REPUBLIQUE ALGERIENNE DEMOCRATIQUE et POPULAIRE.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

UNIVERSITE ABDERRAHMANE MIRA de BEJAIA
Faculté des Sciences
Département de Mathématiques

Mémoire de Magistère

en

Mathématiques

Option

Analyse mathématique et applications

Thème

Modélisation de séries temporelles en présence d'outliers

Présentée par

Hadjila Tabti

Devant le jury composé de :

Ahmed Ait saidi	Maître de conférences	Université de Béjaïa	Président
Hamid Louni	Maître de conférences	UMMTO	Rapporteur
Megdouda Ourbih	Maître de conférences	Université de Béjaïa	Examinateur
Rachid Benabidallah	Maître de conférences	UMMTO	Examinateur

Remerciements

Je tiens à exprimer toute ma gratitude à Hamid Louni, Maître de conférences à l'Université Mouloud Mammeri, sans qui ce mémoire n'aurait pu exister. Ses nombreux conseils et sa grande disponibilité m'ont été un soutien de tous les instants. Je ne saurais lui dire ma reconnaissance pour la confiance et la liberté qu'il m'a accordées dans la réalisation de ce travail.

Ce travail a été réalisé pour une large part à l'UMMTO. D'abord la première année de la post-graduation où j'ai eu la chance d'avoir comme enseignant, le Professeur Mohand Morsli, les maîtres de conférences Fazia Khellas, Rachid Benabidallah et Hamid Louni de qui j'ai appris la rigueur, le gout de l'effort et la tenacité. Aussi la qualité des enseignements prodigués m'ont permis par la suite d'aborder sans trop de difficulté ce sujet de mémoire. Je tiens à les remercier.

Je remercie Ahmed Ait Saidi, Maître de conférences à l'université de Béjaïa, pour m'avoir fait honneur de présider ce jury.

Je remercie vivement mes rapporteurs Megdouda Ourbih, Maître de conférences à l'université de Béjaïa, et Rachid Benabidallah, Maître de conférences à l'UMMTO, d'avoir pris le temps de lire, juger et apprécier ce mémoire et le travail qu'il représente. J'ai grandement apprécié l'intérêt qu'ils ont porté à mon travail ainsi que leurs remarques constructives.

Je veux aussi remercier le Professeur Abdelnasser Dahmani et les maîtres de conférences Louiza et Mohand Bouraïne pour leurs aides. Ainsi que tous les enseignants du département de mathématiques de l'université de Abderrahmane Mira de Béjaïa.

Ensuite, je remercie vivement Ahcen, Amina, Dalila, Djaouida, Fadida, Farida, Nouara et Sonia, qui ont vecu autant que moi les moments de stress, pour leurs conseils.

Grand merci pour Lies et messaoud, pour leurs nobles gestes.

Enfin, Je remercie mes très chèrs parents, mes frères et mes soeurs pour la confiance et le soutien qu'ils m'ont accordé.

Resumé : Les outliers sont de plus en plus étudiés dans la littérature statistique sur les séries temporelles, et cet intérêt va en croissant en économétrie. Une synthèse sur la riche variété des méthodes de traitement des outliers dans les séries temporelles stationnaires en rapport à leur nature et les buts de l'investigation constitue la première préoccupation de ce travail. La deuxième accorde un intérêt particulier aux procédures itératives de modélisation ARMA. Ce sont des techniques de modélisation des outliers basées sur le test de Fox (1972) et sur l'analyse avec intervention de Box et Tiao (1975). Le test qui donne en même temps la position et le type d'intervention est utilisé comme une partie d'une stratégie complète de la modélisation des outliers. Louni (2008) a développé un test de détection de deux types d'outliers (AO et IO) qui semble bien fonctionner dans beaucoup de situations pratiques, notamment quant il est utilisé itérativement. Reprendre la procédure de Chang et al. (1988) pour identifier et corriger les deux types d'outliers en s'appuyant sur ce nouveau test et procéder ensuite à la comparaison des performance reste la principale préoccupation de ce travail.
mot clé :AO et IO outliers, Test de score, Méthode itérative.

abstract :The outliers are increasingly considered in the statistical literature on time series, and this interest is increasing in econometrics. A synthesis of the rich variety of methods for treating outliers in time series stationary in relation to their nature and purposes of the investigation is the first concern of this work. The second gives a particular interest in iterative procedures for ARMA modeling. These are techniques for modeling outliers based on the test of Fox (1972) and analysis with the intervention of Box and Tiao (1975). The test gives in the same time the position and type of intervention is used as part of complete strategy for the modeling of outliers. Louni (2008) developed a test for two types of outliers (AO and IO) that seems to work well in many practical situations, especially when it is used repeatedly. Repeat the procedure of Chang et al. (1988) to identify and correct both types of outliers based on this new test and then proceed to compare the performance remains the main concern of this work.
key words : AO and IO outliers, scor test, iterative procedure

Table des matières

Introduction générale

Dans la constitution de grands ensembles de données les valeurs aberrantes (“outliers ” en anglais) peuvent plus facilement passer inaperçues. La présence de valeurs aberrantes peut alors 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. De nombreux articles parlent pour l'identification des outliers en termes de « valeurs douteuses », « discordantes », « contaminées », « extrêmes », pour ne citer que peu de mots, mais il n'a pas de définition accéptée sur ce que constitue une valeur aberrante. De plus, du fait que les différentes situations et buts exigent différents modèles et descriptions, plusieurs approches existent pour identifier une observation comme valeur aberrante. Ces aspects seront abordés dans la première partie de ce travail avec un éclairage sur le traitement des valeurs aberrantes dans les données statistiques.

Nous commencerons par discuter dans le cadre général des données statistisques d'une variété d'idées générales que nous avons trouvé utiles quand on essaye de trier l'abondante et souvent confuse littérature sur les outliers. L'idée d'un outlier, les causes, les raisons d'étude et les approches générales du problème des outliers sont discutés dans le premier chapitre. Dans le deuxième, des définitions formelles et une classification des outliers dans le contexte des séries temporelles sont décrites. Leurs effets dans le cas des modèles autorégressifs moyennes mobiles (ARMA) sont aussi discutés.

Tests de détection des outliers et estimateurs robustes dans le cas des modèles ARMA feront l'objet du troisième chapitre. Le test du rapport de vraisemblance habituellement utilisé dans la phase l'identification des outliers des procédures itératives avec interventions est exposé dans le détail. Tout comme est exposé une version du test de score pour la détection introduit dans Louni (2008). Les estimateurs des moindres carrés, les M-

estimaeurs et GM-estimateurs seront aussi présentés.

Enfin dans le dernier chapitre, il est question de modélisation de séries temporelles en présence d'outliers. Après l'introduction de l'analyse avec interventions à la base de nombreuses procédures de modélisation des outliers, la procédures itérative de Chang *et al* (1988) pour identifier et corriger deux types d'outliers est considérée. Celle-ci met jeu dans la phase d'identification des outliers le test du rapport de vraisemblance cité ci-haut. Après quoi nous montrons comment on peut améliorer les performances de cette procédure en substituant à ce test le test basé sur les scores. Une autre voie est possible : ignorer les outliers et utiliser des procédures d'estimations qui résistent à leur influence. Pour donner plus de relief à la procédure itérative, comme Chang *et al*, il aurait fallu implémenter de telles procédures robustes pour établir des comparaisons entre les deux méthodes. Cette implémentaion dépasse largement le cadre de ce mémoire. Nous nous sommes alors appuyés sur les résultats donnés dans Chang *et al* pour établir une comparaison entre les deux voies. Cette comparaison montre la supériorité de la procédure itérative sur les procédures d'estimations robustes des modèles ARMA.

Chapitre 1

Outliers dans les données statistiques

1.1 Introduction

Les observations *non représentatives* ou *aberrantes* ont toujours été considérées comme source d'une contamination, déformant l'information obtenue à partir des données brutes. 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 en adoptant des méthodes qui diminuent leur impact au cours des analyses statistiques (Barnett et Lewis, 1994).

L'étude des outliers n'est pas un nouveau phénomène. En effet, elle a une longue histoire, remontant au tout début de l'analyse statistique. Depuis, un large éventail de méthodes d'analyses statistiques de plus en plus précises ont été construites pour tester des hypothèses concernant des paramètres déterminés ou pour estimer la validité de certains modèles. Cette précision dans l'élaboration et l'utilisation des méthodes statistiques nécessite une évaluation fiable de l'intégrité d'ensembles de données. Le problème des valeurs aberrantes est donc incontournable pour toutes personnes qui manipulent des données et doivent juger de la manière de traiter celles-ci.

Les valeurs aberrantes n'induisent pas forcément en erreur, elles ne sont pas forcément mauvaises ou erronées. Dans certains cas, l'expérimentateur peut être tenté de ne pas rejeter la valeur aberrante mais l'accepter comme une indication intérresante. Il n'est pas approprié d'adopter une attitude radicale, soit de rejet, soit d'inclusion systématique des

données aberrantes. La première attitude peut entraîner la perte d'informations réelles tandis que, dans le cas de l'acceptation des valeurs aberrantes, il y a risque de contamination. En fonction des circonstances, il existe des méthodes, dites robustes, qui prennent en compte toutes données mais minimisent l'influence des valeurs aberrantes.

Durant la dernière décennie, on a réellement pris conscience qu'avant tout traitement d'observations anormales, il faut prendre en compte diverses notions directement liées aux valeurs aberrantes (Barnett et Lewis, 1994). La manière d'appréhender ces valeurs est dès lors plus structuré. En effet, des distinctions bien claires entre les objectifs des analyses statistiques et la manière de considérer les données doivent être réalisées. Barnett et Lewis dressent une classification des types de questions auquelles il faut réfléchir lors de l'étude des valeurs aberrantes. D'après ces auteurs, il est nécessaire de faire la distinction entre les causes déterministes ou aléatoires d'apparition de valeurs aberrantes, entre les différents objectifs à atteindre lors de l'étude des valeurs aberrantes, entre les différents modèles de probabilités spécifiques, entre les données univariées et multivariées et enfin les valeurs aberrantes simples ou multiples. Seules les données univariées et les valeurs aberrantes simples seront abordées ici.

Dans la suite de chapitre, nous allons examiner les différentes manières d'aborder le problème des valeurs aberrantes en prenant en considération ces différentes notions.

1.2 Définitions

Avant d'exposer des concepts relatifs aux valeurs aberrantes, il est nécessaire de les définir de manière précise. Qu'appelle- t-on une valeurs aberrantes (“outliers ” en anglais) ? La réponse à cette question n'est pas évidente comme nous le verrons au cours de cette discussion.

Aucune observation ne peut être garantie d'être une manifestation totalement dépendante du phènoméne sous étude. Un événement avec une chance sur un million surviendrait avec la fréquence appropriée sans tenir compte du fait que nous soyons surpris. Intuitivement, la probable validité d'une observation est renvoyée par sa relation aux autres observations obtenues dans des conditions similaires. Les observations qui, dans l'opinion de l'investigateur,

se démarquent du groupe principal de données ont été appelées “outliers ”, “observations discordantes ”, “valeurs douteuses ”, “contaminants ”, “valeurs surprenantes ”, “données rebelles ”, “ valeurs sales ” pour ne mentionner seulement que le peu de termes qui ont été utilisés le long des années. Les chercheurs sont directement concernés quand de telles observations surviennent.

De nombreux auteurs ont cherché à décrire le terme de valeur aberrante et les définitions fournies ont évolué au cours du temps. Par exemple, Grubbs (1968) définit une valeur aberrante comme étant *une observation qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon dans lequel il apparaît.* 20 ans plus tard, Carletti (1988) s'intéresse aux *valeurs anormales* qu'il définit comme étant *une valeur qui paraît suspecte parcequ'elle s'écarte d'une façon marquée des autres valeurs de la variable étudiée ou ne semble pas respecter une norme ou une relation définie.* Ces énoncés, plutôt vagues, illustrent bien la nature subjective de la notion de valeur aberrante.

Barnett et Lewis (1994) 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 observation 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. L'observation ne va pas sembler extrême mais va apparaître comme étant *étonnamment extrême*. L'expression “semble être inconsistante” est cruciale car elle émane d'un jugement subjectif de la part de l'observateur qui s'intéresse aux données. Ce qui est important c'est de savoir si les données font vraiment partie de la population principale. Si ce n'est pas le cas, elles sont alors considérées comme des *contaminants*, définis comme étant des *observations issues d'autres populations*. Les contaminants peuvent poser problème lors de l'application de méthodes inférentielles à partir de la population d'origine. Il est clair que tout contaminant se trouvant au milieu d'un ensemble de données ne va être “visible” et il est improbable qu'il affecte sérieusement le processus d'inférence. Néanmoins, si de telles observations, étrangères à la population principale, sont situées dans les queues des distributions, elles peuvent causer des difficultés dans la tentative de décrire la population et déformer l'estimation des paramètres de la population.

Barnett et Lewis (1994) ont affiné leur définition en faisant intervenir la notion de modèle de probabilité : une valeur aberrante est *une observation qui apparaît douteuse dans le contexte d'un modèle de probabilité, désigné initialement pour expliquer le processus de génération de données.* Everitt (2002) tient également compte des modèles de probabilités 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.*

A partir de ces définitions , on se rend compte qu'il est nécessaire de définir également d'autres termes qui sont utilisés de manière courante et qui ont tendance à semer la confusion dans les esprits. Le terme de valeurs extrêmes est défini par Everitt (2002) comme les valeurs les plus grandes et les plus petites parmi un ensemble d'observations. Barnett et Lewis (1994) ont distingué, dans le cas univarié, les notions de valeurs aberrantes, d'observations extrêmes et de contaminants à l'aide d'une figure dont une adaptation est présentée à la **figure 1.**

Soit $x_1, x_2, \ldots, x_n$, un échantillon aléatoire univarié de taille n, provenant d'une distribution F, et soit $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ les données ordonnées dans l'ordre croissant.

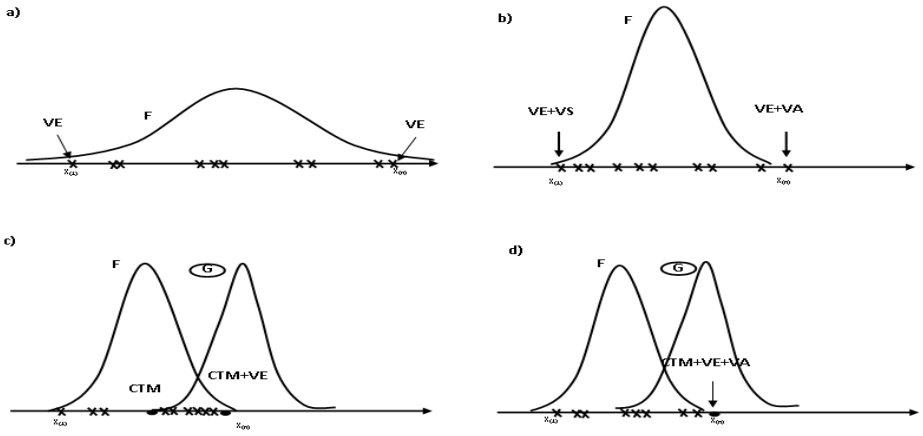

FIG. 1.1 – Les valeurs extrêmes(VE),Les contaminants(VC) et les outliers(valeurs aberrante(VA))

Les observations $x_{(1)}$ et $x_{(n)}$ sont respectivement l'observation extrême inférieure et supérieure.

Le fait de déclarer qu'une observation extrême est une valeur aberrante dépend de la manière par laquelle elle apparaît en fonction du modèle F. En effet dans la **figure 1(a)**, ni la valeur $x_{(1)}$, ni $x_{(n)}$ ne semblent correspondre à une valeur aberrante. Par contre, dans la **figure 1(b)**, $x_{(n)}$ est une valeur aberrante supérieure ou située au niveau de la queue droite de la distribution. La valeur $x_{(1)}$ cause également quelques problèmes et peut être considérée comme suspecte pour la queue gauche de la distribution. Ainsi, on voit que les valeurs extrêmes peuvent être ou ne pas être des valeurs aberrantes. Toute valeur aberrante est par contre toujours une valeur extrême de l'échantillon. Si toute les observations ne proviennent pas de la distribution F mais que l'une ou l'autre est issue de la distribution G, de moyenne plus élevée que F, les observations de G sont considérés comme des contaminants. De tels contaminants peuvent apparaître comme étant extrêmes mais ce n'est pas forcément le cas. La **figure 1(c)** montre deux contaminants indiqués par un rond noir ; celui situé à droite est l'extrême supérieur tandis que celui de gauche se trouve au milieu de l'échantillon. Néanmoins, $x_{(n)}$, bien qu'il soit extrême et contaminant, n'est pas une valeur aberrante. Enfin au niveau de la **figure 1(d)**, la valeur extrême $x_{(n)}$, correspond à un contaminant qui est également une valeur aberrante. Une valeur aberrante peut donc être la manifestation d'un contaminant. Ces diverses situations indiquent la complexité de l'étude des valeurs anormales et la difficulté de définir le type d'observation rencontré de manière précise. Le terme de *valeur suspecte* correspond, selon Barnett et Lewis (1994), à une valeur moins extrême qu'une valeur jugée aberrante de manière statistique. Les définitions de valeurs suspectes et aberrantes sont complétées dans le suite de ce chapitre en liaison avec les *tests de discordance*.

Enfin, il est nécessaire de parler des *observations influentes* qui sont définies par Everitt (2002) comme étant des observations qui ont une influence disproportionnée sur un ou plusieurs aspects de l'estimateur d'un paramètre, en particulier, les coefficient de régression. Selon Cook et Weisberg (1980), les observations influentes sont celles pour lesquelles les caractéristiques de l'analyse sont altérées de manière considérable quand elles sont suprimées.

1.3 Origines des valeurs aberrantes et objectifs pousuivis

Les valeurs aberrantes dans les séries peuvent survenir pour différentes raisons. Une classification de ces différentes origines ont été discutées dans la littérature par divers auteurs, nous avons retenue celle de la monographie Barnett et Lewis (1994) qui reste notre principale référence dans la l'élaboration de ce chapitre.

Lors da la collecte de données, différentes sources de variabilité peuvent être rencontrées dont la variabilité inhérente, l'erreur de mesure et l'erreur d'exécution.

La variabilité inhérente correspond à l'expression de la manière par laquelle les observations varient de manière aléatoire à travers la population. Une telle variation est une caractéristique naturelle de la population. Elle est incontrolable et reflète les propriétés de la distribution d'un modèle de base qui décrit correctement la génération des données.

En ce qui concerne **l'erreur de mesure**, ou l'erreur liée à la méthode de mesure, des inadéquations au niveau des instruments de mesures surimposent un degré plus élevé de variabilité au facteur inhérent. L'arrondi des valeurs obtenues ou les erreurs de saisie correpondent également à des erreurs de mesure. Cette erreur est liée à des circonstances bien déterminées. L'erreur de mesure peut également être de nature aléatoire, cette variabilité correspond alors à l'incertitude de la méthode de mesure. Quelques contrôles de ce type de variabilité sont possible et facilement réalisables.

Une autre source de variabilité apparaît dans la collecte imparfaite des données, c'est l'**erreur d'execution** qui est liée à des circonstances bien déterminées. Par inadvertance, un échantillon peut être biaisé ou peut inclure des individus qui ne sont pas vraiment représentatifs de le population d'intérêt. Des erreurs d'execution de la manipulation ou dans l'assemblage des données peuvent aussi mener à des outliers de nature déterministe. De même, des erreurs lors du traitement ou des erreurs de gestion des données peuvent conduire à des observations erronées. De telles situations se présentent quand les erreurs humaines mènent à l'enregistrement évident de données incorrectes ou quand le manque de critiques vis-à-vis des facteurs pratiques entraîne des interprétations erronées. Le traite-

ment de telles valeurs aberrantes dans ces situations ne relève pas du domaine de l'analyse statistique mais du bon sens tout simplement.

Les diverses sources de variation, qui provoquent l'apparition de valeurs aberrantes de natures différentes, ont montré la complexité de l'examen des valeurs aberrantes. Cependant, le fait d'être capable de préciser ces notions de nature et d'origine des valeurs aberrantes permet actuellement de déterminer de manière plus structurée les objectifs à atteindre lors de l'examen d'observations anormales. Les objectifs des valeurs aberrantes dépendent en effet de l'origine et de la nature de celles-ci, comme le montre la **figure 2**. Cette figure permet de visualiser clairement le schéma de traitement des valeurs aberrantes et des objectifs poursuivis.

Pour 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, de rejeter les valeurs discordantes et procéder à l'analyse à partir de l'échantillon modifié. D'autres possibilités peuvent cependant également être intéressantes. En effet, on peut choisir d'utiliser un autre modèle que celui choisi initialement de sorte à incorporer la valeur aberrante de manière non discordante. On peut également concentrer son attention sur les valeurs aberrantes et identifier des facteurs non pris en compte initialement mais qui ont une grande importance pratique. Dans le cas d'expérimentations, dans le but est de rechercher des effets importants de facteurs expérimentaux, les valeurs aberrantes peuvent permettre d'identifier des caractéristiques importantes du point de vue pratique plutôt que de refléter une possible inadéquation du modèle.

L'analyse des données peut également faire l'objet de l'une ou l'autre forme d'accomodation. Ce choix est réalisé en fonction des objectifs de l'analyse statistique, car si on s'intérresse spécifiquement aux caratéristiques 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 également. 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 des paramètres.

Il faut reconnaitre que le rejet inconsidéré des valeurs aberrantes a des conséquences statistiques non négligeable pour l'analyse ultérieure de l'échantillon qui n'est plus aléatoire mais qui devient un échantillon censuré. Le remplacement des données rejetées par des équivalents statistiques implique des conséquences similaires.

Quant aux valeurs aberrantes dont la nature est déterminée, c'est-à-dire les erreurs de mesure ou d'exécution, elles peuvent être rejetées ou faire l'objet de corrections dans la mesure ou celles-ci sont encore réalisables.

1.3.1 Tests de discordance

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 une déclaration *subjective* de la présence d'une valeur aberrante dans un ensemble de données possède des implications *objectives* importantes pour l'analyse future des données. Dans cette optique, Barnett et Lewis (1994) proposent d'utiliser un *test de discordance*. L'objectif poursuivi lors de l'utilisation de tests de discordance 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. Le 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.

Qu'est ce qui est accompli si une observation est montrée discordante ? Il est établi (à un niveau du test) qu'il n'est pas raisonnable de croire qu'une telle observation est issue du modèle de base F. Si la seule alternative est que l'observation singulière vient de G (un *contaminant*) et le reste de l'échantillon vient de F, alors la discordance de l'observation en question démontrée par le test mène à l'adoption de ce modèle alternatif (et à l'implication que l'observation discordante est un contaminant). Autrement dit, rejet de F comme source homogène et adoption d'un modèle alternatif $\bar{F}$ qui déclare qu'il y a un contaminant. Par ailleurs, comme avec tout test de signification pur, l'inférence ne dépend pas de la forme de $\bar{F}$, seulement de celle de F. Cependant, la forme de $\bar{F}$ est cruciale à toute considération

des propriétés d'un test particulier ou toute comparaison de tests rivaux.

Parmi les tests de discordances, une distinction peut être réalisé en fonction du type de distribution de la population-parent dont 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) donne un classement des tests de discordance en tenant compte du critère (par exemple *rapport excès/étalement*) retenu pour effectuer le test. De même pour les échantillons extraits d'une population non-normale, ces même auteurs présentent une liste de détection de valeurs aberrantes. Les distributions concernées sont les suivantes : exponentielle, exponentielle tronquée, gamma, uniforme, poisson et binomiale.

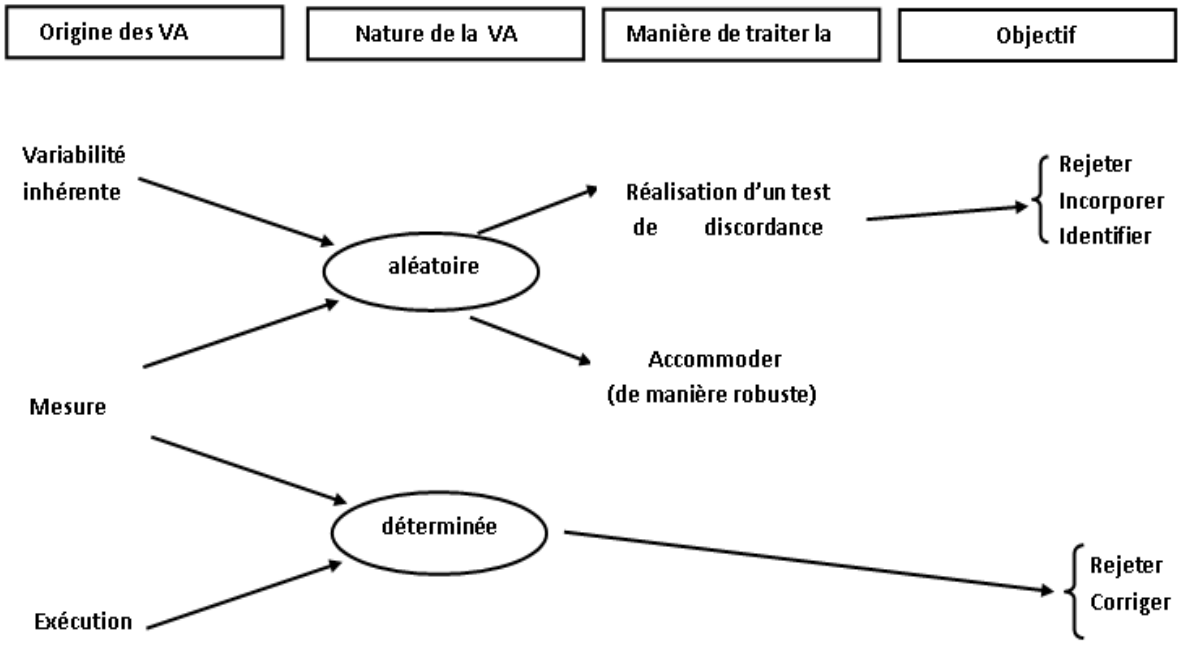

FIG. 1.2 – Schéma de Barnett et Lewis pour l'examen des outliers

Il faut signaler que ces tests sont sujets à l'éffet de masque évoqué plus haut, qui consiste en l'incapacité d'une procédure statistique d'identifier une valeur aberrante en présence de plusieurs valeurs suspectes. Des tests qui evitent cet effet de masque et d'autres contraintes liées aux tests de discordance comme par exemple le nombre de valeurs aberrantes à connaitre *a priori* sont présentés dans la monographie de Barnett et Lewis (1994).

Dans le cas des observations dépendantes comme les séries temporelles deux tests de discordance dans les modèles ARMA seront présentés au chapitre 3. Le très populaire test du rapport de vraisemblance de Fox (1972) est le premier, le second est une variante du test de score introduit par Louni (2008).

1.3.2 Procédures d'accommodation

Les procédures d'accommodation sont des méthodes statistiques destinées à réaliser des inférences sur la population à patir de laquelle l'échantillon aléatoire a été obtenu. Les résultats acquis par l'intermédiaire de ces procédures ne sont pas 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 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. L'utilisation de la médiane de l'échantillon comme estimateur de position en est un exemple très simple.

Les procédures d'accommodation permettent dès lors d'éviter de rejeter les 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. 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. Les méthodes robustes vont bien au delà de cette protection contre les valeurs aberrantes, elles fournissent aussi une protection contre divers types d'incertitude sur le mécanisme de génération des données. Elles comprennent notamment des procédures d'inférences pour lesquelles les estimations retiennent les propriétés statistiques de tout un ensemble de distributions possibles.

Les méthodes robustes peuvent également répondre spécifiquement au problème de valeurs aberrantes losqu'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 parceque le modèle de base ne reflète pas le degré approprié de variabilité , il est nécessaire de s'intéresser à des distributions plus étendues que la distribution normale classiquement

retenue.

L'omission des 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 permet d'exprimer la contamination du modèle initial, l'estimation ou le test des paramètres du modèle initial peuvent être très intéressants et il est alors important d'utiliser des procédures robustes appropriées pour se protéger des composants de faibles probabilité ou contre les valeurs décalées.

A partir de maintenant, nous considérons des données de séries temporelles. Les modèles ARMA sont la structure la plus courante pour la détection et la modélisation des séries temporelles, ce choix peut être justifié par ce qu'on appelle la décomposition de Wold. Les modèles ARMA sont aussi familiers, et leurs propriétés sont bien connues. Ils offrent donc beaucoup de commodités pour examiner les propriétés de données de séries temporelles et aussi pour détecter les outliers notamment. En effet, toutes les définitions des outliers présentées dans le chapitre suivant sont basées sur ces modèles, et ceci a conduit à l'utilisation de définitions similaires aussi bien qu'à d'autres modèles. Les modèles ARMA ne sont en aucun cas la seule option, néanmoins ces modèles constituent la première étape dans l'examen des outliers, et une fois leurs rôle pris en compte, ces considérations peuvent être aussi bien étendues à d'autres modèles.

Chapitre 2

Outliers dans les modèles ARMA

2.1 Introduction

Dans le cas des observations dépendantes comme les séries temporelles, la valeur absolue d'une observation n'est pas nécessairement une mesure d'abbération. Par conséquent, la définition a été basée sur d'autres propriétés des données. Habituellement, les outliers dans les séries temporelles sont définies de manière informelle, soit comme des valeurs imprévues, soit comme des valeurs surprenantes en rapport avec le reste de la série, souvent les valeurs voisines. Les mots tels que « inhabituelle » ou « suspecte » sont régulièrement utilisés quand on définit les outliers, ce qui montre la nature souvent subjective de l'analyse des outliers. Une observation est un outlier parcequ'elle est vu comme telle. Cette subjectivité évoquée au chapitre précédent n'est pas discutée ici. L'idée présentée avant pour les données indépendantes peut cependant être modifiée pour être aussi bien ajustée aux séries temporelles. A la place de la distribution des données, on peut spécifier un modèle préliminaire de données, et de façon similaire pour les outliers, on peut spécifier soit un modèle à paramètres qui évoluent dans le temps pour l'ensemble des données ou un modèle différent pour les outliers.

Les modèles ARMA sont la structure la plus courante pour la détection et la modélisation des séries temporelles, ce choix peut être justifié par ce qu'on appelle la décomposition de Wold. Les modèles ARMA sont aussi familier, et leurs propriétés sont bien connues. Ils offrent donc beaucoup de commodités pour examiner les propriétés de données de séries temporelles et aussi pour détecter les outliers notamment. En effet, toutes les définitions des outliers présentées dans la section suivante sont basées sur ces modèles, et ceci a

conduit à l'utilisation de définitions similaires aussi bien qu'à d'autres modèles. Les modèles ARMA ne sont en aucun cas la seule option, et leur utilisation a été sévèrement critiqué. Cette critique est pertinente à des degrés divers, néanmoins ces modèles ARMA constitue la première étape dans l'examen des outliers, et une fois leurs rôle compris en eux, ces considération peuvent être aussi bien étendues à d'autres modèles.

2.2 Classification des outliers

Les outliers peuvent prendre plusieurs formes dans les séries temporelles. Des définitions formelles et une classification des outliers dans le contexte des séries temporelles ont été proposées en premier par Fox(1972). Avant cet article, l'hypothèse d'indépendance et identiquement distribuée était communément faite dans l'analyse des outliers. Fox note que ceci conduit aux « procédures d'échantillonnage » qui ne sont évidemment pas appropriées pour les séries temporelles. Dans le cas des séries représentées par des modèles autoregressifs (AR) purs, il propose alors une classification des outliers de type I et de type II. Plus tard, ces deux types sont adoptés sous le nom des *additive ouliers* et des *innovational outliers* avec respectivement l'abréviation de AO et IO. Avec d'autres catégories d'outliers nous les définissons ci-dessus dans le cas des modèles ARMA.

La forme d'un modèle autoregressif moyenne mobile (ARMA) d'ordre p et q pour un processus z_t est

$$\phi(B)z_t = \theta(B)a_t,$$

où B est le polynôme retard tel que $Bz_t = z_{t-1}$;

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p \quad \text{et} \quad \theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$$

sont respectivement des polynômes de degrés p et q, et les résidus a_t sont des v.a. gaussiennes *i.i.d.* (indépendantes et identiquement distribuées) de moyenne zéro et de variance σ^2. Les racines des polynômes $\phi(B)$ et $\theta(B)$ sont supposées à l'extérieur du cercle unité pour assurer la stationnarité et l'inversibilité du processus.

Dans la littérature standard sur les outliers (*e.g.* Tsay (1986 et 1988), Chang et *al* (1988)) une série temporelle est représentée par un processus ARMA avec intervention

(*i.e.* outliers).

La série observée y_t est décrite par le modèle

$$y_t = z_t + f(t),$$

où z_t suit un modèle ARMA et $f(t)$ est une fonction paramétrique représentant les perturbations exogènes de z_t tels que les outliers ou les changements de niveau. Elle peut être déterministe ou stochastique selon le type de perturbation. En pratique, $f(t)$ est spécifiée par l'analyse des données basée sur l'information de la perturbation et du processus (y_t). Par exemple, elle peut être une fonction linéaire de quelques variables exogènes tels que les effets du jour de marché ou des vacances dans l'analyse des séries temporelles saisonnières. Pour le modèle déterministe, on suppose $f(t)$ de la forme

$$f(t) = \omega \frac{\omega(B)}{\delta(B)} \mathbf{1}_d(t)$$

où $\mathbf{1}_d(t) = 1$, si $t = d$ et $\mathbf{1}_d(t) = 0$ si $t \neq d$ est un indicateur signifiant l'occurrence de la perturbation à l'instant d,

$$\omega(B) = 1 - \omega_1 B - \cdots - \omega_s B^s \quad \text{et} \quad \delta(B) = 1 - \delta_1 B - \cdots - \delta_r B^r$$

sont respectivement des polynômes de B de degré s et r et ω une constante caractérisant l'impact initial de la perturbation.

Pour le modèle stochastique, $f(t)$ est supposée de la forme

$$f(t) = \omega \frac{\omega(B)}{\delta(B)} e_t(d)$$

où

$$e_t(d) = 0 \ \text{ si } t < d$$

et

$$\{e_t(d) |\ t \geq d\}$$

est une suite de variable aléatoire *i.i.d.* de moyenne 0 et de variance σ_e^2. Ici, la fonction $f(t)$ peut affecter le modèle et la variance de z_t. $e_t(d) = a_t$ pour $t \geq d$ est utilisé dans la suite pour caractériser les *Variance Changes* dans z_t.

Le premier et le plus couramment étudié est l'*Additve Outlier* ou AO (type I de Fox). Un AO affecte seulement une seule observation de la série et non ses valeurs futures. Après la perturbation, la série revient à son cours normal comme si de rien n'était. Formellement un AO se produisant à l'instant $t = d$ est une perturbation d'amplitude $\omega = \omega_A$ affectant seulement l'observation y_d. En termes de polynômes, les AO sont modélisés par

$$\omega(B)/\delta(B) = 1.$$

Un exemple type de AO peut être une erreur de saisie.

En revanche, un *Innovational Outlier* ou un IO (type II de Fox) affecte plusieurs observations. Formellement un IO est une perturbation $\omega = \omega_I$ dans la série des innovations a_t du modèle ARMA. Il affecte la série de manière temporaire au travers la structure dynamique

$$\pi(B) = \phi(B)/\theta(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots .$$

Cette structure est analogue à celle du modèle ARMA de z_t impliquant qu'un IO affecte aussi toutes les observations pour tout $t \geq d$. Dans ce cas

$$\omega(B)/\delta(B) = \theta(B)/\phi(B).$$

En pratique, un IO est une cause externe brutale.

Habituellement, seuls les AO et IO sont considérés dans la littérature, mais l'influent article de Tsay (1988) définit aussi bien trois autres types d'outliers, à savoir, les *Level Shifts*, les *Trancient Changes* et les *Variance Changes*.

Les *Level Shifts* (LS), quelquefois appelé *Level Changes*, changent simplement le niveau (ou la moyenne) de la série par une certaine amplitude, $\omega = \omega_L$, à partir d'une certaine observation y_d. Le modèle de la série change de

$$y_t = z_t \text{ à } y_t = z_t + \omega_L,$$

où ω_L peut être une constante soit positive soit négative. Dans ce cas le polynôme est

$$\omega(B)/\delta(B) = 1/(1 - B).$$

Le changement est par conséquent permanent.

Les *Temporary Changes* (TC) est une généralisation des AO et des LS dans le sens où, comme un AO, il cause un impact initial à la différence que cet effet a aussi des conséquences sur les observations ultérieures. L'impact d'un TC n'est cependant pas permanent, il décroît à une vitesse exponentielle . Formellement, un TC a un effet initial $\omega = \omega_T$ à l'instant d et cette effet décroît graduellement, avec

$$\omega(B)/\delta(B) = 1/(1 - \delta B)$$

où $0 < \delta < 1$ représente la rapidité de retour. C'est pour cette raison que ce modèle fait référence à un modèle (TC) de changement transitoire. Avec $\delta = 0$ cet outlier devient un AO et avec $\delta = 1$ un LS.

En général, les AO et les IO sont considérés comme des points atypiques tandis que les TC et les LS comme des changements structurels. Les TC représentent un changement éphémère dans une série temporelle, et les LS sont plus le reflet d'un choc permanent. Les procédures itératives de détection et de modélisation des outliers présentées dans la littérature peuvent traiter l'ensemble de ces outliers dans un cadre unifié. Une de ces procédure se trouve au centre de nos préoccupations et sera présentée dans le détail dans le chapitre 4.

Les *Variance Changes* (VC) restent encore loins des AO et IO et ne sont pas habituellement considérés en connection avec tous les outliers. Formellement, un outlier de type VC d'amplitude $w = w_V$ est modélisé par

$$\omega(B)/\delta(B) = \theta(B)/\phi(B)$$

et

$$e_t(d) = a_t \text{ pour } t \geq d.$$

En considérant

$$b_t = \phi(B)/\theta(B)y_t,$$

on a

$$b_t = a_t \quad \text{pour} \quad t < d,$$

mais

$$b_t = (1 + \omega_V)a_t \quad \text{pour } t \geq d.$$

Par conséquent, la variance de la série observée change de

$$\sigma_a^2 \quad \text{à} \quad (1 + \omega_V)^2\sigma_a^2$$

à la date $t = d$.

Wu, Hoskin et Ravishanker (1993) proposent un autre type d'outlier intéressant, les *Reallocation Outliers* (RO). Ils sont définis comme un bloc d'outliers de type AO dont la somme total des effets est égale à zéro. Avec les mêmes notations que précédemment, un RO est défini par,

$$f(t) = \sum_{k=0}^{K} \omega_{R,k}\mathbf{1}_{d+k}(t),$$

avec la restriction que la somme $\sum \omega_{R,k}$ est égale à zéro. Dans cette formulation il y a $K + 1$ outliers d'amplitude $\omega_{R,k}$ aux dates $t = d, d + 1, d + 2, \ldots, d + K$.

Intuitivement, les RO peuvent offrir une alternative convenable à un bloc de AO, mais par l'observation il se peut qu'il soit difficile d'identifier laquelle formulation est la mieux appropriée.

Un autre type d'outliers sont les *Volatilité Outliers* en rapport avec les modèles ARCH (Autorégressif à Hétéroscédascité Conditionnelle). Ces modèles ne sont pas abordés dans ce travail.

Comme on l'a déjà souligné, ces définitions ont été adoptées dans la plupart des études sur les outliers dans les séries temporelles. Elles peuvent s'appliquer notamment dans la distinction entre AO et IO, aussi, elles ont été utilisées dans plusieurs familles de modèles.

2.3 Les effets des outliers

Dans le cas des modèles ARMA, certains effets des outliers sur l'identification, l'estimation et la prévision sont connus. Tout d'abord, les outliers affectent la structure d'autocorrétion de la série temporelle, et par conséquent aussi, ils affectent le biais de l'estimateur de la fonction d'autocorrelation (ACF), d'autocorrélation partielle (PACF) et

d'autocorrélation étendue (EACF). Ces biais sont sévères, ils dépendent, à côté d'attributs évidents comme le nombre, le type, l'amplitude et la position des outliers, du modèle sous-jacent et de sa structure d'autocorrélation. Deutsch, Richards et Swain (1990) ont présenté quelques résultats sur les effets des outliers sur l'identification des modèles ARMA. Ces résultats montrent que pour les séries courtes ou de longueur assez modestes, dans la cas ou le vrai modèle est un AR, la présence d'un seul outlier entraîne souvent une fausse identification du modèle comme un modèle moyenne mobile MA ou ARMA, qui plus est avec des ordres p et q non adéquats. De façon similaire, les outliers altèrent les estimateurs des paramètres des modèles ARMA. Par exemple, Chen et Liu (1993b) ont obtenus par simulation numérique quelques résultats suggérant que les AO, les TC et les LS causent de substantiels biais aux estimateurs des paramètres des modèles ARMA, alors que les IO ne produisent que des effets mineurs.

L'identification des modèles ARMA est traditionnellement basée sur l'estimation des autocorrélations (ACF) et des autocorrélations partielles (PACF) qui, à moins que les outliers ne soient pris en compte, sera trompeuse. Glendinning (1998) a discuté sur ce sujet et aussi sur des méthodes de sélection de modèles robustes. Masarotto (1987) propose également une méthode robuste pour l'estimation des ACF et des PACF. Bustos et Yohai (1986) examinent les propriétés de plusieurs procédures d'estimation des modèles ARMA. Les méthodes des moindres carrés et du maximum de vraisemblance sont toutes les deux sensibles à la présence des outliers, notamment les AO, alors que plusieurs estimateurs robustes peuvent contrôler certains problèmes causés par les outliers (certains estimateurs utilisées dans notre étude seront détaillés dans le chapitre 3).

D'un autre côté, les outliers ont aussi des impacts évidents sur la prévision dans les modèles ARMA, particulièrement les outliers proches du début de la prévision de la période peuvent avoir de sérieuses conséquences. La prévision peut n'être altérée que de façon bénigne par les *Additive Outliers* mais les intervalles de prédiction peuvent devenir sérieusement trompeurs, puisque les outliers causent une amplification de l'estimation de la variance de la série (pour des résultats précis voir Ledolter (1989) et Hotta (1993)). Les *Level Shifts* (LS) et les *Temporary Changes* (TC) ont davantage d'impacts sur la prévision (Trivez 1993). Pour plus de détails sur la prévision en présence des outliers voir aussi Chen et Liu (1993a).

En plus des autres difficultés, deux problèmes sont rencontrés dans l'analyse des outliers : l'effet de masque (*masking effect*) et l'effet d'entraînement (*swamping effect*). Ces concepts sont reliés à la détection des outliers et peuvent interagir pour compliquer davantage le problème. L'effet de masque se produit lorsqu'un outlier empêche d'autres observations douteuses d'être déclarées comme outlier. Inversement, l'effet d'entraînement se rencontre quand un outlier affecte la série de sorte que d'autres observations apparaissent à tort comme des outliers. On peut dire que l'effet de masque correspond à une sous-estimation du nombre d'outliers alors que l'effet d'entraînement à une surestimation de ce nombre. Ces notions sont étroitement connectées à des méthodes spécifiques de détection, et ne sont pas des propriétés des données elles mêmes. En d'autres termes, l'effet de cache et l'effet d'entraînement sont seulement des déficiences de certains méthodes, et non des types d'outliers en tant que tels.

Chapitre 3

Tests de détection et estimateurs robustes dans les modèles ARMA

L'examen des échantillons univariés pour ajuster des modèles et estimer les paramètres, bien qu'il constitue une part importante de la pratique statistique, est quelque peu limité dans les buts. Plus souvent, et plus utilement, le besoin de considérer des situations plus structurées s'impose. Par exemple, un intérêt pour la manière dans laquelle les observations de la variable de principal intérêt varie avec les valeurs d'autres variables ou varie avec le temps, mènent respectivement à l'étude des modèles de regression et des modèles de séries temporelles. Dans les cas très structurés, comme les modèles de séries temporelles, il faut aussi s'attendre à rencontrer, de temps en temps, des données non representatives comme étant des outliers. Là encore, il est aussi important que dans le cas d'un simple échantillon d'être capable de reconnaître, interpréter et prendre en considération les outliers en utilisant les techniques statistiques appropriées. Les outliers peuvent avoir, comme avant, un intérêt intrinsèque en eux mêmes, ou peuvent être indicatifs de spécifications incorrectes dans la structure de l'erreur, ou du modèle de base, avec des implications conséquentes dans le choix des procédures d'inférences pertinentes. Dans de telles données très structurées deux complications surviennent : *les outliers tendent à être intuitivement moins apparents*, plus cachés dans la masse de données, et des *méthodes formelles pour leur rejet ou leur "accommodation" sont finement développées.*

3.1 Détection des outliers

La première idée proposée pour détecter les outliers dans les séries temporelles consiste à examiner les moments d'ordre supérieurs de la série, en pratique le *skewness* et le *kurtosis*, ou bien appliquer un processus de lissage (voir Huber (1972)). Ces méthodes sont simples et peuvent être utiles dans beaucoup de cas, mais bien évidemment ne sont pas suffisantes pour une large variété de situations rencontrées dans l'analyse empirique des séries temporelles. Il est donc nécessaire de considérer des méthodes plus élaborées, deux sont présentées dans cette section.

Il y a plusieurs méthodes de détection des différents types d'outliers. Le nombre d'outliers à détecter est variable (un ou plusieurs), et dans les méthodes avec de multiple outliers, il y a une distinction entre les tests selon que le nombre d'outliers à identifier est connu ou inconnu. Dans l'analyse des séries temporelles un modèle bien spécifié est nécessairement exigé dans toutes les méthodes. Les modèles couramment utilisés sont les modèles ARMA et les modèles de régression dynamique.

Il est aussi intéressant de souligner que beaucoup de ces méthodes s'inspirent de l'analyse de la régression. Ceci n'est pas surprenant puisque l'analyse de la régression a constitué le domaine des premiers développements dans la détection des outliers et de la modélisation, et aussi, les méthodes les plus avancées et les plus largement partagées.

Dans la suite deux tests de détection des outliers dans les modèles ARMA seront examinés. Le test du rapport de vraisemblance introduit par Fox (1972) habituellement utilisé dans l'étape de détection des outliers des procédures itératives avec interventions est le premier. Le second, proposé dans Louni (2008), est une variante du test de score qui, nous l'avons dit, sera executé dans la phase de détection à la place du test précédent dans une version modifiée des dites procédures objet du dernier chapitre de ce mémoire.

3.1.1 Test du rapport de vraisemblance

Le premier examen détaillé de la détection des outliers dans les séries temporelles stationnaires a été proposé par Fox (1972). Il distingue deux types (I et II) d'outliers, que nous avons définis avant et, nous l'avons souligné, sont maintenant connus comme *Additive*

Outliers et *Innovation Outliers* respectivement. Seule la situation où tous les outliers sont du même type, connu, est examinée dans le détail sous la supposition que le processus est débarrassé des composantes tendancielle et saisonnière ; le possible effet de leurs corrections sur l'examen des outliers n'as pas été envisagé. Par conséquent, la méthode présentée a clairement des limites, mais elle fournit un point de départ important dans ce domaine d'étude difficile.

Un test pour les *Additive Outliers* est développé en relation avec les modèles d'outliers évoqué plus haut pour les séries temporelles discrète

$$y_t = z_t + \omega_A \mathbf{1}_d(t)$$

où $\mathbf{1}_d(t)$ est précédemment défini (*i.e.* $\mathbf{1}_d(t) = 1$; si $t = d$; 0 sinon) et les z_t satisfont au schéma autoregressif d'ordre p,

$$z_t = \sum_{i=1}^{p} \phi_i z_{t-i} + a_t \qquad (t = p+1, \ldots, n)$$

où les a_t sont indépendants $\mathbf{N}(0, \sigma^2)$. Nous avons donc un ensemble de n observations d'un processus à temps discret, en plus restreint par la supposition que p est connu et que (z_t) est un processus stationnaire, avec un outlier introduit à l'instant d. Les deux cas ou d est connu et inconnu sont considérés et les tests du *maximum du rapport de vraisemblance* de l'hypothèse $H : \omega_A = 0$ contre $\bar{H} : \omega_A \neq 0$ sont développés. Pour le dernier cas, qui reste le plus réaliste (la position d de l'outlier est inconnue), la statistique du *maximum du rapport de vraisemblance* est équivalente à

$$\max_{d=p+1,\ldots,n-p} (k_{d,n})$$

où

$$k_{d,n} = \mathbf{y}' \hat{W}^{-1} \mathbf{y} / (\mathbf{y} - \tilde{\boldsymbol{\omega}}_A)' \tilde{W}^{-1} (\mathbf{y} - \tilde{\boldsymbol{\omega}}_A).$$

Dans cette expression $\tilde{\boldsymbol{\omega}}_A$ est $\tilde{\omega}_A(0, 0, \ldots, 0, 1, 0, \ldots, 0)'$ où 1 apparaît à la position d et $\tilde{\omega}_A$ est l'estimateur du maximum de vraisemblance de ω_A sous le modèle AO ci-dessus ; $\hat{W}^{-1}$ et $\tilde{W}^{-1}$ sont les estimateurs du maximum de vraisemblance de W^{-1} sous H et $\bar{H}$ respectivement, où la matrice de covariance du processus est supposé avoir la forme

$$V = W\sigma^2$$

qui dépend seulement de p et des coefficients autorégressifs ϕ_i $(i = 1, 2, \ldots, p)$. Notons que les éléments de W sont de la forme $w_{t,t'} = w_{|t-t'|}$.

Le test de discordance détecte l'outlier comme l'observation maximisant $k_{d,n}$ et la déclare comme discordante si la valeur maximale est suffisamment grande. La distribution de n variables corrélées de Fisher n'est pas connue. Seuils de significations, calculs de puissance et le comportement des simplifications très commodes du test (voir ci-dessous) sont tous examinés par Fox en utilisant des méthodes de simulation.

Le modèle de l'outlier utilisé par Fox pour un test de discordance pour les *Innovational Outliers* déjà présenté au chapitre précédent a la forme

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \omega_I \mathbf{1}_d(t) + a_t$$

où toutes les quantités sont définies et limitées comme auparavant. Ici encore le test du maximum du rapport de vraisemblance de $H : \omega_I = 0$ contre $\bar{H} : \omega_I \neq 0$ est développé, et étudié par simulation, dans le cas où d est spécifié. Le cas le plus important où la position d n'est pas spécifiée n'est pas envisagé. Comme dans le cas du modèle AO, quelques implications employant le modèle perturbé IO sont examinées par simulation.

Plus tard, ces travaux ont été développés Chang, Tiao et Chen (1988) et Tsay (1986 et 1988) et sont aussi utilisés comme une partie d'une stratégie complète de la modélisation des outliers objet du chapitre suivant.

Tester les cinq types d'outliers présentés dans le chapitre 2 s'appuit sur une simplification de la statistique de test du rapport de vraisemblance décrite ci-dessus dans le cas autorégressif (AR) (voir Chang, Tiao et Chen (1988). Dans la suite de cette section, cette simplification mise en oeuvre dans la modélisation des outliers sera détaillée dans la situation qui nous préoccupe *i.e.* en présence des outliers de type AO et/ou IO.

Rappelons d'abord les écritures des modèles ARMA éventuellement pérturbés par les outliers AO et IO.

Soit z_t le modèle ARMA d'ordre p et q parfaitement donné par

$$\phi(B)z_t = \theta(B)a_t,$$

où B est l'opérateur retard tel que $Bz_t = z_{t-1}$; $\phi(B) = 1 - \phi_1 - \ldots - \phi_p B^p$ $\theta(B) = 1 - \theta_1 - \ldots - \theta_q B^q$ sont deux polynômes en B dont les racines sont suposées en dehors du cercle unité et $\{a_t\}$ une suite de variables indépendantes de loi $\mathbf{N}(0, \sigma_a^2)$.

Si y_t est le processus observé, la présence d'un IO à la date d se traduit sur le modèle par

$$y_t = z_t + \frac{\theta(B)}{\phi(B)}\omega \mathbf{1}_d(t),$$

et pour un AO par

$$y_t = z_t + \omega \mathbf{1}_d(t).$$

Ces modèles peuvent s'écrire en fonction de la suite des innovations a_t,

$$y_t = \frac{\theta(B)}{\phi(B)}\{a_t + \omega \mathbf{1}_d(t)\}, \quad \text{(IO)}, \tag{3.1}$$

et

$$y_t = \frac{\theta(B)}{\phi(B)}a_t + \omega \mathbf{1}_d(t). \qquad \text{(AO)} \tag{3.2}$$

Il est alors possible d'estimer l'amplitude de la perturbation ω. En effet, le filtrage de la série de sorte que

$$e_t = \frac{\phi(B)}{\theta(B)}y_t,$$

et

$$x_t = \frac{\phi(B)}{\theta(B)}\frac{\omega(B)}{\delta(B)}\mathbf{1}_d(t),$$

implique que

$$e_t = \omega x_t + a_t.$$

C'est est une simple équation de régression linéaire. Par conséquent, l'estimateur des moindres carrés ordinaires de ω et sa variance sont donnés par

$$\hat{\omega} = \left(\sum_{t=1}^{n} e_t x_t\right) \Big/ \left(\sum_{t=1}^{n} x_t^2\right)$$

et

$$Var(\hat{\omega}) = \sigma_a^2 \Big/ \left(\sum_{t=1}^{n} x_t^2 \right).$$

Ceci peut être utilisé pour estimer les amplitudes des outliers. L'intuition tient du fait que la série est d'abord filtrée avec le vrai modèle ARMA pour obtenir les résidus.

C'est sans surprise donc, que l'amplitude d'un IO est estimé par le résidu e_d à la date d. En effet, dans ce cas, x_t est simplement l'indicateur $\mathbf{1}_d(t)$ (la seule valeur non nulle est à l'instant d), par conséquent l'estimateur de l'amplitude de l'outlier et sa variance sont données par

$$\hat{\omega}_I = e_d \tag{3.3}$$

et

$$Var(\hat{\omega}_I) = \sigma_a^2.$$

L'estimateur de l'amplitude d'un AO est une combinaison linéaire de $e_t, e_{t+1}, \ldots$, il est donné par

$$\hat{\omega}_A = \rho^2 \left(e_d - \sum_{i=1}^{n-d} \pi_i e_{d+i} \right) \tag{3.4}$$

avec sa variance

$$Var(\hat{\omega}_A) = \rho^2 \sigma_a^2,$$

où les coefficients π_i sont les poids de z_t *i.e.* $\pi(B) = \phi(B)/\theta(B)$, et

$$\rho^2 = (1 + \pi_1^2 + \cdots + \pi_{n-d}^2)^{-1}.$$

D'abord, si les paramètres du modèle ARMA et la postion de l'outlier sont connus. Considérons l'hypothèse $H_0 : \omega = 0$ dans (3.1) et (3.2), H_1 désigne $\omega \neq 0$ dans (3.1), H_2 l'hypothèse $\omega \neq 0$ dans (3.1). Tester la présence d'un IO à la date d *i.e.* H_0 contre H_1, nous l'avons dit ci-dessus, s'appuit sur une simplification de la statistique de vraisemblance (voir Fox (1972)) donnée par

$$\lambda_{I,d} = \frac{\hat{\omega}_I}{\sigma_a}.$$

De façon similaire, la présence d'un AO à la date d *i.e.* H_0 contre H_2, est testée par

$$\lambda_{A,d} = \frac{\hat{\omega}_A}{\rho\sigma_a}.$$

Pour tester la présence d'un IO contre la présence d'un AO à la date d *i.e.* H_1 contre H_2 Chang et *al* (1988) ont proposé la statistique de test

$$\lambda_{IA,d} = \left(\rho^{-2}\hat{\omega}_A^2 - \hat{\omega}_I^2\right) \Big/ \left(2\sigma_a^2(1-\rho^2)^{1/2}\right).$$

Sous l'hypothèse H_0, les statistiques $\lambda_{I,d}$ et $\lambda_{A,d}$ ont des distributions normales.

La situation est compliquée pour la statistique $\lambda_{IA,d}$, néanmoins des seuils critiques du test sont obtenus numériquement.

Pour détecter un IO ou un AO à une position inconnue, cas le plus répandu en pratique, on peut tester toutes les observations respectivement à travers la suite $\lambda_{I,t}$, $t = 1, \ldots, n$, ou la suite $\lambda_{A,t}$, $t = 1, \ldots, n$. En d'autres termes, la possibilité d'un IO ou un AO dans la série d'observations peut être testée par

$$\max_{t=1,\ldots,n} |\lambda_{I,t}|$$

ou

$$\max_{t=1,\ldots,n} |\lambda_{A,t}|.$$

Maintenant, si les paramètres du modèle ARMA et σ_a^2 sont inconnus. A cause de la lourdeur des calculs dû à la nature non linéaire du modèle AO et du cas ARMA, plusieurs statistiques simples surviennent comme possible approximations de ce critère du rapport de vraisemblance, en voici une.

Soient $\hat{\phi}_1, \ldots, \hat{\phi}_p; \hat{\theta}_1, \ldots, \hat{\theta}_q$; et $\hat{\sigma}_a^2$ les estimateurs du maximum de vraisemblance respectifs des paramètres $\phi_1, \ldots, \phi_p; \theta_1, \ldots, \theta_q$; et σ_a^2 obtenus sous l'hypothèse d'absence d'outliers dans la série, *i.e.* sous H_0. De plus, soit $\hat{e}_t$ les résidus calculés à partir d'un tel modèle estimé et $\hat{\pi}(B) = \hat{\phi}(B)/\hat{\theta}(B)$. Considérons

$$\hat{\lambda}_{I,d} = \frac{\hat{\omega}_I}{\hat{\sigma}_a}, \quad \hat{\lambda}_{A,d} = \frac{\hat{\omega}_A}{\hat{\rho}\hat{\sigma}_a},$$

et

$$\hat{\lambda}_{IA,d} = \left(\hat{\rho}^{-2}\hat{\omega}_A^2 - \hat{\omega}_I^2\right) \Big/ \left(2\hat{\sigma}_a^2(1-\hat{\rho}^2)^{1/2}\right),$$

où

$$\hat{\omega}_I = \hat{e}_t, \tag{3.5}$$

$$\hat{\omega}_A = \rho^2 \left(\hat{e}_d - \sum_{i=1}^{n-d} \hat{\pi}_i \hat{e}_{d+i}\right) \tag{3.6}$$

et

$$\hat{\rho}^2 = (1 + \hat{\pi}_1^2 + \hat{\pi}_2^2 - \cdots - \hat{\pi}_{n-d}^2)^{-1}.$$

On peut montrer que $\hat{\lambda}_{I,d}$, $\hat{\lambda}_{A,d}$ et $\hat{\lambda}_{IA,d}$ sont asymptotiquement équivalents respectivement aux statistiques $\lambda_{I,d}$, $\lambda_{A,d}$ et $\lambda_{IA,d}$ pour tester les hypothèses H_0 contre H_1, H_0 contre H_1, et H_1 contre H_2 en un point donné d.

Pour détecter un IO ou un AO à une position inconnue, on peut tester toutes les observations respectivement à travers la suite $\hat{\lambda}_{I,t}$, $t = 1, \ldots, n$ ou la suite $\hat{\lambda}_{A,t}$, $t = 1, \ldots, n$. En d'autres termes, la possibilité d'un IO ou un AO dans la série d'observations peut être testée respectivement par

$$\hat{\eta}_{IO} = \max_{t=1,\ldots,n} |\hat{\lambda}_{I,t}| > C \tag{3.7}$$

où

$$\hat{\eta}_{AO} = \max_{t=1,\ldots,n} |\hat{\lambda}_{A,t}| > C \tag{3.8}$$

où C est une constante positive convenablement choisie. En s'appuyant sur une étude de simulation Chang et *al* recommandent d'utiliser $C = 3$ pour une grande sensibilité aux outliers, $C = 3.5$ pour une moyenne sensibilité et $C = 4$ pour une petite sensibilité lorsque la taille des séries est inférieur à 200.

Distinguer un IO d'un AO est une opération cruciale pour corriger la série. En pratique nous avons peu d'informations disponible sur l'identité de l'outlier éventuel ; donc il n'est pas clair qui des tests de détection, (3.7) ou (3.8), est le plus approprié à une situation donnée. Car, lorsque le test appliqué n'est pas approprié, la puissance de détection pourrait être subtanciellement réduite. En outre, même s'il est connu qu'un outlier est survenu à un instant particulier, la possibilité de son effet adverse peut ne pas être facile à corriger à moins que sa nature soit proprement identifié. La statistique $\hat{\lambda}_{AI,d}$ est destiné à faire la

distinction entre IO et un AO en un point donné. Cependant, quand la position du possible outlier est inconnu, nous pouvons avoir besoin d'executer un tel test itérativement pour chaque point, et ceci peut devenir une tache très lourde.

Pour simplifier le problème, une alternative peut être cette règle simple proposée par Fox(1972) permettant de distinguer un IO d'un AO. Cette règle énonce que, à un instant suspect d, l'éventuel outlier est déclaré comme IO, (resp. AO), si $|\hat{\lambda}_{I,d}| > |\hat{\lambda}_{A,d}|$, (resp. $|\hat{\lambda}_{I,d}| \leq |\hat{\lambda}_{A,d}|$).

Une étude de simulation que, pour des séries de tailles modestes, $n = 50$, cette règle donne des performances comparables à celle du test statistique de $\hat{\lambda}_{IA,d}$ quand l'amplitude de l'outlier ω croit à $5\sigma_a$. En revanche, la comparaison est défavorable quand l'amplitude $\omega = 3\sigma_a$.

Pour toutes ces considérations, dans leurs procédure itérative de modélisation ARMA en présence d'outliers, Chang et *al* (1988) adoptent pour l'étape de détection des ouliers des deux types, le test statistique

$$\hat{\eta}_t = \max_{1,\ldots,n} \left\{ |\hat{\lambda}_{I,t}|, |\hat{\lambda}_{A,t}| \right\}.$$

3.1.2 Test basé sur les scores

Louni (2008) propose un test destiné à détecter et identifier les deux types d'outliers simultanément. Il précise ainsi l'identité de l'outlier sans éxigé d'autres calculs qu'impliquerait une procédure de distinction entre AO et IO. Une distribution asymptotique du test statistique a été obtenue. Ce qui constitue un avantage certain sur le test du rapport de vraisemblance où le reglage des seuils est obtenu numériquement. Par ailleurs, il est flexible, plus synthétique et facile interpréter. Ce test peut être appliqué itérativement pour identifier la position de tous les outliers. Il est appelé *test séquentiel modifié* dans la lignée d'un *test séquentiel* dû à Abraham et Yattawara (1988). Celui-ci comporte deux étapes basées sur les scores où les multiplicateurs de Lagrange. La première est une procédure de test pour la détection et la seconde est une procédure décision pour la détection et l'identification des outliers des deux types. Commençons par le décrire pour mieux comprendre le nouveau test.

Les mêmes modèles que précédemment sont utilisés pour décrire les outliers IO et AO. Nous les rappelons encore une fois pour la simplicité de l'exposé.

La présence d'un *Additive Outlier* à l'instant $t = d$ se traduit sur le modèle par

$$y_t = z_t + \omega_A \mathbf{1}_d(t) = \pi^{-1}(B)a_t + \omega_A \mathbf{1}_d(t)$$

où y_t est la série observée, z_t le processus ARMA, $\mathbf{1}_d(t)$ définie par $\mathbf{1}_d(t) = 1$ si $t = d$ et 0 sinon et ω_A l'amplitude de la perturbation.

Alternativement, la présence d'un *Innovational Outlier* à l'instant $t = d$ se traduit sur le modèle par

$$y_t = \pi^{-1}(B)[a_t + \omega_I \mathbf{1}_d(t)] = z_t + \pi^{-1}(B)\omega_I \mathbf{1}_d(t)$$

où ω_I est l'amplitude de la perturbation.

Dans le cas d'un modèle AR(p) où les deux types d'outliers peuvent être envisagés à l'instant $t = d$, l'observation s'écrit

$$y_t = z_t + \omega_A \mathbf{1}_d(t) \quad \text{avec} \quad \phi(B)z_t = a_t + \omega_I \mathbf{1}_d(t).$$

Lorsque les paramètres ϕ_i et σ^2 sont connus *a priori* et qu'un outlier est suspecté à $t = d$, le test de score permettant de tester $H_0 : \omega_I = \omega_A = 0$ (absence d'outlier) est

$$T_d = \frac{a_k^2}{\sigma^2} + \left(\sum_{i=1}^{p} \phi_i a_{d+i}\right)^2 \Big/ \left(\sigma^2 \sum_{i=1}^{p} \phi_i^2\right).$$

Puisque d est inconnu on considère la statistique

$$\max_{p+1 \le t \le n-p} T_t$$

La distribution asymptotique obtenue dépend d'un paramètre $\mu \in]0, 1]$ (quelquefois appelé *index extremal* de la suite). Empiriquement $\mu \simeq 0.8$.

Lorsque les paramètres sont inconnus, ils sont remplacés par les emv. $\max T_t$ est modifié en $\max \hat{T}_t$. On montre que $\max \hat{T}_t - \max T_t$ converge en probabilité vers zéro. Ainsi, pour

des échantillons assez grands, la procédure peut être utilisée.

Une fois la position de l'outlier déterminée, Abraham et Yatawara recommandent pour la distinction entre les deux types AO et IO l'utilisation de la statistique

$$S = \sigma^2(T_{1d} - T_{2d}),$$

où T_{1d} et T_{2d} sont les statistiques de scores testant respectivement les sous-hypothèses $H_{10} : \omega_A = 0$ et $H_{20} : \omega_I = 0$ données dans le cas d'un AR(p) par

$$T_{1d} = \left(a_d - \sum_{i=1}^{p} \phi_i a_{d+i} \right)^2 \Big/ \sigma^2 \left(1 + \sum_{i=1}^{p} \phi_i^2 \right) \quad \text{et} \quad T_{2d} = \frac{a_d^2}{\sigma^2}.$$

Si $S > 0$, (resp. $S < 0$), on conclut que l'outlier est AO (resp. IO).

Ces procédures de test et de décision peuvent être étendues au modèle ARMA en utilisant le représentation AR(∞).

Plutôt qu'une statistique basée sur la "somme" des 2 statistiques T_{1t} et T_{2t} Louni(2008) propose une statistique basée sur le maximum. Soit

$$T^* = \max_t T_t^* \tag{3.9}$$

où

$$T_t^* = \max\{T_{1t}, T_{2t}\}.$$

La distribution exacte de $T^* = \max T_t^*$ est difficile à obtenir à cause de la corrélation des T_t^*. Cependant, en utilisant des outils de la théorie des valeurs extrêmes, la distribution asymptotique du maximum d'une suite stationnaire avec une structure de dépendance qui n'est "trop forte" peut être obtenue. L'auteur a prouvé que la queue de la distribution de $\{T_t^*\}$ est de type exponentielle et l'*index extremal* de la suite $\{T_t^*\}$ est égale à 1. Ceci à cause du fait que la statistique $T_t^* = \max\{T_{1t}, T_{2t}\}$ où T_{1t} and T_{2t} sont des statistiques de scores à 1 *d.d.l.* plutôt que la statistique de score à 2 *d.d.l.* considérée par les auteurs précédents. Il en résulte la distribution asymptotique,

$$\lim_{n\to\infty} P\{a_n(T^* - b_n) \leq x\} = \exp\{-e^{-x}\},$$

avec les constantes de normalisation

$$a_n = 1/2 \quad \text{et}$$
$$b_n = 2\log(n-2p) + \log(8/\pi) - \log\{2\log(n-2p)\}.$$

Une approximation du seuil de signification $t_n(\alpha)$ à un niveau α, dans le cas d'un modèle AR(p), est donnée par

$$t_n(\alpha) = -2\log\{-\log(1-\alpha)\} + 2\log(n-2p) + \log(8/\pi) - \log\{2\log(n-2p)\}. \quad (3.10)$$

Quand les paramètres de l'AR et σ^2 ne sont pas connus ils sont estimés à partir des données. Comme pour les tests évoqués ci-haut, nous considérons l'estimateur $\hat{T}_t^*$ de T_t^*. Il est obtenu en remplaçant ϕ et σ^2 par les estimateurs du maximum de vraisemblance $\hat{\phi}$ et $\hat{\sigma}^2$ et a_t par les résidus récurssifs $\hat{a}_t = y_t - \sum_{l=1}^{p} \hat{\phi}_l y_{t-l}$ dans les équations (2) et (3). Ainsi, sous H_0, $\hat{T}^* - T^*$ converge en probabilité vers zéro. Après quoi,

$$\lim_{n\to\infty} P\{a_n(\hat{T}^* - b_n) \leq x\} = \exp\{-e^{-x}\}.$$

Donc pour des échantillons suffisamment grand, la procédure de test peut encore s'appliquer.

En faisant appel à d'autres outils de la même théorie, il montre encore que dans le cas ARMA la distribution asymptotique reste la même avec des constantes de normalisations légèrement modifiées.

Comme dans le cas AR, l'estimation des paramètres des modèles ARMA modifie les résidus et donc, les statistiques de tests, de quantités de l'ordre $O_p(\frac{1}{\sqrt{n}})$. Par conséquent, ces dernières n'induisent pas de modifications du test. Donc, pour des échantillons suffisamment grand, le test en question s'applique encore.

3.2 Méthodes robustes

Selon Barnett et Lewis (1994) deux méthodes générales de traitement des outliers se distinguent. La première méthode consiste à simplement identifier les outliers pour des études suplémentaires. Les tests de détection joue alors un rôle important. La deuxième méthode implique que les outliers en tant que tels ne sont pas le centre d'étude, le but consiste à travailler en toute sécurité malgré eux. Ces procédures sont considérées comme s'*adaptant*

aux outliers ou les accommodant. Les techniques d'accomodation sont dites *robustes* face à la pésence des outliers, cependant la *robustesse* n'est pas spécifique à l'examen des outliers. L'estimation des modèles ARMA a fait l'objet de nombreuses propositions dans la littérature. Certains estimateurs robustes nécessaire à notre étude sont décrits dans la section qui suit.

3.2.1 Estimation des modèles ARMA parfaitement observés

Considérons le processus autoregressif moyenne mobile ARMA d'ordre p et q *i.e.* une suite z_t satsifaisant à l'équation aux différences

$$z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \tag{3.11}$$

où les innovations a_t sont des variables aléatoires *i.i.d.* de distribution F. Posons $\phi = (\phi_1, \phi_2, \ldots, \phi_p)'$, $\theta = (\theta_1, \theta_2, \ldots, \theta_q)'$, et $\lambda = (\phi', \theta')'$, où prime désigne la transposée.

Rappelons que (3.11) peut aussi s'écrire comme

$$\phi(B) z_t = \theta(B) a_t,$$

où B est le polynôme retard tel que $Bz_t = z_{t-1}$ et $\phi(B)$, $\theta(B)$ sont les polynômes opérateurs

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p, \theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q.$$

Les racines des polynômes $\phi(B)$ et $\theta(B)$ sont supposées à l'extérieur du cercle unité pour assurer la stationnarité et l'inversibilité du processus.

Supposons maintenant que le processus observé y_t satisfait au modèle ARMA(p, q) donné par (3.11) mais les innovations à la place d'être, disons de loi normale, ont « une plus grande queue » que la loi normale. De telles distributions génèrent des outliers qui ont, on va le voir dans la suite, des effets considérables sur les estimateurs des moindres carrés des paramètres d'un modèle de regression. Par exemple, F peut être une distribution normale contaminée de la forme

$$F = (1 - \varepsilon) N(0, \sigma^2) + \varepsilon G, \tag{3.12}$$

où ε est petit et G est une loi arbitraire de dispersion plus grande que σ, par exemple $G = N(0, \xi^2)$ avec $\xi^2 \gg \sigma^2$. Ceci exprime que les innovations proviennent de $\mathbf{N}(0, \sigma^2)$ avec

une probabilité de $1-\varepsilon$ et de G avec une probabilité ε. Les innovations a_t provenant de G sont les *Innovationals Outliers*. Un tel modèle est dit ARMA IO.

Soulignons le fait important que le processus obsérvé y_t vérifie l'équation aux différences (3.11) bien que les innovations a_t contiennent des outliers. Ce qui justifie la terminologie de modèle ARMA parfaitement observé.

Estimateurs des moindres carrés

Soit $y_1, \ldots, y_n$ la série parfaitement observée correspondant au modèle ARMA(p,q). Les estimateurs des moindres carrés notés par $\hat{\lambda}_{MC}$, conditionnellement aux observations $y_1, \ldots, y_p$ et et aux valeurs initiales $a_p = a_{p-1} = a_{p-2} = \ldots = a_{p-q+1} = 0$, sont obtenus en résolvant le problème

$$\min_{\lambda} \sum_{t=p+1}^{n} \hat{a}_t^2(\lambda), \tag{3.13}$$

où

$$\hat{a}_t(\lambda) = \theta(B)^{-1}\phi(B)y_t, \quad \text{avec} \quad y_t = 0 \quad \text{pour} \quad t \leq 0.$$

Ces résidus sont calculés itérativement à partir de la formule de récurrence

$$\begin{aligned} \hat{a}_t(\lambda) = \; & y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} \\ & + \theta_1 \hat{a}_{t-1}(\lambda) + \cdots + \theta_q \hat{a}_{t-q}(\lambda) \end{aligned}$$

le point de départ de la récurrence étant

$$\hat{a}_p(\lambda) = \hat{a}_{p-1}(\lambda) = \cdots = \hat{a}_{p-q+1}(\lambda) = 0.$$

Sous l'hypothèse de normalité des innovations, ces estimateurs sont asymptotiquement efficaces. Malheuresement, les estimateurs des moindres carrés sont très sensibles à cette hypothèse de normalité et perdent beaucoup d'efficacité si la distribution F des innovations est une loi normale contaminée de la forme (3.12).

Les M-estimateurs

Si on pose

$$r_{t-j}(\lambda) = \frac{\hat{a}_t(\lambda)}{\partial \phi_i} = \theta^{-1}(B)y_{t-j} = \phi^{-1}(B)\hat{a}_{t-j}(\lambda)$$

et

$$s_{t-j}(\lambda) = \frac{\hat{a}_t(\lambda)}{\partial\theta_j} = \theta^{-2}(B)\phi(B)y_{t-i} = \theta^{-1}(B)\hat{a}_{t-i}(\lambda)$$
$$d_t(\lambda) = (r_{t-1}(\lambda), \ldots, r_{t-p}(\lambda), s_{t-1}(\lambda), \ldots, s_{t-q}(\lambda))',$$

la solution λ de (3.13) satisfait à l'équation vectorielle

$$\sum_{t=p+1}^{n} \hat{a}_t(\lambda) d_t(\lambda) = 0.$$

Par conséquent un M-estimateur $\hat{\lambda}_M$ de λ (et $\hat{\sigma}$ pour σ_a) est donné par la solution simultanée de

$$\sum_{t=p+1}^{n} \psi\left(\frac{\hat{a}_t(\lambda)}{\sigma}\right) d_t(\lambda) = 0 \tag{3.14}$$

$$\sum_{t=p+1}^{n} \chi\left(\frac{\hat{a}_t(\lambda)}{\sigma}\right) = 0 \tag{3.15}$$

avec les fonctions ψ et χ convenablement choisies continues et bornées (la dernière équation sert comme estimation simultanée du paramètre d'echelle).

Un bon choix de ψ est la fonction de Huber

$$\psi_{H,c} = \begin{cases} u & \text{si} \quad |u| \leq c, \\ \\ c\ \text{sign}(u) & \text{si} \quad |u| > c. \end{cases}$$

Pour se protéger contre des distributions à grande queue (*i.e.* loi de Cauchy) qui génèrent des outliers d'innovations, on peut utiliser une fonction ψ tel que $\psi(u) = 0$ si $|u| \geq c$, c étant une constante fixée *a priori*. Les fonctions ψ possédant cette propriété sont appelées fonctions redescendantes. Un type de fonction redescendante fréquemment utilisée est la fonction bicarrée de Tukey

$$\psi_{B,c} = \begin{cases} u(1 - \frac{u^2}{c^2})^2 & \text{si} \quad |u| \leq c, \\ \\ 0 & \text{si} \quad |u| > c. \end{cases}$$

Huber [19] propose de choisir χ dans la famille

$$\chi_{H,c}(u) = \psi^2_{H,c}(u) - b$$

avec $b = E_\Phi \psi^2_{H,c}(u)$ où Φ est une distribution $\mathbf{N}(0,1)$. Ce choix de b donne un estimateur $\hat{\sigma}$ consistant lorsque les résidus sont $\mathbf{N}(0,\sigma^2)$.

On peut également se donner *a priori* un estimateur robuste dS_n de σ, comme par exemple

$$S_n = \frac{1}{0,6745} Med\{|\hat{a}_t|\}_{p+1\leq t\leq n}.$$

Dans le cas où ψ est une fonction redescendante, le système (3.10) et (3.15) a plusieurs solutions et certaines parmi elles ne convergent pas vers la vraie valeur du paramètre. Par conséquent, un bon choix de l'estimateur initial $\hat{\lambda}_0^{(0)}$ dans l'algorithme de résolution du système est essentiel.

Là encore, malheureusement, ces M-estimateurs ne seront pas robustes vis à vis d'outliers de type AO. Ce fait est reporté par Denby et Martin (1979) pour les processus autoregressifs.

Nous allons donc considérer une classe d'estimateurs robustes pour les perturbations additives.

3.2.2 Estimation dans les modèles ARMA perturbés

Supposons maintenant que le processus observé y_t n'est pas un processus ARMA, à la place nous avons.

$$y_t = z_t + v_t,$$

où z_t est un processus ARMA(p,q) qui satisfait à l'équation (3.11) avec les résidus a_t de distribution $N(0,\sigma^2)$ et v_t une suite de variable aléatoire indépendante, indépendante de z_t. v_t a une distribution H, donnée par

$$H = (1-\varepsilon)\delta_0 + \varepsilon G,$$

où δ_0 est la masse de Dirac en zéro et G une distribution quelconque. Ainsi, avec une probabilité $1-\varepsilon$, z_t est un processus ARMA parfaitement observé, et avec une probabilité ε, l'observation est la réalisation d'un ARMA(p,q) plus une erreur de distribution G. ε est la proportion d'*Additives Outliers*. Ce modèle est dit ARMA AO.

GM-Estimateurs

En choisissant une fonction convenable $\eta : \mathbb{R}^{p+q} \times \mathbb{R} \to \mathbb{R}$, où $\eta(x, .)$ est continue, paire et bornée pour tout $x \in \mathbb{R}^{p+q}$, nous arrivons à un GM-estimateur pour λ par la solution simultanée des équations

$$\sum_{t=p+1}^{n} \eta\left(\frac{\hat{a}_t(\lambda)}{\sigma}\right) d_t(\lambda) = 0$$
$$\sum_{t=p+1}^{n} \chi\left(\frac{\hat{a}_t(\lambda)}{\sigma}\right) = 0.$$

Remarquons que le choix de $\eta(r, x) = r$ et $\chi(r) = r^2 - 1$ conduit estimateurs MC, et avec $\eta(x, r) = \psi(r)$ nous avons les M-estimateurs.

Un exemple type de $\eta(., .)$ comme l'ont suggérés Martin et Yohai (1986) est

$$\eta(u, v) = \psi(u)\psi(v)$$

avec ψ étant la fonction ψ_H de Huber ou la fonction bicarrée de Tukey ψ_T. Martin et Yohai (1986) montrent que les GM-estimateurs peuvent contrôler aussi bien les outliers IO que les outliers AO dans le cas d'un modèle AR(1). Le problème avec le GM-estimateur est sa performance quand l'ordre de la structure AR croit.

Plusieurs classes d'estimateurs plus perfomants sont proposés dans la littérature. Par exemple, Bustos et Yohai (1986) présentent deux nouvelles classes d'estimateurs basée sur l'autocovariance des résidus (RA estimateurs) et basée sur l'autocovariance des résidus tronqués (TRA estimateurs) qu'ils comparent favorablement aux estimateurs des moindres carrés, M et GM-estimateurs. Mais nous n'allons pas les étudier ici, notre présentation des estimateurs se limite uniquement à ceux que nous utiliserons dans la suite.

Chapitre 4

Procédure itérative avec intervention dans les modèles ARMA

4.1 Introduction

La technique de modélisation des outliers que nous examinerons dans la suite est basée sur le test du rapport de vraisemblance de Fox (1972) et sur l'analyse avec intervention de Box et Tiao (1975). Leur combinaison donne une solution au problème avec l'analyse avec intervention qui suppose que les interventions doivent être connues au préalable. La recherche des outliers donne aussi en même temps la position et le type d'intervention à modéliser.

Les premières propositions supposent que le modèle ARMA et le mécanisme générant les outliers connus, ce qui est la situation que traite d'ordinaire l'analyse avec intervention de Box et Tiao. Le développement de ces procédures itératives a progressé avec un gain certain de flexibilité et de réalisme. L'ultime étape, la plus réaliste, traite de modèles où le processus ARMA et l'instant correspondant (éventuellement plusieurs) de l'outlier sont inconnus.

Plusieurs propositions sur les procédures itératives avec intervention dans le cas ARMA existent, nous accordons un intérêt particulier à celle développée dans Chang et *al* (1988). C'est une procédure itérative comportant deux étapes : l'identification des outliers IO et AO et l'estimation des paramètres du processus autoregressif et moyenne mobile (ARMA) en présence des outliers. La puissance de la procédure itérative de détection des outliers est ob-

tenue numériquement. Aussi ladite procédure est comparée favorablement aux procédures d'estimations robustes décrites dans la dernière section du chapitre 3.

Dans leur procédure, Chang et *al* font appel dans la phase de détection et identification des outliers au test du rapport de vraisemblance décrit précédemment. L'objectif ici est d'executer cette phase en faisant appel cette fois au test basé sur les scores décrit dans le même chapitre. Cette description des deux tests fait ressortir que le deuxième test présente l'avantage d'avoir une loi asymptotique qui permet un reglage des seuils plus précis à l'inverse du premier test obtenu par des calculs de simulation. Cette précision, comme on devait s y attendre, enmène une meilleure puissance de détection. Bien qu'elle constitue une suite naturelle, la deuxième phase que constitue l'estimation des paramètres du modèle ARMA n'a pas été réalisée dans ce travail, les calculs trop lourds qu'elle implique dépasse largement le cadre de ce travail. Elle est inscrite dans nos perspectives.

La suite de ce chapitre s'organise comme suit. L'analyse avec intervention à l'origine des méthodes itératives est décrite dans la section 4.2. La section 4.3 présente une version modifiée de méthode itérative de Chang et *al*. La puissance du nouveau test (Louni (2008)) de détection sera donnée. Les résultats seront confrontées à ceux obtenus par Chang *et al*.

4.2 L'analyse avec intervention

L'analyse avec intervention forme la base de nombreuses procédures de modélisation des outliers. Elle a été introduite en premier par Box et Tiao (1975). Elle commence par se demander si une intervention connue d'un quelconque phénomène a des conséquences attendues sur les réalisations de ses séries temporelles, et si oui, comment cette intervention peut-elle être mesurée. Ils présentent pour les effets d'une intervention le modèle dynamique de la forme

$$y_t = z_t + f(\kappa, \xi, t),$$

$$\phi(B) z_t = \theta(B) a_t$$

et

$$f(\delta, \omega, I_t(d), t) = \sum_{d=1}^{K} \frac{\omega_d(L)}{\delta_d(L)} I_t(d),$$

où y_t est la série sous étude, z_t une série ARMA et $f(.)$ une fonction déterministe représentant les effets des variables exogènes, ξ, en particulier les interventions. Ici $\delta_d(L)$ et $\omega_d(L)$ $(d = 1, 2, \ldots, K)$ sont des polynômes, $I_t(d)$ est un indicateur qui dénote l'occurrence ou la non occurrence de l'intervention et K le nombre d'interventions.

L'indicateur $I_t(d)$ peut être une variable « impulsion »définie par

$$Q_t(d) = \begin{cases} 0, & \text{si} \quad t \neq d \\ 1, & \text{si} \quad t = d, \end{cases}$$

elle est alors destinée à rendre compte de l'influence sur y_t d'un phénomène ayant eu lieu à la date d uniquement, par exemple une grève.

Tout comme il peut être une variable de « saut »définie par

$$S_t(d) = \begin{cases} 0, & \text{si} \quad t < d \\ 1, & \text{si} \quad t \geq d, \end{cases}$$

dans ce cas elle rend compte de l'influence d'un phénomène commençant à la date d, par exemple le changement de réglementation.

Ou bien une variable « palier »définie par

$$P_t(d) = \begin{cases} 0, & \text{si} \quad t < d_1 \text{ ou } t > d_2, \\ 1, & \text{si} \quad d_1 \leq t \leq d_2, \end{cases}$$

qui est destinée à rendre compte d'un phénomène transitoire ayant lieu entre les dates d_1 et d_2, par exemple une modification provisoire de réglementation.

Lorsque la forme des polynômes ω_d et δ_d a été choisie, c'est à dire lorsqu'on a choisi leur degré, et éventuellement les racines de module 1 de δ_d, on peut estimer les divers paramètres par une méthode de type moindre carrés ou maximum de vraisemblance.

Cette formulation est flexible. Transformer la forme des polynômes $\delta_d(L)$ et $\omega_d(L)$, donne lieu à un ensemble important de structures dynamiques, incluant les outliers de types AO, IO, TC et LS, dans lequel on choisira selon le contexte. (Pour des exemples, voir Box et Tiao (1975), Figure B.).

4.3 Procédure itérative pour la détection des outliers et l'estimation des paramètres

La procédure itérative en question traite deux modèles avec intervention assez simples qui représentent une large proportion d'outliers rencontrés dans la pratique. Ces deux modèles, présentés précédemment, font références à l'outlier IO et l'outlier AO que nous rappelons en termes d'innovations a_t.

$$y_t = \frac{\theta(B)}{\phi(B)}\{a_t + \omega \mathbf{1}_d(t)\}, \quad \text{(IO)},$$

et

$$y_t = \frac{\theta(B)}{\phi(B)} a_t + \omega \mathbf{1}_d(t), \qquad \text{(AO)}$$

où y_t est la série observée et z_t le processus ARMA.

Le nouveau test pour la détection et l'identification simultanée des outliers des deux types mis en oeuvre pour la procédure itérative modifiée dans sa phase de détection est décrit dans le détail dans le chapitre précédent. Rappelons la définition du test.

$$\hat{T} = \max_t \{\hat{T}_{1t}, \hat{T}_{2t}\}, \quad t = 1, \ldots, n.$$

où les statistiques de tests $\hat{T}_{1t}$ et $\hat{T}_{2t}$ sont données respectivement par

$$\hat{T}_{1t} = \left(\hat{a}_t - \sum_{i=1}^{p} \hat{\phi}_i \hat{a}_{t+i}\right)^2 \Big/ \hat{\sigma}^2 \left(1 + \sum_{i=1}^{p} \phi_i^2\right),$$

et

$$\hat{T}_{2t} = a_t^2 / \hat{\sigma}^2.$$

Une approximation du seuil de signification $t_n(\alpha)$ à un niveau α, dans le cas d'un modèle AR(p), est donnée par

$$t_n(\alpha) = -2\log\{-\log(1-\alpha)\} + 2\log(n-2p) + \log(8/\pi) - \log\{2\log(n-2p)\}.$$

Dans le cas du modèle ARMA, grace à l'écriture AR(∞) de ce dernier, les modications du test sont données par,

$$\hat{T}_{1t} = \left(\hat{a}_t - \sum_{i=1}^{\infty} \hat{\pi}_i \hat{a}_{t+i} \right)^2 \Big/ \left[\hat{\sigma}^2 \left(1 + \sum_{i=1}^{\infty} \hat{\pi}_i^2 \right) \right],$$

$\hat{T}_{2t} = \hat{a}_t^2/\hat{\sigma}^2$ et le seuil de signification

$$t_n(\alpha) = -2\log\{-\log(1-\alpha)\} + 2\log n - 2\log\log n + \log(4/\pi).$$

Le test ci-dessus est éxecuté dans la procédure itérative pour traiter les situations où les IO et AO peuvent exister en nombre inconnu. La procédure commence avec la modélisation de la série observée y_t sous la supposition qu'elle ne contient pas d'outliers. Après quoi, l'étape de détection des outliers et celle de l'estimation des paramètres sont alternativement effectuées. Voici les détails de la procédure itérative.

4.3.1 Phase de détection des outliers

1. A partir du modèle estimé, calculer les résidus $\hat{e}_t$, et considérons $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n} \hat{e}_t^2$ comme estimateur de σ_a^2. Une possible alternative robuste de cet estimateur peut être basée sur la médiance de la valeur absolue des résidus.

2. Calculer les statistiques de scores $\hat{T}_{1t}$ et $\hat{T}_{2t}$ ci-dessus et poser $\hat{T}_t = \max_k\{\hat{T}_{1t}, \hat{T}_{2t}\}$ pour $t = 1, \ldots, n$. Si $\max \hat{T}_t = \hat{T}_{2d} > t_n(\alpha)$, alors la possibilté d'un IO au temps $t = d$ est significative. L'impact ω de cet outlier est estimé par $\hat{\omega}_I$ donné dans (3.5). Il est ainsi éléminé en définissant un nouveau résidu $\check{e}_d = \hat{e}_d - \hat{\omega}_I$ au temps d. Si $\max \hat{T}_t = \hat{T}_{1d} > t_n(\alpha)$, alors c'est la possibilté d'un AO au temps $t = d$ qui est significative et son impact est estimé alors par $\hat{\omega}_A$ donné dans (3.6). L'effet de cet AO peut être corrigé en introduisant de nouveaux résidus $\check{e}_t = \hat{e}_t - \hat{\omega}_A\hat{\pi}(B)\mathbf{1}_d(t)$ pour $t \geq d$. En accord avec le type d'outlier un nouvel estimateur $\check{\sigma}_a^2$ est calculé à partir des résidus modifiés.

3. Si un IO ou un AO est capturé dans l'étape 2, recalculer les statistiques $\hat{T}_{1t}$ et $\hat{T}_{2t}$, pour $t = 1, \ldots, n$ à partir des mêmes estimateurs initiaux des paramètres de la série, mais en utilisant les résidus modifiés $\check{e}_t$ et l'estimateur $\check{\sigma}_a$, puis répéter l'étape 2.

4. Continuer de répéter les étapes 2 et 3 jusqu'à ce que plus aucun outlier ne soit détecté.

4.3.2 Phase d'estimation des paramètres

5. Supposons que k temps $d_1, d_2, \ldots, d_k$ sont déclarés comme les occurrences des éventuels outliers IO et AO. Traiter ces instants comme connus, et estimer simultanément les impacts de ces outliers $\omega_1, \omega_1, \ldots, \omega_k$ et les paramètres de la série en utilisant, comme décrit dans Box et Tiao (1975), le modèle de la forme

$$y_t = \sum_{j=1}^{k} \omega_j L_j(B) \mathbf{1}_{k_j}(t) + \frac{\theta(B)}{\phi(B)} a_t, \tag{4.1}$$

où $L_j(B) = 1$ en présence d'un AO et $L_j(B) = \theta(B)/\phi(B)$ dans le cas d'un IO à l'instant $t = k_j$.

En traitant le modèle (4.1) ci-dessus comme le modèle intermédiaire, nous executons de nouveau l'étape de détection. Les notations $\hat{\pi}_j$, $\hat{\omega}_j$ et $\hat{e}_t$ représentent les valeurs estimées obtenues à partir de l'estimation conjointe de tous les paramètres du modèle (4.1). Si aucun outlier n'est détecté, nous arrétons. Sinon, l'étape d'estimation est répétée, avec les nouveaux outliers identifiés incorporés dans le modèle (4.1), jusqu'à ce que plus aucun outlier ne soit identifié et tous les impacts ont été simultanément estimés avec les paramètres de la série.

4.4 Performance de la procédure itérative

Nous avons conduit une étude de simulation dans le but d'obtenir des informations sur la performance de la procédure ci-dessus. Les résultats sur la puissance de détection des outliers sont donnés dans la section qui suit. L'estimation robuste dépasse largement le cadre de ce travail, cependant, bien que les calculs ne soient pas réalisés sur les mêmes données que la puissance de la détection, nous avons reporté ici ceux obtenus par Chang *et al* pour montrer la supériorité de la procédure itérative de modélisation ARMA.

4.4.1 Puissance de la procédure itérative de détection des outliers

La première étape de la procédure itérative est destinée à la détection des outliers et l'identification de leurs type. Comme Chang *et al.* (1988), pour évaluer la puissance des deux tests, on devrait estimer la probabilité d'une détection correcte de la position de l'outlier et la probabilité d'une identification correcte du type de l'outlier. Dans cette étude

nous avons considéré la situation (a) d'un ou deux outliers, (b) la série x_t est un AR(1), (c) les deux types AO et IO, (d) taille des outliers, et (e) la taille des échantillons. Pour établir une comparaison des performance des deux procédures nous avons aussi refait les calculs des puissances de Chang *et al* à partir des mêmes données, ils sont donnés dans la table 1.

Dans le cas d'un seul outlier, nous avons examiné 24 situations en considérant une série temporelle générée par par un AR(1) de paramètre $\phi = .6$ et $\sigma_a^2 = 1$, deux types d'outliers IO et AO, deux tailles d'outliers $\omega = 3\sigma_a$ et $\omega = 5\sigma_a$, et trois tailles d'échantillon, $n = 50$,$n = 100$ et $n = 150$. Les occurrences des outliers se trouvent au milieu des échantillons, soit $d = 26$ pour $n = 50$, $d = 51$ pour $n = 100$ et $d = 76$ pour $n = 150$.

Dans le cas de deux outliers, nous considérons le même modèle $AR(1)$ de paramètre ϕ et $\sigma_a^2 = 1$, trois types d'outliers, 2AO, 2IO et un AO, un IO. Deux amplitudes de perturbation $\omega = 3\sigma_a$ et $\omega = 5\sigma_a$ pour les deux outliers et trois tailles $n = 50$, $n = 100$ et $n = 150$ pour les échantillons. On suppose que les deux outliers surviennent aux instants $d_1 = 17$ et $d_2 = 34$ pour n=50, $d_1 = 34$ et $d_2 = 66$ pour $n = 100$ et $d_1 = 51$ et $d_2 = 101$ pour n=150.

Dans chacun des cas précédents, les données de séries temporelles sont générées en accord avec les spécifications citées. En supposant que le modèle est connu et qu'il n y a pas d'outliers dans l'échantillon, à l'aide de la méthode du maximum de vraisemblance exacte, les paramètres du modèle sont estimés et par suite les résidus. Partant de ces paramètres et résidus, la première itération est executée avec une valeur critique prédéterminée $C = 4$ pour le premier test et celle donnée dans pour le second. Les résultas sont alors comparés avec la spécification des alternatives. En répétant ces opérations (1000), nous pouvons estimer la probabilité de détecter correctement la position d'un outlier et la position d'identifier correctement le type AO ou IO sachant que la position est était correctement détectée. La table 1 et 2 donnent les résultats sur 1000 répétitions des opérations précédentes.

Les nombres entre parenthèses sont les pourcentages d'identification correcte sachant que la position est était correctement détectée. Dans le cas de deux outliers, les lignes désignées par 1^{er} et 2^{er} montre les résultas de détection et identification du premier et second outlier respectivement.

Les résultats montrent que la performance du nouveau test est comparée favorablement au test du rapport de vraisemblance.

Table 1. Fréquences de la détection correcte de la position des outliers (pourcentage d'identification correcte du type) dans le test du rapport de vraisemblance : Valeur critique $C = 4$; 1000 répétitions

	$\omega = 3\sigma_a$			$\omega = 5\sigma_a$		
n	50	100	150	50	100	150
AR, 1 AO	97	203	255	690	903	937
	(.76)	(.84)	(.85)	(.85)	(.92)	(.93)
AR, 1 IO	61	105	122	557	721	791
	(.95)	(.92)	(.88)	(.97)	(.95)	(.92)
AR, 2 AO	10	35	49	312	649	708
	(.50)	(.56)	(.81)	(.48)	(.70)	(.77)
1^{er}outlier	74	169	218	539	810	837
	(.75)	(.81)	(.86)	(.77)	(.88)	(.91)
2^{eme}outlier	77	172	223	527	804	862
	(.74)	(.81)	(.88)	(.77)	(.87)	(.90)
AR, 2 IO	14	22	28	402	615	640
	(.99)	(.93)	(.89)	(.98)	(.94)	(.93)
1^{er} outlier	71	130	152	610	783	834
	(.96)	(.97)	(.90)	(.98)	(.96)	(.96)
2^{eme} outlier	79	139	160	643	782	830
	(.98)	(.98)	(.93)	(.97)	(.97)	(.97)
AR, IO AO	11	28	35	330	563	607
	(.90)	(.76)	(.88)	(.83)	(.88)	(.88)
1^{er} outlier	48	122	125	414	605	651
	(.97)	(.95)	(.95)	(.99)	(.98)	(.97)
2^{eme} outlier	112	200	267	747	914	947
	(.78)	(.85)	(.88)	(.87)	(.92)	(.94)

Table 2. Fréquences de la détection correcte de la position des outliers (pourcentage d'identification correcte du type) dans le test basé sur les scores : le seuil critique = .01, 1000 répétitions

	$\omega = 3\sigma_a$			$\omega = 5\sigma_a$		
n	50	100	150	50	100	150
AR, 1 AO	102	165	181	704	865	905
	(.80)	(.83)	(.84)	(.82)	(.92)	(.92)
AR, 1 IO	65	80	88	570	675	684
	(.92)	(.90)	(.86)	(.96)	(.94)	(.94)
AR, 2 AO	14	20	21	342	559	629
	(.43)	(.53)	(.73)	(.51)	(.70)	(.81)
1^{er}outlier	78	130	149	543	762	798
	(.75)	(.85)	(.85)	(.77)	(.87)	(.91)
2^{eme}outlier	72	130	156	574	763	802
	(.73)	(.84)	(.85)	(.78)	(.87)	(.91)
AR, 2 IO	16	20	22	498	555	557
	(.94)	(.84)	(.90)	(.96)	(.97)	(.96)
1^{er} outlier	133	147	150	784	791	797
	(.96)	(.96)	(.96)	(.98)	(.98)	(.96)
2^{eme} outlier	80	108	111	623	714	717
	(.92)	(.92)	(.92)	(.97)	(.97)	(.95)
AR, IO AO	15	15	21	341	477	519
	(.80)	(.86)	(.76)	(.80)	(.86)	(.89)
1^{er} outlier	53	78	75	431	577	599
	(.96)	(.95)	(.92)	(.98)	(.97)	(.96)
2^{eme} outlier	118	185	185	746	876	905
	(.75)	(.84)	(.84)	(.88)	(.92)	(.94)

4.4.2 Estimation en présence des outliers

Pour mettre en perspective la procédure itérative, dans cette section, nous présentons l'étude comparative réalisée par Chang *et al* sur l'estimation des paramètres de la série temporelle. Ces auteurs opposent les estimations obtenues par la méthode itérative à celles données par les M-estimateurs, les GM-estimateurs proposés par Denby et Martin (1979).

La série temporelle a été générée par un modèle $AR(1)$ sous différentes situations. En plus du modèle nul, le cas IO et le cas AO, ils ont aussi considérés le modèle contaminé $100(\gamma_a + \gamma_v - \gamma_a\gamma_v)\%$ dans laquel le nombre d'outliers est déterminé par mécanisme aléatoire :

$$y_t = \phi y_{t-1} + a_t^*, \quad z_t = y_t + v_t, \tag{4.2}$$

où z_t sont les observations, les a_t^* sont *i.i.d.* de densité normale contaminée

$$(1-\gamma_a)N(0,\sigma_a^2) + \gamma_a N(0,\kappa_a^2\sigma_a^2),$$

et les v_t sont *i.i.d.* avec une densité

$$(1-\gamma_a)\delta(0) + \gamma_v N(0,\kappa_a^2\sigma_a^2),$$

où γ_0 représente la densité dégénérée en 0 et v_t et a_t^* sont toutes indépendantes. Ce modèle, considéré aussi par Denby et Martin(1979), peut générer les IO ou AO ou les deux. Dans la suite ce modèle est dit *modèle mixte*.

Pour chaque série simulée, les estimateurs suivants du paramètre du modèle AR(1) sont calculés :

- Estimateur des moindres carrés (MC).
- M-estimateur défini à partir de la fonction de Huber (M-H) où la constante C_{aH} est fixée à 1.5.

- M-estimateur défini à partir de la fonction bicarré (M-B) avec la constante C_{aH} fixée à 6.0.

- GM-estimateur défini par deux fonctions de poids de type Huber pour les résidus et les observations respectivement (GM-H). Les constantes sont $C_{aH} = 1.5$ et $C_{ZH} = 1.0$.

- GM-estimateur donné par deux fonctions de poids de type bicarrée pour les résidus et les observations respectivement (GM-B). Les constantes sont $C_{aB} = 6.0$ et $C_{ZB} = 3.9$.
- Estimateur calculé par la méthode itérative décrite précédemment avec une valeurs critique $C = 3$ (P).

A partir de l'algorithme IWLS (*iterated weighted least squares algorithm*) de (Beaton et Tukey 1974) les M-H et GM-H estimateurs sont obtenus avec les estimateurs MC comme valeur initiale (voir Denby et Martin (1979)). Les M-H et GM-H estimateurs sont alors utilisés comme valeurs initiales pour d'autres itérations de l'algorithme IWLS pour obtenir les M-B et GM-B estimateurs respectivement. Dans ces itérations, le paramètre d'échelle σ_a est estimé par la médiane des valeurs absolues des résidus divisée par .6745 et σ_z par la médiane de la valeur absolue des déviations des observations de la médiane de leurs échantillon divisée par .6745.

Comme les estimateurs cités, excépté l'estimateur MC, sont calculés à partir de l'algorithme IWLS, un critère de convergence est nécessaire. L'algorithme a été est considéré comme convergeant lorsque les estimateurs donnés par les deux plus récentes itérations différent des autres par un seuil inférieur à .0001

Pour chaque modèle en présence d'outliers, 500 échantillons de même taille sont générés avec $\sigma_a^2 = 1$ et, pour le modèle mixte (4.2), $\sigma_\varepsilon = 1$. Les tailles sont modérées, $n = 50$ ou $n = 75$. Les moyennes et les écarts types aussi bien que les erreurs quadratiques moyennes des estimateurs du paramètre ϕ sont calculés à partir de 500 répétitions. Le tableau 3 présente les résultats trouvés dans le cas où $\phi = .6$. Des résultas similaires sont obtenus dans le cas où $\phi = .9$. Les approximations communément adoptées, $(1-\phi^2)/n)^{1/2}$, pour les écarts types des estimateurs des moindres carrés (MC) de ϕ sous le modèle nul sont aussi donnés en bas de la table pour référence.

TAB. 4.1 – La moyenne, L'écart type et la moyenne quadratiques des erreurs des estimateur des paramètres de $AR(1)$

Modèle des outliers	*LS*	*M-H*	*M-B*	*GM-H*	*GM-B*	*P*
Nul	.5885	.5883	.5885	.5947	.5940	.5933
n=50	(.1143)	(.1157)	(.1147)	(.1223)	(.1225)	(.1165)
	.0132	.0135	.0133	.0150	.0150	.0136
1 IO fixé	.5804	.5808	.5814	.5813	.5756	.5932
T=24, n=50,	(.1110)	(.0989)	(.0978)	(.1125)	(.1251)	.0968
$\omega = 5$	.127	.0101	.0099	.0130	.0162	.0094
1 AO fixé	.4013	.4746	.4721	.5441	.5724	.5802
T=24, n=50,	(.1270)	(.1294)	(.1332)	(.1298)	(.1283)	(.1304)
$\omega = 5$	.0399	.0324	.0341	.0199	.0172	.0174
2 AO fixé	.4013	.4239	.4202	.5123	.5513	.5496
T=17, n=34,	(.1369)	(.1386)	(.1418)	(.1356)	(.1343)	(.1428)
$\omega = -3, 5, n = 50$	.0130	.0104	.0105	.0124	.0144	.0102
2 IO fixé	.5809	.5847	.5849	.5886	.5854	.5998
T=24, n=50,	(.1123)	(.1008)	(.1014)	(.1110)	(.1192)	(.1011)
$\omega = -3, 5, n = 50$	.0130	.0104	.0105	.0124	.0144	.0102
3 AO fixé	.3265	.3350	.3283	.4505	.5151	.4971
T=14,26,38,	(.1354)	(.1300)	(.1319)	(.1393)	(.1524)	(.1686)
$\omega = -4, 5, 4, n = 50$	.0931	.0871	.0912	.0417	.0304	.390
3 AO fixé	.3863	.4060	.3990	.5075	.5541	5649
T=20,42,61,	(.1112)	(.1117)	(.1109)	(.1064)	(.1054)	(.0141)
$\omega = -4, 5, 4, n = 75$	.0580	.0501	.0527	.0199	.0132	.0141
mixte	.4929	.5056	.5094	.5470	.5604	5606
$\gamma_a = .02$, $\gamma_v = .02$,	(.1559)	(.1460)	(.1455)	(.1241)	(.1246)	(.1242)
$\kappa_a^2 = 25$,$\kappa_v^2 = 25$, n=50	.0357	0302	.0293	.0182	.0171	.0170
mixte	.4540	4720	4727	.5183	.5360	5448
$\gamma_a = .02$, $\gamma_v = .05$,	(.1617)	(.1478)	(.1491)	(.1308)	(.1354)	(.1289)
$\kappa_a^2 = 16$,$\kappa_v^2 = 16$, n=50	.0474	.0382	.0384	.0237	.0224	.0196

Note : $((1 - \phi^2)/n)^{1/2} \geq .1131$ quand n=50. la première entrée dans chaque cas est la moyenne ; la seconde entrée (entre parenthèses) est l'écart type et la troisième est les erreurs quadratiques moyennes

Ces résultats révèlent que le biais des M-estimateurs dans le cas AO ont un biais plus sévère que que ceux des estimateurs des moindres carrés. Ce résultat a déjà été établit par Denby et Martin (1979). D'un autre coté les GM-estimateurs sont moins efficaces sous le modèle nul et le modèle IO. L'estimateur calculé à partir de la procédure itérative de Chang *et al.* donne cependant la plus petite moyenne quadratique dans la plupart des cas étudiés. Il est comparé favorablement avec chacune des les méthodes robustes M-H, M-B, et GM-H, et seulement dans le cas AO avec $\phi = .6$ la procédure itérative est quelquefois mauvaise relativement au GM-H.

Les erreurs quadratiques moyennes de l' estimateur de la procédure en question est, dans la plupart des cas, légèrement proche de ceux de l'estimateur des moindres carrés sous le modèle nul, probablement à cause du fait de la qualité de détection dans la procédure. De plus, la plupart des GM-B estimateurs sont obtenus après avoir effectuer huit itérations avec l'algorithme IWLS, alors que dans la procédure itérative au plus quatre pour le cycle estimation.

Chapitre 5

Conclusion

Plus aucune questions ne se pose sur les effets négatifs qu'occasionnent les outliers sur les méthodes traditionneles. La présence des outliers peut en effet conduire à des estimations biaisées de paramètres et, suite à la réalisation de tests statistiques, à une interprétation des résultats qui peut être altérée. La première partie de ce travail a permis de mettre l'accent sur les diverses notions à prendre en compte lors de l'examen des outliers. Les termes principaux de la définition d'un outlier ont été développés. Celui-ci correspond à une valeur particulièrement surprenante et, en fonction de l'objectif fixé, est statistiquement discordante dans le contexte d'une modèle de probabilité désigné initialement. La nature des outliers (caractère aléatoire ou déterministe) détermine clairement la manière de traiter ceux-ci ultérieurement. Les objectifs poursuivis lors de l'examen des outliers sont le rejet, l'incorporation, l'identification, l'accomodation ou la correction.

En fonction de l'objectif à atteindre et de la nature de la valeur anormale, le traitement des données est très différent. Il est donc primordiale de déterminer au préalable la nature et les objectifs à poursuivre lors de toute étude de valeurs qui semblent suspectes.

Losque l'objectif est de rejeter ou d'identifier une valeur anormale, la discordance de cette valeur est évaluée par des tests statistiques. Des méthodes spécifiques liées à l'accomodation ont été développées plus tard. Celles-ci permettent de minimmiser l'influence des outliers lorsque toutes les données sont prises en compte lors de l'analyses des données.

L'examen des échantillons univariés pour ajuster des modèles et estimer les paramètres, bien qu'il constitue une part importante de la pratique statistique, est quelque peu limité

dans les buts. Plus souvent, et plus utilement, le besoin de considérer des situations plus structurées s'impose. Dans les cas très structurés, comme les modèles de séries temporelles, il faut aussi s'attendre à rencontrer des données non representatives comme étant des outliers. Tout comme précédemment des méthodes formelles pour leur rejet ou leur accommodation sont développées.

Les méthodes robustes sont quelquefois vues comme plus objectives et préférables à la détection des outliers et procédures de modélisation. Mais les méthodes d'estimation robustes exigent qu'une fonction de poids soit sélectionnée parmi un nombre d'alternatives, et ce choix implique lui aussi un élément de subjectivité. De plus les méthodes robustes souffrent d'un certain manque d'efficacité, notamment en l'absence des outliers.

Dans beaucoup de cas nous pensons que, le moyen le plus efficace de manier les outliers est d'utiliser une méthode pour leur détection et identification après quoi, modéliser la série en utilisant une des variantes des méthodes avec intervention. De cette manière toute l'information contenue dans les données sera utilisée à la différence des méthodes robustes. Aussi, en utilisant les méthodes robustes, l'analyse perd de précieuses informations sur les données qui elles restent disponibles dans les procédure de détection. Et un argument de taille, ce mémoire le montre, la méthode itérative de modélisation donne de meilleur perfomance que l'estimation robuste du modèle.

Maintenant une perspective naturelle à ce travail peut être l'implémentation de la procédure itérative modifiée et l'estimation robuste. Comparer les deux méthodes et confronter les résultats à ceux de la procédure itérative originale.

Bibliographie

Abraham, B. and Box, G.E.P. (1979) Bayesian analysis of some outlier problems in time series. *Biometrika*, 66, 229-36.

Abraham, B. and Chuang, A. (1989) Outliers detection and time series modeling. *Technometrics*, 31, 241-48.

Barnett, V. and Lewis, T. (1994) *Outliers in statistical data.* 3rd ed. John Wiley, Chichester.

Battaglia, F. and Orfei, L. (2005) Outlier detection and estimation in nonlinear time series. *Journal of Time Series Analysis*, 26, 107-21.

Box, G. E. P., and Tiao, G. C. (1975) Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70, 70-79.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series : Theory and Methods,* 2nd Edn. New-York : Springer.

Bustos, O. H. and Yohai, V. J (1986) Robust estimates for ARMA models. *Journal of the American Statistical Association*, 81, 155-168.

Chang, I., Tiao, G. C. and Chen, C. (1988) Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.

Deutsch, S. J., Richards, J. E. and Swain, J. J. (1990) Effects of a single outlier on ARMA identification. *Communications in Statistics - Theory and methods*, 19, 2207-27.

Fox, A. J. (1972) Outliers in time series. *Journal of the Royal Statistical Society*, Series B, 34, 350-63.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes.* Springer, New-York.

Louni, H. (2008) Outliers detection in times series. *Journal of Time Series Analysis*, 9, 109-19.

Tsay, R. S. (1988) Outliers, level shifts and variance changes in time series. *Journal of Forecasting*, 7, 1-20.

Yohai, V. J. (1987) High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15, 642-656.

Wu, L.S-Y, Hosking, J.R.M. and N. Ravinshanker (1993) Reallocation outliers in time series. *Journal of the Royal Statistical Society*, Series C (Apllied Statistics), 42, 301-313.