les nouvelles données sont sans dimension.

2. L'utilisation de la métrique $D_{\frac{1}{\delta^2}}$ est équivalent à l'utilisation de la métrique $M = I_p$ sur les données centrées et réduites.

3. L'utilisation de $M = I_p$ conduirait à privilégier les variables les plus dispersées, pour les quelles les différences entre individus sont les plus forte, et à négliger les différences entre les autres variables. La métrique $D_{\frac{1}{\delta^2}}$ rétablit alors l'équilibre entre les variables en donnant à toute la variance 1.

1.3.2 L'inertie

On appelle inertie totale de nuage de points la moyenne pondérée des carrés des distances des points au centre de gravité:

$$I_g = \sum_{i=1}^{n} p_i \|x_i - g\|_M^2 = \sum_{i=1}^{n} p_i (x_i - g)' M (x_i - g).$$

I_g mesure la dispersion du nuage des points autour de g, elle généralise la variance qui est une mesure de la dispersion des points sur un axe autour de la moyenne.

L'inertie en un point x_0 quelconque est définie par :

$$I_{x_0} = I_g + (g - x_0)' M (g - x_0) = I_g + \|g - x_0\|_M^2.$$

Si $g = 0$, on aura:

$$I_g = \sum_{i=1}^{n} p_i x_i^{'} M x_i.$$

Dans ce cas, on remarque que:

$$\begin{aligned} I_g &= Trace(M\Sigma_e) \\ &= Trace(\Sigma_e M). \end{aligned}$$

En effet, grâce à la commutativité sous la trace, on obtient:

$$\begin{aligned} I_g &= Trace(\sum_{i=1}^{n} p_i x_i^{'} M x_i) \\ &= Trace(\sum_{i=1}^{n} p_i M x_i x_i^{'}) \\ &= Trace(M \sum_{i=1}^{n} p_i x_i x_i^{'}) \\ &= Trace(M\Sigma_e) \\ &= Trace(\Sigma_e M). \end{aligned}$$

Remarques 1.2.

1. Si $M = I_p$, l'inertie est égale à la somme des variances des p variables.

2. Si M=$D_{\frac{1}{\delta^2}}$:

$$\begin{aligned} Trace(M\Sigma_e) &= Trace(MD_{\frac{1}{\delta^2}}) \\ &= Trace(D_{\frac{1}{\delta}}\Sigma_e D_{\frac{1}{\delta}}) \\ &= Trace(R_e), \end{aligned}$$

   dans ce cas, l'inertie est égale au nombre de variables et ne dépend pas de leur valeurs.

1.4 Espace des variables

Chaque variable X^j est une suite de valeurs numériques:

$$X^j = (X^j(x_1), X^j(x_2), \cdots, X^j(x_n))' = (X_1^j, X_2^j, \cdots, X_n^j)'$$

de l'espace $\mathbb{R}^n$ considéré comme l'espace des variables.

1.4.1 La métrique des poids

Pour étudier les proximités des variables, il faut munir cet espace d'une métrique, c'est-à-dire trouver une matrice $\mathbb{R}^n$ d'ordre n symétrique définie positive. Pour l'espace des variables, le choix se porte sur la matrice diagonale des poids D pour les raisons suivantes:

1. Le produit scalaire entre deux variables X^j et X^k qui vaut:

$$\prec X^j, X^k \succ_D = (X^j)' D X^k = \sum_{i=1}^{n} p_i X_i^j X_i^k,$$

   n'est rien d'autre que la covariance δ_{jk} si les deux variables sont centrées.

2. La norme d' une variable $\|X^j\|_D$ est alors $\|X^j\|_D^2 = \delta_j^2$, ce qui veut dire que la longueur d'une variable est égale à son écart-type.

3. Le cosinus de l'angle entre deux variables centrée X^j et X^k dans $\mathbb{R}^n$ est donné par:

$$\cos\theta_{jk} = \frac{\prec X^j, X^k \succ_D}{\|X^j\|_D \|X^k\|_D} = \frac{\delta_{jk}}{\delta_j \delta_k}.$$

Le cosinus de l'angle entre deux variables centrées n'est rien d'autre que le coefficient de corrélation linéaire .

Si dans l'espace des individus $\mathbb{R}^p$ on s'intéresse aux distances entre points, dans l'espace des variables $\mathbb{R}^n$, on s'intéresse plutôt aux angles en raison de la propriété précédente.

1.4.2 Variables engendrées par un tableau de données

Dans tous ce qui suit, on suppose que les données sont centrées (i.e, $g = 0_{\mathbb{R}^p}$).

A une variable X^j, on peut associer un axe de l'espace des individus $\mathbb{R}^p$ et un vecteur de l'espace des variables $\mathbb{R}^n$.

On peut également déduire de $X^1, X^2, \cdots, X^p$ de nouvelles variables par combinaison linéaire, ce qui revient à projeter les individus sur de nouveaux axes de $\mathbb{R}^p$.

Considérons un axe Δ de l'espace des individus engendré par un vecteur unitaire a.

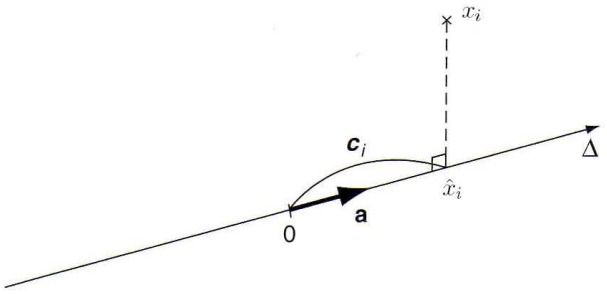

FIG. 1.1 – *La projection d'un individu sur un axe*

La liste des coordonnées c_i $(i = 1, \cdots, n)$ des individus sur Δ forme une nouvelle variable ou composante C.

Comme $c_i =\prec x_i, a \succ_M = x_i^{'} Ma$, alors :

$$C = \begin{pmatrix} c_1 \\ . \\ . \\ . \\ . \\ c_n \end{pmatrix} = XMa.$$

En posant $f = Ma$, on aura alors:

$$C = Xf = \sum_{j=1}^{p} X^j f_j.$$

Le vecteur f est appelé facteur.

Remarque 1.1. Lorsque $M = I_p$, ces distinctions disparaissent et on peut identifier totalement axes et facteur).

La variance de C vaut alors:

$$Var(C) = f^{'} \Sigma_e f.$$

En effet:

$$\begin{aligned} Var(C) &= C^{'} DC \\ &= (Xf)^{'} D(Xf) \\ &= f^{'}(X^{'} DX)f \\ &= f' \Sigma_e f. \end{aligned}$$

1.5 L'analyse

1.5.1 Projection des individus sur un sous-espace

Le principe de la méthode est d'obtenir une représentation approchée du nuage des n individus dans un sous-espace E_q de dimension faible. Ceci s'effectue par la projection ainsi que l'illustre la figure Fig 1.2.

Le choix de l'espace de projection s'effectue selon le critère suivant qui revient à déformer le moins possible les distances en projection. En effet, en projection les distances ne peuvent que diminuer. En d'autre termes , il faut que l'inertie du nuage projeté sur le sous-espace E_q $(q < p)$ soit maximale.

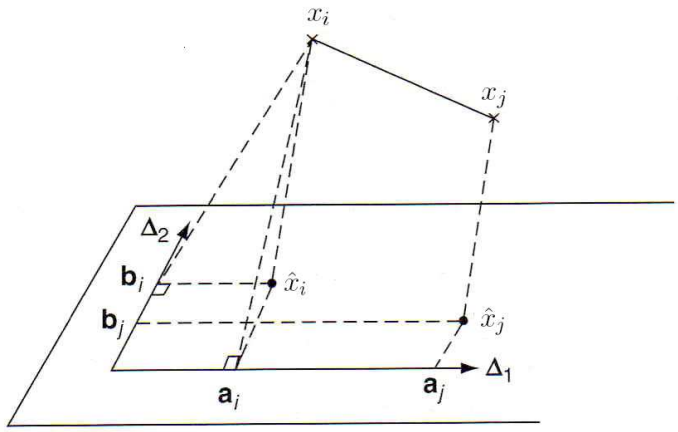

FIG. 1.2 – *La projection des individus sur un sous espace*

Calculons d'abord l'inertie du nuage projeté sur E_q

Soit P l'opérateur de projection M-orthogonale sur E_q :

P vérifient: $P^2 = P$ et $P'M = MP$, (i.e, P est idempotent et M symétrique).

Le nuage projeté est alors associé au tableau de données XP', car chaque individu x_i (ou ligne de X) se projette sur E_q selon un vecteur colonne Px_i, ou un vecteur ligne x_iP' (voir Fig 1.3).
La matrice de variance du tableau XP' $(Var(XP'))$ pour des variables centrées est donnée comme suit:

$$\begin{aligned} Var(XP') &= (XP')'D(XP') \\ &= P(X'DX)P' \\ &= P\Sigma_e P'. \end{aligned}$$

Notons par $I_{E_q}(\mathcal{N}')$ l'inertie du nuage projeté sur E_q.

Par définition:

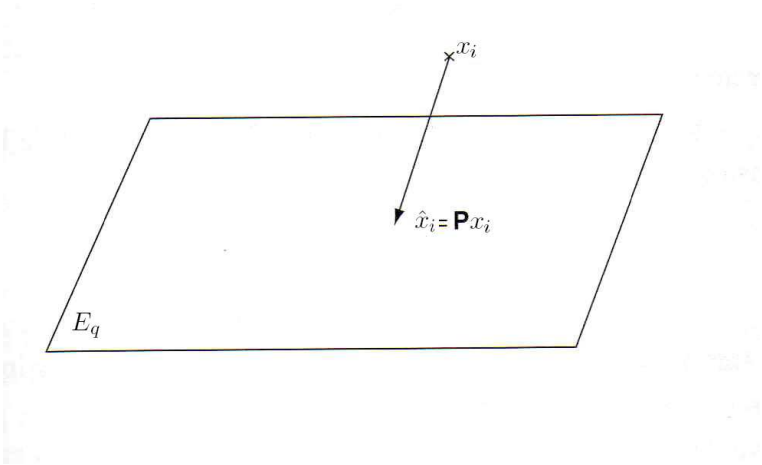

FIG. 1.3 – *Projection d'un individu sur un sous espace E_q*

$$I_{E_q}(\mathcal{N}') = Trace(P\Sigma_e P' M).$$

Par des opérations élémentaires, on déduit:

$$\begin{aligned} Trace(P\Sigma_e P' M) &= Trace(P\Sigma_e MP), \text{ car } P'M = MP \\ &= Trace(\Sigma_e MP^2) \text{ car } Trace(AB) = Trace(BA) \\ &= Trace(\Sigma_e MP) \text{ car } P \text{ est idempotent (i.e. } P^2 = P) \end{aligned}$$

Finalement:

$$I_{E_q}(\mathcal{N}') = Trace(\Sigma_e MP). \tag{1.13}$$

Le problème est donc de trouver le sous espace E_q tel que $Trace(\Sigma_e MP)$ soit maximale.

Pour résoudre ce problème, on utilise la conséquence du théorème suivant.

Théorème 1.1 (Saporta [54]). *Soit E_q un sous-espace portant l'inertie maximale, alors le sous-espace de dimension $q+1$ portant l' inertie maximale est la somme directe de E_q et du sous espace de dimension 1 M-orthogonal à E_q portant l'inertie maximale : les solutions sont emboîtées.*

Une conséquence du théorème (Saporta [54])

La recherche du sous espace sous E_q tel que $I_{E_q}(\mathcal{N}')$ soit maximale, revient donc aux recherches successives suivantes:

1. Rechercher la droite Δ_{a_1}, telle que l'inertie du nuage projeté sur Δ_{a_1}, notée $I_{\Delta_{a_1}}(\mathcal{N}')$

soit maximale.

2. Rechercher la droite Δ_{a_2}, $a_1 \perp_M a_2$ telle que $I_{\Delta_{a_2}}(\mathcal{N}')$ soit maximale.

3. Rechercher la droite Δ_{a_k}, $a_k \perp_M a_j$ $\forall j = 1, \cdots, k-1$, telle que $I_{\Delta_{a_k}}(\mathcal{N}')$ soit maximale.

Dans ce cas, le sous espace E_q s'écrit comme suit:

$$E_q = \Delta_{a_1} \oplus \Delta_{a_2} \oplus \cdots \oplus \Delta_{a_q}.$$

Dans ce qui suit, on se propose de déterminer le sous espace E_q.

1.6 Éléments principaux

1.6.1 Axes principaux

Dans une première étape, nous devons chercher une droite Δ_a engendrée par a maximisant l'inertie du nuage projeté sur cette droite $(I_{\Delta_a}(\mathcal{N}'))$.

Soit a un vecteur porté par cette droite; le projecteur M-orthogonal sur la droite est alors :

$$P = a(a'Ma)^{-1}a'M.$$

L'inertie du nuage projeté sur cette droite vaut, d'après (1.13) :

$$\begin{aligned} I_{\Delta_a}(\mathcal{N}')) &= Trace(\Sigma_e MP) \\ &= Trace(\Sigma_e Ma(a'Ma)^{-1}a'M) \\ &= \frac{1}{a'Ma} Trace(\Sigma_e Maa'M), \ car \ a'M\Sigma_e Ma \ est \ un \ scalaire. \\ &= \frac{Trace(a'M\Sigma_e Ma)}{a'Ma} \\ &= \frac{a'M\Sigma_e Ma}{a'Ma}. \end{aligned}$$

Par définition,

$$\frac{a'M\Sigma_e Ma}{a'Ma} = \frac{\prec a, \ \Sigma_e Ma \succ_M}{\|a\|_M^2}.$$

Donc, chercher $\max I_{\Delta_a}(\mathcal{N}')$ revient à chercher:

$$\max_a \prec a, \ \Sigma_e Ma \succ_M .$$

Or, on sait que d'après la proposition 2 de l'Annexe 1:

$\max_{a} \prec a,\ \Sigma_e M a \succ_M$, avec $\|a\|_M = 1$ est atteint pour $a = a_1$, vecteur propre de $\Sigma_e M$ associé à la plus grande valeur propre de $\Sigma_e M$, et ce maximum vaut λ_1.

La matrice $\Sigma_e M$ étant M-symétrique possède des vecteurs propres M-orthogonaux deux à deux.

Dans la deuxième étape, on cherche une droite Δ_a, tel que $\|a\|_M = 1$ et $a \perp_M a_1$ maximisant l'inertie du nuage projeté sur cette droite $I_{\Delta_a}(\mathcal{N}')$.

En utilisant le même raisonnement que la première étape, on retrouve que le maximum de $I_{\Delta_a}(\mathcal{N}')$ est atteint pour $a = a_2$, vecteur propre de $\Sigma_e M$, associée à la deuxième valeur propre de $\Sigma_e M$, et ce maximum vaut λ_2.

D'où le résultat suivant :

Théorème 1.2 (Saporta [54]). *Le sous-espace E_q de $\mathbb{R}^p$ de dimension q $(q < p)$ tel que $I_{E_q}(\mathcal{N}')$ soit maximale est engendré par les vecteurs propres de $\Sigma_e M$ associés aux q plus grandes valeurs propres.*

Autrement dit:

$$E_q = \Delta_{a_1} \oplus \Delta_{a_2} \oplus \cdots \oplus \Delta_{a_q},$$

où a_1,$a_2, \cdots, a_q$ sont les vecteurs propres de $\Sigma_e M$, associés respectivement à ses q plus grandes valeurs propres: $\lambda_1 > \lambda_2 > \cdots > \lambda_q$.

Remarque 1.2. On appelle axes principaux les vecteurs propres de $\Sigma_e M$, M-normés à 1 (i.e, $\|a_k\|_M^2 = 1$, $k = 1, \cdots, q$).

1.6.2 Facteurs principaux

A l'axe principal a_k, M-normé à 1, est associé le facteur principal $f_k = M a_k$.

Puisque a_k est vecteur-propre de $\Sigma_e M$, donc:

$$\Sigma_e M a_k = \lambda a_k \Rightarrow M \Sigma_e M a_k = \lambda M a_k,$$

or $M a_k = f_k$, par suite :

$$M \Sigma_e f_k = \lambda f_k.$$

Les facteurs principaux sont donc les vecteurs propres de $M\Sigma_e$, M^{-1} normés à 1 , i.e , $\|f_k\|^2_{M^{-1}} = 1$, $(k = 1 \cdots ,q)$. En effet:

On sait que: $a'_k M a_k = 1 \Longleftrightarrow a'_k M(M^{-1}M)a_k = 1$, car $M^{-1}M = I_q$

$\Longleftrightarrow a'_k M'(M^{-1}M)a_k = 1$, car M est symétrique.

$\Longleftrightarrow (Ma_k)'M^{-1}(Ma_k) = 1$, car $(Ma_k)' = a'_k M'$

$\Longleftrightarrow \prec Ma_k, Ma_k \succ_{M^{-1}} = 1$

$\Longleftrightarrow \prec f_k, f_k \succ = 1$, car $f_k = Ma_k$

$\Longleftrightarrow \|f_k\|^2_{M^{-1}} = 1.$

1.7 Composantes principales

Les composante principales se sont les variables C^k (éléments de $\mathbb{R}^n$) définies, par les facteurs principaux, comme suit :

$$C^k = Xf_k.$$

La variable C^k est le vecteur renfermant les coordonnées des projections M-orthogonales des individus sur l'axe défini par a_k avec a_k unitaire.

La variance de la composante principale C^k ($Var(C^k)$) est égale à la valeur propre λ_k, autrement dit :

$$Var(C^k) = \lambda_k.$$

En effet :

$$\begin{aligned} Var(C^k) &= (C^k)'DC^k \\ &= (Xf_k)'D(Xf_k) \\ &= f'_k X'DXf_k \\ &= f'_k \Sigma_e f_k, \end{aligned} \tag{1.14}$$

or, on sait que f_k est vecteur propre de $M\Sigma_e$, donc:

$$(M\Sigma_e)f_k = \lambda_k f_k \Longleftrightarrow \Sigma_e f_k = \lambda_k M^{-1} f_k, \tag{1.15}$$

en remplaçant $\Sigma_e f_k$ donnée par (1.15) dans la formule (1.14), on aura :

$$Var(C^k) = \lambda_k f_k' M^{-1} f_k.$$

Comme:

$$f_k' M^{-1} f_k = \|f_k\|^2_{M^{-1}} = 1,$$

on déduit que:

$$Var(C^k) = \lambda_k.$$

Remarques 1.3.

1. Les composantes C^k, $(k = 1, \cdots, p)$, sont les combinaisons linéaires des variables $X^1, X^2, \cdots, X^p$, de variance maximale sous la contrainte $f_i' M^{-1} f_i = 1$.

2. Les composantes principales sont elles-mêmes vecteurs propres d'une matrice de taille n. En effet, on sait que: $M\Sigma_e f_k = \lambda_k f_k$, or $\Sigma_e = X'DX$, d'où:

   $$M(X'DX) f_k = \lambda_k f_k.$$

   En multipliant à gauche par X et en remplaçant $X f_k$ par C^k, on aura :

   $$(XMX')DC^k = \lambda_k C^k.$$

   Notons la matrice XMX' par U.

Conclusion:

L'A.C.P. est une méthode qui consiste à remplacer les variables $X^1, X^2, \cdots, X^p$ qui sont corrélées, par des nouvelles variables , ce sont les composantes principales $C^1, C^2, \cdots, C^p$ combinaisons linéaires des variables X^j $(j = 1, \cdots, p)$, non corrélées entre elles, et de variance maximale.

La Table 1.1 résume les propriétés vérifiées par les vecteurs principaux, les facteurs principaux, et les composantes principales.

Vecteurs principaux	Facteurs principaux	Composantes principales
$a_k \in \mathbb{R}^p$	$f_k = Ma_k$	$C^k = Xf_k = XMa_k$ dans $\mathbb{R}^n$
Vecteurs propres de $\Sigma_e M$	Vecteurs propres de $M\Sigma_e$	Vecteurs propres de $UD = XMX'D$
$\Sigma_e M a_k = \lambda_k a_k$	$M\Sigma_e f_k = \lambda f_k$	$UDC^k = \lambda_k C^k$
$\|a_k\|_M = 1$	$\|f_k\|_{M^{-1}} = 1$	$\|C^k\|_D = \sqrt{\lambda_k}$
$\prec a_k, a_j \succ_M = 0, \ \forall k \neq j$	$\prec f_k, f_j \succ_{M^{-1}} = 0, \forall k \neq j$	$\prec C^k, C^j \succ_D = 0 \ , \ \forall k \neq j$

Table 1.1: Les propriétés vérifiées par les vecteurs principaux, les facteurs principaux, et les composantes principales.

Chapitre 2

Les éléments de la statistique robuste

2.1 Introduction

Le terme "robuste" a été introduit en 1953, dans un article de G. Box [7] sur l'estimation de la variance dans le cas non gaussien, au sens de résistance à une déviation par rapport à la loi normale, mais le premier travail mathématique sur l'estimation robuste semble remonter à 1818 avec Pierre Simon de Laplace [45] qui a tenté d'utiliser l'ordre des données dans un problème d'estimation du coefficient d'une régression linéaire dans son livre " Deuxième supplément à la théorie analytique des probabilités ", on y trouve en particulier la distribution de la médiane.

L'émergence de la statistique robuste moderne ne date que des années soixante du siècle dernier avec les travaux pionniers de Tukey [58], Huber [29] et Hampel [25]. Depuis cette période, de nombreux modèles et méthodes ont été réexaminés sous l'angle de la robustesse. La prise en compte de l'impact sur les méthodes statistiques de valeurs atypiques (aberrantes) ou de toute autre structure minoritaire présente dans les données, est d'autant plus importante à l'heure actuelle que l'on dispose de bases de données de plus en plus grandes et dont la fiabilité et la qualité sont relativement inégales.

Comme l'a souligné le statisticien G. Box [7] en écrivant "All models are wrong, some are useful", il est évident que tout modèle sous-jacent n'est qu'un reflet simplifié de la réalité. Or, l'estimation des paramètres d'un modèle n'est valable que sous certaines hypothèses qui sont bien trop souvent passées sous silence dans la pratique. Une hypothèse fondamentale pour les estimateurs dits classiques suppose que tous les individus examinés ont un comportement compatible avec le modèle sous-jacent.

Or, la présence dans la population de plusieurs types de comportement non identifiés par le modèle ou l'existence de valeurs aberrantes va à l'encontre de cette hypothèse. Les recherches intégrant des méthodes statistiques robustes destinées à surmonter ces difficultés sont intéressantes tant au niveau théorique qu'au niveau pratique.

Ces méthodes robustes sont également essentielles dans la détection des valeurs aberrantes.

Ce chapitre sera consacré à la présentation des concepts de base de la théorie de la robustesse. Il s'organise en trois parties: la première partie comporte les rappels et les définitions nécessaires pour le thème traité (valeurs aberrantes, estimateur robuste, L-estimateurs, fonctionnelle statistique, maximum du biais, fonction d'influence et ses propriétés, point de rupture). Dans la deuxième partie, nous présentons les différents estimateurs du paramètre de localisation. La troisième partie, traite les différents estimateurs robustes de la matrice de covariance; nous donnons les définitions ainsi que les algorithmes de calcul des estimateurs: Le MCD estimateur (minimum covariance determinant), le MCD repondéré (MCD[1]) et le S-estimateur.

2.2 Définitions

Définition 2.1 (Données aberrantes). Les données aberrantes (outliers) sont des observations atypiques bien éloignées de la masse des données et sont des points isolés ou en petit groupes de points. Elles sont dues à des erreurs de copie, de calcul, de changement d'unités, ou des données n'obéissant pas au même modèle (présence de plusieurs classes).

Les données aberrantes sont les plus "dangereuses" pour l'estimateur.

Définition 2.2 (Méthode statistique robuste). Une méthode statistique sera dite "robuste" lorsqu'elle est insensible à une petite déviation du modèle initial: En pratique une petite partie des données est remplacée par des nouvelles qui peuvent être très différentes. De même, un estimateur sera dit robuste s'il ne perd pas trop de ses qualités optimales lorsqu'on s'éloigne des hypothèses sous lesquelles il a été conçu.

2.3 Les estimateurs de paramètre de localisation

2.3.1 Le problème

Un des travaux essentiels d'un analyseur de données est de connaître, à partir d'un échantillon de n observations, la localisation (la tendance centrale) de ces valeurs.

Le problème de l'analyseur de données reviendra alors à trouver un bon estimateur de la localisation des valeurs; s'il a la chance de savoir que les données suivent une loi Gaussienne (cas extrêmement rare), il utilisera la moyenne comme estimateur. En effet, nous savons que la moyenne dans ce cas précis est le meilleur estimateur au sens de plusieurs critères, mais à priori, il ne connaît pas la distribution exacte de ces données. Il lui faudra alors utiliser des estimateurs robustes. Ces estimateurs sont basés sur la statistique d'ordre.

2.3.2 Statistiques d'ordre

Définition 2.3 (Statistiques d'ordre). Soit X_1,X_2,$\cdots$,X_n un échantillon de taille n, les observations réarrangées dans un ordre de magnitudes croissantes, notées $X_{(1)}$, $X_{(2)}$,$\cdots$,$X_{(n)}$ sont appelées statistiques d'ordre de l'échantillon et $X_{(i)}$ est appelée la i-ième statistique d'ordre.

On sait que la médiane, qui est basée sur les statistiques d'ordre, a l'avantage d'être robuste. C'est pourquoi, nous allons définir des estimateurs qui feront intervenir ces statistiques d'ordre.

2.3.3 Les L-estimateurs

Définition 2.4. Soit $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ les statistiques d'ordre d'un échantillon de taille n. Soit a_1,a_2,$\cdots$,a_n des nombres réels vérifiant $0 \leq a_i \leq 1$, $i = 1,\cdots,n$ et $\sum_{i=1}^{n} a_i = 1$; un L-estimateur T de poids a_1,a_2,$\cdots$,a_n est défini par l'expression suivante:

$$T = \sum_{i=1}^{n} a_i X_{(i)}.$$

En d'autres termes, T est la combinaison convexe des $X_{(i)}$, $i = 1,\cdots,n$.

Définition 2.5 (La médiane). La médiane est le L-estimateur qui ne fait intervenir que la statistique d'ordre centrale si n est impair et qui donne la moyenne des deux statistiques d'ordre centrales si n est pair.

$$Si \;\; n = 2p+1, \;\; \text{alors} \;\; a_i = \begin{cases} 1 & si \;\; i = p+1 \\ 0 & sinon \end{cases}$$

$$Si \;\; n = 2p, \;\; \text{alors} \;\; a_i = \begin{cases} 1/2 & si \;\; i = p \;\; ou \;\; i = p+1 \\ 0 & sinon \end{cases}$$

Définition 2.6 (Moyenne α-censurée). Soit α un réel vérifiant $0 < \alpha < 1/2$, la moyenne α-censurée, notée $T(\alpha)$, est un L-estimateur avec des poids a_i tels que:

$$a_i = \begin{cases} \frac{1}{n-2[n\alpha]} & si \;\; [n\alpha]+1 \leq i \leq n-[n\alpha] \\ \\ 0 & sinon, \end{cases}$$

où $[x]$ désigne la partie entière de x.

Le principe général de cet estimateur est d'omettre une proportion α des plus petites valeurs et une autre proportion α des plus grandes valeurs, puis de calculer la moyenne de l'ensemble des valeurs restantes.

Définition 2.7 (Mimoyenne). La mimoyenne est définie comme étant la moyenne α-censurée pour $\alpha = 0.25$. Dans ce calcul, on ne fait intervenir que la moitié centrale des observations triées.

2.4 Les mesures de la robustesse

Définition 2.8 (Fonctionnelle statistique). Soit $\mathfrak{X}$ l'espace d'échantillonnage (en général $\mathfrak{X} = \mathbb{R}$ *ou* $\mathfrak{X} = \mathbb{R}^p$), et notons par η l'ensemble des lois de probabilités sur $\mathfrak{X}$. Supposons que P(ou F) dépend d'un paramètre $\theta \in \Theta$ qu'on cherche à estimer. Dans de nombreux cas, θ est une fonction de loi de P ou de la fonction de répartition F.

$$\theta = T(P) = T(F),$$

où T est une fonctionnelle définie sur η à valeurs dans Θ.

Étant donnée un échantillon X_1,X_2,...,X_n l'estimateur naturel de θ est alors

$$\hat{\theta_n} = T(X_1,X_2,...,X_n) = T(P_n) = T(F_n).$$

Toute statistique s'écrivant sous la forme $T_n = T(F_n)$ est appelée *fonctionnelle statistique*.

Exemples 2.1.

1. La fonctionnelle associée à la moyenne est:

   $\theta = T(P) = \int X dP = \int x dF(x) = E(X) = T(F)$.

   Son estimateur associé est:

   $$\hat{\theta_n} = \overline{X} = \int X dP_n = \frac{1}{n}\sum_{i=1}^{n} X_i = T(F_n).$$

2. La fonctionnelle associée à la moyenne censurée au taux $\alpha \in]0,1/2[$ est:

   $$\theta = T(P) = T(F) = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(t)dt.$$

$(x = F^{-1}(t))$

Son estimateur associé est :

$$\hat{\theta_n} = \frac{1}{1-2\alpha} \int_{F_n^{-1}(\alpha)}^{F_n^{-1}(1-\alpha)} x dF_n(x) = \frac{1}{n-2n\alpha} \sum_{i=F_n^{-1}(\alpha)}^{F_n^{-1}(1-\alpha)} X_{(i)},$$

où

$$F^{-1}(t) = \inf\{x/F(x) \geq t)\} = \sup\{x/F(x) \leq t\},$$

et

$$F_n^{-1}(t) = \begin{cases} X_{(i)}, & si \quad \frac{i-1}{n} < t \leq \frac{i}{n} \\ \\ X_{(n)}, & si \quad 1 - \frac{1}{n} < t \leq 1. \end{cases}$$

2.4.1 Maximum du biais

Soit $\mathcal{F}$ l'ensemble convexe des lois de probabilités sur ($\mathbb{R}^p$,$B_{\mathbb{R}^p}$), où $B_{\mathbb{R}^p}$ représente la tribu borélienne de $\mathbb{R}^p$, F une loi de $\mathcal{F}$ et pour
ε $\in$[0,1], le voisinage de contamination est défini par :

$$\mathcal{P}_\varepsilon(F) = \{G \in \mathcal{F}/G = (1-\varepsilon)F + \varepsilon H, H \in \mathcal{F}\}.$$

Le "maximum du biais asymptotique" (Huber, [31]) est donné par :

$$b(\varepsilon) = \sup_{G \in \mathcal{P}_\varepsilon(F)} \|T(G) - T(F)\|,$$

où $\|.\|$ désigne une norme convenable. Pour un vecteur, elle représente la norme euclidienne , et pour une matrice A, elle représente la norme matricielle suivante:

$$\|A\| = \sup_{\|u\|=1} \|Au\|.$$

b(ε) mesure la robustesse globale de T en F.

2.4.2 Point de rupture

Définition 2.9 (point de rupture asymptotique). La fraction minimale ε^* de contamination qui rend b(ε) non borné est appelé "point de rupture asymptotique "de T en F (Hampel, [26]):

$$\varepsilon^* = min\{\varepsilon : b(\varepsilon) = +\infty\}.$$

ε^* représente aussi la proportion limite de données aberrantes que peut tolérer l'estimateur.

Soit X=$\{x_1, \cdots, x_n\}$ un échantillon dans $\mathbb{R}^p$. On considère tous les échantillons "corrompus" X_m obtenus en remplaçant n'importe quels m points des données originales par des points quelconques.

Définition 2.10 (Point de rupture empirique d'un paramètre de position). Pour un estimateur d'un paramètre de position $T(X)$, on définit le biais maximal par:

$$biais(m; T,X) = \sup_{X_m} \|T(X) - T(X_m)\|,$$

et le point de rupture par:

$$\varepsilon_n^*(T,X) = min\Big\{m/n; biais(m; T,X) = +\infty\Big\}.$$

Le point de rupture empirique est donc la fraction minimale de contamination telle que T peut prendre des valeurs arbitrairement loin.

L'estimateur le plus utilisé d'un paramètre de position est la moyenne $\bar{x}$. Cependant, cet estimateur n'est pas robuste. En effet, il suffit de prendre une seule observation, de l'envoyer arbitrairement loin pour que la moyenne prenne une valeur arbitrairement éloignée. On dit que la moyenne a un point de rupture $1/n$, donc son point de rupture asymptotique est 0 i.e:

$$\varepsilon_n^*(\bar{x}) = \frac{1}{n} \quad \text{et} \quad \varepsilon^*(\bar{x}) = 0.$$

Remarques 2.1.

i. Le point de rupture appartient à l'intervalle [0,1/2].
ii. La valeur d'un point de rupture est une indication de la robustesse d'une statistique.

2.4.3 Fonction d'influence

Que se passe t-il pour un estimateur T_n=$T(F_n)$, si l'on ajoute une observation supplémentaire x à un grand échantillon?

Pour mesurer l'influence relative de cette observation sur l'estimateur, Hampel [26] a proposé de considérer la fonction suivante :

$$IF(x;T,F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F) + \varepsilon\delta_x) - T(F)}{\varepsilon} = \frac{dT(\tilde{F})}{d\varepsilon}|_{\varepsilon=0},$$

où δ_x est la mesure de Dirac au point x de $\mathbb{R}^p$ et $\tilde{F} = (1-\varepsilon)F + \varepsilon\delta_x$.

La quantité $IF(x;T,F)$, si elle existe, est connue sous le nom de "courbe d'influence" ou "fonction d'influence" de T en F. Elle est sans doute l'outil le plus important en statistique robuste; on peut en rappeler brièvement quelques propriétés :

a) $E(IF(X;T,F)) = \int_{\mathbb{R}^p} IF(x;T,F)\, dF(x) = 0$, i.e. l'influence moyenne est nulle pour tout point supplémentaire x réalisation de F.

b) $T(G) - T(F) = \int_{\mathbb{R}^p} IF(x;T,F)\, dG$ +reste, où G$\in \mathcal{P}_\varepsilon(F)$, et le reste est une quantité suffisamment petite.

c) En remplaçant G par F_n dans b) on aura:

$$T_n = T(F) + \frac{1}{n}\sum_{i=1}^{n} IF(x_i;T,F) + reste.$$

Donc

$$\sqrt{n}(T(F_n) - T(F)) \simeq \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(x_i;T,F).$$

La suite $\{IF(x_i;T,F)\}_{i=1,\cdots,n}$ est une suite de v.a.i.i.d. En appliquant le théorème central limite, on aura:

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow[n\to\infty]{\mathcal{L}} N(0,A(T,F)),$$

où

$$A(T,F) = \int_{\mathbb{R}^p} IF(x;T,F)IF(x;T,F)'\, dF(x).$$

Où A' désigne la transposée de A.

d) La borne supérieure de IF, donnée par :

$$\gamma^* = \sup_x \|IF(x;T,F)\|,$$

γ^* est appelée par Hampel [26] "la sensibilité aux grosses erreurs". Elle est liée au maximum du biais $b(\varepsilon)$ par la relation :

$$b(\varepsilon) = \varepsilon\gamma^*.$$

En effet: d'après b) on déduit:

$$\begin{aligned} T(G) - T(F) &\simeq \int_{\mathbb{R}^p} IF(x;T,F)d[(1-\varepsilon)F + \varepsilon H](x) \\ &\simeq \int_{\mathbb{R}^p} \varepsilon IF(x;T,F)\, dH(x) \\ &\leq \varepsilon \sup_x \|IF(x;T,F)\| \end{aligned}$$

d'où $b(\varepsilon) = \varepsilon\gamma^*$.

On voit ainsi que la fonction d'influence présente deux aspects très importants :

D'une part, par définition, elle nous renseigne sur l'influence relative de chaque observation sur l'estimateur.

La propriété d) montre que si IF n'est pas continue et bornée, une observation aberrante peut causer une grande perturbation de l'estimateur. D'autre part, elle permet une évaluation simple et immédiate des propriétés asymptotiques d'un estimateur.

La propriété c) montre que la variance asymptotique est connue dès que l'on connaît la fonction d'influence IF. En conséquence, on peut déduire une propriété fondamentale de IF :

Un estimateur T_n est robuste si sa fonction d'influence est continue et bornée.

2.5 Les estimateurs robustes de la matrice de covariance

Dans toute la suite, on suppose qu'on observe un échantillon $(x_1, \cdots, x_n)$ d'observations dans $\mathbb{R}^p$ de distribution $F_{\mu,\Sigma}$, où $\mu \in \mathbb{R}^p$ et $\Sigma \in \text{SDP}(p)$.
$\text{SDP}(p)$ désigne l'ensemble des matrices carrées d'ordre p symétriques, définies positives.

2.5.1 Le M- estimateur

Définition 2.11 (Définition du M-estimateur). Les M-estimateurs (Maronna, [42]) du vecteur moyen et de la matrice de covariance sont les solutions $\hat{\mu}_M$ et $\hat{\Sigma}_M$ des équations suivantes:

$$\frac{1}{n}\sum_{i=1}^{n} w_1(d_i)(x_i - \mu) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} \Big\{ w_2(d_i^2)(x_i - \mu)(x_i - \mu)' - w_3(d_i)\Sigma \Big\} = 0. \tag{2.1}$$

Le système (2.1) peut s'écrire d'une manière équivalente sous la forme implicite:

$$\hat{\mu}_M = \frac{\sum_{i=1}^{n} w_1(\hat{d}_i)x_i}{\sum_{i=1}^{n} w_1(\hat{d}_i)},$$

$$\hat{\Sigma}_M = \frac{\sum_{i=1}^{n} w_2(\hat{d}_i^2)(x_i - \hat{\mu}_M)(x_i - \hat{\mu}_M)'}{\sum_{i=1}^{n} w_3(\hat{d}_i)}, \tag{2.2}$$

où $\hat{d}_i = \Big\{ (x_i - \hat{\mu}_M)'\hat{\Sigma}_M^{-1}(x_i - \hat{\mu}_M) \Big\}^{1/2}$, $w_1(.)$, $w_2(.)$ et $w_3(.)$ sont des "fonctions poids".

Exemple 2.1 (Un exemple des fonctions w_1, w_2,w_3). *Huber, en 1981 [31], a proposé les fonctions suivantes:*

$$w_3(y) = 1,\ w_1(y) = \frac{\Psi_H(y,b)}{y},\ w_2(y) = \frac{\Psi_H(y,b^2)}{y}, \tag{2.3}$$

avec $\Psi_H(y,b) = max\Big\{ -b,\ \ min(y,b) \Big\}$

$$= \begin{cases} -b & \text{si} \quad y < -b \\ y & \text{si} \quad -b \leq y \leq b \\ b & \text{si} \quad y > b \end{cases}$$

où $b = \sqrt{q_\tau}$, avec $q_\tau = \chi^2_{p,0.9}$.

Quelques propriétés:

1. La méthode de résolution du système (2.2) est itérative. Pour calculer les estimateurs, on utilise l'algorithme de Devlin et al [21].

2. Huber [31] a montré que toute solution de (2.1) a un point de rupture qui ne dépasse pas $\dfrac{1}{p+1}$. Pour plus de détails sur le point de rupture d'un M-estimateur (voir Tyler, [59]).

3. Le point de rupture d'un M-estimateur est très petit lorsque p est assez grand, ce qui constitue un handicap: le M-estimateur sera sensible aux outliers dans le cas de données de grande dimension.

4. On dit que le M-estimateur est robuste dans le sens où sa fonction d'influence est bornée lorsque w_1, w_2, w_3 sont choisis judicieusement.

L'algorithme de calcul du M-estimateur

On sait que d'après (2.1), les M-estimateurs du vecteur moyen et de la matrice de covariance sont respectivement:

$$\hat{\mu}_M = \frac{\sum_{i=1}^n w_1(\hat{d}_i) x_i}{\sum_{i=1}^n w_1(\hat{d}_i)},$$

$$\hat{\Sigma}_M = \frac{\sum_{i=1}^n w_2(\hat{d}_i^2)(x_i - \hat{\mu}_M)(x_i - \hat{\mu}_M)'}{\sum_{i=1}^n w_3(\hat{d}_i)},$$

où $\hat{d}_i = \left\{ (x_i - \hat{\mu}_M)' \hat{\Sigma}_M^{-1} (x_i - \hat{\mu}_M) \right\}^{1/2}$.

On adoptera les notations suivantes :

1. On pose $p_{\mu_i} = w_1(\hat{d}_i)/\sum_{i=1}^{n} w_1(\hat{d}_i)$, $\hat{\mu_M}$ s'écrit:

$$\hat{\mu_M} = \sum_{i=1}^{n} p_{\mu_i} x_i.$$

2. On pose $P_{\Sigma_i} = w_2(\hat{d}_i)/\hat{d}_i^2 \sum_{i=1}^{n} w_3(\hat{d}_i)$, $\hat{\Sigma_M}$ s'écrit:

$$\hat{\Sigma_M} = \sum_{i=1}^{n} p_{\Sigma_i}(x_i - \hat{\mu_M})(x_i - \hat{\mu_M})'.$$

Dans la pratique, la méthode de calcul de $\hat{\mu_M}$ et $\hat{\Sigma}_M$ est basée sur l'algorithme itératif suivant :

Algorithme:

(1) Définir les fonctions $w_1(.)$, $w_2(.)$, $w_3(.)$.

(2) Au départ, on attribue à chaque observation les poids initiaux: $P_{\mu_o} = P_{\Sigma_o} = \frac{1}{n}$, à partir desquels, on calcule μ_o et Σ_o:

$$\mu_o = \sum_{i=1}^{n} P_{\mu_o} x_i, \Sigma_o = \sum_{i=1}^{n} P_{\Sigma_o}(x_i - \mu_o)(x_i - \mu_o)'. \tag{2.4}$$

A partir de μ_o et Σ_o, on calcule pour chaque $x_i (i = 1, \cdots, n)$ la quantité:

$$d_{1,i}^2 = (x_i - \mu_o)' \Sigma_o^{-1} (x_i - \mu_o). \tag{2.5}$$

(3) En fonction des distances $d_{1,i}$, on peut alors attribuer aux observations de nouveaux poids $P_{\mu_{1,i}}$ et $P_{\Sigma_{1,i}}$, qui permettent le calcul d'une nouvelle moyenne μ_1 et d'une nouvelle matrice de covariance Σ_1 où

$$P_{\mu_{1,i}} = \frac{w_1(d_{1,i})}{\sum_{i=1}^{n} w_1(d_{1,i})} \quad , \quad P_{\Sigma_{1,i}} = \frac{w_2(d_{1,i})}{\sum_{i=1}^{n} w_3(d_{1,i})}, \quad (i = 1, \cdots, n)$$

et

$$\mu_1 = \sum_{i=1}^{n} P_{\mu_{1,i}} x_i \quad , \quad \Sigma_1 = \sum_{i=1}^{n} P_{\Sigma_{1,i}}(x_i - \mu_1)(x_i - \mu_1)'. \tag{2.6}$$

Ces dernières formules déterminent alors pour chaque observation, de nouvelles distances $d_{2,i}$ qui définissent à leurs tours de nouveaux poids $P_{\mu_{2,i}}$ et $P_{\Sigma_{2,i}}$. La méthode est itérative et une observation voit son poids diminuer petit à petit au fur et à mesure que la moyenne et la matrice des variances covariances se robustifient.

(4) On arrête l'itération dès que les poids obtenus au $j^{éme}$ passage vérifient:

$$(\sum_{i=1}^{n}(P_{\Sigma_{j,i}} - P_{\Sigma_{j+1,i}})^2)^{\frac{1}{2}} < \beta_o,$$

où $\beta_o = 10^{-5}$.

2.5.2 L'estimateur MCD

Définition 2.12 (Définition du MCD). Le MCD (*minimum covariance déterminant*, Rousseeuw [50]) est déterminé en sélectionnant un sous échantillon (x_{i_1},x_{i2},$\cdots$,x_{ik}) de taille k ($1 \leq k \leq n$), qui minimise la variance généralisée (c-à-d le déterminant de la matrice de covariance calculée sur un sous échantillon) sur tous les sous échantillons choisis de taille k.

Les estimateurs de la moyenne et de la matrice de covariance sont alors respectivement:

$$\hat{\mu}_{MCD} = \frac{1}{k}\sum_{j=1}^{k} x_{ij}, \tag{2.7}$$

$$\hat{\Sigma}_{MCD} = c_p \frac{1}{k}\sum_{j=1}^{k}(x_{ij} - \hat{\mu}_{MCD})(x_{ij} - \hat{\mu}_{MCD})', \tag{2.8}$$

où c_p est un facteur choisi de façon à rendre l'estimateur Fisher-consistant.

Pour déduire la fonction d'influence du MCD, il faut l'écrire sous forme d'une fonctionnelle. Cherchons donc cette fonctionnelle.

La fonctionnelle statistique du MCD

Soit α la proportion de données qui ne détermine pas le MCD. On considère :

$$D_F(\alpha) = \Big\{A \subseteq \mathbb{R}^p, P_F(A) = 1 - \alpha\Big\}.$$

Pour chaque ensemble A$\in D_F(\alpha)$, la moyenne et la matrice de covariance calculées sur A sont respectivement:

$$\mu_A(F) = \frac{1}{1-\alpha}\int_A x dF(x),$$

$$\Sigma_A(F) = \frac{1}{1-\alpha}\int_A (x-\mu_A(F))(x-\mu_A(F))' dF(x).$$

Le sous échantillon A est appelé solution du MCD si:

$$\det(\Sigma_A(F)) \leq \det(\Sigma_{\tilde{A}}(F)) \quad \forall \tilde{A} \in D_F(\alpha).$$

Le MCD théorique de la moyenne et de la matrice de covariance sont alors:

$$\mu_{MCD}(F) = \frac{1}{1-\alpha}\int_A x dF(x), \tag{2.9}$$

$$\Sigma_{MCD}(F) = \frac{c_\alpha}{1-\alpha}\int_A (x-\mu_{MCD}(F))(x-\mu_{MCD}(F))' dF(x). \tag{2.10}$$

c_α étant la constante assurant la consistance au sens de Fisher de Σ.

Quelques propriétés:

1. Le point de rupture du MCD est $\min(\alpha, 1-\alpha)$.

2. Pour le choix de k, Rousseeuw et Lopuhaä [38] proposent de prendre $k = (n+p+1)/2$ pour obtenir un plus grand point de rupture.

3. En pratique, $k = 0.75n$

4. Le MCD est implémenté dans , Matlab, S+, SAS, . . .

Remarque 2.1. L'estimateur MCD vise à minimiser le déterminant de la matrice de covariance. En effet, la présence des valeurs aberrantes augmente la variance des données, donc permet d'isoler le déterminant de la matrice de covariance le plus petit qui permet de rejeter ces valeurs. Dans son principe, cette méthode cherche k observations, avec k le

nombre d'observations considérées comme *'saines'*, qui minimisent le déterminant de la matrice de covariance.
L'algorithme qui découle de cette méthode (Rousseuw & Van Driessen [51]) est le suivant:

Algorithme approximatif: FASTMCD

1. Effectuer 500 fois:

(a) Sélectionner un échantillon aléatoire H_0 contenant k observations. La valeur de k est par défaut:

$$k = \frac{1}{2}(p + n + 1).$$

(b) Calculer la moyenne $\mu_0 = ave(H_0)$ et la covariance $\Sigma_0 = cov(H_0)$ de l'échantillon sélectionné, ainsi que la distance de Mahalanobis $d_0(i)$ pour $i = 1, \cdots, n$:

$$d_0(i) = \sqrt{(x(i) - \mu_0)' \Sigma_0^{-1} (x(i) - \mu_0)}.$$

On classe les distances de la plus petite à la plus élevée, et on choisit les k observations associées aux distances les plus petites.

(c) A partir des k observations, on calcule les moyennes et les variances correspondantes ainsi que les distances de Mahalanobis.

(d) On classe à nouveau les différentes distances de Mahalanobis la plus petite à la plus élevée, et on choisit les k observations avec les distances les plus petites. A partir des k observations, on calcule le déterminant de la matrice de covariance.

2. Parmi les 500, choisir les 10 échantillons pour lesquels les valeurs du déterminant de la matrice de covariance sont les plus petites.

3. Effectuer les points 1.b et 1.d jusqu'à convergence du déterminant de la matrice de covariance.

Cet algorithme est illustré par le schéma suivant:

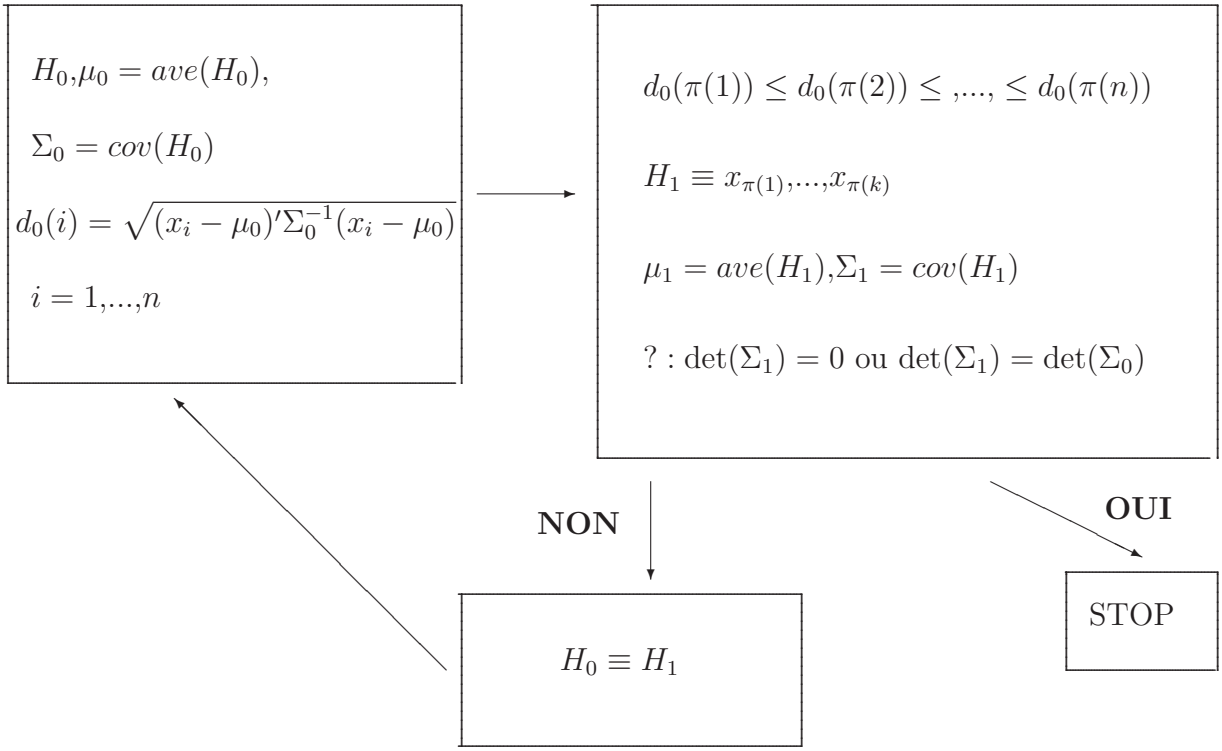

FIG. 2.1 -Algorithme FASTMCD

MCD d'une distribution elliptique

On suppose qu'on cherche à estimer les paramètres μ et Σ d'une distribution $F_{\mu,\Sigma}$, de densité :

$$f_{\mu,\Sigma}(x) = \frac{g((x-\mu)'\Sigma^{-1}(x-\mu))}{\sqrt{\det(\Sigma)}}.$$

La fonction g est supposée connue et sa dérivée négative.

$F_{\mu,\Sigma}$, telle qu'elle est définie, représente la classe des distributions elliptiques symétriques et unimodales.

D'après B.D.G [8], pour ces distributions le problème du MCD possède une unique solution donnée par l'éllipsoide:

$$A(F_{\mu,\Sigma}) = \Big\{x \in \mathbb{R}^p, (x-\mu)'\Sigma^{-1}(x-\mu) \leq q_\alpha\Big\}, \tag{2.11}$$

avec $G(t) = P_{F_{0,I_p}}(x'x \leq t) \quad et \quad q_\alpha = G^{-1}(1-\alpha)$.

Les fonctionnelles du MCD sont alors:

$$\mu_{MCD} = \frac{1}{1-\alpha}\int_{A(F_{\mu,\Sigma})} x dF_{\mu,\Sigma}(x), \tag{2.12}$$

$$\Sigma_{MCD} = \frac{c_\alpha}{1-\alpha}\int_{A(F_{\mu,\Sigma})} (x-\mu_{MCD})(x-\mu_{MCD})' dF_{\mu,\Sigma}(x) \tag{2.13}$$

En faisant un changement de variable $z = \Sigma^{-\frac{1}{2}}(x-\mu)$, $\Sigma(F_{\mu,\Sigma})$ peut s'écrire :

$$\Sigma_{MCD} = \Big(\frac{c_\alpha}{1-\alpha}\int_{z'z\leq q_\alpha} z_1^2 dF_{0,I_p}(z)\Big)\Sigma.$$

Pour que Σ soit consistante au sens de Fisher (i.e: $\Sigma = \Sigma(F_{\mu,\Sigma})$), il suffit de prendre :

$$\frac{c_\alpha}{1-\alpha}\int_{z'z\leq q_\alpha} z_1^2 dF_{0,I_p}(z) = 1$$

Donc

$$c_\alpha = \frac{1-\alpha}{\int_{z'z\leq q_\alpha} z_1^2 dF_{0,I_p}(z)}.$$

En utilisant le lemme de Lopuhaä (voir Annexe 2), alors :

$$\int_{z'z\leq q_\alpha} z_1^2 dF_{0,I_p}(z) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+1} g(r^2) dr.$$

Par suite:

$$c_\alpha = (1-\alpha)\Big\{\frac{\pi^{p/2}}{\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+1} g(r^2) dr\Big\}^{-1}.$$

Remarque 2.2. Si F est la loi multinormale de $\mathbb{R}^p$, alors: $(x-\mu)'\Sigma^{-1}(x-\mu) \rightsquigarrow \chi_p^2$, donc $q_\alpha = \chi^2_{p,(1-\alpha)}$, quantile d'ordre $(1-\alpha)$ d'une loi χ_p^2.

Fonction d'influence du MCD de la matrice de covariance

Puisque le MCD est affine équivariant, il suffit de déduire la fonction d'influence pour $F = F_{0,I_p}$, puisque:

$$IF(x;\Sigma_{MCD},F) = B.IF(B^{-1}(x-\mu);\Sigma_{MCD},F_{0,I_p}).B',$$

où B vérifie $\Sigma_{MCD} = B^2$.

Théorème 2.1 (Croux & Haesbroeck [15]). *La fonction d'influence de la matrice de covariance du MCD est donnée par:*

$$IF(x,\Sigma_{MCD},F) = \frac{-1}{2c_3} I(\|x\|^2 \leq q_\alpha) xx' + \gamma(\|x\|) I_p, \tag{2.14}$$

où :

$$\gamma(\|x\|) = -1 + \frac{1}{2}\mathrm{tr}(IF(x,\Sigma_{MCD},F)) + \frac{c_\alpha}{1-\alpha}\frac{q_\alpha}{p}\Big\{(1-\alpha) I(\|x\|^2 \leq q_\alpha)$$

$$-\mathrm{tr}(IF(x,\Sigma_{MCD},F))\Big(c_2 + \frac{1-\alpha}{2}\Big)\Big\},$$

où

$$\mathrm{tr}(IF(x,\Sigma,F)) = (b_1 - pb_2)^{-1}\Big\{(c_\alpha \|x\|^2/(1-\alpha)) I(\|x\|^2 \leq q_\alpha)$$

$$+p\Big((c_\alpha/(1-\alpha))(q_\alpha/p)(1-\alpha - I(\|x\|^2 \leq q_\alpha)) - 1\Big)\Big\},$$

I désigne la fonction indicatrice et les constantes b_1, b_2, c_2, c_3, et c_4 sont déterminées par les relations suivantes:

$$c_2 = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+1} g'(r^2) dr.$$

$$c_3 = \begin{cases} \dfrac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \displaystyle\int_0^{\sqrt{q_\alpha}} r^{p+3} g'(r^2) dr & \text{si} \quad p \geq 2 \\ 0 & \text{sinon.} \end{cases}$$

$$c_4 = \frac{3\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+3} g'(r^2) dr. \qquad (2.15)$$

$$b_1 = \frac{c_\alpha(c_3 - c_4)}{1-\alpha}.$$

$$b_2 = \frac{1}{2} + \frac{c_\alpha}{1-\alpha}\Big[c_3 - \frac{q_\alpha}{p}\Big(c_2 + \frac{1-\alpha}{2}\Big)\Big].$$

Fonctions d'influence des éléments de la matrice de covariance du MCD

Notons par:

Σ_{MCD}^{ij}: l'élément (i,j) de Σ_{MCD}.

Σ_{MCD}^{ij}: l'élément i de la diagonale de Σ_{MCD}.

Les fonctions d'influence des éléments diagonaux et non diagonaux du MCD de la matrice de covariance Σ, sont respectivement:

$$IF(x,\Sigma_{ii},F) = \frac{1}{b_1}\Big\{\frac{c_\alpha}{1-\alpha} x_i^2 I(\|x\|^2 \leq q_\alpha) + \frac{b_2}{b_1 - pb_2}\frac{c_\alpha}{1-\alpha}\|x\|^2 I(\|x\|^2 \leq q_\alpha)$$
$$+ \frac{b_1}{b_1 - pb_2}\Big[\frac{c_\alpha}{1-\alpha}\frac{q_\alpha}{p}(1 - \alpha - I(\|x\|^2 \leq q_\alpha)) - 1\Big]\Big\}. \qquad (2.16)$$

$$IF(x,\Sigma_{ij},F) = \frac{x_i x_j}{-2c_3} I(\|x\|^2 \leq q_\alpha) \quad \text{si} \quad i \neq j. \qquad (2.17)$$

Variance asymptotique

On note par:

$\hat{\Sigma}_{ij}$: l'élément (i,j) d'un estimateur $\hat{\Sigma}$ de Σ.

$\hat{\Sigma}_{MCD}^{ij}$: l'élément (i,j) de $\hat{\Sigma}_{MCD}$.

$IF^2(x,\Sigma,F)_{ij}$: l'élément (i,j) de $IF^2(x,\Sigma,F)$.

Φ_p: la loi normale multivariée centrée et réduite.

Définition 2.13. La variance asymptotique de $\hat{\Sigma}_{ij}$ est donnée par:

$$V_{AS}(\hat{\Sigma}_{ij},F) = \int_{\mathbb{R}^p} IF^2(x,\Sigma,F)_{ij}\, dF(x) \qquad \forall \quad 1 \leq i,j \leq p. \tag{2.18}$$

Proposition 2.1 (Croux & Haesbroeck[15]). *Les expressions explicites de la variance asymptotique des éléments diagonaux et non diagonaux de $\hat{\Sigma}_{MCD}$ sont respectivement:*

$$V_{AS}(\hat{\Sigma}^{ii}_{MCD},\Phi_p) = \Big\{b_1(b_1-pb_2)(1-\alpha)\Big\}^{-2}\Big\{(1-\alpha)b_1^2(\alpha((c_\alpha q_\alpha/p)-1)^2-1)-2c_3c_\alpha^2(3(b_1-pb_2)^2$$

$$+(p+2)b_2(2b_1-pb_2))\Big\}. \tag{2.19}$$

$$V_{AS}(\hat{\Sigma}^{ij}_{MCD},\Phi_p) = -\frac{1}{2c_3}. \qquad si \quad i \neq j \tag{2.20}$$

Démonstration:

On va démontrer seulement la formule (2.20).

Si $F = \Phi_p$, donc la fonction de densité est donnée par:

$f(x) = g(x'x)$ avec $g(t) = \dfrac{1}{(2\pi)^{p/2}} e^{-t/2}$.

En remplaçant $IF(x,\Sigma_{ij},F)$ dans (2.18), on obtient:

$$V_{AS}(\hat{\Sigma}^{ij}_{MCD},\Phi_p) = \int_{x'x\leq q_\alpha} \frac{x_i^2 x_j^2}{4c_3^2} g(x'x)dx.$$

En utilisant le lemme de Lopuhaä, on aura:

$$\int_{x'x\leq q_\alpha} x_i^2 x_j^2 g(x'x)dx = \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+3} g(r^2)dr.$$

Puisque $g(r^2) = -2g'(r^2)$, donc:

$$V_{AS}(\hat{\Sigma}^{ij}_{MCD},\Phi_p) = \frac{-1}{2c_3^2}\frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)}\int_0^{\sqrt{q_\alpha}} r^{p+3}g'(r^2)dr.$$

or

$$c_3 = \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)}\int_0^{\sqrt{q_\alpha}} r^{p+3}g'(r^2)dr.$$

Ce qui donne:

$$V_{AS}(\Sigma^{ij}_{MCD},\Phi_p) = \frac{-1}{2c_3}.$$

Efficacité asymptotique

La variance asymptotique est utilisée pour calculer l'efficacité asymptotique d'un estimateur sous le modèle de distribution F.

Définition 2.14. L'efficacité asymptotique de $\hat{\Sigma}_{ij}$ est définie comme suit:

$$Eff(\hat{\Sigma}_{ij},F) = \frac{1}{V_{AS}(\hat{\Sigma}_{ij},F)\mathcal{I}(\hat{\Sigma}_{ij},F)} \qquad \forall \quad 1 \leq i,j \leq p, \tag{2.21}$$

où $\mathcal{I}(\hat{\Sigma}_{ij},F)$ représente l'information de Fisher de $\hat{\Sigma}_{ij}$.

L'information de Fisher sous le modèle normal est donnée comme suit:

$$\mathcal{I}(\hat{\Sigma}_{ij},\Phi_p) = \begin{cases} \dfrac{1}{2} & \text{si} \quad i=j \\ \\ 1 & \text{sinon.} \end{cases} \tag{2.22}$$

La Figure 2.2 (a) (respectivement 2.2 (b)) représente les valeurs des efficacités asymptotiques des éléments diagonaux (respectivement non diagonaux) de $\hat{\Sigma}_{MCD}$ pour $p \in \{2,3,5,10\}$ et $\alpha \in [0,0.5]$.

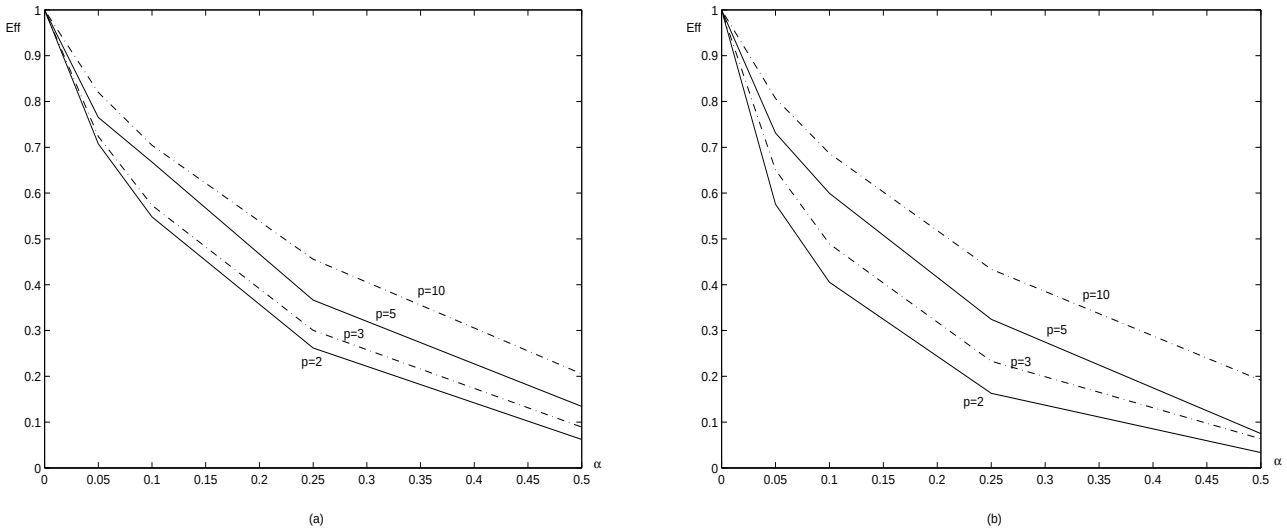

Figure 2.2: Efficacité asymptotique des éléments du MCD
(a): élément diagonal, (b): élément non diagonal

D'après la Figure 2.2, on remarque que lorsque le point de rupture α augmente, l'efficacité diminue rapidement.

Pour améliorer l'efficacité du MCD sous le modèle normal, le MCDR (MCD repondéré) a été introduit (Rousseeuw [50]).

La première étape du MCDR est notée par MCD^1.

2.5.3 L'estimateur MCD^1

Définition 2.15 (Définition du MCD^1). Le MCD^1 du vecteur moyen et de la matrice de covariance qu'on note respectivement $\hat{\mu}^1$ et $\hat{\Sigma}^1$, sont définis comme suit:

$$\hat{\mu}^1 = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad et \quad \hat{\Sigma}^1 = c_1 \frac{\sum_{i=1}^n w_i (x_i - \hat{\mu}^1)(x_i - \hat{\mu}^1)'}{\sum_{i=1}^n w_i}. \tag{2.23}$$

On note par: ($\hat{\mu}^0$,$\hat{\Sigma}^0$) le MCD initial du vecteur moyen et de la matrice de covariance.

Les poids w_i, $(i = 1, \cdots, n)$ sont calculés à partir des estimateurs initiaux, en effet:

$$w_i = w[(x_i - \hat{\mu}^0)'(\hat{\Sigma}^0)^{-1}(x_i - \hat{\mu}^0)],$$

où w:$[0, +\infty[\to \mathbb{R}$ est une fonction "poids" qui est définie par:

$$w(t) = I_{[0,q_\beta]}(t) \quad avec \quad q_\beta = G^{-1}(1-\beta) \quad \text{et} \quad G(u) = P_F(x'x \leq u).$$

Si F est la loi normale multivariée, alors : $q_\beta = \chi^2_{p,(1-\beta)}$.

Le facteur c_1 assurant la consistance de Σ^1 est donnée par:

$$c_1 = (1-\beta)\Big\{\frac{\pi^{p/2}}{\Gamma(p/2+1)}\int_0^{\sqrt{q_\beta}} r^{p+1}g(r^2)dr\Big\}^{-1}.$$

Quelques propriétés:

i. Le MCD[1] et le MCD ont le même point de rupture.

ii. Le MCD[1] s'obtient en utilisant le logiciel S-plus grâce à la fonction **cov.mcd(X)**, X étant la matrice des données.

Algorithme de calcul

Le poids de chaque observation est donnée comme suit:

$$P_{\mu_i} = P_{\Sigma_i} = \frac{w_i}{\sum_{i=}^n w_i}.$$

(1) On calcule le MCD du vecteur moyen $\hat{\mu_0}$ et celui de la matrice de covariance $\hat{\Sigma_0}$.

(2) On calcule les poids $P_i = \frac{w_i}{\sum_{i=1}^n w_i}$, $i = 1,..,n$.

(3) On calcule les estimateurs:

$$\hat{\mu}^1 = \sum_{i=1}^n P_i x_i \quad et \quad \hat{\Sigma}^1 = c_1 \sum_{i=1}^n P_i (x_i - \hat{\mu}^1)(x_i - \hat{\mu}^1)'.$$

Fonction d'influence du MCD[1] de la matrice de covariance

Théorème 2.2 (Lopuhaä[40]). *La fonction d'influence de Σ^1 sous le modèle de distribution $F = F_{0,I_p}$, a été déduite par Lopuhaä [40] comme suit:*

$$IF(x,\Sigma^1,F) = \frac{d_2+2d_3}{d_2}\Big(IF(x,\Sigma^0,F) + \frac{1}{2}tr(IF(x,\Sigma^0,F))I_p\Big) + \frac{1}{d_2}I(x'x \le q_\beta)xx' - I_p. \tag{2.24}$$

Les constantes d_2 et d_3 sont déterminées par les relations suivantes:

$$d_2 = \frac{(1-\beta)}{c_1} \quad , \quad d_3 = \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)}\int_0^{\sqrt{q_\beta}} r^{p+3}g'(r^2)dr.$$

Fonctions d'influence des éléments du MCD[1] de la matrice de covariance

En prenant le $i^{ème}$ élément de la diagonale dans la matrice de $IF(x,\Sigma^1,F)$, on trouve:

$$IF(x,\Sigma^1,F)_{ii} = \frac{d_2+2d_3}{d_2}\Big(IF(x,\Sigma^0,F)_{ii} + \frac{1}{2}\mathrm{tr}(IF(x,\Sigma^0,F))\Big) + \frac{1}{d_2}I(x'x \le q_\beta)x_i^2 - 1 \tag{2.25}$$

En prenant l'élément (i,j) dans la matrice de $IF(x,\Sigma^1,F)$, on a:

$$IF(x,\Sigma^1,F)_{ij} = \frac{d_2+2d_3}{d_2}IF(x,\Sigma^0,F)_{ij} + \frac{1}{d_2}I(x'x \le q_\beta)x_ix_j \quad \text{si} \quad i \ne j \tag{2.26}$$

où $IF(x,\Sigma^0,F)_{ii}$ et $IF(x,\Sigma^0,F)_{ij}$ sont déjà données par les formules (2.16) et (2.17) respectivement,

$$\text{et } \mathrm{tr}(IF(x,\Sigma^0,F)) = (b_1-pb_2)^{-1}\Big\{(c_\alpha\|x\|^2/(1-\alpha))I(\|x\|^2 \le q_\alpha) + p\Big((c_\alpha/(1-\alpha))(q_\alpha/p)(1-\alpha-I(\|x\|^2 \le q_\alpha)) - 1\Big)\Big\}.$$

Efficacité asymptotique

Les tables 2.1 et 2.2 donnent les efficacité asymptotiques pour les éléments des estimateurs de la matrice de covariance: MCD et MCD[1]. On considère deux cas pour le point de rupture ($\alpha = 0.25$ et $\alpha = 0.5$).

α			$p=2$	$p=3$	$p=5$	$p=10$
0.25	Φ_p	MCD	0.262	0.300	0.366	0.459
		MCD^1	0.599	0.680	0.753	0.836
0.5	Φ_p	MCD	0.062	0.089	0.134	0.205
		MCD^1	0.455	0.595	0.720	0.820

Table 2.1: Efficacité asymptotique d'un élément diagonale du MCD et MCD^1 (Croux & Haesbroeck [15])

α			$p=2$	$p=3$	$p=5$	$p=10$
0.25	Φ_p	MCD	0.163	0.233	0.324	0.438
		MCD^1	0.637	0.736	0.814	0.878
0.5	Φ_p	MCD	0.033	0.063	0.113	0.191
		MCD^1	0.401	0.618	0.783	0.873

Table 2.2: Efficacité asymptotique d'un élément non diagonale du MCD et MCD^1 (Croux & Haesbroeck [15])

On conclut que, l'efficacité asymptotique du MCD^1 est meilleure que celle du MCD.

2.5.4 Le S-estimateur

Définition 2.16 (Davies [20]). Soit ξ: $\mathbb{R} \rightarrow [0,\infty[$ une fonction qui satisfait les conditions suivantes:
(R_1) ξ est symétrique, sa dérivée ψ est continue et $\xi(0)=0$.
(R_2) Il existe une constante $c_0 > 0$ telle que ξ soit croissante sur un intervalle $[0,\infty[$.
On pose $a_0 = sup(\xi)$.

Le S-estimateur (Davies [20]) du vecteur moyen et de la matrice de covariance est défini comme la solution $(\hat{\mu}_S,\hat{\Sigma}_S)$ du problème de minimisation du déterminant de S sur tous les vecteurs $t \in \mathbb{R}^p$ et $S \in SPD(p)$ sujets à la contrainte :

$$\frac{1}{n}\sum_{i=1}^{n} \xi\Big\{(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\Big\}^{1/2} = b_0, \tag{2.27}$$

où $0 < b_0 < a_0$, la constante c_0 doit vérifier $\xi(c_0) = \dfrac{b_0}{r_0}$, et $o < r_o \leq \dfrac{(n-p)}{2n}$.

Si la distribution F est elliptique, alors $b_0 = E[\xi(\|x\|)] = E[\xi(\sqrt{x'x})]$.

L'espérance E est prise par rapport à la loi F_{0,I_p}, de densité : $f_{0,I_p}(x) = g(x'x)$.

Exemple 2.2 (Un exemple de fonction ξ:). *(fonction "biweight" de Tukey)*

$$\xi(y,c_0) = \begin{cases} \dfrac{y^2}{2} - \dfrac{y^4}{2c_0^2} + \dfrac{y^6}{6c_0^4} & si \quad |y| \leq c_0 \\ \\ \dfrac{c_0^2}{6} & si \quad |y| \geq c_0 \end{cases} \tag{2.28}$$

Dans ce cas:

1. $\psi(y,c_0) = \xi'(y,c_0) = y(1-(y/c_0)^2)^2 I_{[-c_0,c_0]}(y)$, $\xi(c_0) = \dfrac{c_0^2}{6}$.

2. Si la distribution F est elliptique, c_0 doit vérifier $E[\xi(\sqrt{x'x})] = r_0\dfrac{c_0^2}{6}$, avec :

$$E[\xi(\sqrt{x'x})] = \int_0^\infty \xi(\sqrt{x'x})g(x'x)d(x)$$

D'après le lemme de Lopuhaä:

$$\int_0^\infty \xi(\sqrt{x'x})g(x'x)d(x) = \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty \xi(r)g(r^2)r^{p-1}dr$$

Donc c_0 est solution de :

$$\frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty \xi(r)g(r^2)r^{p-1}dr = r_0\frac{c_0^2}{6}$$

Quelques propriétés:

1. Le S-estimateur représente la classe des estimateurs ayant un plus grand point de rupture, il peut atteindre $(n-p+1)/n$. De plus il est asymptotiquement normal (voir Davies [20]).

2. Le S-estimateur peut se calculer en utilisant le "Fast algorithm "donné dans Ruppert [53].

3. Le point de rupture asymptotique d'un S-estimateur est: $\varepsilon* = r_0$. (Lopuhaä et Rousseeuw, [38]).

Fonction d'influence d'un S-estimateur de la matrice de covariance

Puisque le S-estimateur est affine équivariant, on va seulement donner $IF(x;\Sigma_S,F_{0,I_p})$. Le théorème suivant donne l' expression de la fonction d'influence d'un S-estimateur de la matrice de covariance d'une distribution elliptique.

Théorème 2.3 (Lopuhaä [37]). *Soit $\xi :\rightarrow [o,\infty[$ qui satisfait (R_1) et (R_2). La fonction d'influence $IF(x;\Sigma_S,F_{0,I_p})$ existe, elle est définie comme suit:*

$$IF(x;\Sigma_S,F_{0,I_p}) = \frac{1}{p}tr[IF(x,\Sigma_S,F_{0,I_p})]I_p + \frac{1}{\gamma_1}p\psi(\|x\|)\|x\|\Big(\frac{xx'}{\|x\|^2} - \frac{1}{p}I_p\Big), \tag{2.29}$$

où $\frac{1}{p}\text{tr}[IF(x,\Sigma_S,F_{0,I_p})] = \frac{2}{\gamma_2}(\xi(\|x\|) - b_0)$.

Les constantesγ_1, γ_2, b_0 sont données comme suit:

$$\gamma_2 = E[\psi(\|x\|)]\|x\|,$$

$$\gamma_1 = \frac{E[\psi'(\|x\|)\|x\|^2 + (p+1)\psi(\|x\|)\|x\|]}{p+2}, \text{ et } b_0 = E[\xi(\|x\|)].$$

En utilisant le lemme de Lopuhaä [40], les constantes γ_1, γ_2 peuvent s'écrire comme suit:

$$\gamma_2 = \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty \psi(r)r^p g(r^2)dr,$$

$$\gamma_1 = \frac{2\pi^{p/2}}{\Gamma(p/2)(p+2)}\int_0^\infty [\psi'(r)r^2 + (p+1)\psi(r)r]r^{p-1}g(r^2)dr.$$

Chapitre 3

Les mesures d'influence en A.C.P.

3.1 Introduction

Au cours de la dernière décennie, de nombreux travaux ont été consacrés à la sensibilité des méthodes statistiques descriptives multivariées, particulièrement, dans le domaine de l' analyse en composantes principales. Dans ce contexte, les fonctions d'influence des valeurs propres et vecteurs propres utilisées par Radhakishnan et Kshirsagar[48], Critchley [13], Jolliffe et Morgan [33] et d'autres sont révélées être des outils convenables.

Mais, en A.C.P., on est souvent intéressé par le sous espace E_q engendré par les q premières composantes principales que par les composantes séparément. D'où l'utilité d'étudier l'influence sur ce sous espace. En effet, une observation peut être influente sur des vecteurs principaux mais pas sur le sous espace engendré par ces vecteurs. Cet aspect a été abordé par Tanaka [57]. Une autre approche pour évaluer la sensibilité sur le sous espace E_q est basée sur la fonction d'influence d'un certain coefficient et de quelques mesures d'influence.

Ce chapitre est composé de trois parties. Pour mieux comprendre l'utilisation des différentes mesures d'influence dans la pratique, chaque partie est illustrée par un exemple numérique.

Dans la première partie, nous exposerons l'approche de Critchley [13] où la fonction d'influence des valeurs propres et vecteurs propres est utilisée pour la détection des observations influentes sur les éléments propres de la matrice de covariance et de corrélation classiques.

La seconde partie sera consacrée à l'étude de l'influence sur le sous espace E_q engendré par les q dominantes composantes principales. Concernant cet aspect, les approches dues à Tanaka [57], Bénasséni [2], Prendergast [46] et Prendergast & Li Wai Suen [47] sont développées. Dans la troisième partie, une comparaison entre les mesures

citées dans la partie précédente est effectuée à travers un exemple pratique.

3.2 Détection des observations influentes sur les éléments propres des matrices de covariance et de corrélation classiques

Pour détecter les observations influentes sur les éléments propres de la matrice de covariance classique (ou de corrélation), Critchley [13] propose de calculer les fonctions d'influence des éléments propres.

3.2.1 Fonctions d'influence des éléments propres de la matrice de covariance classique

Soit X un vecteur aléatoire dans $\mathbb{R}^p$, on note F sa fonction de répartition. Son espérance mathématique et sa matrice de covariance sont respectivement :

$$\mu = \mu(F) = \int_{\mathbb{R}^p} x\, dF(x),$$

$$\Sigma = \Sigma(F) = \int_{\mathbb{R}^p} (x-\mu)(x-\mu)'\, dF(x).$$

On suppose que Σ admet des valeurs propres distinctes $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.
Notons v_1,v_2,$\cdots$,v_p les vecteurs propres orthonormés correspondants.

Pour calculer la fonction d'influence de chaque élément propre de Σ, on introduit la contaminée de F par δ_x.

Définition 3.1 (Fonction d'influence). On appelle contaminée de F par δ_x la fonction $\tilde{F}$ donnée par:

$$\tilde{F} = (1-\varepsilon)F + \varepsilon\delta_x, \tag{3.1}$$

où $x \in \mathbb{R}^p$ et ε un élément de [0,1].

Proposition 3.1. *La matrice de covariance associée à la loi $\tilde{F}$, notée par $\tilde{\Sigma}$, est donnée comme suit:*

$$\tilde{\Sigma} = \Sigma(\tilde{F}) = \Sigma + \varepsilon\Big[(x-\mu)(x-\mu)' - \Sigma\Big] - \varepsilon^2(x-\mu)(x-\mu)'. \tag{3.2}$$

Démonstration En effet, par définition:

$$\tilde{\mu} = \mu(\tilde{F}) = \int_{\mathbb{R}^p} z d\tilde{F}(z), \tag{3.3}$$

$$\tilde{\Sigma} = \Sigma(\tilde{F}) = \int_{\mathbb{R}^p} (z - \mu(\tilde{F}))(z - \mu(\tilde{F}))' d\tilde{F}(z). \tag{3.4}$$

En remplaçant (3.1) dans (3.3), on obtient:

$$\tilde{\mu} = (1 - \varepsilon)\mu + \varepsilon x \tag{3.5}$$

Ensuite, en remplaçant (3.1) et (3.5) dans (3.4), on aura:

$$\tilde{\Sigma} = (1 - \varepsilon) \int_{\mathbb{R}^p} \Big[(z - \mu) - \varepsilon(x - \mu)\Big] \Big[(z - \mu) - \varepsilon(x - \mu)\Big]' dF(z)$$
$$+\varepsilon(1 - \varepsilon)^2 (x - \mu)(x - \mu)'.$$

On pose:

$$J = \int_{\mathbb{R}^p} \Big[(z - \mu) - \varepsilon(x - \mu)\Big] \Big[(z - \mu) - \varepsilon(x - \mu)\Big]' dF(z)$$

Calculons J

$$J = \int_{\mathbb{R}^p} \Big[(z - \mu) - \varepsilon(x - \mu)\Big] \Big[(z - \mu) - \varepsilon(x - \mu)\Big]' dF(z)$$

$$= \int_{\mathbb{R}^p} (z - \mu)(z - \mu)' dF(z) - \varepsilon \int_{\mathbb{R}^p} (z - \mu)(x - \mu)' dF(z)$$

$$-\varepsilon \int_{\mathbb{R}^p} (x - \mu)(z - \mu)' dF(x) + \varepsilon^2 (x - \mu)(x - \mu)' \int_{\mathbb{R}^p} dF(z)$$

$$= \Sigma + \varepsilon^2 (x - \mu)(x - \mu)'.$$

Par suite,

$$\Sigma(\tilde{F}) = \Sigma + \varepsilon \Big[(x - \mu)(x - \mu)' - \Sigma\Big] - \varepsilon^2 (x - \mu)(x - \mu)'.$$

D'aprés Rellich [49], les éléments propres de $\tilde{\Sigma}$ s'écrivent alors pour chaque $j \in \{1, \cdots, p\}$:

$$\tilde{\lambda}_j = \lambda_j + \varepsilon\lambda_j^{(1)} + \frac{1}{2}\varepsilon^2\pi_j + O(\varepsilon^3),$$
$$\tilde{v}_j = v_j + \varepsilon v_j^{(1)} + \frac{1}{2}\varepsilon^2\gamma_j + O(\varepsilon^3).$$

Or, on a pour chaque j:

$$IF(x;\lambda_j,F) = \lim_{\varepsilon\to 0}\frac{\lambda_j(\varepsilon) - \lambda j}{\varepsilon} = \frac{d\tilde{\lambda}_j}{d\varepsilon}|_{\varepsilon=0} = \lambda_j^{(1)}$$
$$IF(x;v_j,F) = \lim_{\varepsilon\to 0}\frac{v_j(\varepsilon) - vj}{\varepsilon} = \frac{d\tilde{v}_j}{d\varepsilon}|_{\varepsilon=0} = v_j^{(1)} \tag{3.6}$$

Les nombres $\lambda_j^{(1)}$ et $v_j^{(1)}$ sont calculés en utilisant le lemme de R. Sibson [56].

Lemme 3.1 (R. Sibson [56]). *Soient B,C,D des matrices symétriques, λ_B une valeur propre simple de B et e le vecteur propre orthonormé correspondant. On considère la perturbation suivante:*

$$B(\varepsilon) = B + \varepsilon C + \frac{1}{2}\varepsilon^2 D + O(\varepsilon^3)$$

Supposons que les éléments propres de $B(\varepsilon)$ sont:

$$\lambda(\varepsilon) = \lambda + \varepsilon\mu + \frac{1}{2}\varepsilon^2 v + O(\varepsilon^3),$$
$$e(\varepsilon) = e + \varepsilon f + \frac{1}{2}\varepsilon^2 g + O(\varepsilon^3).$$

Alors :

$$\mu = e'Ce, \quad f = -(B-\lambda I)^+Ce, \quad v = e'(D - 2C(B-\lambda I)^+C)e, \tag{3.7}$$

où $(B-\lambda I)^+$ est l'inverse généralisé de $(B-\lambda I)$ au sens suivant:

$$(B-\lambda I)(B-\lambda I)^+(B-\lambda I) = (B-\lambda I), \quad (B-\lambda I)^+(B-\lambda I)(B-\lambda I)^+ = (B-\lambda I)^+.$$

De plu, sa décomposition spectrale est donnée par :

$$(B-\lambda I)^+ = \sum_{\lambda_k\neq\lambda}(\lambda_k - \lambda)^{-1}e_k e_k',$$

où $\{e_1,e_2,\cdots,e_k\}$ est une base orthonormale de vecteurs propres de B associés aux valeurs propres $\lambda_1 > \lambda_2 > \cdots > \lambda_k$.

Ce lemme donne les expressions explicites de $\lambda_j^{(1)}$ et $v_j^{(1)}$. Ainsi, on établit les formules explicites des fonctions d'influence des éléments propres de la matrice de covariance Σ . Ces formules sont données par le théorème suivant:
D'après les relations (3.6) et (3.7):

Théorème 3.1 (Critchley [13]). *Les fonctions d'influence des valeurs propres et vecteurs propres de* Σ *pour chaque* $j \in \{1,2,\cdots,p\}$ *sont respectivement:*

$$IF(x;\lambda_j,F) = \prec x-\mu,v_j \succ^2 -\lambda_j, \tag{3.8}$$

$$IF(x;v_j,F) = - \prec x-\mu,v_j \succ \sum_{k\neq j}(\lambda_k-\lambda_j)^{-1}v_k \prec x-\mu,v_k \succ , \tag{3.9}$$

où $\prec x-\mu,v_j \succ = (x-\mu)'v_j$.

Remarque 3.1. En utilisant la formule (3.2), on déduit la fonction d'influence de la matrice de covariance classique Σ :

$$IF(x;\Sigma,F) = \frac{d\Sigma(\tilde{F})}{d\varepsilon}|_{\varepsilon=0} = (x-\mu)(x-\mu)' - \Sigma \tag{3.10}$$

Donc $\Sigma(\tilde{F})$ peut s'écrire :

$$\Sigma(\tilde{F}) = \Sigma + \varepsilon IF(x;\Sigma,F) + O(\varepsilon).$$

Fonctions d'influence empiriques

Dans la pratique F est estimée par F_n la fonction de répartition empirique associée à l'échantillon $(x_1,x_2,\cdots,x_n)$ issu de F. On note alors:

$$\mu(F_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \quad , \quad \Sigma(F_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' = \hat{\Sigma}_{cov}.$$

Et, pour chaque $j \in \{1,2,\cdots,p\}$,

$$\lambda_j(F_n) = \hat{\lambda}_j \quad , \quad v_j(F_n) = \hat{v}_j.$$

Les fonctions d'influences empiriques (en anglais, Empirical Influence Function, notée EIF) de la $j^{ème}$ valeur propre et du $j^{ème}$ vecteur propre (notées EIF) sont alors:

$$EIF(x;\lambda_j,F) = IF(x,\hat{\lambda}_j) = \prec x - \bar{x},\hat{v}_j \succ^2 - \hat{\lambda}_j,$$

$$EIF(x;v_j,F) = IF(x,\hat{\alpha}_j) = - \prec x - \bar{x},\hat{v}_j \succ \sum_{k\neq j} (\hat{\lambda_k} - \hat{\lambda}_j)^{-1} \hat{v}_k \prec x - \bar{x},\hat{v}_k \succ .$$

En s'intéressant plus particulièrement au point $x = x_i$, pour chaque $i \in \{1,2,\cdots,n\}$, il vient pour chaque $j \in \{1,2,\cdots,p\}$:

$$IF(x_i,\hat{\lambda}_j) = \prec x_i - \bar{x},\hat{v}_j \succ^2 - \hat{\lambda}_j, \tag{3.11}$$

$$IF(x_i,\hat{v}_j) = - \prec x_i - \bar{x},\hat{v}_j \succ \sum_{k\neq j} (\hat{\lambda_k} - \hat{\lambda}_j)^{-1} \hat{v}_k \prec x_i - \bar{x},\hat{v}_k \succ . \tag{3.12}$$

Remarques 3.1.

i. Pour déterminer les observations influentes sur les valeurs propres, il faut calculer $EIF(x_i,\lambda_j,F)$ pour chaque $i = 1,\cdots,n$.

ii. Pour détecter les observations influentes sur les vecteurs propres, il faut calculer $\|EIF(x_i,v_j,F)\|$ pour chaque $i = 1,\cdots,n$.

3.2.2 Fonctions d'influence des éléments propres de la matrice de corrélation classique

Notons par Σ^{ii} l'élément (i,i) de Σ. La matrice de corrélation est:

$$R = L^{-\frac{1}{2}} \Sigma L^{-\frac{1}{2}},$$

où

$$L^{-\frac{1}{2}} = diag((\Sigma^{11})^{-\frac{1}{2}},\cdots,(\Sigma^{pp})^{-\frac{1}{2}}).$$

: On suppose que R admet des valeurs propres distinctes $\beta_1 > \beta_2 \cdots > \beta_p$ les valeurs propres de R, et $\nu_1,\cdots,\nu_p$ les vecteurs propres orthonormés correspondants.

Par analogie avec ce qui a été fait pour la matrice de covariance, les formules des fonctions d'influence des éléments propres de R pour chaque $i \in \{1, \cdots, n\}$ et pour chaque $j \in \{1, \cdots, p\}$ sont données par le théorème suivant :

Théorème 3.2 (Critcheley [13]). *Les fonctions d'influence des éléments propres de R sont données comme suit:*

$$IF(x_i, \beta_j) = \prec z_i, \nu_j \succ^2 - \beta_j \nu_j' D_{z_i} \nu_j, \tag{3.13}$$

$$IF(x_i, \nu_j) = \sum_{k \neq j} \Big(\prec z_i, \nu_j \succ \prec z_i, \nu_k \succ - \frac{\beta_k + \beta_j}{2} \nu_j' D_{z_i} \nu_k \Big) \frac{\nu_k}{\beta_j - \beta_k}, \tag{3.14}$$

où

$$z_i = L^{-\frac{1}{2}}(x_i - \bar{x}), \quad D_{z_i} = diag(z_i z_i'), \quad et \prec z_i, \nu_j \succ = z_i' \nu_j.$$

Exemple numérique

Pour illustrer les résultats théoriques, Critchley [13] a utilisé les données de Kendall [35] sur lesquelles:

1. Il calcule les fonctions d'influence empiriques des valeurs propres de la matrice de covariance classique $IF(x_i, \hat{\lambda}_j)$, et les normes des fonctions d'influence empiriques des vecteurs propres associés $\|IF(x_i, \hat{v}_j)\|$, pour $j = 1{,}2$ et $i = 1, \cdots, n$.

2. Il calcule les fonctions d'influence empiriques des valeurs propres de la matrice de corrélation classique $IF(x_i, \beta_j)$, et les normes des fonctions d'influence empiriques des vecteurs propres associés $\|IF(x_i, \nu_j)\|$, pour $j = 1{,}2$ et $i = 1, \cdots, n$.

La Table 3.1 représente les données de Kendall [35] concernant vingt observations sous forme de vingt différents sols sur lesquels on observe les quatre variables suivantes:

X^1: contenu de vase, X^2: contenu d'argile, X^3: matière organique, X^4: acidité.

i	X^1	X^2	X^3	X^4	i	X^1	X^2	X^3	X^4
1	13	9.7	1.5	6.4	11	26.5	14.9	2.4	6.7
2	10	7.5	1.5	6.5	12	22.3	8.4	4	7
3	20.6	12.5	2.3	7	13	30.8	7.4	2.7	6.4
4	33.8	19	2.8	5.8	14	25.3	7	4.8	7.3
5	20.5	14.2	1.9	6.9	15	31.2	11.6	2.4	6.5
6	10	6.7	2.2	7	16	22.7	10.1	3.3	6.2
7	12.7	5.7	2.9	6.7	17	31.2	9.6	2.4	6
8	36.5	15.7	2.3	7.2	18	13.2	6.6	2	5.8
9	37.1	14.3	2.1	7.2	19	11.1	6.7	2.2	7.2
10	25.5	12.9	1.9	7.3	20	20.7	9.6	3.1	5.9

Table 3.1: Tableau des données de Kendall.

La Figure 3.1 (respectivement 3.2) représente les graphiques associés aux fonctions d'influences empiriques (pour les deux premières valeurs propres) et aux normes des fonctions d'influence empiriques (pour les vecteurs propres) des éléments propres de la matrice de covariance classique (respectivement de la matrice de corrélation classique).

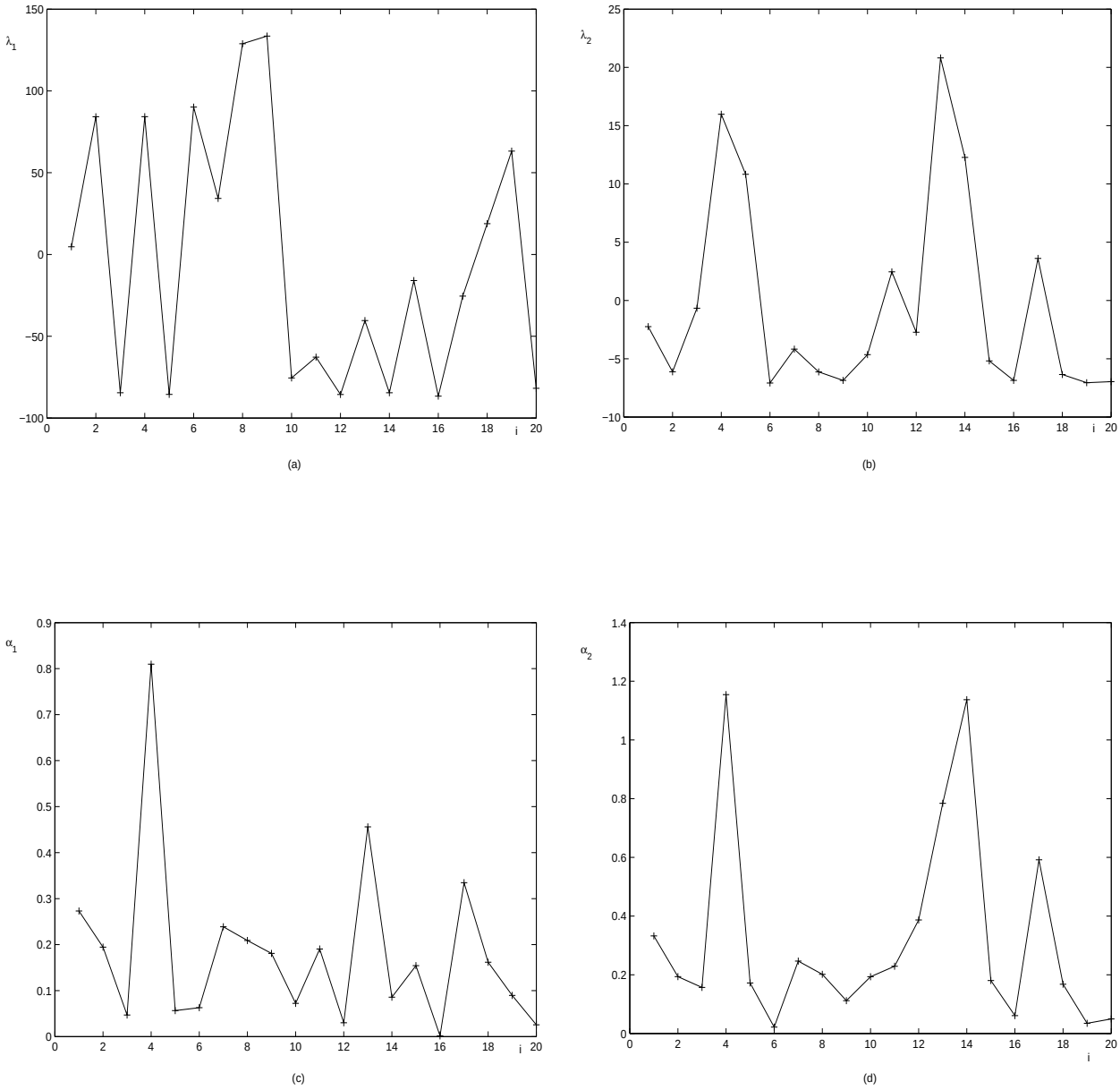

Figure 3.1: Graphiques des fonctions d'influence des valeurs propres et normes des fonctions d'influence des vecteurs propres de la matrice de covariance classique.
(a): $IF(x_i,\hat{\lambda}_1)$, (b): $IF(x_i,\hat{\lambda}_2)$, (c): $\|IF(x_i,\hat{v}_1)\|$, (d): $\|IF(x_i,\hat{v}_2)\|$

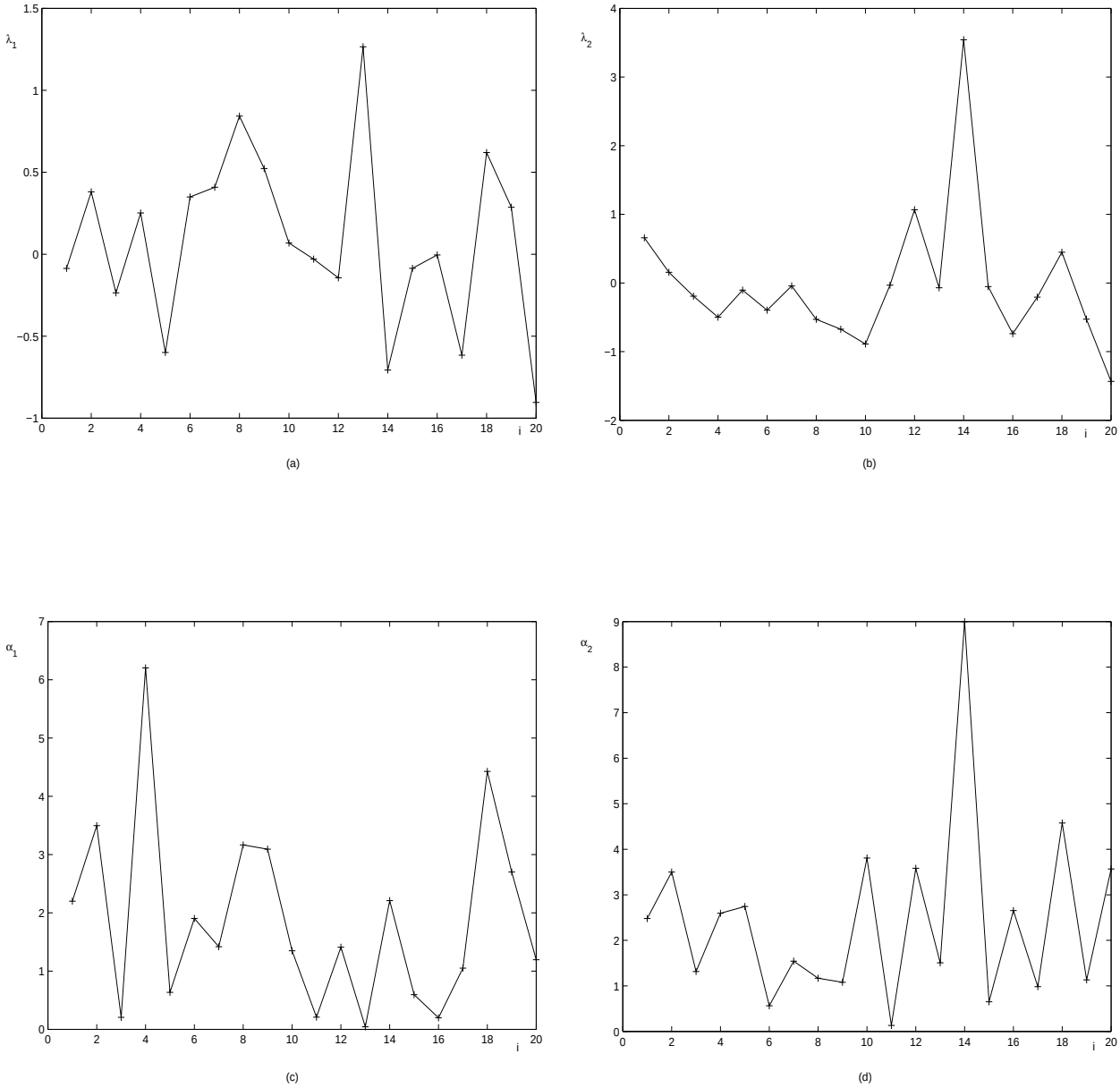

Figure 3.2: Graphiques des fonctions d'influence des valeurs propres et normes des fonctions d'influence des vecteurs propres de la matrice de corrélation classique.
(a): $IF(x_i,\beta_1)$, (b): $IF(x_i,\beta_2)$, (c): $\|IF(x_i,\nu_1)\|$, (d): $\|IF(x_i,\nu_2)\|$

Les figures 3.1 et 3.2 mettent en évidence les observations influentes sur les éléments propres des matrices de covariance et de corrélation classiques. Elles sont résumées par la Table 3.2 .

covariance		corrélation	
éléments propres	observations influentes	éléments propres	observations influentes
$\hat{\lambda}_1$	—	β_1	13
$\hat{\lambda}_2$	4,13	β_2	14
$\hat{v}_1$	4,13	ν_1	4
$\hat{v}_2$	4,13,14	ν_2	14

Table 3.2: Observations influentes détectées par $EIF(x_i,\lambda_j)$ et $\|EIF(x_i,v_j)\|$.

3.3 Influence sur le sous espace engendré par les dominantes composantes principales

3.3.1 Approche de Tanaka

Pour étudier la stabilité du sous espace engendré par les premières composantes principales, Tanaka [57] a proposé de calculer les fonctions d'influence de l'opérateur de projection orthogonale sur ce sous espace, et de la décomposition spectrale de la matrice de covariance ou de corrélation.

Quelques définitions et notations

Soit W une matrice symétrique $(p \times p)$ ayant p valeurs propres $\alpha_1 > \alpha_2 \cdots > \alpha_p$. Notons par $u_1, \cdots, u_p$ les vecteurs propres orthonormés correspondants.

On sait que la matrice perturbée (associé à $\tilde{F}$) $\tilde{W}$ s'écrit comme suit:

$$\tilde{W} = W + \varepsilon W^{(1)} + \varepsilon^2 W^{(2)} + O(\varepsilon^3) \tag{3.15}$$

Les éléments propres de $\tilde{W}$ pour chaque $j \in \{1, \cdots, p\}$ sont:

$$\tilde{\alpha}_j = \alpha_j + \varepsilon \alpha_j^{(1)} + \varepsilon^2 \alpha_j^{(2)} + O(\varepsilon^3), \tag{3.16}$$

$$\tilde{u}_j = u_j + \varepsilon u_j^{(1)} + \varepsilon^2 u_j^{(2)} + O(\varepsilon^3). \tag{3.17}$$

D'après le lemme de Sibson [56], les valeurs de $\alpha_j^{(1)}$ et $u_j^{(1)}$ pour chaque s$\in \{1, \cdots, p\}$ sont alors:

$$\alpha_j^{(1)} = u_j' W^{(1)} u_j, \tag{3.18}$$

$$u_j^{(1)} = \sum_{k \neq j} (\alpha_j - \alpha_k)^{-1} (u_k' W^{(1)} u_j) u_k. \tag{3.19}$$

Remarque 3.2. La matrice W représente la matrice de covariance ou de corrélation, et $W^{(1)}$ est la fonction d'influence de W.

On considère la décomposition spectrale de la matrice W:

$$W = V_1 \Theta_1 V_1' + V_2 \Theta_2 V_2', \tag{3.20}$$

avec Θ_1=diag$(\alpha_1, \cdots, \alpha_q)$ et Θ_2= diag$(\alpha_{q+1}, \cdots, \alpha_p)$,
et
$V1 = (u_1, \cdots, u_q)$ et $V2 = (u_{q+1}, \cdots, u_p)$.

Pour étudier la stabilité de la décomposition (3.20), on considère les deux matrices suivantes:

$$Q_1 = V_1 \Theta_1 V_1' = \sum_{j=1}^{q} \alpha_j u_j u_j', \tag{3.21}$$

$$P_1 = V_1 V_1' = \sum_{j=1}^{q} u_j u_j'. \tag{3.22}$$

Puisque Q_1 est le premier terme de la décomposition (3.20), alors la stabilité de Q_1 impliquera directement la stabilité de la décomposition (3.20).

Définition 3.2. La matrice P_1 est l'opérateur de projection orthogonale sur le sous espace $L(V_1)$ engendré par les vecteurs propres associés aux q plus grandes valeurs propres.

Donc la stabilité de P_1 impliquera directement la stabilité du sous espace $L(V_1)$.

Ce qui nous amène à étudier les fonctions d'influence de P_1 et Q_1.

Fonction d'influence de Q_1

Théorème 3.3 (Tanaka (1988)). *Supposons que W est une matrice symétrique $(p \times p)$ ayant une décomposition spectrale de la forme (3.20). Sa fonction d'influence est $W^{(1)}$. La fonction d'influence de $Q_1 = V_1 \Theta_1 V_1'$ est définie comme suit:*

$$\begin{aligned} IF(x; Q_1) = & \sum_{j=1}^{q} \sum_{k=1}^{q} (u_j' W^{(1)} u_k) u_j u_k' \\ & + \sum_{j=1}^{q} \sum_{k=q+1}^{p} \alpha_j (\alpha_j - \alpha_k)^{-1} (u_j' W^{(1)} u_k)(u_j u_k' + u_k u_j'). \end{aligned} \tag{3.23}$$

Démonstration:

On a: $Q_1 = \sum_{j=1}^{q} \alpha_j u_j u'_j$, donc $Q_1(\varepsilon)$ s'écrit:

$$Q_1(\varepsilon) = \sum_{j=1}^{q} \alpha_j(\varepsilon) u_j(\varepsilon) u'_j(\varepsilon). \tag{3.24}$$

Posons:

$$A(\varepsilon) = \alpha_j(\varepsilon) u_j(\varepsilon) u'_j(\varepsilon). \tag{3.25}$$

D'après (3.16) et (3.17), la relation (3.25) s'écrit:

$$A(\varepsilon) = \alpha_j u_j u'_j + \varepsilon \alpha_j u_j u_j^{\prime(1)} + \varepsilon \alpha_j u_j^{(1)} u'_j + \varepsilon \alpha_j^{(1)} u_j u'_j + O(\varepsilon^2).$$

Par suite:

$$Q_1(\tilde{F}) = \sum_{j=1}^{q} \alpha_j u_j u'_j + \varepsilon \sum_{j=1}^{q} \alpha_j u_j u_j^{\prime(1)} + \varepsilon \sum_{j=1}^{q} \alpha_j u_j^{(1)} u'_j + \varepsilon \sum_{j=1}^{q} \alpha_j^{(1)} u_j u'_j + O(\varepsilon^2).$$

En utilisant la définition de la fonction d'influence, on trouve:

$$IF(x; Q_1) = \sum_{j=1}^{q} \alpha_j u_j u_j^{\prime(1)} + \sum_{j=1}^{q} \alpha_j u_j^{(1)} u'_j + \sum_{j=1}^{q} \alpha_j^{(1)} u_j u'_j. \tag{3.26}$$

En remplaçant (3.18) et (3.19) dans (3.26), on obtient :

$$IF(x; Q_1) = \sum_{j=1}^{q} \sum_{k=1}^{q} (u'_j W^{(1)} u_k) u_j u'_k + \sum_{j=1}^{q} \sum_{k=q+1}^{p} \alpha_j (\alpha_j - \alpha_k)^{-1} (u'_j W^{(1)} u_k)(u_j u'_k + u_k u'_j).$$

Fonction d'influence de P_1

Théorème 3.4 (Tanaka [57]). *En utilisant les mêmes notations du théorème 3.1, la fonction d'influence de $P_1 = V_1 V_1'$ est donnée par:*

$$IF(x; P_1) = \sum_{j=1}^{q} \sum_{k=q+1}^{p} (\alpha_j - \alpha_k)^{-1} (u'_j W^{(1)} u_k)(u_j u'_k + u_k u'_j). \tag{3.27}$$

Démonstration:

On a $P_1 = \sum_{j=1}^{q} u_j u_j'$, donc $P_1(\varepsilon)$ s'écrit:

$$P_1(\varepsilon) = \sum_{j=1}^{q} u_j(\varepsilon) u_j'(\varepsilon). \tag{3.28}$$

En remplaçant (3.17) dans (3.28), on obtient:

$$P_1(F_\varepsilon) = \sum_{j=1}^{q} u_j u_j' + \varepsilon \sum_{j=1}^{q} u_j u_j'^{(1)} + \varepsilon \sum_{j=1}^{q} u_j^{(1)} u_j' + O(\varepsilon^2). \tag{3.29}$$

En utilisant la définition de la fonction d'influence, on trouve:

$$IF(x; P_1) = \sum_{j=1}^{q} u_j u_j'^{(1)} + \sum_{j=1}^{q} u_j'^{(1)} u_j'. \tag{3.30}$$

En remplaçant (3.19) dans (3.30), on obtient:

$$IF(x; P_1) = \sum_{j=1}^{q} \sum_{k=q+1}^{p} (\alpha_j - \alpha_k)^{-1} (u_j' W^{(1)} u_k)(u_j u_k' + u_k u_j').$$

Exemple pratique

Pour illustrer les résultats théoriques, Tanaka a repris les données de Kendall (Table 3.1) sur lesquelles:

1. Il calcule $IF(x_i, \hat{Q}_1)$ et $IF(x_i, \hat{P}_1)$ pour chaque $i = 1, \cdots, 20$.
2. Il trace les graphiques associés à $\|IF(x_i, \hat{Q}_1\|)$ et $\|IF(x_i, \hat{P}_1)\|$.
3. A partir des graphiques, il déduit les observations influentes de $\hat{P}_1$ et $\hat{Q}_1$.

La matrice de covariance empirique utilisée est $\hat{\Sigma}_{cov} = \frac{1}{n} \sum_{i=1}^{20} (x_i - \bar{x})(x_i - \bar{x})'$, ses valeurs propres sont: $\hat{\lambda}_1 = 82.308$, $\hat{\lambda}_2 = 6.739$, $\hat{\lambda}_3 = 0.448$ et $\hat{\lambda}_4 = 0.246$.
et $q = 2$, dans ce cas le sous espace $L(V_1)$ est engendré par les vecteurs propres associés au deux premières valeurs propres $\hat{\lambda}_1$ et $\hat{\lambda}_2$ respectivement.

La Figure 3.4 (a) (respectivement 3.4 (b)) représente le graphique associé à la norme de la fonction d'influence de $\hat{Q}_1$ (respectivement à la norme de la fonction d'influence de $\hat{P}_1$).

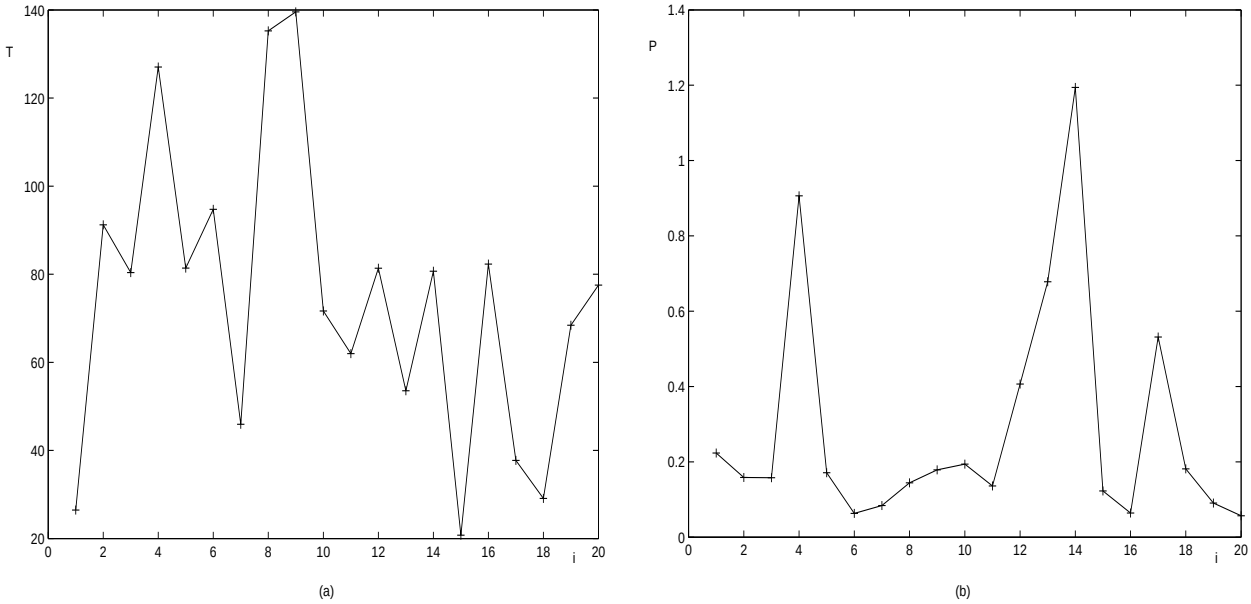

Figure 3.4: Graphiques des normes des fonctions d'influence de $\hat{Q}_1$ et$\hat{P}_1$
(a): $\|IF(x_i,\hat{Q}_1\|)$ (b): $\|IF(x_i,\hat{P}_1)\|$

La Figure 3.4 met en évidence les observations influentes pour $\hat{P}_1$ et $\hat{Q}_1$:

Pour $\hat{Q}_1$, les observations influentes sont: x_4, x_8 et x_9.
Pour $\hat{P}_1$, les observations influentes sont: x_4 et x_{14}.

On rappelle que d'après Critchley [13]:

Pour $\hat{\lambda}_1$, il n'y a pas d'observations influentes.
Pour $\hat{\lambda}_2$, les observations influentes sont: x_4 et x_{13}.

Remarques 3.2.

i. x_{13} est influente sur $\hat{\lambda}_2$, mais pas sur $L(V_1)$.

ii. x_{14} est influente sur $L(V_1)$, mais pas sur $\hat{\lambda}_2$.

On conclut qu'une observation peut être influente sur une valeur propre et non pas sur le sous espace engendré par un vecteur propre associé à cette valeur propre, et inversement.

3.3.2 Le coefficient de Bénasséni

Bénasséni [2], a défini un coefficient qui mesure la proximité entre les sous espaces engendrés par les vecteurs initiaux et leurs versions correspondantes après perturbation.

Quelques notations et définitions

On note par :

E_q: le sous espace engendré par les q premiers vecteurs propres $v_1, \cdots, v_q$ de Σ.

$\tilde{E}_q$: le sous espace engendré par les q vecteurs propres $\tilde{v}_1, \cdots, \tilde{v}_q$ de $\tilde{\Sigma}$.

L'idée principale est d'utiliser la distance entre E_q et $\tilde{E}_q$ qui est définie par:

$$\sum_{j=1}^{q} \|v_j - \tilde{P}v_j\| \quad \text{ou} \quad \sum_{j=1}^{q} \|\tilde{v}_j - P\tilde{v}_j\|, \tag{3.31}$$

où P et $\tilde{P}$ sont les opérateurs de projections sur E_q et $\tilde{E}_q$ respectivement, qui sont définis comme suit:

$$P = VV' \quad \text{et} \quad \tilde{P} = \tilde{V}\tilde{V}',$$

avec:

$$V = (v_1, \cdots, v_q), \ \ et \ \ \tilde{V} = (\tilde{v}_1, \cdots, \tilde{v}_q). \tag{3.32}$$

$\tilde{P}v_j$ est la projection de v_j sur $\tilde{E}_q$ et $P\tilde{v}_j$ est la projection de $\tilde{v}_j$ sur E_q.

Le coefficient de sensibilité

A cause de la normalisation des vecteurs v_j et $\tilde{v}_j$, on peut noter que la norme dans la somme de (3.31) n'est rien d'autre que le sinus de l'angle entre le vecteur et sa projection donc:

$$\frac{1}{q}\sum_{j=1}^{q} \|v_j - \tilde{P}v_j\| \leq 1.$$

Définition 3.3 (Bénasséni [2]). Pour étudier la proximité entre E_q et $\tilde{E}_q$, on utilise le coefficient suivant:

$$\rho = 1 - (\frac{1}{q}\sum_{j=1}^{q} \|v_j - \tilde{P}v_j\|). \tag{3.33}$$

L'outil ρ est considéré comme un critère mesurant la sensibilité du sous espace E_q engendré par les q premiers vecteurs propres lorsque la loi F des observations est contaminée.

Cas particuliers:

i. Si $E_q = \tilde{E}_q$, i.e., il n'y a pas de contamination alors: $\tilde{P}v_j = v_j$ et, dans ce cas :

$$\rho = 1.$$

ii. Si E_q est orthogonal à $\tilde{E}_q$ alors: $\tilde{P}v_j = 0$, dans ce cas :

$$\rho = 0.$$

Remarque 3.3. Le coefficient ρ peut être aussi noté par$\tilde{\rho}$ pour dire que c'est le coefficient associé à $\tilde{F}$.

Pour étudier la stabilité du sous espace engendré par les axes principaux, il suffit de calculer la fonctions d'influence de ρ.

Cherchons donc la fonction d'influence de ρ.

Fonction d'influence du coefficient de sensibilité: Cas classique

Théorème 3.5 (Bénasséni [2]). *La fonction d'influence du premier coefficient de sensibilité ρ est définie comme suit:*

$$IF(x,\rho) = -\frac{1}{q}\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j - \lambda_k)^{-2}y_j^2 y_k^2\Big]^{1/2}\Big\}, \tag{3.34}$$

où $y_j = v_j'(x - \mu)$.

Démonstration :

Supposons que le projecteur $\tilde{P}$ s'écrit sous la forme suivante:

$$\tilde{P} = P + \varepsilon P^{(1)} + O(\varepsilon^2), \tag{3.35}$$

où $P^{(1)}$ est la fonction d'influence de P qui a été déduite par Tanaka (1988) comme suit:

$$P^{(1)} = \sum_{l=1}^{q} \sum_{k=q+1}^{p} (\lambda_l - \lambda_k)^{-1} (v_l' \Sigma^{(1)} v_k)(v_l v_k' + v_k v_l'), \tag{3.36}$$

où $\Sigma^{(1)}$ est la fonction d'influence de Σ, donnée par (3.10).

En utilisant (3.35), $\tilde{P} v_j$ peut s'écrire:

$$\tilde{P} v_j = v_j + \varepsilon P^{(1)} v_j + O(\varepsilon^2). \tag{3.37}$$

En insérant (3.36) dans (3.37), on obtient:

$$\tilde{P} v_j = v_j + \varepsilon \sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-1} (v_j' \Sigma^{(1)} v_k) v_k. \tag{3.38}$$

Puisque $v_j' \Sigma^{(1)} v_k = y_j y_k$, où $y_j = (x - \mu)' v_j$, (3.38) peut s'écrire:

$$\tilde{P} v_j = v_j + \varepsilon \sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-1} y_j y_k v_k. \tag{3.39}$$

La prochaine étape consiste à calculer $\|v_j - \tilde{P} v_j\|$.

Calculons $\|v_j - \tilde{P} v_j\|$.

En utilisant (3.38):

$$\begin{aligned} \|v_j - \tilde{P} v_j\| &= \varepsilon \| \textstyle\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-1} y_j y_k v_k \| \\ &= \varepsilon \Big[\Big(\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-1} y_j y_k v_k \Big)' \Big(\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-1} y_j y_k v_k \Big) \Big]^{1/2}. \end{aligned}$$

Après quelques calculs, on obtient :

$$\|v_j - \tilde{P} v_j\| = \varepsilon \Big[\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-2} y_j^2 y_k^2 \Big]^{1/2}.$$

Par suite:

$$\rho = 1 - \frac{\varepsilon}{q} \Big\{ \sum_{j=1}^{q} \Big[\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-2} y_j^2 y_k^2 \Big]^{1/2} \Big\}.$$

Puisque: $\rho = T(\tilde{F})$ et $T(F) = 1$, alors:

$$\lim_{\varepsilon \to 0} \frac{T(\tilde{F}) - T(F)}{\varepsilon} = -\frac{1}{q}\Big\{ \sum_{j=1}^{q} \Big[\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-2} y_j^2 y_k^2 \Big]^{1/2} \Big\}.$$

Donc :

$$IF(x,\rho) = -\frac{1}{q}\Big\{ \sum_{j=1}^{q} \Big[\sum_{k=q+1}^{p} (\lambda_j - \lambda_k)^{-2} y_j^2 y_k^2 \Big]^{1/2} \Big\}.$$

Exemple pratique

Pour illustrer les résultats théoriques, Bénasséni a repris les données de Kendall (Table 1.1). Les table 3.3 et 3.4 représentent les valeurs de $|EIF(x_i,\rho)|$ pour chaque $i = 1, \cdots, 20$, calculées sur chaque sous espace E_A engendré par un, deux ou trois vecteurs principaux. A étant un ensemble d'indices.

Les valeurs les plus grandes sont soulignées.

$i\backslash A$	{1}	{2}	{3}	{4}	{1,2}	{1,3}	{1,4}
1	0.287	0.350	0.496	0.450	0.141	0.385	0.368
2	0.205	0.204	0.198	0.119	0.112	0.167	0.162
3	0.049	0.165	0.379	0.390	0.082	0.214	0.219
4	0.852	1.215	3.335	3.371	0.539	2.089	2.103
5	0.059	0.181	0.096	0.185	0.087	0.078	0.122
6	0.066	0.023	0.126	0.140	0.035	0.095	0.075
7	0.251	0.260	0.089	0.033	0.056	0.165	0.142
8	0.220	0.212	0.654	0.665	0.102	0.434	0.426
9	0.190	0.118	2.049	2.049	0.117	1.097	1.099
10	0.076	0.203	1.243	1.252	0.112	0.659	0.661
11	0.201	0.241	0.137	0.017	0.076	0.168	0.109
12	0.032	0.407	1.165	1.095	0.211	0.596	0.563
13	0.480	0.825	0.918	0.647	0.371	0.696	0.563
14	0.090	1.197	3.600	3.419	0.612	1.842	1.754
15	0.162	0.190	0.266	0.238	0.082	0.207	0.200
16	0.002	0.064	1.695	1.694	0.033	0.848	0.848
17	0.352	0.623	2.373	2.348	0.310	1.357	1.348
18	0.170	0.177	2.367	2.368	0.128	1.258	1.254
19	0.094	0.036	0.302	0.315	0.053	0.197	0.169
20	0.027	0.052	1.783	1.783	0.038	0.904	0.899

Table 3.3: Valeurs de $|EIF(x_i,\rho)|$ pour le cas classique.

$i\backslash A$	{2,3}	{2,4}	{3,4}	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}
1	0.368	0.396	0.136	0.172	0.240	0.179	0.122
2	0.167	0.161	0.083	0.043	0.114	0.096	0.096
3	0.253	0.236	0.110	0.167	0.154	0.089	0.020
4	2.194	2.140	0.634	1.385	1.310	0.690	0.365
5	0.130	0.080	0.105	0.084	0.044	0.090	0.021
6	0.074	0.081	0.036	0.065	0.046	0.010	0.031
7	0.149	0.146	0.045	0.014	0.048	0.110	0.096
8	0.431	0.427	0.094	0.280	0.243	0.104	0.111
9	1.078	1.078	0.126	0.735	0.738	0.064	0.109
10	0.712	0.676	0.131	0.479	0.448	0.109	0.037
11	0.115	0.129	0.071	0.007	0.056	0.113	0.072
12	0.580	0.748	0.230	0.385	0.503	0.163	0.015
13	0.565	0.724	0.399	0.260	0.453	0.423	0.185
14	1.852	2.285	0.721	1.232	1.532	0.509	0.040
15	0.199	0.213	0.072	0.089	0.132	0.094	0.072
16	0.868	0.872	0.045	0.578	0.581	0.030	0.001
17	1.386	1.444	0.367	0.888	0.951	0.353	0.158
18	1.260	1.256	0.128	0.853	0.843	0.101	0.097
19	0.169	0.173	0.052	0.136	0.107	0.017	0.045
20	0.915	0.905	0.039	0.616	0.606	0.027	0.015

Table 3.4: Valeurs de $|EIF(x_i,\rho)|$ pour le cas classique.

Remarques 3.3. D'après les résultats des tables 3.3 et 3.4, on remarque que :

1. les observations x_4 et x_{14} sont les plus influentes sur la plus part des sous espaces E_A choisis. Ce résultat est attendu puisque ces deux observations sont aussi les plus influentes pour l'opérateur de projection de Tanaka (1988).

2. Une observation peut être influente sur le sous espace Δv_i engendré par un seul vecteur v_i ($i \in I$) mais pas sur le sous espace engendré par ces vecteurs $Vect\{v_i, i \in I\}$. Par exemple, x_4 est influente sur les sous espaces engendrés par v_1, par v_3 et par v_4, mais pas sur le sous espace engendré par $\{v_2,v_3,v_4\}$.

3. Lorsque la dimension de E_A augmente, $|EIF(x_i,\rho)|$ diminue pour chaque $i = 1, \cdots, 20$.

3.3.3 Mesure d'influence de Prendergast (2008)

Prendergast [46] a introduit une mesure basée sur les coefficients RV d'Escoufier [23] et GCD (Generalised Coefficient of Determination) de Yonaï [60].

Le coefficient RV d'Escoufier et le coefficient GCD de Yanai

On considère deux matrices de même dimension $A \in \mathbb{R}^{q\times p}$ et $B \in \mathbb{R}^{q\times p}$.

Définition 3.4. Le coefficient RV d'Escoufier [23] est définie comme suit :

$$RV(A,B) = \frac{trace(B'AA'B)}{\sqrt{trace(A'AA'A)trace(B'BB'B')}}.$$

Définition 3.5. Le coefficient GCD (Generalised Coefficient of Determination) de Yanai [60] est donné par la formule suivante:

$$GCD(A,B) = \frac{trace\{(A(A'A)^{-1}A')(B(B'B)^{-1}B')\}}{q}.$$

On sait que la matrice de projecteur orthogonal de A (resp.B) est $P_A = A(A'A)^{-1}A'$ (resp.$P_B = B(B'B)^{-1}B$), dans ce cas:

$$GCD(A,B) = \frac{trace(P_AP_B)}{q}.$$

Remarque 3.4. Si les colonnes de A sont orthogonaux de même les colonnes de B sont aussi orthogonaux, dans ce cas ces deux coefficients sont identiques.

Notons par: $V_1 = (u_1,u_2,\cdots,u_q)$ la matrice des q premiers vecteurs de W (W représente la matrice de covariance ou de corrélation) , et $\tilde{V_1} = (\tilde{u}_1,\tilde{u}_2,\cdots,\tilde{u}_q)$, la matrice des q premiers vecteurs propres associées à $\tilde{W}$.

Posons: $S = \{1,\cdots,q\}$ et $\bar{S}$ le complémentaire de S. On suppose pour $j \in S$, $r \in \bar{S}$, et $\lambda_j \neq \lambda_r$.

Prendergast [46] a considéré $RV(A,B)$ et $GCD(A,B)$ dans le cas où $A = V_1$ et $B = \tilde{V_1}$. Comme les deux matrices V_1 et $\tilde{V_1}$ sont orthogonales i.e ($V_1'V_1 = I_q$ et $\tilde{V_1}'\tilde{V_1} = I_q$), dans ce cas:

$$RV(V_1,\tilde{V_1}) = GCD(V_1,\tilde{V_1}) = \frac{1}{q}trace(P_1\tilde{P}_1),$$

où: $P_1 = V_1V_1'$ et $\tilde{P}_1 = \tilde{V_1}\tilde{V_1}'$.

Le théorème suivant propose une nouvelle mesure d'influence.

Théorème 3.6 (Prendergast [46]). *On considère le coefficient suivant:*

$$\rho_S(W,F) = \frac{1}{\varepsilon^2}[1 - \frac{1}{q}trace\{P_1\tilde{P}_1\}].$$

Alors ρ_S peut s'écrire comme suit:

$$\rho_S(W,F) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{\{u_j' IF(W,F,x)u_r\}^2}{(\lambda_j-\lambda_r)^2} + O(\varepsilon),$$

où $IF(W,F,x)$ est la fonction d'influence de W.

Démonstration:

On sait que $\tilde{P}_1$ s'exprime comme suit:

$$\tilde{P}_1 = P_1 + \varepsilon P_1^{(1)} + \frac{1}{2}\varepsilon^2 P_1^{(2)} + O(\varepsilon^3),$$

avec $P_1^{(1)}$ est donné par Tanaka [57] et $P_1^{(2)}$ par Tanaka & Castano-Tostado [9].

Puisque $(I - P_1)$ est une matrice de projection, donc:

$$\begin{aligned} q - trace[P_1\tilde{P}_1] &= trace[(I-P_1)\tilde{P}_1] \\ &= trace[(I-P_1)\tilde{P}_1(I-P_1)] \\ &= \varepsilon trace\Big[(I-P_1)\Big(P_1^{(1)} + \frac{1}{2}\varepsilon P_1^{(2)}\Big)(I-P_1)\Big] + O(\varepsilon^3). \end{aligned}$$

Comme $(I-P_1)P_1 = 0$, et d'après Tanaka & Castano-Tostado [9], on a aussi

$$(I-P_1)P_1^{(1)}(I-P_1) = 0.$$

D'où, le fait que $u_r'u_t = 0$ pour $(r \neq t)$, et $u_r'u_r = 1$:

$$\begin{aligned} trace[(I-P_1)P_1^{(2)}(I-P_1)] &= trace\Big[\sum_{r\in\bar{S}}\sum_{t\in\bar{S}}\sum_{j\in S}\frac{u_j'W^{(1)}u_t}{\alpha_j-\alpha_t}\frac{u_j'W^{(1)}u_r}{\alpha_j-\alpha_r}\Big] \\ &= 2\sum_{r\in\bar{S}}\sum_{j\in S}\frac{(u_j'W^{(1)}u_r)^2}{(\alpha_j-\alpha_r)^2}. \end{aligned}$$

Remarque 3.5. Pour ε suffisamment petit, on remarque que $\rho_S(W,F)$ est approximativement égale au coefficient de second ordre en ε de l'expression $1 - \frac{1}{q}trace\{P_1\tilde{P}_1\}$, donc on peut utiliser la mesure suivante:

$$\tilde{\rho}_S(W,F) = \lim_{\varepsilon\to 0}\rho_S(W,F) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{\{u_j' IF(W,F,x)u_r\}^2}{(\alpha_j-\alpha_r)^2}. \tag{3.40}$$

Dans ce qui suit, on donnera la formules de $\tilde{\rho}_S(W,F)$ lorsque:

i. W est la matrice de covariance.

ii. W est la matrice de corrélation.

Proposition 3.2 (Prendergast [46]). *Soit C_0 la fonctionnelle associée à l'estimateur classique de la matrice de covariance Σ, ses valeurs propres $\lambda_1 > \lambda_2 > \cdots ,\lambda_p$ et les vecteurs propres correspondants, $v_1,v_2,\cdots ,v_p$. La fonction d'influence de cet estimateur est $IF(C_0,F,x) = (x-\mu)(x-\mu)'-\Sigma$, en remplaçant cette fonction dans (3.40), on obtient:*

$$\tilde{\rho}_S(C_0,F) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{y_j^2 y_r^2}{(\lambda_j-\lambda_r)^2},$$

où $y_j = v_j'(x-\mu)$.

Remarque 3.6. On sait que la fonction d'influence du coefficient de sensibilité de Bénasséni est donnée par:

$$IF(\rho,F,x) = -\frac{1}{q}\sum_{j\in S}\Big\{\sum_{r\in\bar{S}}\frac{y_j^2 y_r^2}{(\lambda_j-\lambda_r)^2}\Big\}^{1/2}.$$

On conclut que la mesure d'influence de Bénasséni [2] contient la même information de sensibilité que la mesure $\tilde{\rho}_S(C,F)$ de Prendergast [46].

Proposition 3.3 (Prendergast [46]). *Soit R_0 la fonctionnelle associée à l'estimateur classique de la matrice de corrélation R, soient $\beta_1,\beta_2,\cdots ,\beta_p$, et $\nu_1,\nu_2,\cdots ,\nu_p$ les valeurs propres et vecteurs propres de R.*

Notons par x_i le i^{eme} élément x, μ_i le i^{eme} élément de μ et σ_{ii} l'élément i de la diagonale de Σ.
Soit $D = diag(z_1^2,\cdots ,z_p^2)$, avec $z = [z_1,\cdots ,z_p] = [(x_i-\mu_i)/\sigma_{ii},\cdots ,(x_p-\mu_p)/\sigma_{pp}]$. La fonction d'influence de l'estimateur classique de R est donnée par Critchley [13] comme suit:

$$IF(R_0,F,x) = zz' - (DR+RD)/2,$$

en remplaçant cette fonction dans (3.40), on aura:

$$\tilde{\rho}_S(R_0,F) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{1}{(\beta_j-\beta_r)^2}\Big\{z_j\tilde{z}_r - \frac{1}{2}(\beta_j+\beta_r)\nu_j' D\nu_r'\Big\}.$$

Remarques 3.4. Dans la pratique la mesure de Prendergast [46] comme suit:

1. Pour détecter les observations influentes en ACP basée sur la matrice de covariance empirique, on utilise la mesure suivante:

$$\tilde{\rho}_S(C_0,F_n,x) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{y_j^2 y_r^2}{(\hat{\alpha_j}-\hat{\alpha_r})^2}, \tag{3.41}$$

avec $y_j = \hat{v_j}'(x-\hat{\mu})$, où v_j et $\hat{\alpha_j}$, $j = 1,\cdots,p$ sont les éléments propres de la matrice de covariance empirique.

2. Et pour l'ACP basée sur la matrice de corrélation empirique, on utilise la mesure suivante:

$$\tilde{\rho}_S(R_0,F_n,x) = \frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{1}{(\hat{\beta}_j-\hat{\beta}_r)^2}\Big\{z_j z_r - \frac{1}{2}(\hat{\beta}_j+\hat{\beta}_r)\hat{\nu_j}'D\hat{\nu_r}'\Big\}, \tag{3.42}$$

avec $\hat{\nu_j}$, $\hat{\beta}_j$ sont les éléments propres de la matrice de corrélation empirique .

3.3.4 Mesure d'influence de Prendergast & Li Wai Suen (2011)

Prendergast & Li Wai Suen [47] proposent une autre mesure qui est basée sur la moyenne des carées des corrélations canoniques. Cette mesure est utilisée pour la détection des valeurs influentes en grande dimension.

On considère $V'X = [v_j'X]_{j\in S}$ qui est la matrice des variables engendrée par les q premières composantes principales. De même, on considère $\tilde{V}'X=[\tilde{v_j'}X]_{j\in S}$ la matrice des variables obtenue après avoir perturbé $V'X = [v_j'X]_{j\in S}$ par la loi $\tilde{F}$ qui est donnée par la formule (3.1).

Définition 3.6. Soient $R_1,\cdots,R_q$ les corrélation canoniques entre $V'X$ et $\tilde{V}'X$. La moyenne des carrées de ces corrélations est définie comme suit:

$$\frac{1}{q}\sum_{j=1}^{q}R_j^2 = \frac{1}{q}trace\Big[\tilde{V}'\Sigma V\{V'\Sigma V\}^{-1}V'\Sigma\tilde{V}'\{\tilde{V}'\Sigma\tilde{V}\}^{-1}\Big], \tag{3.43}$$

où V et $\tilde{V}$ sont donnés par (3.32).

Le théorème suivant définit une autre mesure d'influence.

Théorème 3.7 (Prendergast & Li Wai Suen [47]). *Une mesure d'influence basée sur la moyenne des carrées des corrélations canoniques est donnée par:*

$$\frac{1-\frac{1}{q}\sum_{j=1}^{q}R_j^2}{\varepsilon^2}\Big|_{\varepsilon=0}=\frac{1}{q}\sum_{j\in S}\sum_{r\in\bar{S}}\frac{\lambda_r}{\lambda_j}\frac{y_j^2y_r^2}{(\lambda_j-\lambda_r)^2}.$$

Preuve:

Par définition, $\tilde{V}$ s'écrit comme suit:

$$\tilde{V}=V+\varepsilon V^{(1)}+\frac{1}{2}\varepsilon^2V^2+O(\varepsilon^3), \qquad (A.1)$$

avec

$$V^{(1)}=\{\partial\tilde{V}/\partial\varepsilon\}|_{\varepsilon=0} \quad \text{et} \quad V^{(2)}=\{\partial^2\tilde{V}/\partial^2\varepsilon\}|_{\varepsilon=0}.$$

Posons:

$$\tilde{R}^2=\frac{1}{q}\sum_{j=1}^{q}R_j^2.$$

Par ailleurs, on a:

$$\tilde{R}^2=\frac{1}{q}trace\Big[\tilde{V}'\Sigma V\{V'\Sigma V\}^{-1}V'\Sigma\tilde{V}'\{\tilde{V}'\Sigma\tilde{V}\}^{-1}\Big].$$

En utilisant (A.1) et $V=\tilde{V}-\varepsilon V^{(1)}-\frac{1}{2}\varepsilon^2V^{(2)}$, et après avoir réécrit $\varepsilon^2\{\tilde{V}'\Sigma\tilde{V}\}^{-1}$ sous cette forme:

$$\varepsilon^2\{\tilde{V}'\Sigma\tilde{V}\}^{-1}=\varepsilon^2\{V'\Sigma V\}^{-1}+O(\varepsilon^3),$$

on aura:

$$\begin{aligned}\tilde{R}^2 &= \frac{1}{L}trace\Big[I_L-\varepsilon V'^{(1)}\Sigma V\{\tilde{V}'\Sigma\tilde{V}\}^{-1}-\varepsilon^2V'^{(1)}\Sigma V^{(1)}\{V'\Sigma V\}^{-1}\\ &+ \varepsilon V'^{(1)}\Sigma V\{V'\Sigma V\}^{-1}-\varepsilon^2V'^{(1)}\Sigma V\{V'\Sigma V\}^{-1}V'^{(1)}\Sigma V\{V'\Sigma V\}^{-1}\Big]+O(\varepsilon^3). \qquad (A.2)\end{aligned}$$

Étant donné que:

$$\{\tilde{V}'\Sigma\tilde{V}\}^{-1}\tilde{V}'\Sigma\tilde{V}=I_q,$$

et en utilisant:

$$\frac{\partial}{\partial\varepsilon}\{\tilde{V}'\Sigma\tilde{V}\}^{-1}\Big|_{\varepsilon=0} = -(V'\Sigma V)^{-1}\Big\{\frac{\partial}{\partial\varepsilon}\tilde{V}'\Sigma\tilde{V}\Big\}\Big|_{\varepsilon=0}(V'\Sigma V)^{-1},$$

donc le premier ordre de la série $\{\tilde{V}'\Sigma\tilde{V}\}^{-1}$ est donné comme suit:

$$(V'\Sigma V)^{-1} - \varepsilon(V'\Sigma V)^{-1}(V'^{(1)}\Sigma V + V\Sigma V^{(1)})(V'\Sigma V)^{-1} + O(\varepsilon^3). \qquad (A.3)$$

En remplaçant (A.3) de $\{\tilde{V}'\Sigma\tilde{V}\}^{-1}$ dans (A.2), on aura:

$$\tilde{R}^2 = \frac{1}{q}trace\Big[I_L - \varepsilon^2 V'^{(1)}\Sigma^{1/2}P_{L'}\Sigma^{1/2}V^{(1)}(V'\Sigma V)^{-1}\Big] + O(\varepsilon^3), \qquad (A.4)$$

où $P_{q'} = I_q - \Sigma^{1/2}V(V'\Sigma V)^{-1}V'\Sigma^{1/2}$ est la matrice de projection sur le complémentaire de sous espace $\Sigma^{1/2}V$.

Rappelons que v_j est le j^{eme} vecteur propre de Σ qui correspond à la valeur propre λ_j. Puisque $(V'\Sigma V)^{-1}$ est une matrice diagonale $(q \times q)$, le j^{eme} élément de sa diagonale est λ_j^{-1}, la forme (A.4) s'écrit :

$$\tilde{R}^2 = 1 - \varepsilon^2\frac{1}{q}\sum_{j\in S}\frac{1}{\lambda_j}\gamma_j'\Sigma^{1/2}P_{q'}\gamma^{1/2}\gamma_j + O(\varepsilon^3), \qquad (A.5)$$

où γ_j représente la fonction d'influence du j^{eme} vecteur propre de la matrice de covariance Σ. Elle est donné par Critchley [13] comme suit:

$$\gamma_j = IF(v_j,F,x) = y_j\sum_{k=1,k\neq j}\frac{y_k}{\lambda_j - \lambda_k}v_k, \qquad (A.6)$$

avec $y_k = v_k'(x - \mu)$.

Puisque $P_{q'}$ est la matrice de projection sur le complémentaire de l'espace colonne de $\Sigma^{1/2}V$, on a $P_{q'}\Sigma^{1/2}v_j = 0$ pour chaque $j \in S$ et $P_{q'}\Sigma^{1/2}v_r = \lambda_r^{1/2}v_r$ pour chaque $r \in \bar{S}$. De (A.6), on déduit:

$$P_{q'}\Sigma^{1/2}\gamma_j = y_j\sum_{r\in\bar{S},r\neq j}\frac{\lambda_r^{1/2}y_r}{\lambda_j - \lambda_r}v_r, \quad j \in S. \qquad (A.7)$$

Pour achever la démonstration, il suffit d'insérer (A.7) dans (A.5).

Remarque 3.7. Dans la pratique, pour détecter les observations influentes en ACP basée sur la matrice de covariance empirique, on utilise la mesure (SCI) (squared canonical influence) définie par:

$$SCI = \frac{1}{q} \sum_{j \in S} \sum_{r \in \bar{S}} \frac{\hat{\lambda}_r}{\hat{\lambda}_j} \frac{y_j^2 y_r^2}{(\hat{\lambda}_j - \hat{\lambda}_r)^2}, \tag{3.44}$$

avec $\hat{\lambda}_j, \quad j = 1, \cdots, p$ sont les valeurs propres de la matrice de covariance empirique, et $y_j = v_k'(x - \mu)$.

3.3.5 Comparaison entre les mesures d'influence

Pour faire une comparaison entre le coefficient de Bénasséni [2], la mesure de Prendergast [46] et celle de Prendergast & Li Wai Suen [47], ces derniers ont utilisé les données de logements de Boston (Boston housing data, Harrison and Rubinfield, [28]) composées de 14 variables socio-économiques pour les 506 secteurs de recensement de Boston.

Les résultats obtenus sont illustrés par la Figure (3.5).
La ligne continue (–) représente la mesure SCI (Prendergast & Li Wai Suen [47]).
La ligne discontinue (- -) représente la mesure de Prendergast [46].
La ligne pointillée (...) représente le coefficient de Bénasséni [2].

Interprétation

D'après la figure, on peut remarquer que certaines observations sont détectées par les anciennes mesures (Bénasséni [2] et Prendergast [46]) comme étant influentes mais elles ne sont pas influentes sur le sous espace engendré par les dominantes composantes principales comme le montre la nouvelle mesure de Prendergast & Li Wai Suen [47]. Par exemple, les mesures de Bénasséni [2] et Prendergast [46] détectent l'observation 419 comme étant la plus influente sur le sous espace, comparativement avec la mesure SCI, cette observation a peu d'influence sur le sous espace par rapport à l'influence des autres observations par exemple l'observation 427. SCI montre que l'observation 427 est la plus influente sur le sous espace, mais les autres mesures suggèrent que son influence est relativement modérée. Même remarque pour l'observation 380.

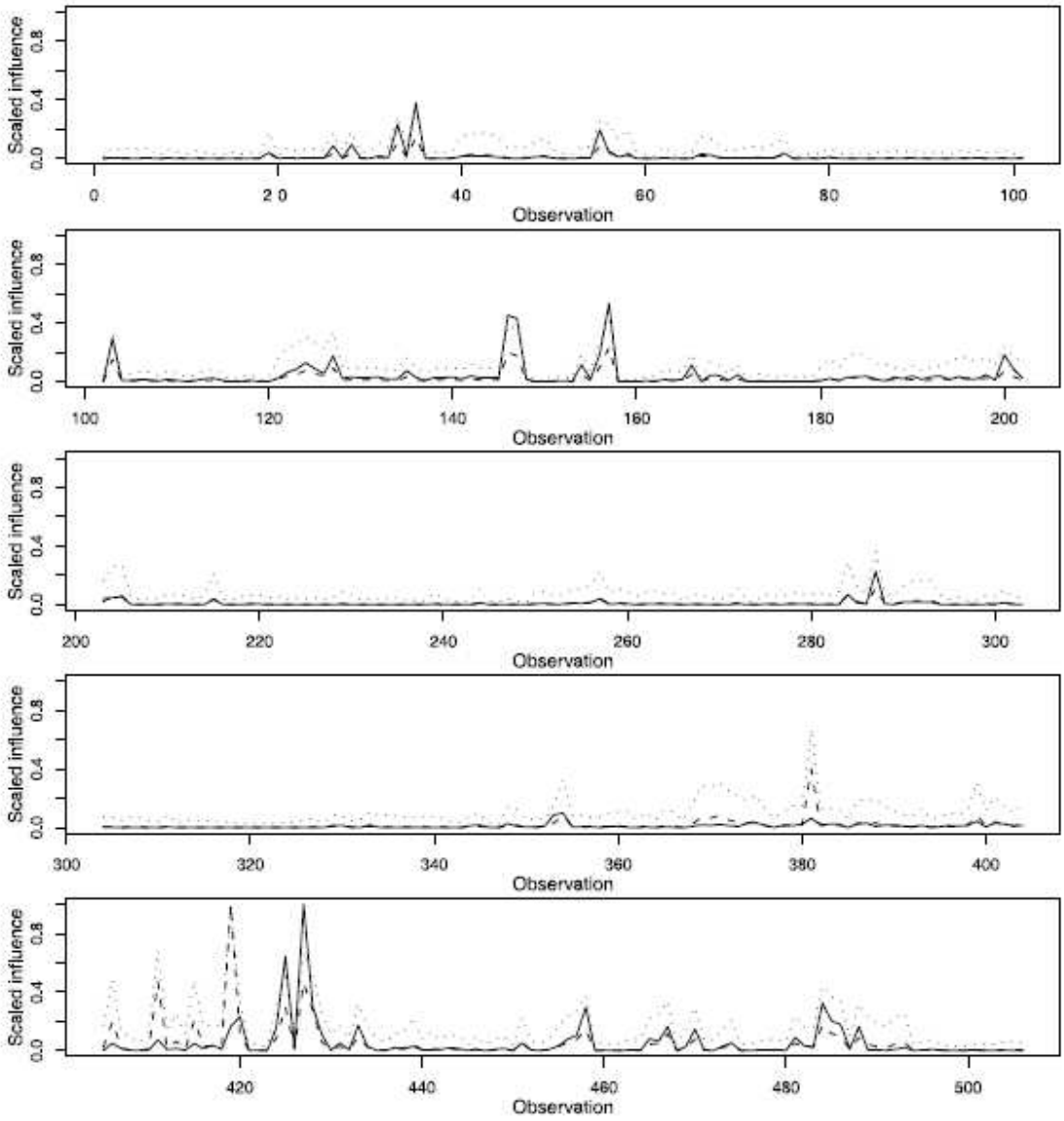

Figure 3.5: Comparaison des mesures, (–) la mesure SCI (Prendergast & Li Wai Suen, 2011), (- -) représente la mesure de Prendergast (2008), (...) la mesure de Bénasséni (1990).

Chapitre 4

Le coefficient de sensibilité en A.C.P: cas robuste

4.1 Introduction

Bénasséni [2] a étudié l'influence du sous espace engendré par les axes principaux d'une A.C.P classique i.e, basée sur l'estimateur classique de la matrice de covariance Σ en calculant la fonction d'influence du coefficient de sensibilité ρ.

Nous allons adopter la même démarche en introduisant un estimateur quelconque C_n de Σ. Ce chapitre, consacré à notre première contribution [11], est constitué de deux étapes. Dans une première étape, nous caractérisons la formule théorique de la fonction d'influence de ρ basée sur C_n, en particulier lorsque C_n est le MCD[1] estimateur. La seconde étape est consacrée à l'étude comparative des fonctions d'influence de ρ dans le cas classique et robuste à travers deux exemples numériques.

4.2 Fonction d'influence de ρ basée sur un estimateur C_n de Σ

Notons par C la fonctionnelle qui correspond à un estimateur C_n de Σ, où $C_n = C(F_n)$, F_n étant la fonction de répartition associée à l'échantillon, et par ρ_C la fonctionnelle associée à un estimateur ρ_{C_n} de ρ.

Pour déterminer la fonction d'influence de ρ_C, on utilisera le lemme suivant qui donne la formule générale de la fonction d'influence associée à un estimateur C_n de la matrice de covariance:

Lemme 4.1 (Croux & Haesbroeck [16]). *Pour toute fonctionnelle C affine équivariante d'une matrice de covariance Σ , il existe deux fonctions α_C, β_C :* $[0,\infty[\rightarrow \mathbb{R}$

telles que:

$$IF(x,C;F) = \alpha_C\{d(x)\}(x-\mu)(x-\mu)' - \beta_C\{d(x)\}\Sigma, \tag{4.1}$$

où $d^2(x) = (x-\mu)'\Sigma^{-1}(x-\mu)$, *et* $F = N_p(\mu,\Sigma)$.

Théorème 4.1 (Cheikh & Ibazizen [11]). *Soit $F = N_p(\mu,\Sigma)$. La fonction d'influence de coefficient de sensibilité ρ basée sur un estimateur C_n de Σ est donnée par:*

$$IF(x,\rho_C;F) = -\frac{|\alpha_C\{d(x)\}|}{q}\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-2}y_j^2y_k^2\Big]^{1/2}\Big\}, \tag{4.2}$$

où $y_j = v_j'(x-\mu)$ *et* $d^2(x) = (x-\mu)'\Sigma^{-1}(x-\mu)$.

Avec λ_j, v_j pour $(j = 1,\cdots,p)$, sont respectivement les valeurs propres et vecteurs propres de Σ et μ représente le vecteur moyen de F.

Démonstration :

Par définition :

$$\rho = \rho_C(\tilde{F}) = 1 - (\frac{1}{q}\sum_{j=1}^{q}\|v_j - \tilde{P}v_j\|). \tag{4.3}$$

D'après (3.38), $\tilde{P}v_j$ s'écrit comme suit:

$$\tilde{P}v_j = v_j + \varepsilon\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-1}(v_j'IF(x;C,F)v_k)v_k, \tag{4.4}$$

où $IF(x;C,F)$ est donnée par (4.1).

Calculons $v_j'IF(x;C,F)v_k$:

$$\begin{aligned} v_j'IF(x;C,F)v_k &= v_j'\Big[\alpha_C\{d(x)\}(x-\mu)(x-\mu)' - \beta_C\{d(x)\}\Sigma\Big]v_k \\ &= \alpha_C\{d(x)\}y_jy_k - v_j'\beta_C\{d(x)\}\Sigma v_k \\ &= \alpha_C\{d(x)\}y_jy_k - \lambda_k\beta_C\{d(x)\}v_j'v_k, \quad \text{car } \Sigma v_k = \lambda_k v_k. \end{aligned}$$

Puisque $k \neq j$, donc $v_j'v_k = 0$, par la suite, on obtient:

$$v_j'IF(x;C,F)v_k = \alpha_C\{d(x)\}y_jy_k. \tag{4.5}$$

En remplaçant (4.5) dans (4.4), $\tilde{P}v_j$ s'écrit sous la forme suivante:

$$\tilde{P}v_j = v_j + \varepsilon\alpha_C\{d(x)\}\sum_{k=q+1}^{p}(\lambda_j-\lambda_j)^{-1}y_jy_kv_k.$$

Calculons $\|v_j - \tilde{P}v_j\|$:

$$\|v_j - \tilde{P}v_j\| = \varepsilon|\alpha_C\{d(x)\}|\Big[\Big(\sum_{k=q+1}^{p}(\lambda_j-\lambda_j)^{-1}y_jy_kv_k\Big)'\Big(\sum_{k=q+1}^{p}(\lambda_j-\lambda_j)^{-1}y_jy_kv_k\Big)\Big]^{1/2}.$$

Après quelques développements, on obtient :

$$\|v_j - \tilde{P}v_j\| = \varepsilon|\alpha_C\{d(x)\}|\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_j)^{-2}y_j^2y_k^2\Big]^{1/2}. \tag{4.6}$$

En insérant (4.6) dans (4.3), ρ s'écrit comme suit:

$$\rho = 1 - \varepsilon\frac{|\alpha_C\{d(x)\}|}{q}\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-2}y_j^2y_k^2\Big]^{1/2}\Big\}.$$

En dérivant par rapport à $\varepsilon = 0$, on obtient:

$$IF(x,\rho_C,F) = -\frac{|\alpha_C\{d(x)\}|}{q}\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-2}y_j^2y_k^2\Big]^{1/2}\Big\}.$$

Remarque 4.1. Si $\alpha_C\{d(x)\} = 1$, on retrouve la formule de Bénasséni [2] correspondante au cas classique.

4.2.1 Cas particulier d'une A.C.P robuste: utilisation du MCD[1]

On note par C^1 la fonctionnelle correspondante au MCD[1] estimateur. Lorsque C_n représente le MCD[1] estimateur, dans ce cas, $IF(x,\rho_{C^1};F)$ est donnée par la proposition suivante :

Proposition 4.1 (Cheikh & Ibazizen, [11]). *La fonction d'influence $IF(x,\rho_{C^1},F)$ est donnée par la formule suivante:*

$$IF(x,\rho_{C^1;F},F) = -\frac{|\alpha_{MCD^1}\{d(x)\}|}{q}\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-2}y_j^2y_k^2\Big]^{1/2}\Big\},$$

où

$$\alpha_{MCD^1}(d(x)) = \frac{1}{d_2} I(d(x)^2 \leq q_\beta) + \frac{d_2 + 2d_3}{d_2} \alpha_{MCD}(d(x)),$$

avec

$$\alpha_{MCD}(d(x)) = \frac{-1}{2c_3} I(d(x)^2 \leq q_\alpha). \tag{4.7}$$

Les constantes d_2, d_3,c_3 sont données par les relations suivantes:

$$c_3 = \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+3} g'(r^2) dr,$$
$$d_2 = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \int_0^{\sqrt{q_\beta}} r^{p+1} g(r^2) dr,$$
$$d_3 = \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\beta}} r^{p+3} g'(r^2) dr,$$

où $\quad q_\beta = \chi^2_{p,(1-\beta)}, \quad q_\alpha = \chi^2_{p,(1-\alpha)}, \quad g(t) = \frac{1}{(2\pi)^{p/2}} e^{-t/2}, \quad \alpha = 0{,}25$ et $\beta = 0{,}025$.

Démonstration:

L'expression de la fonction d'influence du MCD^1 pour la loi elliptique $F = F_{0,I_p}$, de densité $f_{0,I_p}(x) = g(x'x)$, est donnée par Lopuhaä [40] comme suit :

$$IF(x,MCD^1;F) = \frac{d_2 + 2d_3}{d_2}\Big(IF(x,MCD;F) + \frac{1}{2}\mathrm{tr}(IF(x,C^0;F))I_p\Big) + \frac{1}{d_2} I(x'x \leq q_\beta) xx' - I_p. \tag{4.8}$$

En réécrivant (4.8) sous cette forme:

$$IF(x,C^1;F) = \alpha_{C^1}(\|x\|)xx' - \beta_{C^1}(\|x\|)I_p,$$

avec C^0 est la fonctionnelle correspondante au MCD , on déduit que:

$$\alpha_{C^1}(\|x\|) = \frac{1}{d_2} I(\|x\|^2 \leq q_\beta) + \frac{d_2 + 2d_3}{d_2} \alpha_{C^0}(\|x\|),$$

où

$$\alpha_{C^0}(\|x\|) = \frac{-1}{2c_3} I(\|x\|^2 \leq q_\alpha).$$

4.3 Exemples pratiques

4.3.1 Exemple 1: Données de Kendall.

On reprend les données de Kendall (Table 3.1) sur lesquelles on calcule le MCD[1] du vecteur moyen ($\hat{\mu}^1$) et de la matrice de covariance ($\hat{\Sigma}^1$), puis on calcule $EIF(x_i,\rho)$ pour chaque $i = 1, \cdots ,20$.

En utilisant le logiciel S-plus, on retrouve $\hat{\mu}^1$ et $\hat{\Sigma}^1$ comme suit:

$\hat{\mu}^1 = (18.65,10.41,2.5,6.60)'$.

$$\hat{\Sigma}^1 = \begin{pmatrix} 55.11 & 25.77 & 2.63 & -1.65 \\ 25.77 & 16 & -0.03 & -0.75 \\ 2.63 & -0.03 & 0.55 & -0.05 \\ -1.65 & -0.75 & -0.05 & 0.20 \end{pmatrix}.$$

Les tables 4.1 et 4.2 représentent les valeurs de $|EIF(x_i,\rho)|$ pour chaque $i = 1, \cdots ,20$, calculées sur chaque sous espace E_A engendré par un, deux ou trois vecteurs principaux.

$i\backslash A$	{1}	{2}	{3}	{4}	{1,2}	{1,3}	{1,4}
1	0.223	0.364	0.043	0.148	0.09	0.122	0.183
2	0.262	0.329	0.421	0.234	0.08	0.235	0.242
3	0.06	0.187	0.130	0.110	0.05	0.06	0.08
4	0.314	0.318	0.348	0.185	0.07	0.242	0.239
5	0.178	0.440	0.514	0.312	0.113	0.217	0.244
6	0.1001	0.09	0.272	0.136	0.03	0.115	0.115
7	0.253	0.259	0.062	0.027	0.021	0.142	0.140
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0.114	0.126	0.587	0.293	0.04	0.200	0.200
12	0.153	0.636	0.002	0.309	0.165	0.077	0.230
13	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0.155	0.277	0.06	0.119	0.06	0.09	0.136
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0.06	0.0078	0.282	0.143	0.03	0.100	0.0779
20	0.055	0.41	0.521	0.328	0.110	0.158	0.190

Table 4.1: Valeurs de $|EIF(x_i,\rho)|$ pour le cas robuste.

$i\backslash A$	{2,3}	{2,4}	{3,4}	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}
1	0.203	0.151	0.151	0.124	0.017	0.17	0.08
2	0.370	0.372	0.138	0.229	0.164	0.171	0.113
3	0.158	0.105	0.095	0.110	0.047	0.08	0.026
4	0.330	0.368	0.098	0.180	0.140	0.147	0.143
5	0.457	0.371	0.257	0.300	0.213	0.230	0.069
6	0.183	0.198	0.043	0.111	0.109	0.050	0.050
7	0.148	0.152	0.042	0.027	0.032	0.112	0.090
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0.350	0.368	0.06	0.227	0.225	0.070	0.060
12	0.319	0.087	0.309	0.213	0.0009	0.256	0.057
13	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0.170	0.119	0.125	0.104	0.028	0.134	0.058
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0.144	0.173	0.036	0.115	0.098	0.004	0.026
20	0.462	0.306	0.229	0.313	0.192	0.170	0.024

Table 4.2: Valeurs de $|EIF(x_i,\rho)|$ pour le cas robuste.

D'après les résultats des tables 4.1 et 4.2, on remarque que:

1. $|EIF(x_i,\rho)|$ a diminué presque pour toutes les observations par rapport au cas classique, et beaucoup d'observations ont une influence nulle.

2. Il n'y a pas une observation qui est influente par rapport aux autres.

3. Les observations x_{14} et x_4 n'apparaissent plus comme influentes.

La Figure 4.1 représente les graphiques de $|EIF(x_i,\rho)|$ pour le cas classique et robuste, pour les différents choix de E_A.
Dans le graphe (a): $A = \{1\}$, dans le graphe (b): $A = \{3\}$, dans le graphe (c): $A = \{1,2\}$, dans le graphe (d): $A = \{1,2,3\}$.

La ligne discontinue (respectivement continue) représente le cas classique (respectivement le cas robuste).

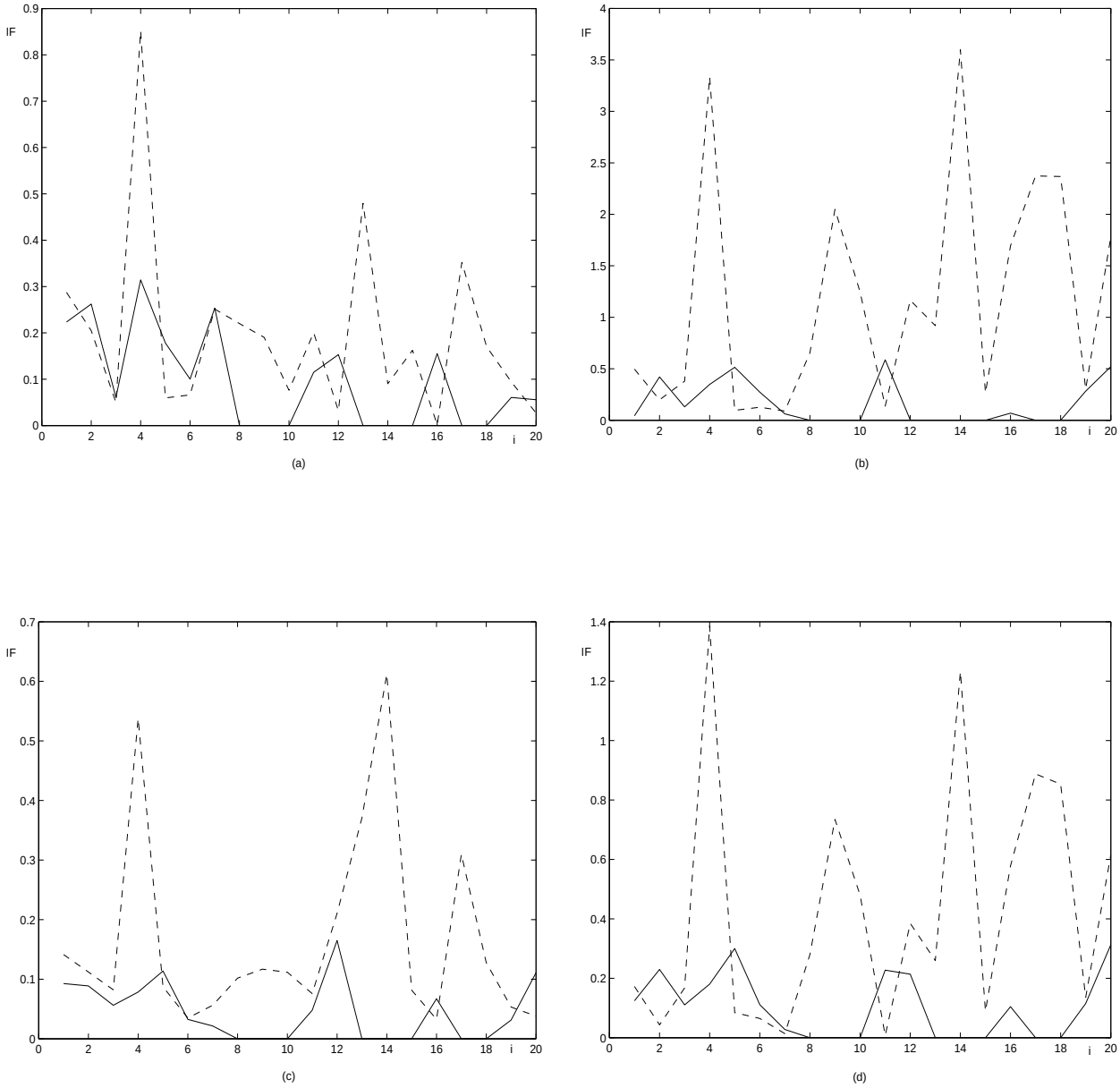

Figure 4.1: Graphiques de $|EIF(x_i,\rho)|$ pour le cas classique (ligne discontinue) et robuste (ligne continue) pour les différents choix de E_A.
(a): $A = \{1\}$, (b): $A = \{3\}$, (c): $A = \{1,2\}$, d): $A = \{1,2,3\}$.

Les graphiques de la Figure 4.1, illustrent la différence entre le cas classique et robuste. En effet, pour le cas classique, on voit que x_4 et x_{14} sont très influentes, contrairement au cas robuste, il n'y a pratiquement aucune observation influente.

4.3.2 Exemple 2: Simulations

On simule un échantillon de 20 individus dans $\mathbb{R}^6$ issu de la loi:

$$F = 0.9N_6(0,\Sigma) + 0.1N_6(\mu,\Sigma).$$

avec $\mu = (5,0,0,0,0,0)'$ et Σ=diag(1,4,4,4,4,4).

Cet échantillon est donné par la table 4.3.

i	X^1	X^2	X^3	X^4	X^5	X^6
1	0.9344	2.4572	-0.4990	-1.4152	-1.1876	-0.5246
2	6.2428	-3.0978	-0.7735	0.5502	1.6523	-1.9586
3	-0.1043	0.2557	0.1251	0.7432	-0.2079	-1.3936
4	-0.3868	0.0323	2.7384	0.8338	0.1375	0.5884
5	0.4726	3.5655	0.5212	3.0215	0.6339	1.6069
6	0.5809	3.5572	-1.8761	-1.8333	0.7520	1.8196
7	5.1542	-0.4045	2.9774	-1.2433	1.6191	3.8576
8	0.3961	-1.7228	4.8639	-1.6810	0.5609	1.6408
9	1.2278	-0.1271	1.2905	-3.5426	0.1192	-1.5204
10	-1.6909	2.2073	2.9250	0.4724	-2.1954	4.8305
11	-0.4021	1.8283	-0.2719	2.6284	0.6447	-0.9530
12	0.0762	-0.2103	2.8340	1.4158	0.7358	-1.2057
13	-0.8521	1.3102	2.9404	-1.6208	-2.5523	3.4445
14	0.1019	-1.6040	-2.5017	2.4754	3.0563	3.5538
15	0.6312	0.1665	4.2801	2.5269	-3.5013	-0.0289
16	2.4681	-1.3383	0.5199	-0.7447	2.6371	-1.3062
17	0.0622	-1.4717	-0.3586	2.1693	0.2783	-0.0312
18	-0.9385	-2.9562	0.7239	0.9556	0.6435	-3.7551
19	0.6805	0.4668	2.4790	0.2513	0.3594	-1.2101
20	-1.0369	-0.5906	2.9122	3.6051	-2.6672	0.7745

Table 4.3: Données issues de $F = 0.9N_6(0,\Sigma) + 0.1N_6(\mu,\Sigma)$ avec $\mu = (5,0,0,0,0,0)'$ et $\Sigma = \text{diag}(1,4,4,4,4,4)$.

Le MCD[1] du vecteur moyen ($\hat{\mu}^1$) et de la matrice de covariance ($\hat{\Sigma}^1$), associés aux données de la Table 4.3, sont respectivement:

$\hat{\mu}^1 = (0.22, 0.58, 1.55, 0.19, -0.003)'$.

$$\hat{\Sigma}^1 = \begin{pmatrix} 1.04 & -0.60 & -0.53 & -0.65 & 0.99 & -1.47 \\ -0.59 & 2.65 & -0.73 & 0.89 & -1.05 & 1.11 \\ -0.53 & -0.73 & 2.84 & -1.05 & -0.47 & 1.32 \\ -0.65 & 0.89 & -1.05 & 3.69 & 0.56 & 0.13 \\ 0.99 & -1.05 & -0.47 & 0.56 & 1.80 & -1.88 \\ -1.47 & 1.11 & 1.32 & 0.12 & -1.88 & 3.94 \end{pmatrix}.$$

La Table 4.4 donne les valeurs de $|EIF(x_i,\rho)|$ pour le cas classique et robuste.

	Cas classique				Cas robuste			
	{1,2}	{1,3}	{1,2,3}	{1,2,4}	{1,2}	{1,3}	{1,2,3}	{1,2,4}
1	1.79	1.954	1.6401	2.177	0.707	0.75	0.325	0.378
2	3.388	3.512	2.05	2.93	0	0	0	0
3	1.824	1.888	0.755	1.48	0.503	0.927	0.267	0.578
4	0.74	0.782	0.247	0.572	0.172	0.342	0.20	0.269
5	2.437	2.509	1.4	2.165	1.085	1.435	0.276	1.685
6	12.853	13.026	3.865	10.298	0	0	0	0
7	15.782	15.829	4.484	12.242	0	0	0	0
8	4.025	4.114	1.597	3.533	2.177	3	0.379	0.714
9	3.86	4.5722	2.932	4.128	2.463	4.06	0.281	0.777
10	1.00	1.702	0.617	0.584	0.265	0.526	0.072	0.113
11	5.264	5.288	0.847	3.96	0.519	1.485	0.138	0.411
12	1.66	1.706	0.39	1.124	0.420	1.167	0.696	1.291
13	3.715	3.943	1.116	2.783	1.09	3.121	0.193	0.415
14	8.718	10.25	5.639	9.877	0	0	0	0
15	7.81	8.305	1.863	5.669	0	0	0	0
16	1.162	1.368	0.353	0.818	0.951	1.78	0.299	1.34
17	2.816	2.899	0.998	2.346	1.08	1.982	0.87	1.893
18	2.85	4.042	1.539	2.101	0	0	0	0
19	1.057	1.096	0.486	0.975	0.315	0.502	0.216	0.763
20	2.486	2.529	1.649	1.109	0	0	0	0

Table 4.4: Valeurs de $|EIF(x_i,\rho)|$ pour le cas classique et robuste.

La Figure 4.2 représente les graphiques de $|EIF(x_i,\rho)|$ pour le cas classique et robuste pour les différents choix de E_A.
Dans le graphe (a): $A = \{1,2\}$, dans le graphe (b): $A = \{1,3\}$, dans le graphe (c): $A = \{1,2,3\}$, dans le graphe (d): $A = \{1,2,4\}$.

La ligne discontinue (respectivement continue) représente le cas classique (respectivement le cas robuste).

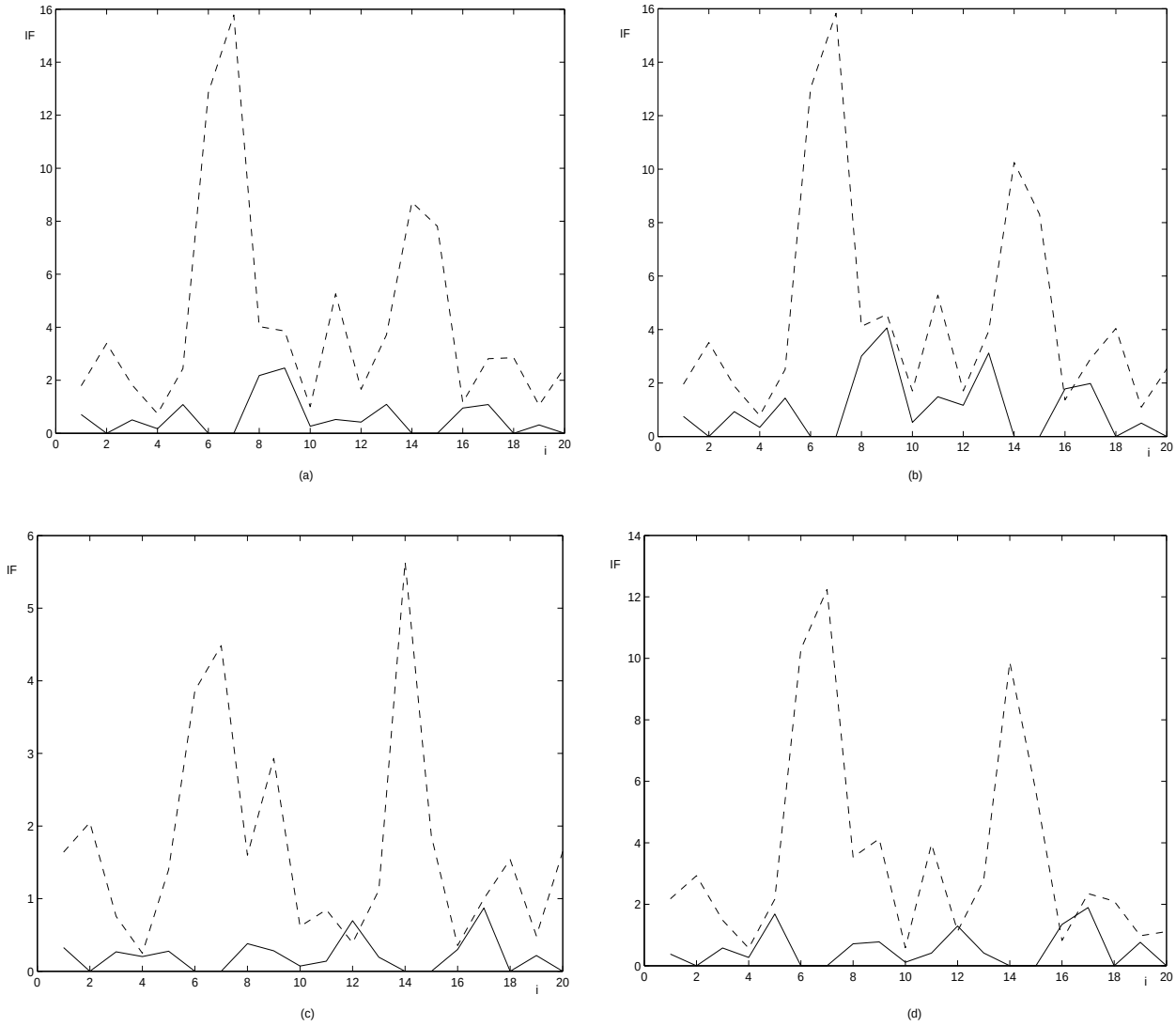

Figure 4.2: Graphiques de $|EIF(x_i,\rho)|$ pour le cas classique (ligne discontinue) et robuste (la ligne continue) pour les différents choix de E_A.
(a): $A = \{1,2\}$, (b): $A = \{1,3\}$, (c): $A = \{1,2,3\}$, d): $A = \{1,2,4\}$.

Pour cet exemple aussi, la différence entre le cas classique et robuste est nette.
En effet, d'après les graphiques de la Figure 4.2, on remarque que:

1. Pour le cas classique, les observations les plus influentes sont x_6 et x_7, sauf pour $A = \{1,2,3\}$ x_{14} est la plus influente.

2. Pour le cas robuste, il n'y a pratiquement aucune observation influente et les valeurs de $|EIF(x_i,\rho)|$ ont diminué par rapport au cas classique.

On peut conclure que:

Le MCD[1] réduit le poids des observations influentes. Donc l'utilisation d'un estimateur robuste en A.C.P. éliminera les observations aberrantes qui conduisent à une fausse analyse.

Chapitre 5

Etude comparative des estimateurs robustes basée sur le coefficient de sensibilité en A.C.P.

5.1 Introduction

La fonction d'influence du coefficient de sensibilité de Bénasséni ρ lorsqu'on utilise le MCD[1] estimateur a été déjà établie par Cheikh & Ibazizen [11] (voir chapitre 4). Nous allons étendre ce travail aux estimateurs MCD et le S-estimateur.

Ce chapitre consacré à notre deuxième contribution (Cheikh [12]), est composé de deux parties. Dans la première partie:

1. Nous caractérisons les fonctions d'influences de ρ lorsque l'ACP est basée sur le MCD estimateur et le S-estimateur.

2. Grâce à ces fonctions, nous établirons une étude comparative de ces estimateurs (au sens de la robustesse).

3. Pour confirmer les résultats de cette étude dans le cas général, on considérera un autre critère de comparaison qui est la sensibilité aux grosses erreurs.

La seconde partie est consacrée à l'étude empirique du MSE des différents estimateurs de ρ basés sur l'estimateur classique et les estimateurs robustes suivants: le MCD, le MCD[1], et le S-estimateur. Cette étude, nous permettra aussi de comparer ces estimateurs .

5.2 Fonction d'influence de ρ: utilisation du MCD et du S-estimateur

Notons par C^0, S les fonctionnelles correspondantes aux MCD, et le S estimateur respectivement.

Proposition 5.1 (Cheikh [12]). *Les fonctions d'influence du coefficient de sensibilité correspondantes aux MCD et au S-estimateur, sont respectivement données par les formules suivantes :*

i)

$$IF(x,\rho_{C^0};F) = -\frac{1}{q}\Big|\frac{-1}{2c_3}I(\|d(x)\|^2 \le q_\alpha)\Big|\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-2}y_j^2y_k^2\Big]^{1/2}\Big\}, \quad (5.1)$$

avec

$$c_3 = \begin{cases} \dfrac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)}\displaystyle\int_0^{\sqrt{q_\alpha}} r^{p+3}g'(r^2)dr & si \quad p \ge 2 \\ 0 & ailleurs, \end{cases} \quad (5.2)$$

et

$$q_\alpha = \chi^2_{p,(1-\alpha)}, \quad g(t) = \frac{1}{(2\pi)^{p/2}}e^{-t/2}.$$

ii)

$$IF(x,\rho_S;F) = -\frac{1}{q}\Big|\frac{p}{\gamma_1}\frac{\psi(\|d(x)\|)}{\|d(x)\|}\Big|\Big\{\sum_{j=1}^{q}\Big[\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-2}y_j^2y_k^2\Big]^{1/2}\Big\}. \quad (5.3)$$

La constante γ_1 est déterminée par la relation:

$$\gamma_1 = \frac{2\pi^{p/2}}{\Gamma(p/2)(p+2)}\int_0^{\infty}[\psi'(r)r^2 + (p+1)\psi(r)r]r^{p-1}g(r^2)dr,$$

où ψ' désigne la dérivée de ψ.

Démonstration:

On démontre d'abord *ii).*

La fonction d'influence du S-estimateur sous le modèle de distribution $F = F_{0,I_p}$, ayant une fonction de densité $f_{0,I_p} = g(x'x)$, est déduite par Lopuhaä [37] comme suit:

$$IF(x;S,F) = \frac{1}{p}\text{tr}[IF(x,S,F)]I_p + \frac{1}{\gamma_1}p\psi(\|x\|)\|x\|\Big(\frac{xx'}{\|x\|^2} - \frac{1}{p}I_p\Big). \tag{5.4}$$

En utilisant le Lemme de Hampel et al [27], la formule(5.4) se réécrit sous cette forme:

$$IF(x,S;F) = \alpha_S(\|x\|)xx' - \beta_S(\|x\|)I_p, \tag{5.5}$$

d'où, on déduit:

$$\alpha_S(\|x\|) = \frac{p}{\gamma_1}\frac{\psi(\|x\|)}{\|x\|},$$

où:

$$\gamma_1 = \frac{2\pi^{p/2}}{\Gamma(p/2)(p+2)}\int_0^\infty [\psi'(r)r^2 + (p+1)\psi(r)r]r^{p-1}g(r^2)dr.$$

Pour *i)*, on utilise le même raisonement que *ii)*. Pour obtenir $\alpha_{C^0}\{d(x)\}$, il suffit juste d'écrire la fonction d'influence $IF(x;C^0,F)$ donnée par Croux & Haesbroeck [15] sous la forme (5.5) comme suit:

$$IF(x,\Sigma_{MCD},F) = \frac{-1}{2c_3}I(\|x\|^2 \le q_\alpha)xx' + \beta_{C^0}(\|x\|)I_p,$$

où :

$$\beta_{C^0}(\|x\|) = -1 + \frac{1}{2}\text{tr}(IF(x,\Sigma_{MCD},F)) + \frac{c_\alpha}{1-\alpha}\frac{q_\alpha}{p}\Big\{(1-\alpha)I(\|x\|^2 \le q_\alpha)$$

$$-\text{tr}(IF(x,\Sigma_{MCD},F))\Big(c_2 + \frac{1-\alpha}{2}\Big)\Big\},$$

ce qui donne:

$$\alpha_{C^0}(\|x\|) = \frac{-1}{2c_3}I(\|x\|^2 \le q_\alpha).$$

5.2.1 Etude Comparative des fonctions d'influence de ρ

Pour illustrer les fonctions d'influence dans le cas classique et robuste, on trace les graphiques de IF sous le modèle $F = N_2\{0,diag(2,1)\}$.

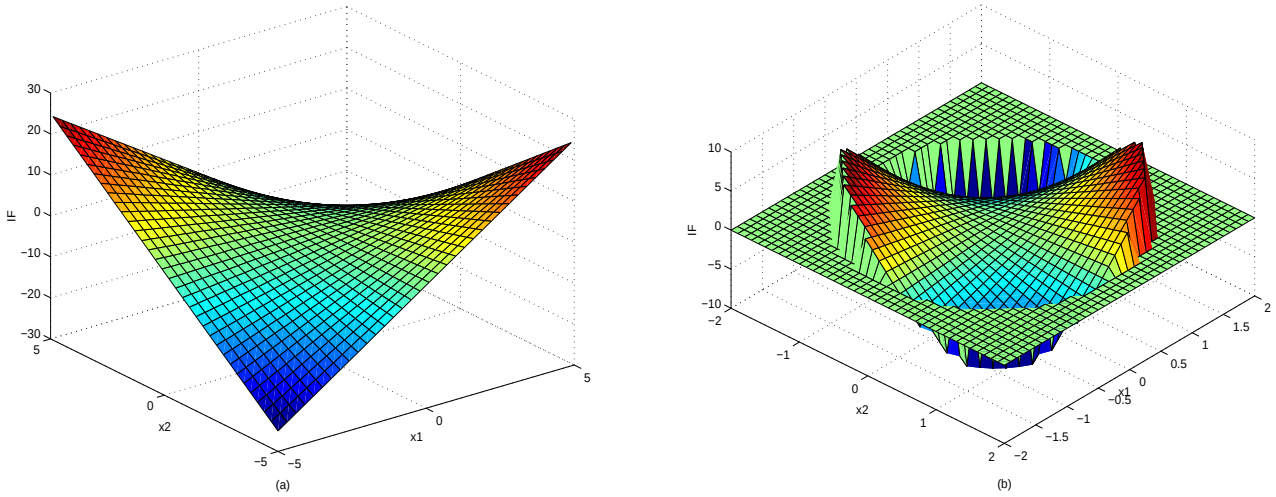

Figure 5.1: Fonction d'influence de ρ pour $p = 2$, et $F = N_2\{0,diag(2,1)\}$
(a): Matrice de covariance classique , (b): MCD estimateur

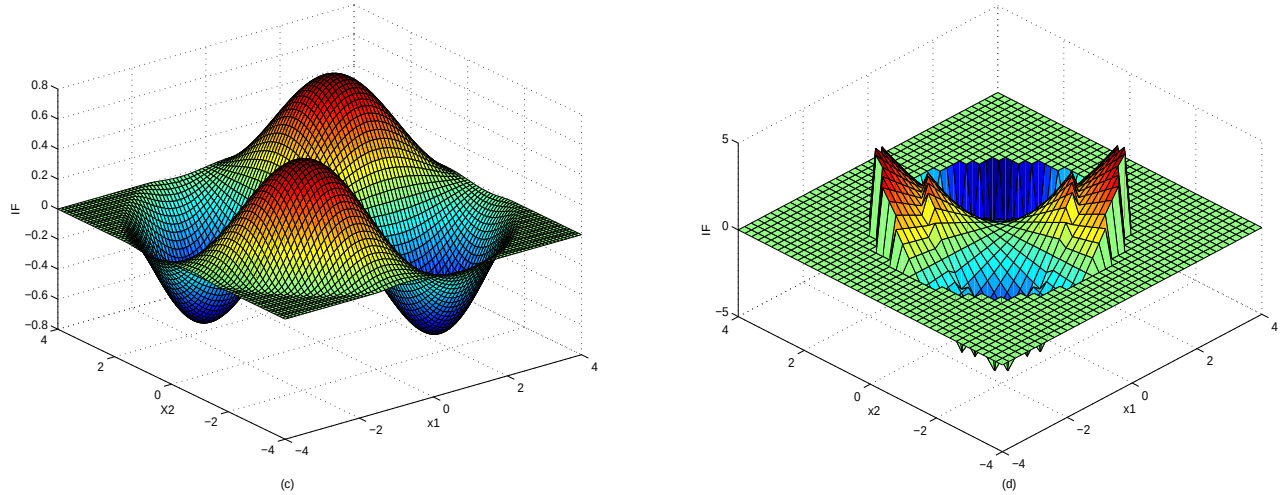

Figure 5.2: fonction d'influence de ρ pour $p = 2$, et $F = N_2\{0,diag(2,1)\}$
(c): S-estimateur , (d): MCD[1] estimateur

Interprétation

C'est intéressant de remarquer que, dans le cas classique, IF n'est pas bornée. En conséquence, un point bien éloigné peut avoir une grande influence sur la matrice de covariance usuelle, par contre, pour les cas robuste IF est bornée, un tel point a une influence très petite.
Le grand avantage du S-estimateur est que sa fonction d'influence est très régulière. C'est ce qui rend cet estimateur le plus attrayant du point de vue de sa fonction d'influence.

Pour confirmer cette conclusion dans le cas général, on calcule la sensibilité aux grosses erreurs (gross error sensitivity) de ρ, notée par GES(ρ_C,F).

5.2.2 Etude comparative de la sensibilité aux grosses erreurs de ρ

On sait que, GES(ρ_C,F) $= \sup_x |IF(x,\rho_C,F)|$. Cette quantité mesure la plus mauvaise influence possible qu'une petite fraction de contamination peut provoquer sur la valeur d'un estimateur.

Les valeurs de GES(ρ_C,F) sont données par la Table 5.1:

p	$p=2$	$p=3$	$p=5$	$p=10$
F / Estimator	$N_2\{0,diag(v_1)\}$	$N_3\{0,diag(v_2)\}$	$N_5\{0,diag(v_3)\}$	$N_{10}\{0,diag(v_4)\}$
Cov	$+\infty$	$+\infty$	$+\infty$	$+\infty$
MCD	11.7620	16.7336	21.2843	23.1950
MCD^1	6.4536	15.5456	16.3250	17.2140
S	3.3159	6.0032	7.0215	8.3210

Table 5.1: Valueurs de GES(ρ_C,F) sous le modèle normal pour $p=2$, $p=3$, $p=5$ et $p=10$.

Où les estimateurs sont:

Cov: l'estimateur classique, MCD: le MCD estimateur, MCD^1: le MCD^1 estimateur, S: le S-estimateurr,

et $v_1 = (1,2)'$, $v_2 = (1,2,3)'$, $v_3 = (1,2,3,4,5)'$, et $v_4 = (1,2,3,4,5,6,7,8,9,10)'$.

Remarques 5.1. D'après la Table 5.1, on remarque que:

1. GES(ρ_{Cov},F) $= +\infty$, on conclut que l'estimateur classique n'est pas robuste.

2. GES(ρ_{MCD^1},F) $<$ GES(ρ_{MCD},F), ceci montre que l'influence des observations extrêmes (outliers) sur le MCD^1 est plus petite que celle exercée sur le MCD ordinaire.

3. GES(ρ_C,F) a la plus petite valeur pour le S-estimateur, ceci confirme la grande robustesse du S-estimateur.

5.3 Etude empirique des estimateurs de ρ

On note par ρ_{C_n} l'estimator de ρ basé sur un estimateur C_n de Σ, et $\tilde{C_n}$ est l'estimateur de $\tilde{\Sigma}$.
Rappelons la formule de ρ_{C_n} :

$$\rho_{C_n} = 1 - \Big(\sum_{j=1}^{q} \| v_j(C_n) - \tilde{P}(\tilde{C_n}) v_j(C_n) \| / q \Big), \tag{5.6}$$

avec:

$\tilde{P}(\tilde{C_n}) = \sum_{j=1}^{q} \tilde{v}_j(\tilde{C_n}) \tilde{v}'_j(\tilde{C_n})$, et les vecteurs $v_j(C_n)$ (resp. $\tilde{v}_j(\tilde{C_n})$), $j = 1, \cdots, q$ sont les vecteurs propres de C_n (resp. $\tilde{C_n}$).

Dans cette section, on effectuera des simulations pour comparer les performances de certains estimateurs robustes ρ_{C_n} pour estimer le coefficient de sensibilité ρ. Nous simulons $m = 1000$ échantillons de taille n=20, 50, 100, 200 issus de 2 différentes distributions avec $\rho = 0.8$ (simulations d'autres valeurs de ρ donnent des même conclusions).

Pour calculer le MCD estimateur on utilise le "FAST-MCD algorithm" de Rousseeuw & Van Driessen [51], quant au S-estimateur on applique le "SURREAL algorithm" de Ruppert [53].

Pour évaluer la performance des estimateurs de ρ, on calcule l'erreur quadratique moyenne empirique. Pour chaque échantillon j, le coefficient de sensibilité est estimé par $\hat{\rho_j}$, et l'erreur quadratique moyenne (MSE) est calculée comme suit:

$$MSE(\rho_{C_n}) = \frac{1}{m} \sum_{j=1}^{m} (\hat{\rho_j} - \rho)^2. \tag{5.7}$$

Les distributions qu'on a utilisé sont données comme suit:

1. La loi normale $N_6(0,\Sigma)$.

2. La loi normale contaminée:

$$(1-\epsilon)N_6(0,\Sigma)+\epsilon N_6(\mu,\Sigma), \tag{5.8}$$

où

$$\Sigma = \text{diag}(1,4,4,4,4,4),\ \ \mu=(5,0,0,0,0,0)',\ \ \epsilon\in[0,1].$$

Les valeurs de $MSE(\rho_{C_n})$ pour les différents estimateurs sont reportées dans la Table 5.2 et Table 5.3 .

Estimateur	$n=20$	$n=50$	$n=100$	$n=200$
Cov	0.1019	0.1008	0.0510	0.0467
MCD	0.1051	0.1045	0.0960	0.0951
MCD^1	0.1309	0.1245	0.1179	0.1046
S	0.1040	0.1030	0.0620	0.0540

Table 5.2: Valeurs du MSE des différents estimateurs de ρ sous la distribution normale $N_6(0,\Sigma)$ pour les tailles: $n=20$, 50 , 100 et 200.

n	Estimateur	$\epsilon=0\%$	$\epsilon=1\%$	$\epsilon=5\%$	$\epsilon=10\%$
50	Cov	0.1006	0.2056	0.6383	0.6400
	MCD	0.1044	0.1045	0.1350	0.1525
	MCD^1	0.1242	0.1240	0.1239	0.1430
	S	0.1032	0.1030	0.1029	0.1027
100	Cov	0.0510	0.1153	0.1389	0.6177
	MCD	0.0962	0.1960	0.2090	0.3652
	MCD^1	0.1176	0.1170	0.1165	0.1260
	S	0.0620	0.0620	0.0619	0.0617
200	Cov	0.0467	0.1478	0.1950	0.2031
	MCD	0.0951	0.0955	0.1400	0.1894
	MCD^1	0.1044	0.1040	0.1039	0.1222
	S	0.0540	0.0540	0.0539	0.0538

Table 5.3: Valeurs du MSE des différents estimateurs de ρ sous la distribution normale symétrique contaminée : $(1-\epsilon)N_6(0,\Sigma)+\epsilon N_6(\mu,\Sigma)$, pour les tailles $n=20$, 50, 100 et 200.

5.3.1 Remarques et conclusion

Pour la loi normale noncontaminée (voir Table 5.2), le MSE a la plus petite valeur pour l'estimateur classique du coefficient de sensibilité. Cela n'est plus le cas en présence des valeurs aberrantes (outliers), comme on le voit dans la Table 5.3, et déjà pour 1% d'outliers,

le MSE de l'estimateur classique est le plus grand de tous les estimateurs considérés, ceci confirme la non robustesse de l'estimateur classique. Pour 1% de contamination, le MSE du MCD et du MCD^1 restent stables, mais pour une grandes quantité de contamination une augmentation légère de MSE est observée pour ces deux estimateurs. Pour $\epsilon = 5\%$, le MCD^1 est plus performant que le MCD.Finalement, on constate la grande robustesse du S-estimateur, où le MSE reste faible même pour 10% de contamination.

On conclut que l'estimateur de coefficient de sensibilité associé à un estimateur très robuste de la matrice de covariance est le plus résistant en présence d'une grande quantité d'outliers.

Conclusion générale et perspectives de recherche

Par ce travail, nous estimons avoir contribué à l'étude de la fonction d'influence du coefficient de sensibilité de Bénasséni ρ en A.C.P. robuste. Nous avons caractérisé cette fonction lorsqu'on utilise les estimateurs robustes de la matrice de covariance le MCD, le MCD[1] et le S-estimateur. Grâce à ces fonctions, nous avons pu établir une comparaison (au sens de la robustesse) entre ces estimateurs, l'étude a montré la grande robustesse du S-estimateur.

Pour confirmer ce résultat obtenu grâce aux fonctions d'influence dans le cas général, nous avons calculé la sensibilité aux grosses erreurs de ρ, et nous avons obtenu les valeurs les plus petites pour le S-estimateur.

Dans le but de connaître le comportement de ces estimateurs en présence des valeurs aberrantes, nous avons effectué une étude empirique en calculant le MSE des estimateurs de ρ, pour cela nous avons utilisé deux lois, la loi multinormale et la loi multinormale contaminée, et nous avons considéré plusieurs pourcentage de contamination.

Les résultats de simulation ont montré que les valeurs du MSE lorsqu'on utilise le S-estimateur sont les plus petites et aussi restent stable même si le pourcentage de contamination est grand. Donc, nous avons conclu que le S-estimateur est le plus résistant en présence de plusieurs valeurs aberrantes, donc le plus robuste.

Comme perspectives de recherche, nous proposons:

- Étendre ce travail à d'autres méthodes statistiques multivariées (l'analyse des correspondance, la régression multiple,...).

- Étendre les travaux de Prendergast (2008) et de Prendergast & Li Wai Suen (2011) à d'autres estimateurs, par exemple, le S-estimateur.

Annexe 1

Cette partie est consacrée à quelques rappels d'algèbre bilinéaires. Pour plus de détails, le lecteur peut se référer aux ouvrages suivants: Saporta [54](Annexe 5, page 480) et Doneddu [22].

A1.1: Produit scalaire, norme

A1.1.1 Produit scalaire

Définition 1. (Produit scalaire). On appelle produit scalaire sur E toute application φ de $E \times E$ dans $\mathbb{R}$ ayant les propriétés suivantes :
1. $\forall(x_1,x_2,y) \in E^3$, $\varphi\ (x_1 + x_2,y) = \varphi\ (x_1,y)+\varphi\ (x_2,y)$.
2. $\forall\alpha \in \mathbb{R},\forall(x,y) \in E^2$, $\varphi\ (\alpha x,y) = \alpha\ \varphi\ (x,y)$.
3. $\forall(x,y_1,y_2) \in E^3$, $\varphi\ (x,y_1 + y_2)= \varphi\ (x,y_1) +\varphi\ (x,y_2)$.
4. $\forall(x,y) \in E^2$, $\varphi\ (x,y) = \varphi\ (y,x)$, ($\varphi$ est symétrique).
5. $\forall x \in E$, $\varphi\ (x,x) = 0 \Longleftrightarrow x = 0$, ($\varphi$ est définie).
6. $\forall x \in E$, $\varphi\ (x,x) \geq 0$ (φ est positive).

L'application φ est appelée *forme bilinéaire symétrique définie positive.*

Définition 2. (Espace euclidien). Lorsque E est muni d'un produit scalaire, on dit que E est *un espace euclidien.*

Notation: on trouvera dans la littérature les différentes notations suivantes pour le produit scalaire :

$$\varphi(x,y) = x.y = \prec x,y \succ .$$

Exemple 1 :

Si $E = \mathbb{R}^n$, on a :

$$\prec x,y \succ = \sum_{i=1}^{n} x_i y_i = x'y = y'x,$$

où, $x = \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_n \end{pmatrix}$ et $y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}$.

Remarque 1. Ce produit scalaire est appelé *produit scalaire canonique* .

Théorème 1. φ est un produit scalaire sur $\mathbb{R}^n$ si et seulement si il existe une matrice M carrée d'ordre n symétrique définie positive (i.e. qui admet n valeurs propres réelles strictement positives), telle que :

$$\varphi(x,y) = x'My.$$

Remarque 2. La matrice associée au produit scalaire de l'Exemple 1 est $M = I_n$.

Pour toute la suite, M est la matrice associée au produit scalaire.

A1.1.2 Norme

Définition 3.(Norme). Une *norme* sur E est une application: $\|.\|$: $\mathbb{E} \to \mathbb{R}_+$, qui vérifie:

1. $\forall\lambda \in \mathbb{R}$,$\forall x \in E$ $\|\lambda x\| = \mid \lambda \mid \times \|x\|$.
2. $\forall(x,y) \in E^2$, $\|x + y\| \leq \|x\|+\|y\|$, (inégalité triangulaire).
3. $\forall x \in E$ $\|x\|=0 \Leftrightarrow x = 0_{\mathbb{E}}$.

Définition 4. (Distance ou métrique). Soient x, $y \in E$. *Une distance* ou *une métrique* sur E est une application d: $E \times E \to \mathbb{R}_+$, qui vérifie:

1. $d\ (x,y) = d\ (y,x)$.
2. $d\ (x,y) \leq d\ (x,z)+d\ (z,y)$.
3. $d\ (x,y) = 0 \Longleftrightarrow x = y$.

(E,d) est appelé *espace métrique.*

Définition 5. (Norme associée à un produit scalaire). Soit E un espace euclidien , $x \in E$.

l'application: $\|.\|_M : E \to \mathbb{R}^+$
$x \to \|x\|_M$,

où $\|x\|_M = \sqrt{\prec x,x \succ}_M = \sqrt{x'Mx}$ est la norme associé au produit scalaire $\prec , \succ_M$.

A1.2: Projecteurs, Matrice M-symétrique

Définition 6. (M-orthogonalité). Soit $x \in E$, on dit que x est M-orthogonal à y (on note $x \perp_M y$) si:

$$\prec x,y \succ_M = x^{'} My = 0,$$

où $x^{'}$ est le transposé de x.

Définition 7. (projecteur M-orthogonal). Soit $x \in E$, E muni d'un produit scalaire, et W_1 un sous espace de E.

l'application P: $E \to E$
$x \to P_x = \hat{x}$,
est un projecteur M-orthogonal sur W_1 si :

1. $\forall x \in E$, $P_x \in W_1$.
2. $P_x \perp_M (x - P_x)$, i.e. $\prec P_x,\ (x - P_x) \succ_M = 0$.

Proposition 1. (Écriture explicite d'un projecteur)

Supposons que W_1 est engendré par p ($p \leq n$) vecteurs linéairement indépendants x_1,x_2,$\cdots$,x_p, et soit X la matrice (n,p) ayant les x_i pour vecteurs colonnes.
Le projecteur M-orthogonal sur W_1 s'écrit comme suit:

$$P = X(X'MX)^{-1}X'M.$$

En particulier, le projecteur M-orthogonale sur un vecteur x s'écrit:

$$P = x(x'Mx)^{-1}x'M = \frac{xx'M}{x'Mx},$$

car $x'Mx$ est un scalaire.

Définition 7. (Matrice M-Symétrique)

Soit E un espace euclidien de dimension fini, et A une matrice carrée d'ordre n. A est dite *M-Symétrique* si $\forall x,\ y \in E$:

$$\prec Ax,y \succ_M = \prec x,Ay \succ_M \Longleftrightarrow A'M = MA.$$

Remarque 2. Si $M = I_n$, on retrouve la définition d'une matrice symétrique.

Propriétés des matrices M-Symétriques

Soit A une matrice carrée d'ordre n, M-Symétrique, alors:

1. les valeurs propres de A sont réelles.

2. Deux vecteurs propres de A associés à deux valeurs propres distinctes sont M-orthogonaux.

3. L'ordre de multiplicité d'une valeur propre λ_i de A est la dimension du sous espace propre E_{λ_i} associé.

Conséquence 1. Toute matrice M-symétrique est diagonalisable.

Proposition 2.

Soit E muni d'un produit scalaire, et A est une matrice symétrique carrée d'ordre n. On note par $v_1 v_2, \cdots, v_n$ les vecteurs propres de A, M-orthonomés.

Alors le problème:

Max $\prec Ax, x \succ_M$ sur $x \in E$, tel que $\|x\|_M$=1, a pour solution x=v_1, et ce maximum vaut λ_1.

Soit F_i le sous espace vectoriel de E , engendré par $\{v_1, v_2, ..., v_i\}$ $(i \leq n)$, alors le problème :

Max $\prec Ax, x \succ_M$ sur $x \in F_i^{\perp}$, a pour solution x=v_{i+1}, et ce maximum vaut λ_{i+1}.

Annexes 2

Cette partie est consacrée à quelques définitions utilisées dans le chapitre 2.

A2.1: Distribution elliptique

Soit $X = (X_1, \cdots, X_p)$ un vecteur aléatoire dans $\mathbb{R}^p$ ayant une moyenne μ et une matrice de covariance Σ.

Définition 1. On dit que la distribution de X est elliptique si sa fonction de densité est de la forme:

$$f(x) = |\Sigma|^{-1/2} g_p\Big((x-\mu)'\Sigma^{-1}(x-\mu)\Big), \tag{1}$$

où $\mu \in \mathbb{R}^p$, $\Sigma \in SDP(p)$. La fonction g_p est appelé "générateur de densité".

Soit $\lambda_1 > , \cdots > \lambda_p$ les valeurs propres de Σ. Notons par $h_1, \cdots, h_p$ les vecteurs propres correspondants. Les courbes d'isodensité sont:

$$|A^{-1/2} H'(x-\mu)| = c, \tag{2}$$

où $H = (h_1, \cdots, h_p)$, $A^{1/2}$=diag$(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$ et c est une constante positive.

L'égalité (2) représente l'équation d'une ellipsoide centrée à μ, dont les axes sont supportés par les vecteurs propres h_i, et dont la longueur des demi axes est proportionnelle à $\sqrt{\lambda_i}$.

Exemples

1. La loi multinormale $N_p(\mu,\Sigma)$ a pour densité :

$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\Big(\frac{-1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big)$$

   $f(x)$ est de la forme (1) avec :

$$g_p(t) = (2\pi)^{-p/2} \exp(\frac{-t}{2}).$$

2. La loi de Student multivariée à ν degrés de liberté $t_{p,\nu}(\mu,\Sigma)$ a pour densité :

$$f_\nu(x) = \frac{\Gamma[\frac{1}{2}(\nu+p)]}{(\nu\pi)^{p/2}\Gamma[\frac{1}{2}\nu]}|\Sigma|^{-1/2}\Big[1+\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{\nu}\Big]^{-(\nu+p)/2}.$$

$f_\nu(x)$ est de la forme (1) avec :

$$g_\nu(t) = \frac{\Gamma[\frac{1}{2}(\nu+p)]}{(\nu\pi)^{p/2}\Gamma[\frac{1}{2}\nu]}\Big(1+\frac{t}{\nu}\Big)^{-(\nu+p)/2},$$

où $\Gamma(a) = \int_0^{+\infty} x^{a-1}exp(-x)dx.$

Cas particulier: Si $\nu = 1$, on retrouve la densité de la loi de Cauchy multivariée :

$$f_1(x) = \frac{\Gamma[\frac{1}{2}(p+1)]}{(\pi)^{p/2}\sqrt{\pi}}|\Sigma|^{-1/2}\Big[1+(x-\mu)'\Sigma^{-1}(x-\mu)\Big]^{-(p+1)/2}.$$

Dans ce cas:

$$g_1(t) = \frac{\Gamma[\frac{1}{2}(p+1)]}{(\pi)^{p/2}\sqrt{\pi}}\Big(1+t\Big)^{-(p+1)/2}.$$

A2.2: Lemme de Lopuhaä [40]

Soit $g : [0,\infty[\to \mathbb{R}$ et $x = (x_1, \cdots, x_p)'$. Alors:

$$\int g(x'x)dx = \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^{+\infty} g(r^2)r^{p-1}dr.$$

$$\int g(x'x)x_i^2dx = \frac{1}{p}\int g(x'x)(x'x)dx.$$

$$\int g(x'x)x_i^2x_j^2dx = \frac{1+2\delta_{ij}}{p(p+2)}\int g(x'x)(x'x)^2dx.$$

où $\delta_{ij} = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{sinon,} \end{cases}$

et $\dfrac{2\pi^{p/2}}{\Gamma(p/2)}$ représente la surface de la sphère unité.

A2.3: Estimateur Fisher-consistant

On rappelle qu'un estimateur T_n de θ est convergent si: $T_n \xrightarrow[n\to\infty]{\mathcal{P}} \theta$.

En particulier si $E(T_n) \to \theta \quad$ et $V(T_n) \to 0$, alors: $T_n \xrightarrow{\mathcal{P}} \theta$.

Fisher, en 1922 [24], a introduit un concept de convergence plus fort que la convergence en probabilité, appelé "consistance".

Définition 2.

Soit F_n la fonction de répartition empirique associée à l'échantillon $(X_1, \cdots, X_n)$ de loi P_θ, de fonction de répartition F. Soit T une statistique définie par : $T(X_1, \cdots, X_n) = T(F_n)$.

T est dite consistante pour un paramètre θ si:

$$T(F) \text{ est exactement égal à } \quad \theta, \quad \text{i.e,} \quad \theta = T(F).$$

Remarque:

La définition d'un estimateur consistant au sens de Fisher est plus forte que celle d'un estimateur convergent, qui n'est qu'une propriété asymptotique. Elle exprime le fait qu'une statistique, lorsqu'elle est calculée sur toute la population, est égale au paramètre à estimer.

A1.4 Estimateur affine équivariant

Définition 3.

Soient μ_n et V_n les estimateurs du vecteur moyen et de la matrice de covariance

respectivement, μ_n et V_n sont dit affines équivariants si:

$$\mu_n(AX^1 + b, AX^2 + b, \cdots, AX^p + b) = A\mu_n(X^1, X^2, \cdots, X^p) + b,$$

$$V_n(AX^1 + b, AX^2 + b, \cdots, AX^p + b) = AV_n(X^1, X^2, \cdots, X^p)A',$$

pour tout $b \in \mathbb{R}^p$, et pour toute matrice A $(p \times p)$ non singulière.

Bibliographie

[1] Anderson, T. W., *Asymptotic theory for principal component analysis*, The Annals of Mathematical Statistics, vol. 34, No. 1, 122-148, (1963).

[2] Bénasséni, J., *Sensitivity coefficients for the subspaces spanned by principal components*, Commun. Statist. Theory Meth., 19(6), 2021-2034, (1990).

[3] Benzécri, J.P., *Leçons sur l'analyse factorielle et la reconnaissance des formes*, Faculté des Sciences de Rennes, (1965-66).

[4] Benzécri, J.P., *Histoire et Préhistoire de l'Analyse des données : Partie 2. Les Cahiers de l'analyse des données*, vol. 1, no 2, 101-120, (1976).

[5] Benzécri, J.P., *Histoire et Préhistoire de l'Analyse des données : Partie 4. Les Cahiers de l'analyse des données*, vol. 1, no 4, 343-366, (1976).

[6] Benzécri, J.P., *Histoire et Préhistoire de l'Analyse des données : Partie 5. Les Cahiers de l'analyse des données*, vol. 2, no 1, 1977, 9-40, (1977).

[7] Box, G. E. P., *Non-normality and tests on variances*, Biometrika, 40, 318-335, (1953).

[8] Bulter, R. W, Davies, P. L. and Jhun, M. (BDJ), *Asymptotics for the minimum covariance determinant estimator*, Ann. Statist. 21, 1385-1400, (1993).

[9] Castano-Tostado, E.and Tanaka, Y., *Some comments on Escoufier's RV-coefficient as a sensitivity measure in principal component analysis*, Comm. Statist. Theory Methods, 19(12), 4619-4626, (1990).

[10] Cheikh, M., *Influence en statistique multidimensionnelle*, Mémoire de Magister, UMMTO, (2007).

[11] Cheikh, M. and Ibazizen, M., Sensitivity coefficient in principal component analysis: robust case, Communication in statistics-Simulation and Computation, 37(8), 1622-1630, (2008).

[12] Cheikh, M., *Comparative study of robust estimators based on a sensitivity coefficient in principal component analysis*, Communication in statistics-Simulation and Computation, DOI: 10.1080/03610918.2012.762390, (2013).

[13] Critchley, F. *Influence in principal components analysis*, Biometrika, 72, 627-636, (1985).

[14] Critchley, F. and al, *Influence functions of two families of robust estimators under proportional scatter matrices*, Stat. Meth. Appl. 15, 295-327, (2007).

[15] Croux, C. and Haesbroeck, G., *Influence function and efficiency of the minimum covariance determinant scatter matrix estimator*,J. Mult. Analysis, 71, 161-190, (1999).

[16] Croux, C. and Haesbroeck, G., *Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies*, Biometrica, 87, 3, 603-618, (2000).

[17] Croux, C. and Joossens, K., *Influence of observations on the misclassification probability in quadratic discriminant analysis*, Journal of Multivariate Analysis, 6, 384-403, (2005).

[18] Croux, C. and Ruiz-Gazen, A., *High breakdown estimators for principal components: the projection-pursuit approach revisited*, J. multivariate. Anal, 95, 206-226, (2005).

[19] Croux, C., Haesbroeck, G. and Joossens, K., *Logistic discrimination using robust estimators: an influence function approach*, The Canadian Journal of Statistics, Vol. 36, No 1, 157-174, (2008).

[20] Davies, P. L., *Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices*, Ann. Statist., 15, 1269-1292, (1987).

[21] Devlin, S. J., Gnanadesikan, R. and Kettanring, J. R., *Robust estimation of dispersion matrices and principal components*, J. Am. Statist. Assoc. 76, 354-362, (1981).

[22] Doneddu, M., *Algèbre et Géométerie*, Librairie Vuibert, Paris, (1986).

[23] Escoufier, Y, *Le traitement des variables vectorielles*, Biometrics, 29, 751-760, (1986).

[24] Fisher, R. A., *On the mathematical foundations of theoretical statistics*, Philos. Trans. Roy. Soc. London Ser., A 222, 309-368, (1922).

[25] Hampel, F. R., *Contributions to the theory of robust estimation*, Ph.D. thesis. Univ. California, Berkeley, (1968).

[26] Hampel, F., *The influence curve and its role in the robust estimation*, Journal of American Statistical Association, vol. 69, No. 341, 383-393, (1974).

[27] Hampel, F. R., Ronchetti, E.M., Rousseeuw, P. J. and Stahel, W.A., *Robust statistics: The approach based on influence functions*, Wiley, New York, (1986).

[28] Harrison, D., Rubinfeld, D.L., *Hedonic housing prices and the demand for clean air*, J. Environ. Econ. Manage., 5, 81-102, (1978).

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

[30] Huber, P. J. *Robust Statistics*, A review, Ann. Math. Statist, 43, 1041-1067, (1972).

[31] Huber, P. J., *Robust Statistics*, Wiley, New-York, (1981).

[32] Ibazizen, M. H., *Contribution à l'étude d'une analyse en composantes principales robuste*, Thèse de 3^{eme} cycle, Laboratoire de Probabilités et Statistique, U.P.S., Toulouse, (1986).

[33] Jollife, I.T , Morgan, B.J.T., *Influence observations in principal component analysis: a case-study*, J. Appl. Statist., 15, 37-50, (1988).

[34] Kamiya, H., *A class of robust principal component vectors*, Journal of Multivariate Analysis, 77, 239-269, (2001).

[35] Kendall, M., *Multivariate Analysis*, Charles Griffin, (1975).

[36] Lecoutre, J. P. et Tassi, P., *Statistique non paramétrique et robustesse*. Economica, Paris, (1987).

[37] Lopuhaä , H. P., *On the relation between S-estimators and M-estimators of multivariate location and covariance*, Ann. Statist., 17, 1662-1683, (1989).

[38] Lopuhaä , H. P. and Rousseeuw, *Properties of affine equivariant estimators of multivariate location and covariance matrices*, Ann. Statist. 19, 229-248, (1991).

[39] Lopuhaä , H. P., *Heighly efficient estimators of multivariate location with heigh breakdown point*, Ann. Statist. 19, 229-248, (1992).

[40] Lopuhaä , H. P., *Asymptotic expansion of S-estimators of location and covariance*, Stat. Neerl. 51, 220–237, (1997).

[41] Ma, Y. and Genton, M., *Highly robust estimation of dispersion matrices*, J. Multivariate Anal. 78(1), 11–36, (2001).

[42] Maronna, R. A., *Robust M-estimators of multivariate location and scatter*, Ann. Statist., 4(1), 51-67, (1976).

[43] Otto, S., R., and Denier, J., P., *An introduction to programming and numerical methods in matlab*, Springer, (2005).

[44] Pearson, K., *On lines and planes of closest fit to systems of points in space*, Philosophical Magazine , 2, 559-572, (1901).

[45] Pierre.S. de Laplace, *Théorie analytique des probabilités: Supplément*, Volumes 1 à 2, Courcier, (1818).

[46] Prendergast, L. A., *A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions*, Electron. J. Stat., 2, 454– 467, (2008).

[47] Prendergast, L. A., Li Wai Suen, C., *New and practical influence measure for subsets of covariance matrix sample principal components with applications to high dimensional datasets*, Comput. Statist. Data Anal. 55(1), 752–764, (2011).

[48] Radhakrishnan, R. and Kshirsagar, A.M., *Influence functions for certain paramameters in multivariate analysis*, Comm. statist. A, 10, 515-529, (1981).

[49] Rellich, F., *Perturbation theory of eigenvalue problems*, Gordon and Breach, (1969).

[50] Rousseeuw, P. J., *Multivariate estimation with high breakdown point*, In Mathematical Statistics and Applications, vol. B, Ed. W. Grossman, G. Pflug, I. Vincze and W. Wertz, 283-297, (1985).

[51] Rousseeuw, P. J. and Van Driessen, K., *A fast algorithm for the minimum covariance determinant estimator*, Technometrics , 41, 212-223, (1999).

[52] Rousseeuw, P. J. and Huber, M., *High-Breakdown Robust Multivariate Methods*, Statistical Science, 1, 92-119, (2008).

[53] Ruppert, D., *Computing S-estimators for regression and multivariate location /dispersion*, J. Compt. Graph. Statist., 1, 253-270, (1992).

[54] Saporta, G., *Probabilités, analyse des données et statistique*. Editions Technip, Paris, (1990).

[55] Shi, L., *Local influence in principal components analysis*, Biometrica, 84, 1, pp. 175-186, (1997).

[56] Sibson, R., *Studies in the robustness of multidimensional Scaling: Perturbation analysis of classical scaling*, J.R. Statist. Soc., B 41, 217-229, (1979).

[57] Tanaka, Y., *Sensitivity analysis in PCA: Influence on the subspace spanned by principal components*, Comm. Statist. Theory-Methods., 17, 3157-3175, (1988).

[58] Tukey, J. W., *A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics* I. Olkin, ed. Stanford Univ. Press, Stanford, CA, (1960).

[59] Tyler,D. E., *Breakdown properties of the M-estimators of multivariate scatter*, Report, Dept. Statistics, Rutgers Univ, (1986).

[60] Yanai, H., *Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (G.C.D.)*, Behaviour metrics, 1, 45-54, (1974).