RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE.
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE.

Université Mouloud Mammeri, Tizi-Ouzou
Faculté des Sciences
Département de Mathématiques

Mémoire de Master en Mathématiques Appliquées

Option : *Processus Aléatoires et Statistique de la Décision*

Thème :

Synthèse des Méthodes sur les Données Aberrantes

Présenté par :

NAIT SAID Henia

Devant le Jury d'Examen composé de :

M^r BOUDIBA Mohand Arezki	Maître de conférence A	Président
M^r MEHIRI Mohamed	Chargé de recherches	Rapporteur
M^r BERKOUN Youcef	Maître de conférence A	Examinateur
M^{elle} ATIL Lynda	Maître de conférence B	Examinatrice

Soutenu le 10 / 10 / 2012

Remerciements

Je tiens à exprimer toute ma reconnaissance à Mr MEHIRI Mohamed qui a proposé et accepté de diriger ce travail. Je le remercie aussi pour sa grande contribution à l'aboutissement de ce travail, et pour sa grande disponibilité malgré son emploi de temps chargé.

Je remercie vivement l'ensemble des membres du jury : M^{elle} ATIL Lynda,M^r BOUDIBA Mohand Arezki ,M^r BERKOUN Youcef pour l'honneur qu'ils me font en acceptant de juger ce travail.

Mes remerciements chaleureux s'adressent également aux camarades et amis : qui m'ont beaucoup aidé dans ce travail.

Je remercie tous ceux qui ont contribué de près ou de loin à la réalisation de ce modeste mémoire.

Je suis également reconnaissant à tous les enseignants qui m'ont formé en général et particulièrement à ceux de l'Université.

Table des matières

Introduction Générale

En raison de l'évolution rapide des moyens de collecte des données et de leur traitement informatique, le problème des valeurs *aberrantes* a pris une importance non négligeable durant ces trois à quatre dernières décennies.
La présence de valeurs anormales peut conduire à des estimations biaisées des paramètres des populations et -suite à la réalisation de tests statistiques- à une interprétation des résultats qui peut être erronée.
Une grande partie de notre travail consiste à synthétiser la diversité des méthodes disponibles pour l'utilisateur et met l'accent sur la manière de traiter les valeurs aberrantes de façon structurée.
Malgré des fondements théoriques très largement développés et une bibliographie très abondante sur le sujet, on constate que la plupart des logiciels statistiques existants sont quelque peu limités quant au traitement de ce type de valeurs.
Les observations aberrantes, appelées en anglais *outliers* ont toujours été considérées comme une source de contamination, déformant l'information obtenue à partir des données brutes. Il est donc naturel de rechercher des moyens d'interpréter ou de caractériser ces valeurs anormales et de mettre au point des méthodes pour les traiter, soit en les rejetant afin de restaurer les propriétés initiales des ensembles de données, soit en adoptant des méthodes qui diminuent leur impact au cours des analyses statistiques (Barnett & Lewis, [1994]), c-à-d. des méthodes robustes, résistantes en présence de telles données afin de diminuer de leur impact sur les conclusions à tirer de l'analyse statistique.

Le mémoire est scindé en trois (3) chapitres. Dans le premier, on donne un bref aperçu des notions et concepts utilisés dans le domaine de la détection univariée ; le chapitre 2 s'attarde sur les méthodes de détection et des outils sous-jacents, en insistant sur les données multivariées ; on traite avec le langage R un exemple de données bivariées portant sur un échantillon de quarante-huit (48) observations. Une conclusion générale termine ce travail, dans laquelle on signale quelques nperspectives concernant le domaine des *outliers*.

Chapitre 1

Généralités, Définitions et Concepts

1.1 Contexte et Problématique

La détection des valeurs aberrantes est un aspect important dans les analyses de données statistiques, car celles qui ne sont pas décelées peuvent avoir un effet indésirable -voire néfaste- sur l'interprétation des résultats. La plupart des méthodes existantes de détection (d'observations aberrantes) ont été conçues pour être appliquées à des données complètes univariées ou bivariées.

Toutefois, les observations aberrantes dans les données réelles sont souvent de nature multivariée. Le problème qu'elles posent devient nettement plus difficile à résoudre en trois dimensions ou plus étant donné qu'on cerne moins bien la notion d'éloignement d'une observation du reste des données.

Dans le contexte unidimensionnel, les observations aberrantes ne peuvent être qu'extrêmement petites ou extrêmement grandes (du moins pour les distributions unimodales), mais, dans le cas d'une dimensionalité plus élevée, la question de la « direction » de l'observation aberrante devient de plus en plus importante. Les observations aberrantes peuvent être relativement proches de la majorité des données ou d'un modèle de base si la distance est mesurée dans une métrique euclidienne, parce que cette dernière ne vérifie que les directions des axes. Par contre, si l'on utilise une métrique appropriée à la structure de corrélation de la majorité des données, l'observation aberrante peut être éloignée. Donc, pour les dimensionalités plus élevées, la forme du nuage de points de la majorité des données doit être bien reflétée par la métrique utilisée pour pouvoir déceler les observations aberrantes.

La détection des observations aberrantes nécessite un modèle pour la majorité des données afin de pouvoir distinguer les observations auxquelles le modèle n'est pas bien ajusté. Donc, leur détection est intrinsèquement liée aux modèles et à leur estimation robuste.

1.2 Problématique de l'étude

Le problème concerne la détection classique des valeurs aberrantes. Bien souvent, les démarches de détection de valeurs aberrantes supposent l'hypothèse de normalité des distributions des populations parentes (Barnett et Lewis, 1994).
L'objectif pratique de toute investigation de données serait de proposer une méthode opérationnelle de détection de valeurs aberrantes, applicable sur de grands ensembles de données avec différents types de contraintes, spatiales ou autres.
Nous cherchons plus particulièrement à cerner, de manière aussi exhaustive que possible, l'ensemble des méthodes de traitement des valeurs aberrantes au sein de divers ensembles de données. Et, pour cela, il est nécessaire de mettre en place des méthodes qui permettent de déterminer, de manière optimale, les valeurs limites à partir desquelles une valeur à intégrer dans une base de données est statistiquement acceptée ou rejetée en suivant et gardant une certaine cohérence dans celles-ci.

1.3 Définitions

De nombreux auteurs ont cherché à définir ce qu'est une *valeur aberrante* et les définitions fournies ont évolué au cours du temps.
Grubbs (1969) définit une valeur *aberrante* comme étant *une observation qui semble dévier de façon très marquée par rapport à l'ensemble des autres membres* de l'échantillon dans lequel elle apparaît. Carletti (1988) s'intéresse aux valeurs anormales qu'il définit comme étant des valeurs qui paraîssent suspectes parce qu'elles s'écartent d'une façon importante des autres valeurs de la variable étudiée ou ne semblent pas respecter une norme ou une relation bien définie. Munoz-Garcia et al. (1990) proposent également une définition du terme valeur aberrante et tentent d'éviter le côté subjectif en ajoutant la condition que l'observation devrait dévier nettement du comportement général par rapport au critère sur lequel l'analyse est réalisée (Barnett et Lewis, 1994). Barnett et Lewis (1994), eux, définissent une valeur aberrante dans un ensemble de données comme étant une observation (ou un ensemble d'observations) qui semble être *inconsistante* avec le reste des données ou d'une autre manière, il y a une valeur aberrante lorsque l'une ou l'autre des observations d'un ensemble de données, détonne ou n'est pas en harmonie avec les autres observations. Ce

qui caractérise la valeur aberrante, c'est son impact sur l'observateur. Selon ces auteurs, l'observation ne va pas sembler extrême mais va apparaître, dans un certain sens, comme étant *étonnamment* extrême.

Everitt (2002) tient également compte des modèles de probabilité sous-jacents dans la définition suivante : les valeurs aberrantes correspondent à des observations qui semblent dévier de manière importante des autres observations de la population de laquelle elles proviennent, ces observations semblent être inconsistantes avec le reste des données, en relation avec un modèle supposé connu. La notion d'observation *influente* est définie par Everitt (2002) comme étant une observation qui a une influence disproportionnée sur un ou plusieurs aspects de l'estimateur d'un paramètre, en particulier, les coefficients de régression. D'après Cook et Weisberg (1980), les observations influentes sont celles pour lesquelles les caractéristiques de l'analyse sont altérées de manière considérable quand elles sont supprimées. Cette influence peut être due à des différences par rapport aux autres observations de la variable explicative, à une valeur extrême pour la variable à expliquer ou à une combinaison des deux. Barnett et Lewis (1994) signalent que les valeurs aberrantes sont souvent des observations influentes. Néanmoins, les notions de valeurs aberrantes et d'observations influentes ne sont pas issues de concepts semblables. Une valeur aberrante peut altérer nettement l'estimation d'un paramètre ou le résultat d'un test spécifique mais cette éventualité n'est pas la base de l'identification d'une valeur aberrante. A la différence des observations influentes, une valeur aberrante tout à fait évidente n'a pas d'effet net sur une estimation particulière ou un test lorsqu'une méthode d'accommodation appropriée est utilisée.

Le terme valeur *suspecte* correspond, selon Barnett et Lewis (1994), à une valeur moins extrême qu'une valeur jugée aberrante de manière statistique.

Il y a des tests correspondant à des méthodes statistiques permettant de tester une valeur aberrante afin de déterminer si elle doit être gardée ou rejetée. Une observation est qualifiée d'aberrante (ou atypique) si elle est "tellement différente" des autres données qu'on en vient à douter de la pertinence de la prendre en compte dans les analyses.

1.4 Origine des Valeurs Aberrantes

Toute étude sur les valeurs aberrantes se doit de prendre en compte la nature et l'origine de celles-ci afin de les identifier et de les traiter de la manière la plus adéquate. De manière générale, les valeurs aberrantes sont de nature aléatoire ou déterministe et peuvent trouver

leur origine à différents niveaux.
L'objectif poursuivi lors de l'étude des valeurs aberrantes est détérminant pour le traitement statistique de celles-ci.
Une classification des différentes manières par lesquelles les valeurs aberrantes peuvent survenir a été discutée dans la littérature par divers auteurs, tels que Beckman et Cook (1983) et Hawkins (1980).
Les objectifs de l'étude des valeurs aberrantes dépendent en fait de l'origine et de la nature de celles-ci, comme le montre la figure. Cette figure permet de visualiser clairement le schéma général de traitement des valeurs aberrantes et des objectifs poursuivis. Pour

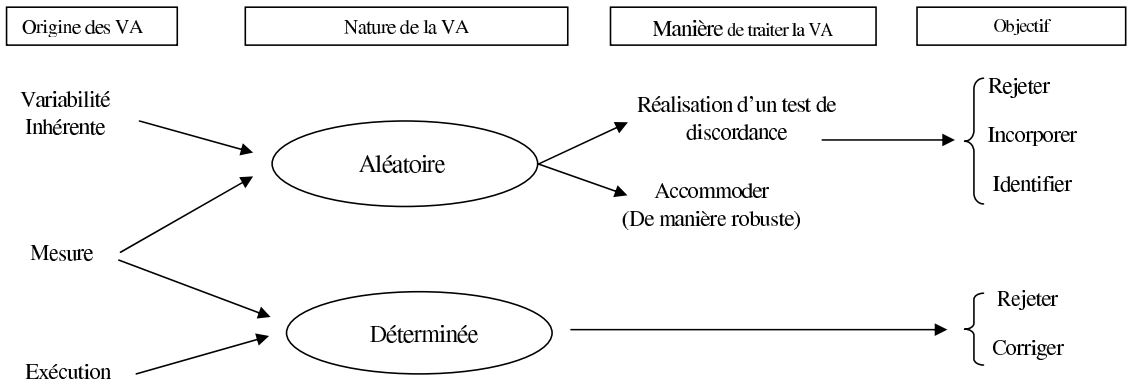

FIG. 1.1 – Schéma général de traitement des valeurs aberrantes et objectifs poursuivis lors de leur examen.

les valeurs aberrantes de nature aléatoire, la réalisation d'un test de discordance doit être perçue uniquement comme la première étape de l'étude de valeurs aberrantes. En effet, en fonction des facteurs étudiés et de l'intérêt pratique de l'étude, il peut être décidé, suite à la réalisation du test, soit de rejeter les valeurs discordantes et de procéder à l'analyse à partir de l'échantillon modifié.
L'analyse des données peut également faire l'objet de l'une ou l'autre forme d'accommodation. Ce choix est réalisé en fonction des objectifs de l'analyse statistique, car si on s'intéresse spécifiquement aux caractéristiques inférentielles d'un modèle de base, quelles que soient la présence et la nature des contaminants, les valeurs aberrantes n'ont qu'un effet de nuisance. Il est alors nécessaire d'utiliser des méthodes *robustes* pour minimiser leur impact. Dans ce cas, l'objectif est l'accommodation en tant que telle et aucun test de discordance n'est approprié. Le but est alors de trouver des procédures statistiques qui ne

recherchent pas les valeurs aberrantes en elles-mêmes mais qui cherchent à les rendre moins importantes quant à leur influence lors de l'estimation de paramètres. Quant à celles dont la nature est déterministe, c'est-à-dire les erreurs de mesure ou d'exécution, elles peuvent être rejetées ou faire l'objet de corrections dans la mesure où celles-ci sont encore réalisables.

1.5 Valeurs Aberrantes et Modèles de Probabilité

Dans les échantillons univariés, les observations susceptibles d'être déclarées comme aberrantes sont identifiées de manière évidente puisqu'il s'agit toujours des valeurs extrêmes de l'échantillon. Néanmoins, des valeurs anormales pour la distribution normale ne le sont pas nécessairement pour une distribution asymétrique. Par définition, les valeurs aberrantes sont inconsistantes avec le reste des données en relation avec un modèle supposé connu. Ainsi, la distribution des données est une notion primordiale lors de l'application de méthodes statistiques car leur traitement est directement lié au choix de cette distribution.

1.6 Valeurs Aberrantes dans le Cas Univarié

1.6.1 Méthodes Statistiques de Traitement des Valeurs Aberrantes

D'une manière générale, l'objectif d'une méthode statistique destinée à l'examen de valeurs aberrantes de nature aléatoire est de fournir des moyens pour vérifier si la présence d'une valeur aberrante dans un ensemble de données possède des implications objectives importantes pour l'analyse future des données. Barnett et Lewis (1994) présentent deux catégories de méthodes statistiques distinctes.
La première méthode est basée sur l'idée de tester une valeur aberrante afin de déterminer si elle doit être gardée ou rejetée en utilisant un test de discordance.
La deuxième méthode est l'accommodation qui consiste en la manière de construire des procédures pour estimer les valeurs des paramètres de la distribution de base de façon relativement libre par rapport à toute influence néfaste d'une valeur aberrante.
Cette dichotomie dans l'approche est fondamentale et se trouve à la base des améliorations récentes dans l'élaboration de méthodes statistiques.
Notons que lors de l'utilisation d'un test statistique, une valeur aberrante peut simplement correspondre à une manifestation de l'erreur de type I, c'est-à-dire qu'il existe une certaine probabilité de mettre en évidence des valeurs aberrantes qui ne le sont pas réellement.

L'erreur de deuxième espèce consisterait par contre à ne pas déceler des valeurs aberrantes qui existent réellement. Une erreur de troisième espèce est liée à un choix inadéquat du modèle de base. Cette dernière erreur est peu connue et est rarement citée de la sorte (Dagnelie, 2003).

1.6.2 Tests de Discordance

L'objectif poursuivi par ce premier type de méthodes statistiques est de tester la valeur aberrante afin de la rejeter de l'ensemble des données ou de l'identifier comme étant une caractéristique d'un intérêt particulier. Un test de discordance correspond à une procédure de détection qui permet de décider si une valeur aberrante peut être considérée comme faisant partie de la population principale.
Soit $x_1, x_2, \ldots, x_n$ l'échantillon dont les valeurs extrêmes sont $x_{(1)}$, $x_{(n)}$. L'une de ces valeurs, par exemple $x_{(n)}$, peut être déclarée aberrante si elle engendre un effet de surprise en fonction de ce qu'on attend de manière informelle du modèle de base F. Supposons que toutes les observations sont bien issues de la distribution F. Un test statistique ou test de discordance peut être réalisé pour examiner si $x_{(n)}$ doit être considérée comme significativement trop grand, c'est-à-dire statistiquement inacceptable, en fonction de la distribution de $X_{(n)}$ sous F. Lorsque le résultat du test indique que $x_{(n)}$ n'est pas acceptable de manière statistique, on peut dire que $x_{(n)}$ est une valeur aberrante supérieure discordante pour le niveau du test. De manière similaire, on peut démontrer des discordances pour les valeurs aberrantes inférieures $x_{(1)}$ ou pour une paire de valeurs aberrantes ($x_{(1)}$,$x_{(n)}$), etc.
Aidé par la notion de test de discordance, il est possible de se rendre compte des différences dans la manière de définir les termes de valeur aberrante et valeur suspecte par les divers auteurs. Une valeur suspecte correspond, selon Barnett et Lewis (1994), à une valeur douteuse qui n'est pas jugée comme aberrante suite à la réalisation d'un test de discordance tandis qu'une valeur aberrante correspond à une valeur étonnamment extrême qui est statistiquement discordante. La valeur suspecte correspond donc à une valeur moins extrême qu'une valeur aberrante.
Parmi les tests de discordance, une distinction peut être réalisée en fonction du type de distribution de la population parente de laquelle provient l'échantillon analysé. On peut distinguer les tests selon qu'ils sont appliqués dans le cas d'une population normale ou d'une autre distribution.
Barnett et Lewis (1994) classent les tests de discordance en sept types différents en tenant compte du critère retenu pour effectuer les tests. Certains tests ont des hypothèses très res-

trictives telles que la connaissance *a priori* du nombre de valeurs anormales ou la position relative de celles-ci (valeur inférieure ou supérieure). Les sept types de tests sont exposés ci-dessous.

1. Les statistiques liées au rapport *excès/étalement*. Ces statistiques correspondent au rapport des différences entre la valeur aberrante et la valeur de l'observation la plus proche, ou toute autre mesure d'étalement de l'échantillon, par rapport à une valeur de dispersion. Tel est le cas par exemple pour le test suivant où le modèle est supposé normal et s correspond à l'écart-type de l'échantillon, celui-ci peut être remplacé par toute mesure de dispersion de l'échantillon :

$$\frac{x_{(n)} - x_{(n-1)}}{S}.$$

   Le test classique de Dixon entre dans cette catégorie (le dénominateur correspond, non pas à s, mais à $x_{(n)} - x_{(1)}$).

2. Les statistiques liées au rapport *amplitude/étalement* pour lesquelles, le numérateur est remplacé par l'amplitude de l'échantillon :

$$\frac{x_{(n)} - x_{(1)}}{S}.$$

3. Les statistiques liées au rapport *écart/étalement*, pour lequel le numérateur correspond à une (mesure de) distance entre une valeur aberrante et une mesure descriptive des données telle que la moyenne. En supposant que les données sont distribuées selon une loi normale, le test classique de Grubbs (1950) -et utilisé par Carletti (1988)- correspond à ce type de test et est destiné à tester soit une valeur aberrante inférieure x(1), soit une valeur aberrante supérieure x(n), soit les deux simultanément :

$$\frac{\bar{x} - x_{(1)}}{S}, \frac{\bar{x} - x_{(n)}}{S}.$$

   La moyenne $\bar{x}$ peut être remplacée par toute autre mesure de position.

4. Les statistiques liées au rapport *extrêmité/position* qui correspondent au rapport de valeurs extrêmes par rapport à des mesures de position. Un exemple de ce type est le suivant :

$$\frac{x_{(n)}}{\bar{x}}.$$

5. Les statistiques liées au rapport de sommes de carrés qui sont légèrement différentes des statistiques précédentes et qui expriment le rapport entre les sommes des carrés pour l'échantillon dont on a extrait la valeur aberrante et l'échantillon global.
   Un test de ce type est proposé par Grubbs (1950) pour tester deux valeurs aberrantes supérieures. Ce test est également utilisé pour tester les valeurs aberrantes supérieures issues d'échantillons de valeurs extrêmes (Fung et Paul, 1985).
6. Les statistiques liées aux moments d'ordre supérieurs correspondent à des rapports de mesures de *symétrie* et d'*aplatissement*. Les tests développés par Ferguson (1961) utilisent ces statistiques.
7. Les statistiques W de Shapiro-Wilks sont également très utilisées pour tester des valeurs aberrantes et sont calculées, par exemple pour une valeur aberrante inférieure $x_{(1)}$, de la manière suivante :

$$W = \frac{n(\bar{x} - x_{(1)})^2}{n - 1 \sum_{i=1}^{n}(x_{(i)} - \bar{x})^2}$$

Ces tests sont principalement utilisés pour tester des observations issues des distributions normale et exponentielle.
Barnett et Lewis (1994) présentent également des tests de discordance spécifiques aux distributions de Gumbel, Fréchet, Weibull, Pareto, Poisson et binomiale.

1.6.3 Accommodation des valeurs aberrantes

Les procédures d'accommodation englobent des méthodes statistiques destinées à réaliser de l'inférence sur la population à partir de laquelle l'échantillon aléatoire a été obtenu. Les résultats acquis par l'intermédiaire de ces procédures ne sont pas sérieusement déformés par la présence des valeurs aberrantes ou par des contaminants. Lorsqu'on suspecte la présence de valeurs aberrantes suite à des erreurs d'exécution ou des mesures aléatoires et que l'objectif de l'étude correspond à l'estimation d'un paramètre du modèle initial, il est intéressant d'utiliser un estimateur qui n'est pas trop sensible à la présence de celles-ci.
Les procédures d'accommodation permettent dès lors d'éviter de rejeter des valeurs aberrantes. Cette manière de travailler implique que les valeurs aberrantes en elles-mêmes ne sont plus le centre d'intérêt de l'étude, le but consiste alors à travailler correctement malgré leur présence. Ceci correspond exactement au concept de *robustesse*. Les techniques d'accommodation sont dites robustes face à la présence de valeurs aberrantes, cependant, le

concept de robustesse, de grande importance, dans le cadre général de l'inférence statistique, n'est pas spécifique à l'examen des valeurs aberrantes.
Des méthodes robustes peuvent également répondre spécifiquement au problème de valeurs aberrantes lorsqu'il y a une contamination et dès lors un décalage par rapport à un modèle de probabilité initial. Il ne faut cependant pas négliger l'importance du modèle de base dans le cas de l'accommodation. Si des valeurs aberrantes sont détectées parce que le modèle initial ne reflète pas le degré approprié de variabilité, il est nécessaire de s'intéresser à des distributions plus étendues que la distribution normale, utilisée classiquement. L'omission de valeurs extrêmes pour se protéger contre les valeurs aberrantes est une manière robuste pour estimer des mesures de dispersion mais si le modèle de base n'est pas correctement choisi, la procédure encourage plutôt la sous-estimation, le but étant de réduire l'effet des valeurs extrêmes. Si d'un autre côté, une hypothèse alternative permet d'exprimer la contamination du modèle initial, l'estimation ou le test des paramètres du modèle initial peut être très intéressant et il est alors important d'utiliser des procédures robustes appropriées pour se protéger des composantes de faible probabilité ou contre les valeurs trop décalées.
Les travaux des trois récentes décenies qui, implicitement ou explicitement, tentent d'accommoder les valeurs aberrantes dans le processus d'inférence se divisent en deux tendances.
La première tendance comprend les méthodes d'estimation qui protègent implicitement contre les valeurs aberrantes en plaçant moins d'importance sur les valeurs extrêmes que sur les autres observations de l'échantillon. Cet accent est une caractéristique de l'ensemble des méthodes robustes développées durant les trente dernières années.
La seconde tendance de l'étude sur la contamination par des valeurs aberrantes est spécifiquement liée à la robustesse lors de la présence de ces valeurs. Les méthodes d'estimation et les tests qui en découlent, portent un regard particulier sur la nature des modèles nécessaires à expliquer la présence des valeurs aberrantes. Ce domaine d'étude est en cours d'expansion et des techniques d'accommodation spécifiques sont développées actuellement.

1.7 Quelques distributions à forte asymétrie

L'objectif de ce chapitre est d'identifier les distributions qui correspondent le mieux aux données observées. Ces distributions doivent permettre de répondre à des situations de mélanges, en s'attachant aux queues des distributions, tout en restant facilement applicables à la détection des valeurs aberrantes. Les distributions à forte asymétrie ont, entre

autres, été étudiées à partir de l'examen des valeurs extrêmes qui consiste en la recherche et la caractérisation des distributions limites vers lesquelles les valeurs extrêmes convergent. Ces études, qualifiées de *Théorie des valeurs extrêmes* ont été développées pour l'estimation d'événements rares.
Elles permettent d'extrapoler le comportement de la queue de la distribution des données à partir des observations les plus élevées (Garrido, 2002), c'est-à-dire de calculer la probabilité qu'un événement se produise même si celui-ci n'a jamais eu lieu. Essenwanger (1986) signale que lors de l'étude d'événements rares, il est intéressant de présenter le risque d'occurrence ou la probabilité d'occurrence de ces événements. C'est pour cette raison que, lors de la présentation des différentes distributions, nous exposons la manière de calculer ce risque. Celui-ci correspond à la probabilité que l'événement se produise et s'exprime de la manière suivante : Probabilité d'occurrence $=1 - F(x)$, $F(x)$ étant la fonction de répartition.
Une notion très fréquemment employée pour l'estimation de certains paramètres correspond à la fonction des quantiles $Q(p)$, qui fournit la plus petite valeur de x pour laquelle $F(x) > p$, avec $0 < p < 1$. Dans le cas d'une distribution théorique quelconque, elle correspond à l'inverse de la fonction de répartition et s'énonce de la manière suivante :

$$Q(p) = inv(x, F(x) \geq p)$$

Lorsque la distribution n'est pas connue, la fonction des quantiles empiriques $\hat{Q}(p)$ est la fonction qui, pour une valeur donnée p $(0 < p < 1)$ fournit la plus petite valeur, vers la gauche, pour laquelle une proportion p des données est rencontrée. Cette fonction correspond à une bonne approximation de la fonction des quantiles correspondante Q(p).

1.7.1 Distribution Exponentielle

La fonction la plus connue de la classe des distributions de Gumbel est la distribution exponentielle, utilisée dans de très nombreuses études statistiques. Elle permet notamment de caractériser le délai d'apparition d'un événement aléatoire tel qu'un accident, la mort d'un individu, la défaillance d'un appareil, au cours d'un intervalle de temps (Johnson et Kotz, 1970 ; Dagnelie, 1998a). Cette distribution a été appliquée lors d'analyses de survie dans le domaine biomédical, le temps de survie étant représenté par une variable aléatoire exponentielle. De nombreuses études sur les distributions exponentielles concernent l'estimation du paramètre de celle-ci et le cas des distributions exponentielles tronquées à

droite ou à gauche (Johnson et Kotz, 1970). La distribution exponentielle de paramètre λ présente une probabilité d'occurrence qui vaut :

$$1 - F(x) = exp(-\lambda x) x > 0, \lambda > 0,$$

et comme fonction de densité de probabilité :

$$f(x) = \lambda exp(\lambda x).$$

L'estimation du paramètre λ peut être réalisée notamment par le maximum de vraisemblance (Johnson et Kotz (1970)). Une méthode d'estimation robuste du paramètre λ, presentée par Brazauskas et Serfling (2001), est basee sur le calcul d'une médiane généralisée, applicable même pour des éffectifs faibles mais dépendant de la taille des échantillons. Les auteurs donnent des conseils pour la sélection de l'un de ces estimateurs, en fonction de son comportement vis-a-vis des valeurs aberrantes supérieures ou inférieures ou en fonction du niveau possible de contamination par des valeurs aberrantes.
En ce qui concerne la recherche de valeurs aberrantes, divers tests de discordance sont développés par Barnett et Lewis (1994). Ces tests sont les suivants : test de la valeur la plus élevée ou supérieure, test de la valeur la plus faible, test des deux valeurs les plus élevées ou les plus faibles, test de la valeur inférieure ou supérieure simultanément (quel que soit le paramètre de position ou celui-ci étant inconnu).
Dans notre cas, ces tests nous semblent limités dans leur utilisation car ils ne permettent de vérifier qu'une ou deux valeurs à la fois. La méthode de détection de valeurs aberrantes que nous devons préférer doit permettre de traiter facilement un nombre variable d'observations plutôt qu'une ou deux observations seulement. Deux tests de détection de valeurs aberrantes d'un nombre k de valeurs aberrantes situées, soit à droite, soit à gauche de la distribution sont présentés par divers auteurs. D'autres tests sont proposés mais concernent le problème d'accommodation des valeurs aberrantes.

1.7.2 Distribution de Weibull

La distribution de Weibull a été développée dans le domaine de la physique, pour la modélisation de la résistance à la rupture de matériaux (Johnson et Kotz, 1970). L'ajustement de la distribution exponentielle n'étant pas assez précise pour décrire ce type de données, une transformation puissance est appliquée aux variables observées et une distribution exponentielle est alors obtenue. La distribution de Weibull apporte ainsi une

certaine flexibilité par rapport aux modèles obtenus à partir de la distribution exponentielle. Rappelons que la queue de la distribution de Weibull présente un étalement vers la droite plus faible que celle de la distribution exponentielle. Réellement, cette distribution est utilisée dans le domaine biomédical (temps de survie), en actuariat et dans tout autre domaine où l'étude des valeurs extrêmes est essentielle. Dans le cas des analyses de survie, le temps de survie est représenté couramment par une variable aléatoire exponentielle ce qui n'est pas toujours idéal. La distribution exponentielle est alors judicieusement remplacée par la distribution de Weibull. Cette distribution présente une probabilité d'occurrence qui s'énonce de la manière suivante :

$$1 - F(x) = exp(-\lambda x^{\tau})$$

où $x > 0$, $\tau > 0$, avec comme fonction densité :

$$f(x) = \lambda\tau x^{\tau-1} exp(-\lambda x^{\tau}).$$

Il est possible de présenter la distribution de Weibull en faisant appel aux notions purement mathématiques sur les transformations de variables (Beirlant et al., [1996]). Pour cela, il est nécessaire de rappeler l'expression théorique de la probabilité d'occurrence et de la fonction densité d'une variable aléatoire transformée :

$$1 - F_y(x) = 1 - F_X(y^{-1}(x)),$$

pour $x \in y(0, \infty)$ et

$$f_y(x) = f_X(y^{-1}(x)) \mid \frac{dy^{-1}(x)}{dx} \mid$$

où y^{-1} correspond à la fonction inverse.
Si les variables x suivent une distribution exponentielle, de parametre $\lambda > 0$:

$$\begin{cases} f_X(x) = \lambda exp(-\lambda x) \\ 1 - F_X(x) = exp(-\lambda x), \quad , \end{cases}$$

En appliquant la transformation $y(x) = x^{\frac{1}{\tau}}$ où τ est positif et à partir des formules présentées pour les variables aléatoires transformées, on obtient :

$$1 - F_X(x) = exp(-\lambda x)$$

$$1 - F_y(x) = 1 - F_X(y^{-1}(x)) = exp(-\lambda y^{-1}(x))$$

Comme $y^{-1}(x) = x^{\tau}$

$$1 - F_y(x) = exp(-\lambda x^{\tau})$$

Pour la fonction de densité de probabilité, on trouve également

$$f(x) = \lambda\tau x^{\tau-1} exp(-\lambda x^{\tau})$$

Cette transformation de puissance de paramètre τ, appliquée a des variables aléatoires qui suivent une distribution exponentielle de paramètre λ positif, mène à la distribution de Weibull de paramètre λ et τ. Le paramètre τ est appelé *l'index de Weibull*.
Il est possible d'imaginer que les tests de détection de valeurs aberrantes, exposés pour la distribution exponentielle, peuvent être appliqués à des données présentant une distribution de Weibull. En effet, si la variable positive X est telle que $Y = X^{\tau}$ possède une distribution exponentielle de paramètre 1, alors X se distribue selon la loi de Weibull de paramètre τ. Si τ est connu, la variable transformée Y est utilisée et les techniques développées pour les distributions exponentielles sont alors facilement applicables.

1.7.3 Distribution de Pareto

La loi de Pareto trouve son origine en économie et établit que, dans une société donnée et pour une période donnée, la distribution du nombre de personnes en relation avec leur revenu a une très forte tendance à décroître dans leur partie supérieure, c'est-à-dire que cette distribution correspond à une fonction puissance décroissante de ce revenu. En d'autres mots, la queue supérieure de la distribution du logarithme de la variable tend vers une loi exponentielle (Barndorff-Nielsen, 1977).
En fait, la distribution est toujours utilisée en économie mais a pris un large développement théorique dans le domaine des assurances et fait actuellement son apparition dans le monde de la finance.
De nombreuses études sur les performances des réseaux de télécommunications font également appel à cette distribution.
La distribution de Pareto, de paramètre α, présente une probabilité d'occurrence qui s'énonce selon l'expression suivante :

$$1 - F(x) = x^{-\alpha} \quad ou \quad 1 - F(x) = x^{-\frac{1}{\gamma}}, \ \ avec \ \ x > 1 \quad et \quad \alpha = \frac{1}{\gamma}$$

La fonction de densité correspond à :

$$f(x) = \alpha x^{-\alpha-1} \quad pour \quad x > 1$$

. Le paramètre α est positif et est appelè *index de Pareto*. Cette distribution (à un paramètre) est aussi appelée *distribution stricte de Pareto*.

La **distribution stricte de Pareto** peut également être présentée en suivant les notions mathématiques sur les transformations de variables, comme exposées pour la distribution de Weibull. Pour obtenir la distribution stricte de Pareto, il faut considérer les variables X, qui suivent une distribution exponentielle de paramètre $\alpha > 0$:

$$\begin{cases} f_X(x) = \alpha exp(-\alpha x) \\ 1 - F_X(x) = exp(-\alpha x) \end{cases} .$$

En appliquant la transformation $y(x) = exp(x)$ et à partir des formules présentées pour les variables aléatoires transformées :

$$1 - F_X(x) = exp(-\alpha x)$$

$$1 - F_Y(x) = 1 - F_x(y^{-1}(x)) = exp(-\alpha y^{-1}(x))$$

comme $y^{-1}(x) = log(x)$, on obtient

$$1 - F_Y(x) = exp(-\alpha log(x)) = exp(logx^{-\alpha}) = x^{-\alpha} \quad pour \quad x > 1$$

Une transformation exponentielle appliquée à des variables aléatoires qui suivent une distribution exponentielle de paramètre positif α, nous mène a la distribution stricte de Pareto de paramètre α. En d'autres termes, les variables aléatoires distribuées selon la distribution stricte de Pareto et transformées par la fonction logarithmique suivent une distribution exponentielle ayant comme parametre α ou $\frac{1}{\gamma}$.
En ce qui concerne la recherche de valeurs aberrantes par des tests de discordance, si une variable aléatoire X suit une distribution de type Pareto, d'index α et si la transformation $Y = logX$ est appliquée, la fonction de répartion de y est alors de la forme suivante (Barnett et Lewis, 1994) :

$$F(x) = 1 - exp(-\alpha y) \quad pour \quad y \geq 1,$$

Y possède ainsi une distribution exponentielle de paramètre α.
Supposons que l'hypothèse de travail considère qu'un échantillon trié par ordre croissant est distribué selon la loi de Pareto $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ et contient une ou plusieurs valeurs aberrantes. Les valeurs transformées $logx_{(1)}, logx_{(2)}, \ldots, logx_{(n)}$ sont alors également triées dans l'ordre croissant et les valeurs $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ sont issues d'une distribution Y exponentielle. Les valeurs $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ de l'échantillon peuvent faire l'objet de tests de discordance à partir des valeurs correspondantes de y qui concernent les distributions exponentielles. Les tests concernent alors les distributions exponentielles et sont présentés par Barnett et Lewis (1994)[1].

Commentaires-Discussion

Le problème de la présence de valeurs aberrantes revêt une importance capitale dans le traitement statistique des données et leur analyse. Les observations contenues dans les bases de données doivent nécessairement faire l'objet d'une validation car l'apparition de valeurs aberrantes est inévitable en raison de la quantité des données traitées et des diverses sources d'erreurs lors de leur recueil ou leur acquisition.

Pour sauvegarder une information de qualité (au sens fidélité au contenu), une recherche de valeurs suspectes ou aberrantes devrait être effectuée avant l'exploitation d'une quelconque base de données, car ces valeurs peuvent conduire à des estimations biaisées de paramètres ou une interprétation altérée des résultats, notamment de certains tests statistiques. Il est donc naturel de rechercher les moyens d'interpréter ou de caractériser ces valeurs anormales et de mettre au point des méthodes pour les traiter, soit en les rejetant afin de restaurer les propriétés initiales des ensembles de données, soit en adoptant des méthodes qui diminuent de leur impact au cours des analyses statistiques.

Les termes liés aux valeurs aberrantes ont été abondamment définis et ont montré l'importance de l'hypothèse d'un modèle de base pour les données traitées. La nature aléatoire ou déterministe d'une observation aberrante et les sources d'apparition de celle-ci ont été exposées en relation directe avec les différents objectifs possibles rencontrés lors de l'examen de valeurs aberrantes.

En fonction de l'objectif à atteindre et de la nature de la valeur aberrante, le traitement des données est très différent. Les deux possibilités pour traiter les données atypiques, dans le cadre d'un objectif donné, sont les tests de discordance et les procédures d'accommodation.

Chapitre 2

Différentes Méthodes de Traitement des Données Aberrantes

Introduction

Tout comme pour le cas univarié, le traitement des données multidimensionnelles est très souvent confronté à diverses questions : contrairement au cas univarié , il n'y a pas d'ordre entre les différentes observations , au-delà de la dimension 3 on ne peut pas avoir une représentation du nuage de points. Dans cette partie , on mettra l'accent sur les différentes questions d'intérêt qui se posent et on envisagera les différentes méthodes de détection des observations aberrantes. Pour ce faire, on aura à parler de notions telles que distances (Mahalanobis) et les tests dont on a parlé au chapitre 1. On tentera une géneralisation des questionnements du cas univarié au cas multivarié ; on essaiera de voir s'il y a des modèles robustes pour mieux résister a la présence de données aberrantes.

2.1 Les différents Outils

2.1.1 Distance de Mahalanobis

C'est une mesure basée sur des corrélations entre les variables par lesquelles différentes configurations peuvent être identifiées et analysées. C'est un moyen très utile pour déterminer la similarité d'un échantillon inconnu à un autre connu. Un trait imporant de cette distance est qu'elle diffère de la distance euclidienne en ce sens qu'elle tient copmte des corrélations entre les observations et elle est invariante par changement d'échelle (elle ne dépend pas de l'échelle de mesure). Pour un vecteur $x = (x_1, x_2 \ldots, x_p)'$ p-dimensionnel d'observations

de *moyenne* $\mu = (\mu_1, \mu_2, \ldots, \mu_p)'$ et de (matrice de) *covariance* Σ, on pose

$$d_M^2(x) = (x - \mu)'\Sigma^{-1}(x - \mu).$$

Comme mesure de dissimilarité entre deux vecteurs x et y provenant d'une même distribution avec Λ comme matrice de covariance, on pose :

$$d^2(x, y) = (x - y)'\Lambda^{-1}(x - y)$$

Dans le cas particulier où la covariance est la matrice Identité, on obtient la distance euclidienne normalisée :

$$d^2(x, y) = \sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}$$

où σ_i est l'écart type de x_i.

Applications

La distance de Mahalanobis trouve beaucoup d'applications, elle est notamment et très souvent

- utilisée dans les techniques de classification,
- liée au T^2de Hotelling (en statistique multivariée),
- employée dans la détection des valeurs aberrantes, spécialement dans les modèles de régression : un point ayant une grande distance de Mahalanobis du reste des points est susceptible d'avoir un grand *levier* puisqu'il a une grande influence sur la pente de l'équation de régression.

2.1.2 L'estimateur du MCD (Minimum Covariance Determinant)

Une méthode d'estimation lors de la détection des valeurs aberrantes qui attire souvent l'attention est celle du " Minimum Covariance Determinant" (MCD). Cette méthode est une des plus robustes mais également une des plus complexes.

En effet, la distance de Mahalanobis définie par :

$$MD_i = \sqrt{(\mathbf{x}_i - T(X))'C(X)^{-1}(\mathbf{x}_i - T(X))}$$

où $T(X) = \frac{1}{n}\sum_i^n \mathbf{x}_i$ et $C(X) = \frac{1}{n}\sum_i^n (\mathbf{x}_i - T(X))'(\mathbf{x}_i - T(X))$, n'est pas robuste puisque, les estimateurs empiriques du vecteur des moyennes arithmétiques et de la matrice de covariance (calculés sur la base de l'échantillon brut) peuvent être contaminés par la présence de valeur aberrantes (à cause de l'effet de masque).

Pour résoudre ce problème, l'estimateur du MCD applique une méthode intuitivement simple mais relativement compliquée à mettre en oeuvre en pratique. Il s'agit de prendre le sous-ensemble de h observations parmi les n de l'échantillon ($\frac{n}{2} \leq h < n$) pour lequel le déterminant de la matrice de covariance est minimal. Les statistiques $T(X)$ et $C(X)$ sont alors estimés par la moyenne arithmétique et la matrice de covariance de l'échantillon.
L'idée est de considérer que l'échantillon qui possède la plus petite variance généralisée (c'est-à-dire la matrice de covariance de plus petit déterminant) ne sera pas contaminé par des valeurs atypiques. Dés lors, nous pourrons nous y fier pour calculer les estimateurs robustes de position T(X) et de dispersion C(X).
Le problème est d'implémenter cette méthode puisque lorsque n est grand, il devient très difficile de vérifier tous les sous-échantillons de taille h. Certains auteurs ont alors proposé des algorithmes qui permettent une approximation. Parmi ces algorithmes, notre attention s'est portée sur le FASTMCD proposé par Rousseeuw & Van Driessen (1999).
Cet algorithme (FASTMCD) commence par prendre un sous-ensemble de $(p+1)$ observations pour lesquelles la moyenne et la matrice de covariance sont calculées et ensuite utilisées pour déterminer les h observations (par défaut, h est égal à 0.75n) les plus proches ; l'opération est réitérée pour quelque cinq cents (500) sous-ensembles de $(p+1)$ observations de l'échantillon. Ensuite, les dix (10) meilleures solutions (c'est-à-dire les 10 sous-ensembles pour lesquels les matrices de covariance ont le plus petit déterminant) sont retenues pour recommencer exactement le même processus. La procédure algorithmique s'arrête lorsque deux déterminants calculés successivement sont égaux (on suppose alors qu'il y a eu convergence), c'est-à-dire lorsque le processus s'est stabilisé.

2.2 Mesures de la robustesse

Les outils de base pour décrire et mesurer la robustesse sont :
- Le point de rupture,
- la fonction d'influence,
- la courbe de sensibilité.

2.2.1 Le point de rupture

Intuitevement, le point de rupture d'un estimateur est la proportion d'observation incohérents (arbitrairement "grandes") qu'un estimateur peut "supporter" avant de produire des résultats arbitrairement grands.

Le plus couramment cité comme exemple type est la *moyenne empirique* : $\overline{X}$. Soit $(x_1, \ldots, x_n)$ les réalisations de n variables $(X_1, \ldots, X_n) \sim N(0, 1)$.On a l'habitude d'estimer la moyenne par $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$: un tel estimateur a un point de rupture nul puisqu'on peut rendre $\overline{X}$ arbitrairement grand en changeant arbitrairement une seule des observations.
Ainsi : plus le point de rupture d'un estimateur est grand, plus celui-ci est robuste.
Remarque

. Les valeurs du point de rupture sont normalement comprises entre 0 et 1/2.

. La valeur max de 1/2 est atteinte par la médiane en tant qu'estimateur de position.

. Des statistiques avec de grands points de rupture sont dites *résistantes*.

2.2.2 La fonction d'Influence

La fonction d'Influence empirique donne une idée sur le comportement d'un estimateur quand on change une observation dans l'échantillon. On suppose le contexte suivant :

1. (Ω, A, P) est un espace probabilisé,
2. (χ, ξ) est un espace mesuré (espace d'états),
3. Θ espace des paramètres de dimension $p \in \mathbb{N}^*$,
4. (Γ, S) est un espace mesuré,
5. $\gamma : \Theta \longrightarrow \Gamma$ est une projection,
6. $F(\Sigma)$ = ensemble de toutes de distributions sur Σ.

Définition (FIE)

Soit $n \in \mathbb{N}^*$ et $(X_1, \ldots, X_n) : (\Omega, A) \longrightarrow (\chi, \xi)$ identiquement indépendamment distribuées (iid) et $(x_1, \ldots, x_n)$ un échantillon issu de ces variables. Soit $T_n : (\chi^n, \xi^n) \longrightarrow (\Gamma, S)$ un estimateur. Soit $i \in (1, \ldots, n)$. La *fonction d'influence empirique* F_i en x est définie par : $x(\in \chi) \longrightarrow T_n(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \in \Gamma$.
Ceci signifie simplement qu'on remplace la i^{eme} valeur dans l'échantillon par une autre arbitraire et on regarde la valeur de l'estimateur. Plutôt que de s'intéresser seulement à l'échantillon, on peut utiliser la distribution des variables. On tente donc de voir ce que devient l'estimateur quand on change la distribution des données légérement : on suppose une distribution donnée et on mesure la sensibilité au changement dans cette distribution. Soit A un ensemble convexe et soit F une distribution dans F de paramètre $\theta \in \Theta$. Soit $T : A \longrightarrow \Gamma$ une fonctionnelle qu'on suppose être la valeur asymptotique d'une

suite $(T_n)_{n\in\mathbb{N}}$ d'estimateurs. On suppose que T est *consistante* au sens de Fisher (i.e. $T(F_\theta) = \theta \ \forall\theta \in \Theta$) (i.e. quand on prend pour modèle F, la suite d'estimateurs (T_n) mesure asymptotiquement la vraie quantité).
Soit G une distribution quelconque dans A. Que se passe-t-il quand les données n'obéissent plus au modèle F exactement mais à un autre, légérement différent, "s'approchant de G" ? on s'intéresse donc à la quantité :

$$dF_{G-F}(F) = \lim_{t\to 0^+} \frac{T((1-t)F + tG) - T(F)}{t}$$

qui est la dérivée de T en F (en fait au sens de Gâteau), dans la direction de G. Soit $x \in \chi$, soit

$$G = \Delta_x = \begin{cases} 1 & en \ \ x \\ 0 & sinon \end{cases}$$

Définition (FI)

La *Fonction d'Influence* décrit l'effet d'une contamination infinitésimale au point x sur l'estimation, standardisée par la masse t de la contamination (le biais asymptotique causé par la contamination des observations) :

La fonction d'influence IF de l'estimateur T, en x, de la distribution F est donnée par :

$$IF(x, T, F) = \lim_{\epsilon\to 0} \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon}$$

Pour un estimateur robuste, on voudrait avoir une fonction d'influence bornée (i.e qui $\nrightarrow \infty$ quand est arbitrairement grand !) ?.
Si nous remplaçons F par $F_{n-1} \approx F$ et prenons $\epsilon = \frac{1}{n}$ nous pouvons voir que IF mesure approximativement n fois le changement en T causé par une observation additionnelle en x quand T est appliqué à un grand échantillon de taille $n-1$.
Comme discuté dans Hampel et al., l'importance de la fonction d'influence se situe dans le fait qu'elle peut décrire l'effet d'une contamination infinitésimale au point x sur l'estimation T sandardisée par la masse (le poids) du contaminant. Elle nous donne une image du biais asymptotique causé par la contamination.
La fonction d'influence a un certain nombre de propriétés très attirantes. Hampel montre que dans des conditions habituellement recontrées en pratique, T_n est asymptotiquement normal de variance asymptotique égale à l'*intégrale du carré de* IF. Huber discute aussi ces conditions regularité.

La distribution de $(\sqrt{n}[T_n - T(F)])$ tends vers $N(0, V(T,F))$, où $V(T,F)$ est la variance asymptotique donnée par :

$$V(T,F) = \int IF(x,T,F)^2 dF(x).$$

Cette formule montre qu'une fonction d'influence bornée a une variance asymptotique bornée, et il y a donc lieu d'être prudent quant à l'utilisation des estimateurs avec des fonctions d'influence non bornées (à l'exemple de la moyenne empirique).
Une autre quantité importante relative à la fonction d'influence est le *maximum de sa valeur absolue*, à partir de laquelle Hampel & al. définissent la sensibilité de l'erreur grossière (Gross-Error sensitivity) γ^* de T en F comme

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

où le supremum est pris pour tout x où $IF(x,T,F)$ existe.
La sensibilité de l'erreur grossière mesure la plus mauvaise influence qu'une petite contamination peut avoir sur la valeur d'un estimateur, et peut être considérée comme une limite supérieure du biais asymptotique de l'estimateur T.
Idéalement,$\gamma^*(T,F)$ est fini. Dans beaucoup de cas, rendre bornée la quantité $\gamma^*(T,F)$ est la première étape pour faire de T une statistique robuste. La fonction d'influence peut aussi nous dire plus concernant T par la sensibilité λ^* du paramètre de position.

2.2.3 Courbe de Sensibilité

Pour un échantillon donné de taille n, Hoaglin et al. discutent de la *courbe de sensibilité* $SC(x|T_n)$, qui est considérée en calculant l'estimateur T_n respectivement avec et sans l'observation au point x ; elle est proportionnelle à la dimension de l'échantillon :

$$SC(x|T_n) = n(T_n(x_1, \ldots, x_{n-1}, x_n) - T_{n-1}(x_1, \ldots, x_{n-1}))$$

. La courbe de sensibilité décrit l'effet d'une observation individuelle sur l'estimation T_n pour un ensemble de données spécifiques. Il n'est pas difficile à calculer empiriquement, mais n'est pas trop utile théoriquement.
Quand on fait tendre la taille n de l'échantillon vers l'infini, la limite résultante s'appelle la *fonction d'influence*. En l'absence de l'ensemble de données, la fonction d'influence est définie en référence à une distribution spécifique, F. Hampel a présenté le concept de

fonction d'influence que lui et ses collaborateurs ont popularisée dans un livre. Bien que l'approche basée sur la fonction d'influence a été utilisée principalement (mais pas exclusivement) sur des données univariées, l'expérience a montré qu'il peut être avantageux d'un point de vue calculatoire d'examiner chaque composante séparément. Nous pouvons alors combiner les estimateurs respectifs pour calculer les distances de Mahalanobis et pour situer les valeurs aberrantes multivariées dans l'espace.
Supposons avoir un échantillon univariée $(x_1, \ldots, x_n)$ avec la distribution empirique F_n. Soit Δ_x le point de masse 1 en x, nous pouvons définir F_n par $F_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_{x_i}$. L'estimation du paramètre de position θ de la distribution sous-jacente F, nous considérons la statistique $T_n = T(x_1, \ldots, x_n) = T(F_n)$. Nous pouvons penser à ceux-ci comme une statistique d'ordre $\{T_n; n \geq 1\}$; un pour chaque échantillon possible de dimension n. Asymptotiquement, nous considérons les estimateurs qui sont des fonctionnelles, c'est-à-dire $\lim_{n\to\infty} T_n(x_1, \ldots, x_n) = T(F)$ où la convergence est en probabilité, et nous dirons que $T(F)$ est la valeur asymptotique de $\{T_n, n \geq 1\}$ en F.

Des changements insignifiants dans quelques observations, qui peuvent être dûs à des arrondis ou des groupements pourraient avoir un effet notable sur l'estimation. Le fait de décaler une observation légèrement du point x à un point voisin y peut être mesuré au moyen de $IF(y; T; F) - IF(x; T; F)$ parce qu'une observation supplémentaire a été enlevée en x et ajoutée en y.

Ceci peut-être approché par la pente de la fonction d'influence λ^* qui est la plus mauvaise de ces différences normalisées :

$$\lambda^* = sup_{x \neq y} \frac{|IF(y; T; F) - IF(x; T; F)|}{|y - x|}$$

En termes de fonction d'influence, ceci signifie que IF s'annule dans une certaine région (habituellement au delà d'une certaine distance de l'estimation de position). En d'autres termes, si IF est nulle dans une certaine région, toutes les observations dans cette région n'ont aucun effet sur T.
Pour F symétrique (quitte à mettre le centre de symétrie en zéro), Hampel et al. définissent le *point de rejet* ρ^* comme

$$\rho^* = inf\{r > 0; IF(x; T; F) = 0; quand\ \ |x| > r\}$$

Si aucun r n'existe, alors $\rho^* = \infty$: par définition toutes les observations plus loin que ρ^* sont rejetés complétement. Pour T égal à la moyenne empirique, il peut être montré

que $IF(x;T;F) = x$, l'efficacite asymptotique $e = 1$ (la variance réciproque). L'efficacité maximale possible, et la sensibilité de position locale $\lambda^* = 1$, aussi la valeur la plus petite possible pour les estimateurs signicatifs. Cependent, la sensibilité de l'erreur grossière $\gamma^* = \infty$ et le point de réjet $\rho^* = \infty$ ont des revers très significatifs en présence des valeurs aberrantes.

Nous déduisons que la moyenne est moins susceptible aux erreurs d'arrondi et est très efficace mais peut "se casser" complètement même en présence d'une seule valeur aberrante lointaine.

L'influence de la médiane de l'échantillon a une discontinuité à l'origine, étant donné qu'elle change de manière discontinue si une des deux valeurs centrales est changée.

La sensibilité de position locale est $\lambda^* = \infty$ à cause de cette discontinuité, tandis que $F = N(0,1)$, la médiane a pour efficacité $e = \frac{2}{\Pi} = 0.637$ et pour $F = N(\mu, \sigma^2)$ une efficacité asymptotique de $e = \frac{2}{\sigma^2 \Pi}$. Hampel et al. montrent que la sensibilité de l'erreur grossière de la médiane est $\gamma^* = 1.253$, qui est en fait la valeur minimale Indépendament de certains inconvénients de la médiane, aucun estimateur ne peut la "battre" du point de vue robustesse. Finalement, le point de rejet $\rho^* = \infty$ peut sembler étonnant à première vue, mais le fait est que la médiane ne rejette pas les valeurs aberrantes, mais elle diminue seulement leur influence, à un degré considérable, ceci s'explique par le fait qu'elle considère une ou les deux observations du milieu -plutôt que toutes les observations- selon que la taille de l'échantillon est impaire ou paire respectivement.

Basé sur la comparaison ci-dessus, nous pouvons facilement conclure que la médiane est beaucoup plus robuste que la moyenne, au regard de l'efficacité et de la sensibilité de position. Il y a encore un moyen d'améliorer la médiane, cependant, en particulier quand on sait que son efficacité vaut $e = \frac{2}{\Pi} = 0.637$.

En se basant sur les différents critères pour apprécier la robustesse d'un estimateur, nous voudrions que celui-ci ait :

- une *petite* sensibilité de l'erreur grossière $\gamma^*(T,F) = sup_{x \in \chi | IF existe} |IF(x,T,F)|$,
- un point de rejet $\rho^* = inf\_r > Or, IF(x;T;F) = 0; quand \ \ |x| > r \ \ fini$,
- une efficacité e *proche de* 1,
- une *petite* sensibilité de position $\lambda^* = sup_{x \neq y} \frac{\|IF(y;T;F) - IF(x;T;F)\|}{\|y - x\|}$.

2.3 Valeurs Aberrantes dans le cas multivarié

Les principes et méthodes de détecion ne sont pas aussi directs dans le cas multivarié que dans le cas univarié, dû au fait que pour $\mathbf{x}$ p-dimensionnel, il n'y a pas d'ordre sur les observations. Si, dand le cas bivarié, un certain éloignement d'une observation, du reste des données est encore perceptible, au dela de la dimension 3, il n'est plus possible de cerner cet "éloignement" d'un point de la masse des données. Il y a, quand même, différents types de sous-ordre qui peuvent être définis.
On rappelle que dans le cas univarié, ou avait besoin de :

- Détecter les valeurs aberrantes dans les données en termes de leur éloignement relativement à un modèle de base F
- Faire des tests de discordances (% modèle de base F) ou utliser des méthodes inférentielles robustes pour l'accommodation.

Dans le cas multivarié, par contre, on a besoin d'une étape antérieure à ces deux options précédentes, à savoir adopter un principe de sous-ordre approprié pour exprimer l'éloignement d'une observation du reste des données, il y en a essentiellement quatre :

1. Sous-ordre marginal,
2. sous-ordre *réduit*,
3. sous-ordre partiel,
4. sous-ordre conditionel.

Le 2^{eme} type de sous-ordre est souvent le seul principe utilisé. Il est basé sur la transformation de toute observation multivariée $\mathbf{x}$ (de dimension p) en une quantité scalaire $R(x)$ et on ordonne l'échantillon en termes de $R_j = R(x_j) \quad (j = 1, \ldots, n)$. Ainsi, le principe d'un test de discordance est le même que dans le cas univarié. Une question pertinente se pose alors : comment choisir la mesure R ?
Dans le cas où F est le modèle normal $N($
$textbfmu, \Sigma)$,on utilise la forme quadratique $R(x, \mu, \Sigma) = (x - \mu)^{'}\Sigma^{-1}(x - \mu)$, qui a visiblement un intérêt certain (dans le cas normal pour F) : cette quantité est pertinente en termes d'ellipsoides de probabilité.
Comment juger de l'"éloignement" d'une observation $\mathbf{x}$ et de son caractère d'aberration ?
Il y a assentiellement deux principes à cet effet :
`Principe A :`
L'observation la plus extrême est celle, $\mathbf{x}_i$, dont l'omission de l'échantillon $\{x_1, \ldots, x_n\}$

donne le plus grand accroissement dans la vraisemblance maximisée sous le modèle F pour le reste des données.
Si cet accroissement est étonnament très grand, on déclare x_i comme observation aberrante.
`Principe B :`
Soit $\overline{F}$ un modèle de contamination pour F (i.e $\overline{F} = (1-\varepsilon)F + \varepsilon F_1 \ \ (F_1 \approx F), \varepsilon$ assez petit,> 0.)
La plus extrême des observations est celle, $\mathbf{x}_i$, dont la considèration en tant que contaminant dans le sens du modèle $\overline{F}$, maximise la différence entre les (logarithmes des) vraisemblances de l'échantillon sous $\overline{F}$ et F.
Si cette difference est étonnament grande, déclarer $\mathbf{x}_i$ comme étant une valeur aberrante.
En analyse multivariée, des formes modifiées d'analyse ont été suggérées qui prodiguent une certaine protection contre la présence de données aberrantes ; on citera essentiellement :

- Campbell [1980] qui propose une analyse en composantes principales (ACP) robuste ; on citera aussi Critchley [1985] qui montre comment les composantes principales peuvent révéler la présence d'une valeur aberrante.
- Campbell [1982], adoptant une même démarche que précédemment mais pour une analyse canonique qui est résistante à la présence d'outliers, ce qui'il illustre à l'aide d'un exemple emprunté aux sciences biologiques ;
- le problème d'accommodation d'outliers en analyse discriminante a aussi été considéré par Campbell [1987], où l'effet d'un contaminant est examiné au moyen d'une distance de Mahalanobis. Les fonctions d'influence sont aussi utilisées pour déterminer des critères de détection d'outliers ;
- une approche à l'analyse discriminante linéaire est considérée par Ganeshanandam et Krzanowski [1989] ;
- une approche robuste de l'analyse de la covariance est décrite par Birch et Myers [1982] ;
- l'influence des outliers sur la stabilité des facteurs en analyse des corespondances, est considérée par Escoffier et LeRoux [1976] tandis que Tanaka et Odaka [1989] proposent une méthode itérative de détection d'observations *influentes* ;
- une procédure d'échelonnement multidimentionnel prouvée comme étant très resistante aux outliers (et examinée par des simulations) a été considérée par Spence et Lewandowski [1989].

2.4 Méthode de Werner [2003]

On décrit succintement les différentes étapes de la méthode due à Werner ([2003]) dans sa thèse.

- Calcul du poids de chaque point en ajustant chaque composante séparément.
- Si le poids d'une observation est inférieur à une valeur "*cutoff*", réaffecter ce poids à 0 sinon lui affecter la valeur 1.
- Option MINWT : Choisir le poids minimum des composantes de chaque observation comme le poids des (p-1) autres composantes.
- Calcul de la moyenne et de la covariance pondérées de toutes les observations qui s'approchent au mieux (au sens d'estimation robuste) de la moyenne et de la matrice de covariance des observations intiales.
- Calcul d'une distance de Mahalanobis robuste pour toutes les n observations, en utilisant la moyenne et la covariance de l'étape précédente.
- Calcul de la densité de ces distances de Mahalanobis, en utilisant pour densité celle de la loi normale et une longueur de fenêtre $h = 0.9\mathtt{a}n^{-1/5}$ où $\mathtt{a}$=min{écart type, rang interquartile/1.34}
- Il y aura un pic décrivant les *inliers*, où la pente a une grande valeur positive suivie d'une grande valeur négative et croît vers 0. Trouver la valeur de la distance de Mahalanobis où cette pente négative entre pour la 1^{ere} fois dans un intervalle de confiance contenant 0. Soit noté M_0 ce point de rejet, la valeur de la distance de Mahalanobis où la pente est suffisamment proche de 0.
- Classer toutes les observations comme *inliers* ou *outliers* selon que leur distance de Mahalanobis est respectivement plus petite ou plus grande que la valeur de M_0 de l'étape précédente.

Chapitre 3

Application en R

3.1 Choix de R pour le calcul pratique

R est un système interactif et convivial destiné aux scientifiques. Il est communément qualifié de langage et de logiciel ; il permet de réaliser différentes analyses statistiques. Initialement, R a été écrit par Ross Ihaka et Robert Gentleman au département de statistique de l'Université d'Auckland (Nouvelle Zélande.) Il est à noter que chacun des prénoms des auteurs commence par la lettre "R", et c'est en leur honneur qu'il a ainsi été baptisé. Par la suite, de nombreuses personnes qui s'intéressent à l'informatique et à la statistique mathématique ont contribué au développement de R.

Notons que R est un logiciel libre, donc son code source est disponible au public. Il est en priorité tourné vers l'utilisateur. L'échange intense entre le nombre très important d'utilisateurs et de développeurs, fait que le code de R est sans cesse amélioré. Ainsi, il est stable, sûr et constamment en avance sur les logiciels commerciaux ayant les mêmes objectifs.

Il fournit également des outils en bioinformatique, statistique multivariée, modèles graphiques, ainsi qu'en finance.
Dans R, on a aussi la possibilité de traiter des données stockés dans des fichiers textes.

Cette énumération non exhaustive des outils de R fait de ce langage un environnement adéquat pour un système statistique. Pour des applications statistiques particulières, des packages de base sont fournis. Notons que le nombre de packages de base est actualisé généralement chaque fois qu'une nouvelle version de R apparaît.

Notons également que pour des applications spécifiques, il est intéressant d'enrichir la librairie de R par des packages développés à cet effet, à recueillir généralement du site Internet : `http ://CRAN.R-project.org/`.

La réunion de toutes ces fonctionnalités fait de R l'environnement idéal des calculs et de diverses analyses scientifiques approfondies.

3.2 Les données

La présence de valeurs aberrantes et leur prise en compte sont des considérations importantes en statistique.
Nous présentons un exemple qui illustre l'influence de quelques données.
Dans l'analyse suivante nous demontrons qu'avec deux valeurs aberrantes incluse dans l'ensemble de données de 48 observation, seulement 15% de la variation dans la variable dépendante est calculé par les différences sur la variable independante.
Cependant, quand les deux valeurs aberrantes sont elvées, 48% de la variation est calculé pour $(r = .69, r^2 = .48, N = 46)$.

Les données proviennent d'une étude portant sur des institutions métropolitaines (universités et collèges), menées par le bureau de l'université du Nord Texas.
Les institutions sont classées avec des critères d'admissions (de celles où les insccriptions sont ouvertes à celles avec des critères sélectifs). La variable indépendante est le score moyen de l'institution (SAT) pour les nouveaux étudiants, et la variable dépendante (GRADRA) est le taux de graduation de l'institution (sur une durée de six ans).
Comme on pouvait anticiper, il y a un *fort* rapport linéaire entre les deux variables (SAT et GRADRA).
L'objectif principal de cette étude est d'appliquer et d'illustrer certaines des différentes méthodologies vues précédemment. (Voir le Tableau 3-1)

Tab. 3.1 – GRADRA vs.SAT

case	SAT	GRADRA	case	SAT	GRADRA
1	1152.00	74.40	25	921.00	28.00
2	1121.00	69.00	26	919.00	44.00
3	1099.00	69.00	27	918.00	36.00
4	1069.00	39.00	28	917.00	46.50
5	1060.00	68.00	29	900.00	50.00
6	1050.00	53.50	30	892.00	51.00
7	1044.00	34.00	31	890.00	29.00
8	1028.00	41.80	32	885.00	25.40
9	1027.00	49.00	33	876.00	31.00
10	1026.00	30.00	34	873.00	44.00
11	1025.00	47.00	35	866.00	41.00
12	1019.00	69.00	36	857.00	23.00
13	1009.00	46.00	37	855.00	39.00
14	1006.00	50.00	38	846.00	37.00
15	1004.00	48.00	39	831.00	23.00
16	1000.00	27.00	40	809.00	32.00
17	1000.00	45.00	41	806.00	12.00
18	998.00	64.00	42	799.00	27.00
19	980.00	53.00	43	795.00	42.40
20	977.00	34.00	44	777.00	41.00
21	968.00	32.00	45	760.00	23.00
22	958.00	45.00	46	677.00	17.00
23	953.00	46.00	47	598.00	72.00
24	927.00	47.00	48	464.00	44.10

Le Score moyenne (SAT) a varié de 464 à 1152. Le taux de graduation rates (GRADRA) à varié de 12% à 74.4%. Dans cet exemple on veut appliquer deux outils importants dans la détection des valeurs aberrantes. La Distance de Mahalanobis et L'utilisation de Minimum Covariance Diterminant (MCD)

3.3 Utilisation de la Distance de Mahalanobis

On commence par présenter les données (avec : X=read.table("data.txt",header=T)) ; on calcule la moyenne (μ) et la matrice de covariance (Σ) qu'on insère dans la distance D^2 et puis on donne une présentation graphique pour situer les valeurs aberrantes.
Pour ce faire, on traite deux situations : la première situation considère les deux points 47 et 48 qui semblent extrêmes, et la seconde n'en tient pas compte.

3.3.1 1^{er} cas : Traitement avec les 2 cas :

Rapplons que la distance de Mahalanobis est donnée par

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

Nous calculons μ avec : mu=colMeans(X), et Σ (Matrice de covariance) avec : sx=cov(X)

```
> mu=colMeans(X)
> mu
      SAT    GRADRA
921.47917  42.48125
> sx=cov(X)
> sx
               SAT   GRADRA
SAT    17291.1059   761.1688
GRADRA   761.1688   219.9641
> D2 <- mahalanobis(X, colMeans(X), Sx)
> D2
 [1]  5.61527253  3.98926523  3.69881232  1.79225351  3.13254954  1.10941424
 [7]  1.90060708  0.81089470  0.66277971  2.19681295  0.61978101  3.19934513
[13] 0.44359457  0.49051284  0.41290515  2.28002153  0.36128859  2.10543481
[19] 0.53639703  0.81843702  0.96706850  0.08158798  0.08181999  0.09981099
[25] 1.12143172  0.01456791  0.21546706  0.09648552  0.41092093  0.56706764
[31] 0.84194858  1.36137679  0.60153059  0.20748350  0.18295989  1.72595569
[37] 0.25724335  0.35447166  1.76166676  0.89568057  4.23072512  1.41353706
[43] 1.08659509  1.33488362  2.32905883  4.61863151  16.32106282 14.64258078
> plot(X,ylab="",xlab="Reported Average SAT")
```

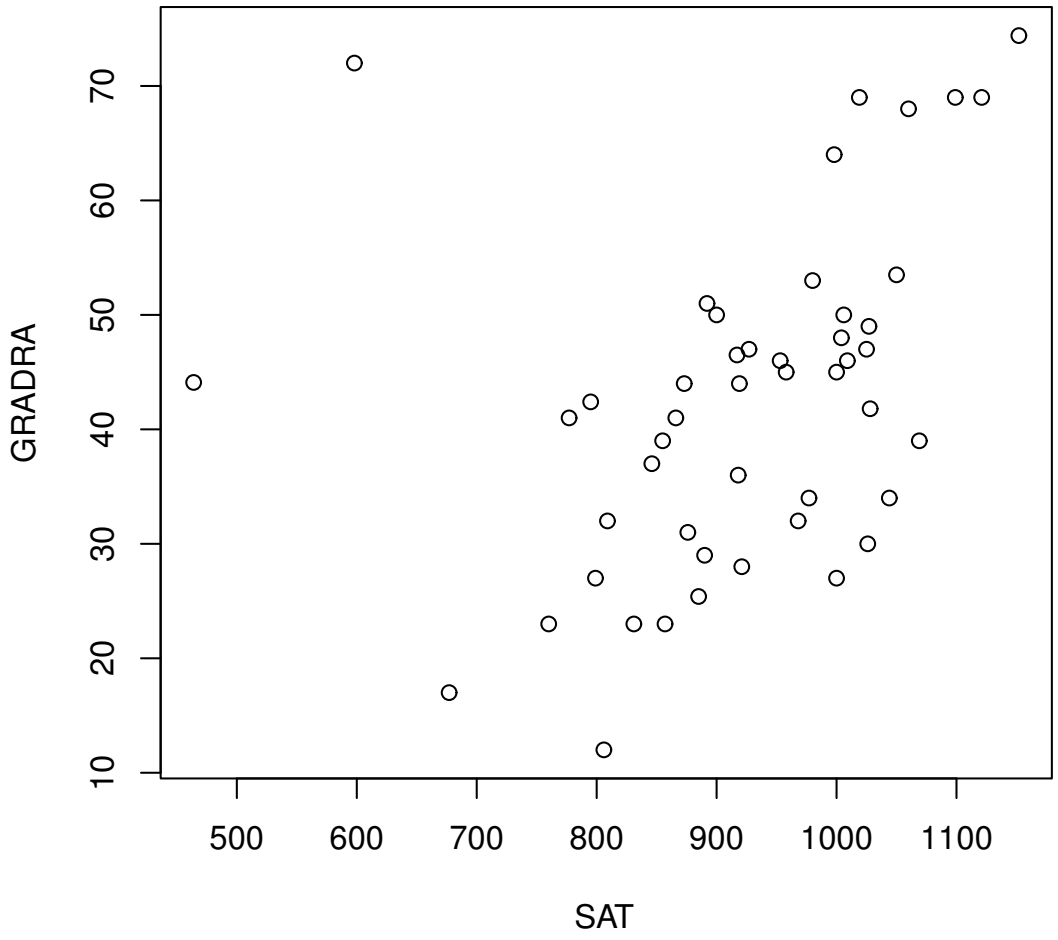

FIG. 3.1 – Nuage de points dans $\mathbb{R}^2$

Fig1 : Basée sur les 48 cas (incluant les deux cas 47 et 48).
Les valeurs de SAT et de GRADRA sont portées sur le graphique. Noter que les valeurs de SAT =464 et GRADRA=44.1% et SAT=598 et GRADRA=72% (cas 47 et 48) semblent être des valeurs aberrantes -dans le coin gauche supérieur du graphe.
Pour voir plus clairement la figure, on dessine un ellipsoide -dit de tolérance, et voici les commandes R qu'il faut utiliser.

```
> tolEllipsePlot(X, m.cov = covMcd(X), cutoff = NULL, id.n = NULL,
+ classic = TRUE, tol = 1e-07,
+ xlab = "", ylab = "",
+ main = "Tolerance ellipse (97.5%)",
+ txt.leg = c("robust", "classical"),
+ col.leg = c("red","blue"),
+ lty.leg = c("solid","dashed"))
```

X : la matrice des données de dimensions $2x2$
m.cov : un objet similaire à ceux de la classe "mcd". Cependant, seules les 2 composantes center (pour la moyenne) et cov (pour la matrice de covariance) sont utilsées.
cutoff : Valeur numerique au dela de laquelle des points-observations sont classés en dehors de l'ellipse.
id.n : Nombre d'observations à identifier par *label*. S'il n'est pas précisé, le nombre d'observations avec une distance de Mahalanobis plus grande que "cutoff" est utlisé.
classic : pour dessiner aussi les distances classiques. Sa valeur par défaut est classic=FALSE
tol : la tolérance à utiliser pour calculer l'inverse, (voir Solve). Par défaut, tol=$1e^{-7}$
txt.leg, col.leg, lty.leg : Vecteurs de caractères de dimension 2 pour la légende, utlisée seulement si classic=TRUE.

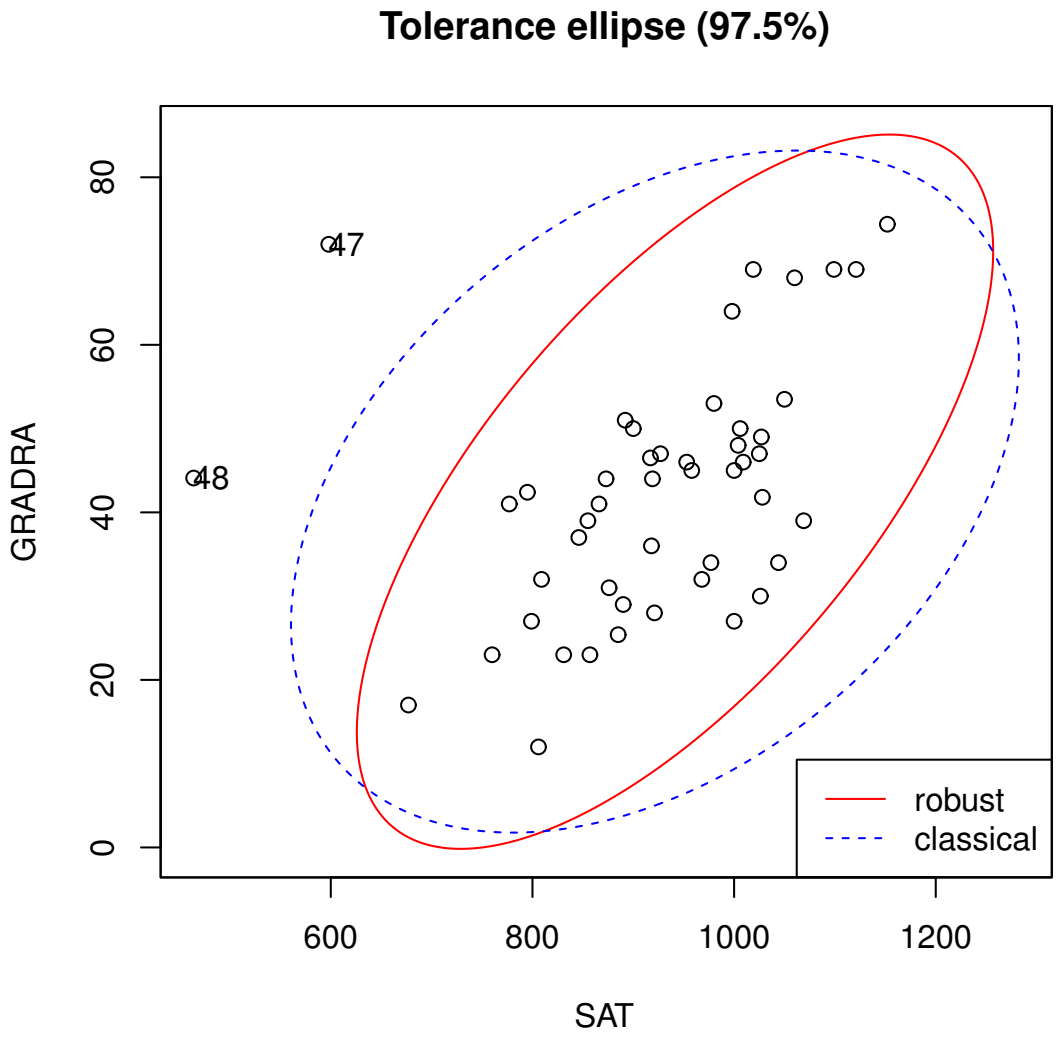

FIG. 3.2 – Ellipsoide de Probabilités

Nous déduisons clairement (comme cela est visible à l'oeil nu) que les observations 47 et 48 sont des valeurs aberrantes.

3.3.2 2^{eme} cas : Traitement sans les deux cas 47 et 48

```
> mu=colMeans(y)
> mu
      SAT    GRADRA
938.45652  41.80435
> sx
             SAT     GRADRA
SAT      10788.831  1035.3957
GRADRA   1035.396   209.8502

> D2 <- mahalanobis(y, colMeans(y), Sx)
> D2
 [1] 5.55228636 3.93617465 3.64677093 3.70733809 3.28046183 1.16211207
 [7] 3.94334341 1.41224908 0.74201117 4.40568425 0.78175041 4.03095932
[13] 0.52123717 0.44943242 0.39826553 4.23334638 0.41756953
2.78719695
[19] 0.63031724 1.33539742 1.52689847 0.05117463 0.09056148 0.37084714\\
[25] 1.35978589 0.18449383 0.17233089 0.45565076 1.41584980 1.88746091 \\
[31] 0.81942310 1.41532040 0.57100598 1.04760619 0.82886017 1.70760220\\
[37] 0.89077769 0.94214961 1.72294831 1.61547102 4.27052882 1.82088765\\
[43] 3.77472261 4.36954430 2.97730929 6.33688467
> plot(y,ylab="", xlab="Reported Average SAT")
```

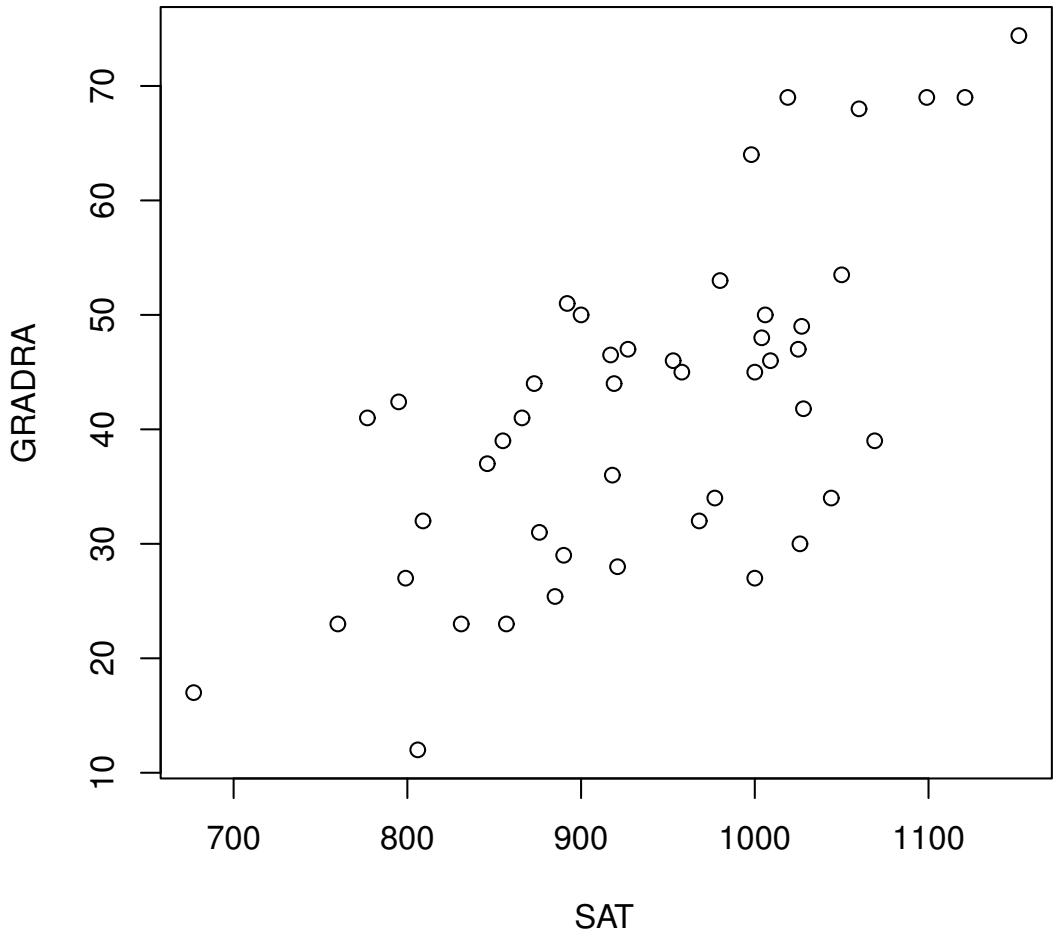

FIG. 3.3 – Nuage de points dans $\mathbb{R}^2$ sans les points 47 et 48

Pour voir plus clairement la figure, on dessine un ellipsoide -dit de tolérance, avec les commandes R suivantes.

```
> tolEllipsePlot(y, m.cov = covMcd(y), cutoff = NULL, id.n = NULL,
+ classic = TRUE, tol = 1e-07,
+ xlab = "", ylab = "",
+ main = "Tolerance ellipse (97.5%)",
+ txt.leg = c("robust", "classical"),
+ col.leg = c("red","blue"),
+ lty.leg = c("solid","dashed"))
```

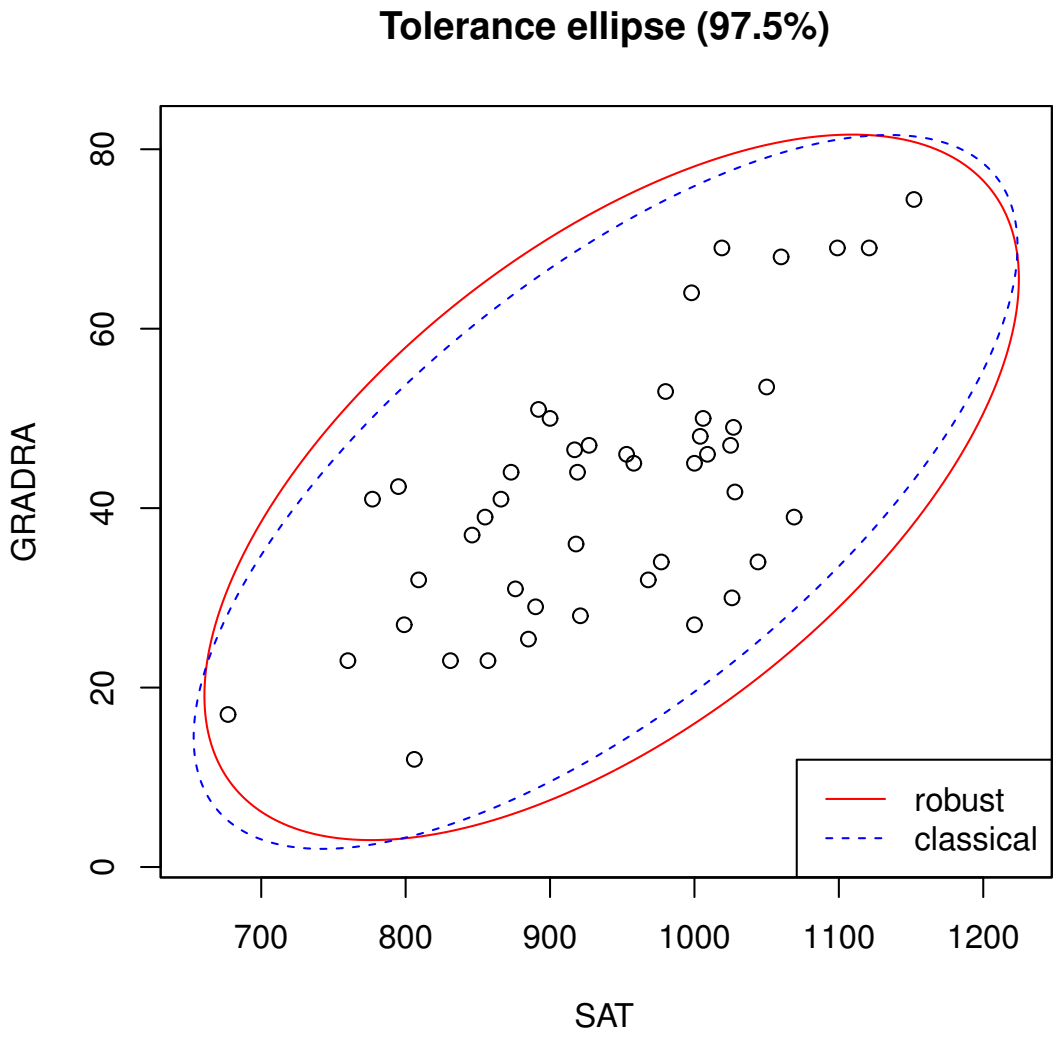

FIG. 3.4 – Ellipsoide de probabilité sans les points 47 et 48

Dans la fig 4, nous voyons qu'aucun point n'est en dehors de l'ellipse, ce qui veut dire qu'il n'y a aucune valeur qui peut être aberrante, donc à détecter.

3.4 Utilisation du MCD

L'objectif est ici de calculer le déterminant de la matrice de covariance à chaque étape d'un processus itératif et choisir la matrice de covariance de plus petit déterminant ; pour cela on considère les deux cas traités précédemment.

3.4.1 1^{er} cas : Traitement avec les 2 points 47 et 48

```
> library(robustbase)
> library(MASS)
> covMcd(X)
Minimum Covariance Determinant (MCD) estimator. Call:
 covMcd(x =X)
 -> Method:  Minimum Covariance Determinant Estimator.

Log(Det.):  11.85
Robust Estimate of Location:
   SAT  GRADRA
941.40   42.47
Robust Estimate of Covariance:
          SAT  GRADRA
SAT     13216  1202.9
 GRADRA   1203   241.2

> mcd <- covMcd(log(X))
> plot(mcd, which = "distance", classic = TRUE)
> plot(mcd, which = "dd")
> plot(mcd, which = "tolEllipsePlot", classic = TRUE)
> op <- par(mfrow = c(2,3))
> plot(mcd)
```

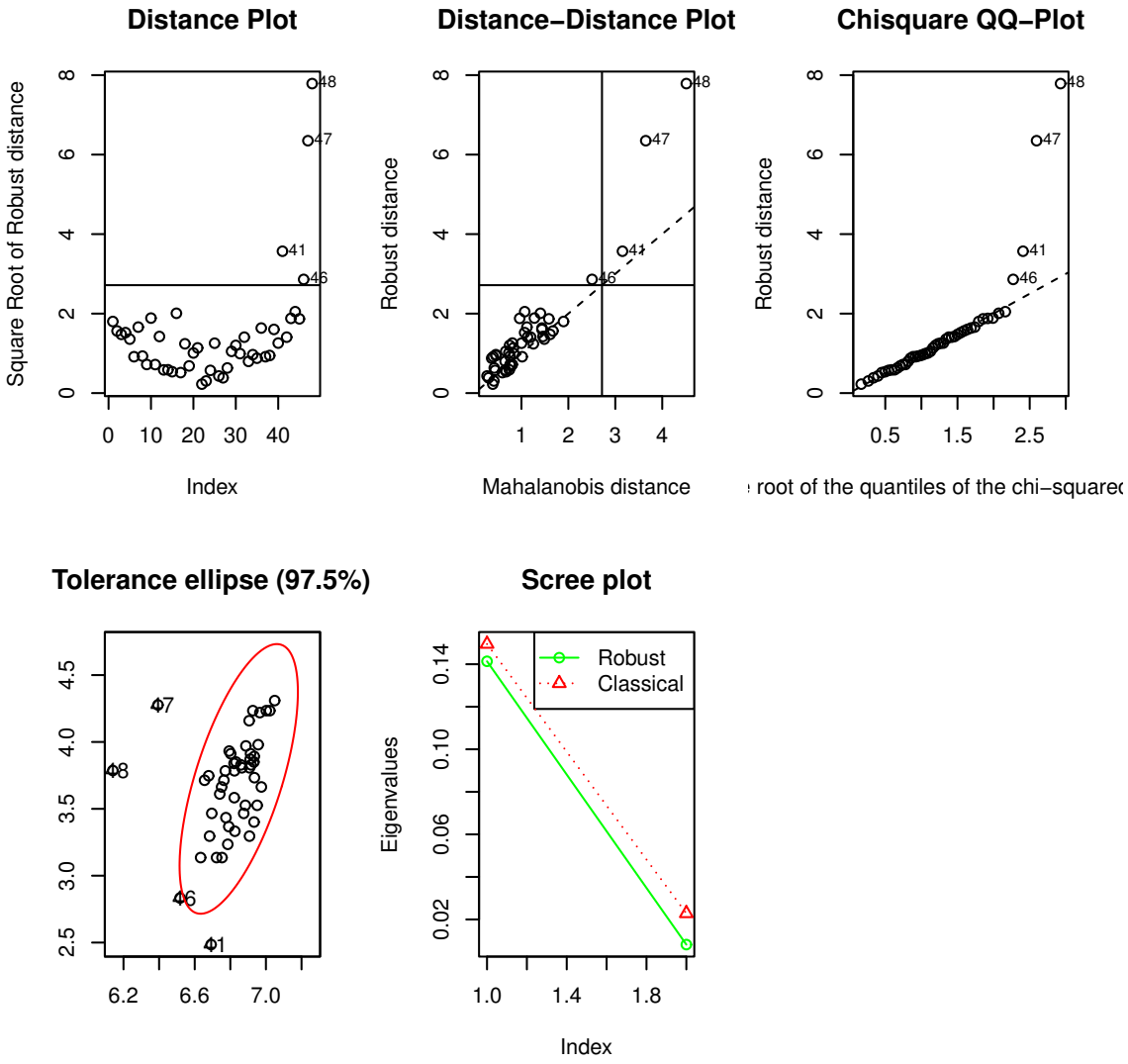

FIG. 3.5 – MCD avec les 2 points 47 et 48

Le DD-plot, qui a été introduite par Rousseeuw et Van Zomeren (1990), montre les distances robuste contre les distances classique de Mahalanobis, les lignes horizontale et verticale sont tracées à la valeur "*cutoff*" dont les valeurs par défaut sont les racines carrées du quantil 97.5% de la distribution chi-deux avec p degrés de liberté. Des points au dela de ces lignes peuvent être considérés comme des valeurs aberrantes.

3.4.2 2^{eme} cas : Traitement sans les deux cas 47 et 48

```
> library(robustbase)
> library(MASS)
> covMcd(y)
Minimum Covariance Determinant (MCD) estimator. Call:
 covMcd(x =y)
-> Method:  Minimum Covariance Determinant Estimator.

Log(Det.):  11.65
 Robust Estimate of Location:
   SAT  GRADRA
942.65   42.32
Robust Estimate of Covariance:
            SAT  GRADRA
SAT     10544.6   868.9
GRADRA    868.9   205.0
```

Après le calcul, on trouve dans le 1^{er} cas que le déterminant minimum est égal à 11.84 et dans le 2^{eme} cas, il est égal à 11.65.

```
> mcd <- covMcd(log(y))
> plot(mcd, which = "distance", classic = TRUE)
> plot(mcd, which = "dd")
> plot(mcd, which = "tolEllipsePlot", classic = TRUE)
> op <- par(mfrow = c(2,3))
> plot(mcd)
```

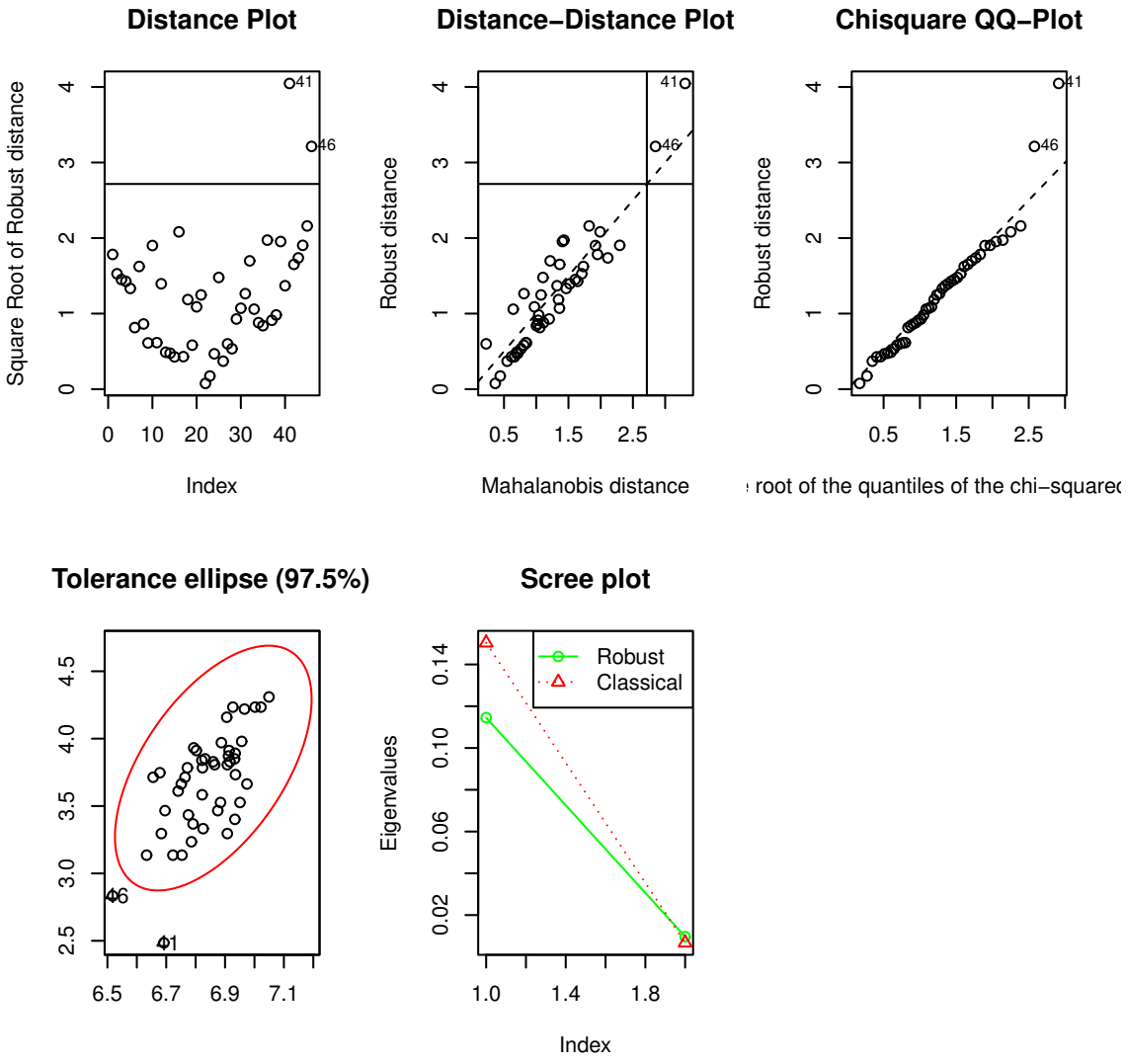

FIG. 3.6 – MCD sans les points 47 et 48

Conclusion Générale

Au terme de ce travail, l'on perçoit bien que l'analyste des données, autant le statisticien que le praticien dans quelque domaine qu'il soit, ne peut pas se passer de la considération de l'une ou l'autre des méthodes de traitement des observations atypiques s'il veut vraiment se prémunir des erreurs qui peuvent survenir de la présence de ces dernières parmi quelques-unes de ses données.

Et, bien sûr l'avancée technologique aidant, via les moyens de calculs auxquels est parvenu le monde d'aujourd'hui, aussi bien en industrie, en ingénierie qu'en médecine ou même les sciences sociales, voire des études touchant directement les activités de l'être humain au quotidien, il y a tout lieu de "filtrer" ses données avant toute tentative de conclusion ou de recommandation pour d'autres éventuelles considérations de données du même type par quelque procédure adéquate.

On signalera quand même qu'il n'y a pas de prévalence d'une méthode ou de quelque outil donnés dans toutes les situations en présence de données aberrantes.

Il reste quand même qu'il y a certaines situations où l'influence de ces valeurs atypiques est réduit grâce à l'utilisation de méthodes ou d'outils statistiques robustes.

A ne pas manquer de signaler la complexité grandissante en présence de données incomplètes ou *manquantes*, le statisticien étant alors confronté au double problème de données *manquantes* et de données *aberrantes* !

Ce modeste travail ne prétend pas faire le tour complet de la question des valeurs atypiques, du fait que le problème des outliers continue de susciter de l'intérêt dans divers domaines où l'homme est appelé à analyser des données concrètes.

On n'ommettra pas de citer l'itéressement du statisticien au cadre des tables de contingence, de l'analyse des séries chronologiques, de méthodes bayésiennes ainsi que les plans d'expérience -très utilisés dans les sciences agronomiques par exemple.

La notion d'*outlier* continue d'occuper une place prépondérante dans le développement de nouvelles méthodologies en statistique, notamment multivariée. Il reste à dire, comme

le rappelle Barnett et Lewis, que le statisticien a besoin de plus de compréhension sur la manière de traiter les observations aberrantes.

Bibliographie

[**1**] Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size.Proc. Roy. Soc. London, Ser. A, 353 (1674), 401-419.

[**2**] Barnett V. & Lewis T. [1994] : "*Outliers in Statistical Data*" New York, John Wiley, 3^{rd} Ed.

[**3**] Becker Claudia & Ursula Gather [2001] : "*The Largest nonidentifiable Outlier : a comparison of Multivariate Simultaneous Outlier Identification rules*" 119-127

[**4**] Beckman,R.J. et Cook, R.D. (1983). Outlier.....s' (with Discussion). Technometrics, 25, 119-163.

[**4**] Beirlant, J., Teugels, J.L. et Vynckier, P. (1996). Practical analysis of extreme values. Leuven, Leuven University Press, 137 pp.

[**5**] Brazauskas, V. et Serfling, R.Small sample performance of robuste estimators of tail parameters for Pareto and exponential models. J. Stat. Comput. Simul., 70, 1-19.

[**6**] Campbell NA. (1978). The influence function as an aid in outlier detection in discriminant analysis. Appl. Stat. 27, p. 251–258.

[**7**] Campbell NA. (1980). Robust procedures in multivariate analysis. I. Robust covariance estimation. Appl. Stat. 29, p. 231–237.

[**8**] Campbell NA. (1982). Robust procedures in multivariate analysis. II. Robust canonical variate analysis. Appl. Stat. 31, p. 1–8.

[**9**] Carletti,G. (1988). Comparaison empirique de méthodes statistiques de détection de valeurs anormales à une et à plusieurs dimensions. Gembloux, Belgique, Fac. Univ. Sci. Agron., Thèse de doctorat, 225 pp.

[**10**] Cédric Béguin & Beat Hulliger [2008] : "*L'algorithme BACON-EEM pour la détection d'observations aberrantes multivariées dans des données d'enquête incomplètes*"

[**11**] Cook,R.D. et Weisberg, S. (1980). Characterisations of an empirical influence function for detecting influential cases in regression. Technometrics, 22, 495-508.

[**12**] Dagnelie,P. (2003). Communication personelle.

[**13**] Dagnelie, P. (1998a). Statistique théorique descriptive et bases de l'inférence statistique. Paris et Bruxelles, De Boeck et Larcier, 508 pp.

[**14**] Essenwanger, O.M. (1986). Elements of statistical analysis. World Survey of Climatology, General Climatology.Amsterdam,Elsevier,vol.1B,424 pp.

[**15**] Everitt, B.S. (2002). The Cambridge dictionnary of statistics. Cambridge, UK, University Press, Second Edition,410 pp.

[**16**] Ferguson, T.S. (1961). Rules for rejection of outliers. Rev. Inst. Int. Stat. 29 (3), 29-43.

[**17**] Garrido, M. (2002). Modélisation des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution. Grenoble, Université Grenoble 1, Mathématiques appliquées, 231 pp.

[**18**] Grubbs,F.E.(1969). Procedures of detecting outlying observations in samples. Technometrics, 11, 1-21.

[**19**] Johnson, N.L. et Kotz, S. (1970). Continuous univariate distributions-1. Distributions in statistics. Boston, Houghton Mifflin company 300 pp.

[**20**] Karl Ho and Jimmie R. Naugher [2000] : Outlier Lies : An Illustrative Example Of Identifying Outliers and APPlying Robust Methods,University of North Texas,vol.26(2)

[**21**] Mark Werner [2003] : Identification Of Multivariate Outliers In Large Data Sets

[**22**] Munoz-Garcia, J., Moreno-Rebollo, J.L. et Pascual-Acosta, A. (1990). Outliers : A formal approach. Int. Statist. Rev. 58, 215-226.

[**23**] Planchon V. [2007] : "*Détection de valeurs aberrantes dans des mélanges de distributions dissymétriques pour des ensembles de données avec contraintes spatiale*",Page 11-89

[**23**] Planchon V. [2005] :Traitement des valeurs aberrantes : concepts actuels et tendances générales Biotechnol. Agron. Soc. Environ. 2005 9 (1), 19–34

[**24**] P.C. Mahalanobis, "`On the generalized distance in statistics`, Proceedings of the National Institute of science of India vol 2 (1936) 49-55

[25] R :Mahalanobis Distance (`stat.ethz.ch/R-manual/R.../mahalanobis.html`)

[**26**] Rousseeuw & Van Driessen [1999] : A fast algorithme of the minimum covariance determinant estimatore.Technometrics, Vol.4,212-223.
(cf. aussi : http ://win-www.uia.ac.be/u/statis/)

Sites Internet utilisés

[**27**] http
www.cran.r-project.org/web/packages/.../robustbase.pdf

[**28**] http
www.r-bloggers.com/mahalanobis-distance-with-r.