1

26

25

24

22

MicroD

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, S.R.C. 1970, c. C-30, et ses amendements subséquents.

DETECTING OVERLY INFLUENTIAL OBSERVATIONS IN TIME SERIES

BY

BERNARD LEFRANÇOIS, B. Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Science

Department of Mathematics and Statistics
Carleton University
Ottawa, Ontario

May 5, 1989

© copyright

1989, Bernard Lefrançois

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0 315 54427 9

The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis

"DETECTING OVERLY INFLUENTIAL OBSERVATIONS IN TIME SERIES"

submitted by

Bernard Lefrançois, B. Sc.

in partial fulfilment of the requirements for
the degree of Master of Science

Thesis Supervisor

Chair, Department of Mathematics and Statistics

Carleton University

June 8, 1989

ABSTRACT

Previous work on the detection of overly influential observations in time series (in the time domain) suggested using measures of influence obtained by analogy with those used in the standard linear model. Furthermore, no criteria were provided to assess the potential discordancy in the influence of an observation. We first review the theory of discordant observations in the linear model. Then we present a simple method for obtaining various measures of influence for the autocorrelation function. We also obtain some of the distributional properties of the measure of sample influence. Hence, thresholds for declaring an observation discordant in its influence can be computed. Finally, three examples of the use of the measure of sample influence with these thresholds are presented.

ACKNOWLEDGEMENTS

First, I would like to thank Darryl Rhoades, Ruth McDougall and Statistics Canada for having given me the time and the financial assistance to complete this thesis.

Next, I would like to thank my thesis supervisor Professor J.N.K. Rao for his careful readings and penetrating comments.

Finally, I dedicate this thesis to the memory of my parents, whose lives were too short to witness its completion.

TABLE OF CONTENTS

	Page
Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
 Chapter I. Introduction	 1
 Chapter II. An Overview of Discordant Observations in the Standard Linear Model	 7
2.1 Discordancies in the Standard Linear Model	12
2.2 Measures of Leverage	16
2.3 Measures of Outlyingness	21
2.4 Measures of Influence	33
2.4.1 The Influence Function	36
2.4.2 Obtaining Measures of Influence	38
2.4.3 Identifying Influential Observations	45
 Chapter III. Influential Observations in Time Series	 56
3.1 Introduction to ARIMA Models	57
3.2 Influence Functions for Time Series	65
3.3 Measures of Influence for the Autocorrelation Function	 67
3.3.1 The Theoretical and Sample ACFs	71
3.3.2 The Sample ACF when One Observation is Missing	 75
3.3.3 Measures of Influence for the Sample ACF	80
3.3.4 Proofs of Propositions 3.4 and 3.5	89

TABLE OF CONTENTS (Cont'd)

	Page
Chapter IV. Identifying Influential Observations in Time Series	103
4.1 The Asymptotic Measure of Sample Influence .	107
4.2 Testing the Influence of a Specific Observation	111
4.3 Testing the Most Influential Observation . .	130
Chapter V. Examples	147
5.1 Proportions of Fatal Road Accidents in the British Isles	151
5.2 Monthly Sales of Variety Stores in the U.S. .	160
5.3 Chemical Process Concentration Readings . . .	177
5.4 Computational Details	184
Chapter VI. Conclusions and Suggestions for Further Research	187
Bibliography	194

LIST OF TABLES

Table	Description	Page
1	Percentage Points of the $\mathcal{W}_1$ Distribution.	123
2	Comparison of the Upper Percentage Points of $\mathcal{W}_1$ With Those From a Log-Normal Approximation.	143
3	Comparison of Chang's(1982) Outlier Analysis with the Influence Analysis of Variety Stores Sales.	169

LIST OF FIGURES

Figure	Description	Page
1	The Elements (+) of a 'Clothes-Pin'.	129
2	Proportions of Fatal Road Accidents in the British Isles.	155
3	Sample ACF and PACF of Fatal Road Accidents.	156
4	Sample ACF and PACF of First Differences of Fatal Road Accidents.	157
5	Influences on the Whole Sample ACF (24 Lags) of Fatal Road Accidents.	158
6	Influences of the Second Autocorrelation Coefficient of the Residuals from Model (5.2).	159
7	Logarithm of Retail Sales of Variety Stores in the U.S..	170
8	Sample ACF and PACF of Variety Stores Sales.	171
9	Sample ACF and PACF of the Differenced Variety Stores Sales.	172
10	Influences on the Whole Sample ACF (24 Lags) of Variety Stores Sales.	173
11	Sample ACF and PACF of the Differenced Variety Stores Sales Excluding Observations 1976/4 and 1977/4.	174
12	Influences on the Whole Sample ACF (24 Lags) of Variety Stores Sales Excluding Observations 1976/4 and 1977/4.	175
13	Comparison of the Seasonally Differenced Variety Stores Sales Series with Chang's Correction and the Correction in Level.	176

LIST OF FIGURES (Cont'd)

Figure	Description	Page
14	Concentration Readings of a Chemical Process (Series A of Box and Jenkins).	181
15	Sample ACF and PACF of the First Differences of Series A.	182
16	Influences on the Whole Sample ACF (20 Lags) of the Differenced Series A.	183

CHAPTER I

INTRODUCTION

The purpose of this thesis is to develop a tool for identifying influential observations in time series when these are viewed in the time domain.

Recall that a time series is a stochastic process $(X_t)_{t \in \mathbb{Z}}$ where X_t is a real random variable observed at the non-random time t . Typical real-life phenomena that may be modelled as time series abound: the concentration of ozone at a specific altitude and location, an electro-cardiogram monitored over a certain period of time, or the unemployment rate as measured monthly by Statistics Canada.

Currently, the most widely used class of models for time series is the class of Autoregressive, Integrated Moving Average (ARIMA) models, which was popularized by Box and Jenkins (1976). To apply these models, it must be assumed that the time series of interest is, possibly after some transformations, weakly stationary. Loosely speaking, this means that the current value of the time series can be expressed as a constant linear combination of its past values. This linear combination may be finite or infinite.

One fundamental quantity in time series analysis is the autocorrelation function (ACF). With it, one can identify a few specific ARIMA models for a given realization of a time series. Then, the parameters of these tentative models can be estimated, and this is accomplished, in a first approximation, by finding the parameters' values generating an ACF which most closely resemble the observed ACF. Finally, the adequacy of these tentative models can be judged by analyzing both the closeness of the theoretical ACF to the observed one, and the ACF of the models' residuals.

When actually observing a phenomenon, a particular observation will sometimes appear to be strikingly different from the others. For instance, while recording the ozone concentration in the upper atmosphere, a cloud of particles from a volcanic eruption affects the reading; while monitoring an electro-cardiogram, the patient gets excited; while measuring the monthly output from a particular sector of the economy, a significant proportion of the work force strikes for one week.

If our knowledge of these phenomena was of sufficient extent so as to have led us in formulating a model (formal or not) of their behaviour, then all these examples have in common that the observation made does not behave in the way we think it should. In other words, these observations lie

outside, or are at the fringe of being outside of the scope of our model(s). Thus the name outliers for such observations. Note that outliers can only be defined with respect to a model.

It is important to detect outliers for at least three reasons. First, they may not be observations from the phenomenon under study, e.g. they may be due to transcription errors. Second, they may reflect a fundamental inadequacy of the entertained model which, if studied, can lead to major improvements, e.g. the discovery of the planet Neptune by Gale in 1846. Finally, they may have an adverse effect on the estimated parameters of the model, either by forcing a specific value for the unknown parameter with (almost) no regard for the (different) value emerging from the bulk of the data, or by altogether changing the model, and thus lead to invalid inferences.

But outliers are only the tip of the iceberg in what has been aptly called case analysis by some authors. For an outlier may influence the estimate of a parameter, or it may not. In fact some observations may not be outliers, but still may be influential in determining the values of the parameters. Such observations clearly need also to be detected.

Some tools have already been developed for detecting outlying observations relative to ARIMA models for time series. The main references here are Abraham and Box(1979), Abraham and Yatawara(1988), Chang(1982), Chang, Tiao and Chen (1988), Fox(1972) and Tsay(1988). All these methods however, suffer from relying on a possibly contaminated sample to identify an appropriate model and to obtain its estimated parameters. Since both the structure and the estimated parameters of the models may have already been overly influenced by some observations, which may also be outliers, the power of these methods for identifying outliers may be seriously reduced.

Specific tools to identify observations influencing the ACF of time series have also been developed by Chernick, Downing and Pike(1982), Lattin(1983), Li and Hui(1987) and by Polasek(1982). However, all of these are found to be wanting in two areas. First, almost all were developed in an ad-hoc fashion, without recourse to a general appropriate theory. Second, none comes with thresholds (critical values) outside which a specific observation can be declared influential or not.

These are the two facets of the study of influential observations in time series that we shall develop in this thesis.

To obtain a proper theoretical framework for the study of influential observations in time series, we needed to review the current theory behind the identification of discordant observations (a generic term that we will precisely define for referring to outliers, influential observations and leverage points) in the standard linear regression model, a situation for which the theory is almost complete. In reviewing this theory, however, we found that most presentations were failing to clearly show the common thread in case analysis.

Thus, in Chapter II we attempt to briefly present all the major aspects of this theory in a way that should prove useful both in averting any confusion in the concepts, and in helping us judge the intrinsic qualities of many, if not all, measures of discordancy that have been suggested for the standard linear model.

In Chapter III, we set the stage for the study of measures of influence in time series by delimiting the framework of time series models that we will consider, and by stating the properties that will be required later on of the ACF computed from a full-sample and those of the ACF computed from a sample with one observation missing at random. We then

review the statistics that have been proposed to measure the influence of one observation on a specific autocorrelation coefficient in the light of a sounder theory. This comparison clearly shows the differences between these various statistics and leads us to adopt the sample measure of influence as an appropriate measure of influence.

In Chapter IV, the statistical properties of the sample measure of influence are studied in order to develop a measure of the influence of one observation on the whole sample ACF. Also, some of the distributional properties of these two measures of influence are derived to obtain thresholds (or critical values) for declaring an observation as overly influential either for a specific autocorrelation coefficient, or for the whole sample ACF.

In Chapter V, three examples are provided comparing the application of these measures of influence and their thresholds with two outlier detection techniques. Even if the measures of influence were not designed for identifying outlying observations, the examples show how useful they are for detecting the overly influential observations, which may also be outliers, at an early stage of the modelling process.

Finally, we conclude with recommendations for further research.

CHAPTER II
AN OVERVIEW OF DISCORDANT OBSERVATIONS IN THE
STANDARD LINEAR MODEL

The purpose of this chapter is to present an overview of the main aspects of the study of discordant observations. As evidenced by the work of Barnett and Lewis(1984), Beckman and Cook(1983), Chatterjee and Hadi(1986,1988), Cook and Weisberg (1982) and Hawkins(1980), it is however a vast field of study. This overview will thus focus only on the study of discordant observations in the standard linear model.

Throughout, we shall only consider the case of at most one discordant observation. Two reasons motivate this choice, apart from the obvious simplification in exposition. First, although most measures of discordancy can easily be extended to handle many discordant observations, the exact number of such observations has to be known in advance to prevent masking and swamping from occurring.

Masking occurs when not all the discordant observations are found, while swamping occurs when non-discordant observations are declared discordant. In practice, it is extremely rare that the exact number of discordant observations is known in advance. We thus have to rely on

brute computational power to determine the most likely subset of discordant observations.

The second reason is that we share the view of Chatterjee and Hadi(1986) that a careful examination of at-most-one discordancy measures will be sufficient in most cases to detect all the discordant observations in a sample. This should not be taken as an elusion, for we recognize the importance of this problem.

The study of discordant observations is one area of statistics which, in the words of Chatterjee and Hadi(1986, p.379), was bound to generate "a bewilderingly large number of statistical quantities", because of the confusion between its causes and effects. The following excerpt from Beckman and Cook(1983) illustrates this confusion in defining what is a discordancy:

"Although much has been written, the notion of outlier seems as vague today as it was 200 years ago. For example, Edgeworth(1887) writes:

"Discordant observations may be defined as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined."

"Eighty-two years later Grubbs(1969) states that

"An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs."

"[...] In an effort to avoid some ambiguity, we adopt the following working notions in the remainder of this article.

Discordant observation - any observation that appears surprising or discrepant to the investigator.

Contaminant - any observation that is not a realization from the target distribution.

Outliers - a collective to refer to either a contaminant or a discordant observation."

Beckman and Cook(1983, pp. 120-121)

Compare now the above definitions with that of Barnett and Lewis(1984):

"We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. The phrase 'appears to be inconsistent' is crucial. It is a matter of subjective judgement on the part of the observer whether or not he picks out some observation (or set of observations) for scrutiny. [...W]hat characterizes the 'outlier' is its impact on the observer (not only will it appear extreme but it will seem, in some sense, surprisingly extreme)."

Barnett and Lewis(1984, pp. 4-5)

Even this definition is not sufficient for the standard linear model, as they later add:

"A crucial distinction must now be recognized in the occurrence of outliers. [...F]or more structured data we would wish to retain the stimulus of 'surprise' but this concept is now far more nebulous. [...] Thus we now need to refine our notion of an outlier. It is not a simple extreme value, but it has a more general pattern-disrupting form. We can, however, no longer rely merely on direct subjective impact but may need [...] to adopt an appropriate outlier detection process [...]"

Barnett and Lewis(1984, pp. 281-282).

Finally, consider the following two recent statements:

"For linear state-space models [...] one can have non-Gaussian distributions [...] Suppose that the non-

Gaussian distributions [...] are moderately Gaussian in the middle but have heavier tails than a Gaussian distribution that fits the non-Gaussian distribution fairly well in the middle. Then the data will contain outliers [...]"

Martin and Raftery(1987, pp. 1045)

"Outlier diagnostics are statistics that focus attention on observations having a large influence on the least squares (LS) estimator [...] Several diagnostic measures have been designed to detect individual cases or groups of cases that may differ from the bulk of the data. [...] One outlying case may cause a poor fit for the majority of the data [...] Therefore, an outlier may have a small LS residual, especially when it is a leverage point [...] Another class of diagnostics is based on the principle of deleting one case at a time [which] give[s] the extent to which the presence of the i-th case affects the regression coefficients [...]"

Rousseeuw and Leroy(1987, p. 216)

Four main factors contribute to this confusion when they are not disentangled. These are:

1. Our purpose in studying the data.
2. The data considered.
3. The model for the data.
4. The metric used for judging discordancy.

Being explicit about these four points should guide one through the myriads of concepts and techniques advocated for dealing with discordant observations. Accordingly, we adopt the following definition: a discordant observation is any observation which, when viewed in an interesting subspace for

the analysis, deviates markedly from the appropriate model for that subspace. As we shall see, this definition forces us to be explicit about the above four points.

The first factor, our purpose in studying the data, is of utmost importance, for it distinguishes between two separate aims in the study of discordant observations: detection or accommodation.

Detecting means measuring each observation for some form of discordancy, leaving temporarily aside any considerations on the possible actions that could be taken as a result of these measurements. In accommodation the aim is to obtain directly the inferences to be drawn from the non-discordant observations.

As such, accommodation techniques belong to the field of robust statistical inference. But there is a close connection between robust estimation techniques and detection methods, since many robust techniques assign a weight to each observation used, and these weights can be interpreted as a measure of accordancy for each observation. The type of discordancy measured may not be interpretable, however.

Our interest mainly resides with detection techniques. The other factors will be illustrated in the following sections.

In section 2.1, we will demonstrate the importance of stating what data are considered by describing the three main types of discordant observations in the standard linear model: outliers, leverage points and influential observations. In the next three sections, we will mostly describe the models and the various measures used for judging these special types of discordancy. In section 2.2, we shall discuss measures of leverage, and in section 2.3, we shall discuss measures of outlyingness. Measures of influence will be discussed in section 2.4.

2.1 Discordancies in the Standard Linear Model

The standard linear model is:

$$Y = X\beta + \epsilon$$

where Y is the $n \times 1$ vector of observed responses, X is a non-stochastic $n \times p$ full-rank matrix of design points, β is a $p \times 1$ vector of unknown parameters, and where ϵ is an $n \times 1$ vector of independent, identically distributed $iid(0, \sigma^2)$ random variables.

Many statistics are of particular interest in the standard linear model, some of which are:

$$H = X(X'X)^{-1}X' \quad \text{the hat matrix,}$$

$$\hat{\beta} = (X'X)^{-1}X'Y \quad \text{the ordinary least squares estimator of } \beta,$$

$\hat{Y} = X\hat{\beta} = HY$ the predicted response values,

$\epsilon = Y - \hat{Y} = (I_n - H)Y$ the vector of residuals,

$\sigma^2 = \frac{\epsilon' \epsilon}{n-p} = \frac{Y'(I_n - H)Y}{n-p}$ the unbiased estimator of σ^2 ,

and

$\rho^2 = \frac{Y'AY - Y'(I_n - H)Y}{Y'AY}$ the squared sample correlation

coefficient, where $A = I_n - 1_n 1_n' / n$, I_n denotes the $n \times n$ identity matrix, and 1_n is the $n \times 1$ column vector of ones.

Consider first the design space $\mathcal{M}(X)$, the space generated by the columns of X . In the standard linear model the design points are considered fixed. However, all the statistics above depend on the design points. Furthermore, inferences made from a linear model become weaker the farther away they are made from the barycenter of the subset of $\mathcal{M}(X)$ formed by the design points, as the variance of the predicted response at $x_0 \in \mathcal{M}(X)$ is given by:

$$\text{Var}[\hat{Y}(x_0)] = \text{Var}[x_0' \hat{\beta}] = \sigma^2 x_0' (X'X)^{-1} x_0.$$

Notice that this variance defines a norm over $\mathcal{M}(X)$. Hence, if x_0 is far from the bulk of the design points according to this norm, the predicted response at x_0 will have a large variance. Clearly then, it is important to identify the design points that may have distorted $\mathcal{M}(X)$: the design points which are discordant with respect to the others. Discordant design points are called leverage points.

But we must distinguish between two different situations: one where we control the explanatory variables, the other when we don't. When we do control the explanatory variables, the points sampled are chosen to be representative of the region over which we want an adequate representation of the relationship studied. Hence, we will not choose a point discordant from the others.

When we do not control the explanatory variables, then some design points may be discordant. Although Barnett and Lewis(1984) recognize the potential influence of remote design points on parameter estimates in the standard linear model, they do not consider these points as 'outliers' (in their terminology) because: "[they] lack the fortuitous (probabilistic) stimulus for [their] extremeness which we have adopted as a characteristic of outlying behaviour".

This is clearly true for controlled design points. But design points that are not controlled have their own probabilistic structure which is not considered relevant in the standard linear model theory. Thus design points may be discordant, and this discordancy can be measured.

Consider now the predicted responses $\hat{Y}$. Some of them may be far from the corresponding observed ones. Hence, discordant responses are those which are far from the entertained model. Accordingly, they are called outliers.

Clearly, outliers need not correspond to leverage points, nor do leverage points correspond to outliers. Both however are discordant: outliers in the response space, leverage points in the design space.

In the above two cases, it was straightforward to define discordancy, for in both we had n observations: n design points and n residuals. It is not so straightforward to judge the discordancy of the statistics $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{\rho}^2$, since we have only one value for each statistic computed from the n observations.

However, each observation contributes to the evaluation of a statistic, and this contribution may be discordant with respect to the contributions of the remaining observations to the point of sensibly determining the value of the statistic. An observation that has too much weight in determining a statistic T is said to be influential for T . It is important to specify for which statistic an observation is influential, for it may be influential for some, and not for others.

This nomenclature is not gratuitous, for it distinguishes between three types of discordancy simply by distinguishing between three different spaces in which each observation can be viewed: the design space, the response space and the 'weight' space. In each space an observation may appear to be inconsistent with the remainder of the observations, an

'outlier' as defined by Barnett and Lewis(1984). Furthermore, as we shall see in the next three sections, the model for the non-discordant observations differs from one space to another as these spaces differ.

2.2 Measures of Leverage

To judge the discordancy of design points we need a model for 'ideal' design points, we need a measure of how the observed points deviate from the idealized ones, and we also need the probabilistic behaviour of this measure. But no probabilistic model is provided by the linear model for the design points, since they are considered fixed. However, when the design can be selected in advance, there are prescriptions to follow for ensuring its optimality for estimating the entertained linear model. We can thus compare the observed design points to an optimal design.

One particular aspect of an optimal design is that no one point should be too remote from the others as to unduly inflate the variance of the predicted response at this point.

Since $\text{Var}[\hat{y}(x_0)] = \sigma^2 h(x_0)$, where $h(x_0) = x_0'(X'X)^{-1}x_0$ defines a norm relative to the space generated by the columns of X , $\mathcal{M}(X)$, the diagonal elements of the hat matrix

$H = X(X'X)^{-1}X'$ provide a measure of leverage for each observation x_i in X :

$$h_{ii} = x_i'(X'X)^{-1}x_i \quad i=1(1)n.$$

This measure of leverage can be interpreted in many ways. For instance, from the equation for the predicted responses $\hat{Y} = HY$, it is seen that h_{ii} is also the weight given to the i -th observed response in determining the predicted one:

$$\hat{y}_i = \sum_{j=1}^n h_{ij}y_j = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j.$$

But h_{ii} is not the sole measure of leverage. The choice of metric used for judging discordancy has been one factor responsible for the growth in the study of discordant observations.

One frequent source of discussion is in the choice of standardization required to make the measure scale-free. Notice that $(X'X)^{-1}$ standardizes our measure of leverage and that it depends on the i -th design point, x_i . Hence h_{ii} is said to be internally standardized. But if x_i is discordant, it distorts $M(X)$ and thus affects the measure of leverage h_{ii} .

So we could use instead

$$h_{ii}^* = x_i'(X_{(i)}'X_{(i)})^{-1}x_i,$$

where $X_{(i)}$ denotes the $(n-1) \times p$ design matrix obtained from X by excluding x_i , in the hope of obtaining a 'robustified' standardization, called external standardization.

[We use the following notation throughout the thesis: a matrix (vector) whose i -th line has been removed will always be indexed by ' (i) '. An externally standardized statistic will always be starred to distinguish it from its internally standardized version. Also, a statistic computed without the i -th observation will always be represented as a function of i , e.g. $\hat{\beta}^{(i)} = (X_{(i)}' X_{(i)})^{-1} X_{(i)}' Y_{(i)}$.]

Now, there is an apparent disadvantage in using h_{ii}^* , instead of h_{ii} . The subspace $\mathcal{M}(X_{(i)})$ from which the length of an observation is measured changes from one observation to another. The measure h_{ii}^* for one observation is then apparently not comparable to that for another.

Both phenomena, the lack of robustness of h_{ii} , and the lack of comparability between the h_{ii}^* 's, are mere appearances however, as it can be shown that:

$$h_{ii} = \frac{h_{ii}^*}{1 + h_{ii}^*} \quad \text{and} \quad h_{ii}^* = \frac{h_{ii}}{1 - h_{ii}} ,$$

i.e. h_{ii}^* is obtained through a monotonic transformation of h_{ii} , and vice versa (both being strictly non-negative). Thus the choice between h_{ii} and h_{ii}^* on the basis of the

'robustness' of the standardization, or on the basis of the comparability between observations, is irrelevant.

Another factor that is more decisive in the choice of a measure of discordancy is the induced scale for discordancy. For we have to determine the threshold above which an observation is declared discordant. For some measures, this threshold will be easier to determine than for others.

For the measures of leverage h_{ii} and h_{ii}^* , we cannot invoke probabilistic arguments derived from the linear model theory since it views the design matrix as fixed.

Some non-probabilistic thresholds were therefore suggested, and are reviewed in Chatterjee and Hadi(1986, 1988). However, probabilistic arguments, independent of those used in the standard linear model, can be used for guidance.

If the model contains an intercept, and if we assume that the rows of the centered design matrix are i.i.d. multivariate Gaussian, Belsley, Kuh and Welsch(1980, pp. 66-67) and Chatterjee and Hadi(1988, pp. 31-34) show that

$$F_i = \frac{(n-p)}{(p-1)} \left(\frac{h_{ii} - \frac{1}{n}}{1 - h_{ii}} \right) \doteq \mathcal{F}_{p-1, n-p}$$

where $\mathcal{F}_{p-1, n-p}$ has Fisher's distribution with $p-1$ and $n-p$ degrees of freedom, and where ' $\doteq$ ' stands for 'approximately distributed as'.

Notice that F_i can be rewritten as a linear function of the externally standardized measure h_{ii}^* :

$$F_i = \left(\frac{n-p}{p-1} \right) \frac{\left(\frac{h_{ii}^*}{1+h_{ii}^*} - \frac{1}{n} \right)}{\left(1 - \frac{h_{ii}^*}{1+h_{ii}^*} \right)} = \left(\frac{n-p}{p-1} \right) \left[\left(1 - \frac{1}{n} \right) h_{ii}^* - \frac{1}{n} \right].$$

It is thus easier to obtain the critical values of h_{ii}^* than those of h_{ii} .

But in practice, it is the largest observed F_i , $F_{(n)}$, which will be most likely tested for discordancy. $F_{(n)}$ does not have the same distribution as F_i , and since the F_i 's are correlated, the exact $100(1-\alpha)\%$ threshold value would be complicated to derive.

Using Bonferroni-like inequalities, which will be more fully discussed in section 2.3, it can be shown that by choosing f_α such that $P(\mathcal{F}_{p-1, n-p} \leq f_\alpha) = (1-\alpha)^{1/n}$ yields a conservative threshold for $F_{(n)}$:

$$P(F_{(n)} \leq f_\alpha) \geq 1-\alpha.$$

2.3 Measures of Outlyingness

In section 2.1, we defined outliers as observed responses (Y) discordant from predicted ones ($\hat{Y}$).

Unlike design points, the non-discordant responses possess a model in the standard theory, namely $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, i.e. the vector of observed responses is a realization of a multivariate normal distribution of expected value $X\beta$ and variance $\sigma^2 I_n$.

Thus, the null hypothesis for testing each component y_i of Y for outlyingness is given. But we also need to specify an alternative hypothesis. Many choices are possible however, each reflecting a particular form of discordancy, thus producing many tests. Various classes of alternatives are reviewed by Barnett and Lewis (1984) and by Hawkins (1980). The class of alternative hypotheses which is most frequently used is that of slippage.

For a sample $\{X_i\}_{i=1}^n$ of independent random variables distributed as F_{μ_i, σ_i^2} , where $E[X_i] = \mu_i$ and $\text{Var}[X_i] = \sigma_i^2$, a slippage alternative states that all the observations, apart from some prescribed small number, arise from the distribution F_{μ, σ^2} , while the remaining few arise from the same family of distributions but with a shifted location or scale.

For the null hypothesis $H_0: X_i \sim F_{\mu, \sigma^2}$ for $i=1(1)n$, the two possible slippage alternatives are that of slippage in the mean:

H_1 : for some unknown i , $E[X_i] = \mu + \Delta$ where $\Delta \neq 0$, and that of slippage in the variance:

H_1' : for some unknown i , $\text{Var}[X_i] = \lambda \sigma^2$ with $\lambda > 1$.

Tests designed to discriminate against slippage in the mean will differ from those designed to discriminate against slippage in the variance, as these alternatives differ in their implications. When it is assumed that an observation has slipped in its mean, we assume that it does not provide any information on the parameters μ and σ^2 . When it is instead assumed to have slipped in its variance, we assume that this observation provides some information on the parameter μ , but none on σ^2 . Note that $\lambda < 1$ is not appropriate for testing discordancy.

The only form of alternatives that we shall consider is that of slippage in the mean.

We can now derive the likelihood ratio test (LRT) against slippage in the mean of the i -th observed response. The residuals $\hat{e} = Y - \hat{Y}$ will clearly play an important role in identifying the outlying observations.

Proposition 2.1: The LRT at level $1-\alpha$ of $H_0: y_j \sim \mathcal{N}(x_j'\beta, \sigma^2)$ for all $j=1(1)n$, against $H_1: y_i \sim \mathcal{N}(x_i'\beta + \delta, \sigma^2)$ with $\delta \neq 0$ for the i -th observed response and $y_j \sim \mathcal{N}(x_j'\beta, \sigma^2)$ for $j \neq i$, is to reject H_0 whenever

$$|r_i| = \frac{|\hat{\epsilon}_i|}{\hat{\sigma}(1-h_{ii})^{1/2}} > b_\alpha$$

where b_α is an appropriately chosen constant.

Proof: Let e_i be the i -th column of the identity matrix I_n , i.e. $I_n = (e_1; e_2; \dots; e_n)$. The linear model may be written as $Y = W\gamma + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $W = (X; e_i)$ is $n \times (p+1)$ and $\gamma' = (\beta', \delta)$.

Under H_0 , the MLE's of β and σ^2 are $\hat{\beta} = (X'X)^{-1}X'Y$ and $\tilde{\sigma}_0^2 = n^{-1}(Y-X\hat{\beta})'(Y-X\hat{\beta})$ respectively. Thus the maximum value of the log-likelihood function of Y under H_0 is:

$$\hat{\ell}_{H_0} = K - \frac{n}{2} \ln \tilde{\sigma}_0^2, \quad K \text{ a constant.}$$

Under H_1 , the MLE's of γ and σ^2 are $\hat{\gamma} = (W'W)^{-1}W'Y$ and $\tilde{\sigma}_1^2 = n^{-1}(Y-W\hat{\gamma})'(Y-W\hat{\gamma})$ respectively. It can be shown however that

$$\tilde{\sigma}_1^2 = \tilde{\sigma}_0^2 - \frac{\hat{\epsilon}_i^2}{1-h_{ii}}$$

where $\hat{\epsilon}_i$ is the i -th residual of the null model, i.e. $\hat{\epsilon} = Y-X\hat{\beta}$, and h_{ii} is the i -th diagonal element of the hat matrix H .

Thus the maximum value of the log-likelihood function of Y under H_1 is:

$$\hat{\ell}_{H_1} = K - \frac{n}{2} \ln \tilde{\sigma}_1^2 = K - \frac{n}{2} \ln \left(\tilde{\sigma}_0^2 - \frac{\hat{\epsilon}_i^2}{1-h_{ii}} \right), \quad K \text{ a constant.}$$

Therefore, the LRT of H_0 against H_1 at level $1-\alpha$ is to reject H_0 whenever

$$T = \hat{\ell}_{H_1} - \hat{\ell}_{H_0} = \frac{n}{2} \ln (\tilde{\sigma}_0^2 / \tilde{\sigma}_1^2) > C_\alpha$$

where C_α is such that $P_{H_0}(T > C_\alpha) = \alpha$.

Now, letting $\hat{\sigma}^2$ denote the unbiased estimator of σ^2 under H_0 , the LRT statistic is equivalent to $|r_i|$.

Q.E.D.

The LRT above only tests the i -th response. In practice, we will test the whole sample of responses for at most one discordancy. Hence we need the LRT for this situation which is given in the next proposition. The proof of Proposition 2.2 follows that of Cook, Holschuh and Weisberg(1982).

Proposition 2.2: The LRT at level $1-\alpha$ of $H_0: y_j \sim \mathcal{N}(x_j' \beta, \sigma^2)$

for all $j=1(1)n$, against $H_1: y_i \sim \mathcal{N}(x_i' \beta + \delta, \sigma^2)$ with $\delta \neq 0$

for at most one $i=1(1)n$, is to reject H_0 whenever

$$|r|_{(n)} = \max_{i=1(1)n} (|r_i|) > b'_\alpha$$

where b'_α is an appropriately chosen constant.

Proof: Let $\Pi = \{\pi_i, i=1(1)n\}$ where each π_i is an $n \times n$ matrix that permutes only the first and the i -th co-ordinate of any vector. I.e., with the e_i 's defined as in the proof of Proposition 2.1 for $i=1(1)n$, $\pi_i e_1 = e_i$.

The linear model may be written as $Y = W\gamma + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $W = (X; \pi e_1)$ is $n \times (p+1)$, $\gamma' = (\beta', \delta)$ and π is an unknown permutation matrix in Π .

The maximum value of the log-likelihood function of Y under H_0 is as in Proposition 2.1:

$$\hat{\ell}_{H_0} = K - \frac{n}{2} \ln \tilde{\sigma}_0^2, \quad K \text{ a constant.}$$

For a fixed permutation matrix π_i , the maximum value of the log-likelihood function of Y under H_1 is also as in Proposition 2.1:

$$\hat{\ell}_{H_1}(\pi_i) = K - \frac{n}{2} \ln \left(\tilde{\sigma}_0^2 - \frac{\hat{\epsilon}_i^2}{1-h_{ii}} \right), \quad K \text{ a constant.}$$

Hence, the maximum value of the log-likelihood function of Y under H_1 over all the permutations $\pi_i \in \Pi$ is:

$$\hat{\ell}_{H_1} = \max_{\pi_i \in \Pi} \{\hat{\ell}_{H_1}(\pi_i)\} = K - \frac{n}{2} \ln \left(\tilde{\sigma}_0^2 - \max_{i=1(1)n} \left\{ \frac{\hat{\epsilon}_i^2}{1-h_{ii}} \right\} \right).$$

Therefore, the LRT of H_0 against H_1 at level $1-\alpha$ is to reject H_0 whenever

$$T' = \hat{\ell}_{H_1} - \hat{\ell}_{H_0} = -\frac{n}{2} \ln \left(1 - \tilde{\sigma}_0^2 \sum_{i=1}^n m_i \lambda_i \left\{ \frac{\epsilon_i^2}{1-h_{i1}} \right\} \right) > C'_\alpha$$

where C'_α is such that $P_{H_0}(T' > C'_\alpha) = \alpha$.

Letting $\hat{\sigma}^2$ denote the unbiased estimator of σ^2 under H_0 , the LRT statistic is thus equivalent to $|r|_{(n)}$, the largest absolute residual.

Q.E.D.

Both statistics r_i and $r_{(n)}$ can be used to test for slippage in the mean; the first to test the i -th response, the second to test for at most one discordancy in the whole sample of responses.

To conclude our discussion of the LRT statistics $|r_i|$ and $|r|_{(n)}$, we need to review their distributional properties under the null hypothesis of no discordancy.

From the model $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ it is deduced that $\hat{\epsilon} \sim \mathcal{N}(0, \sigma^2(I_n - H))$. Note that it is a singular distribution since $I_n - H$ is a symmetric idempotent matrix whose trace is $n - p < n$. Furthermore, the components of $\hat{\epsilon}$ do not necessarily have all the same variance, and they may be correlated. [The net effect of this correlation when there is an outlier is to spread the discordancy over many observations. This shows the potential difficulty in testing

for many outliers using measures derived for the presence of at most one outlier.]

Now the LRT statistic to test the i -th observation is the absolute value of

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}(1-h_{ii})^{1/2}}$$

the i -th internally standardized residual, where $\hat{\sigma}^2$ is the unbiased estimator of σ^2 .

Some authors do not like this standardization, fearing that an outlier will unduly inflate the estimate of $\hat{\sigma}^2$ and hence, decrease the value of $|r_i|$. They prefer the externally standardized residual:

$$r_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}^{(i)}(1-h_{ii})^{1/2}}$$

where $\hat{\sigma}^{(i)2}$ is obtained from the regression of Y on X , leaving the i -th observation aside. But it can be shown that

$$(n-p-1)\hat{\sigma}^{(i)2} = (n-p)\hat{\sigma}^2 - \frac{\hat{\epsilon}_i^2}{(1-h_{ii})}$$

so that

$$r_i^{*2} = \frac{(n-p-1)r_i^2}{n-p - r_i^2} \quad \text{and} \quad r_i^2 = \frac{(n-p)r_i^{*2}}{n-p-1 + r_i^{*2}}.$$

Both transformations, of r_i^2 into r_i^{*2} and of r_i^{*2} into r_i^2 , are monotonic. Thus, $|r_i^*|$ is also an LRT statistic, and the choice between $|r_i|$ and $|r_i^*|$ in terms of their robustness against a possible outlier is irrelevant.

However, $(n-p)^{-1}r_i^2$ has a Beta distribution $B(\frac{1}{2}, \frac{n-p-1}{2})$, whereas r_i has a Student distribution T_{n-p-1} (Cook and Weisberg, 1982, pp. 18-20). Therefore, it is easier to use r_i than r_i^2 in testing for the discordancy of the i -th response.

Because of the correlation between the residuals, which depends on the hat matrix H , expressions for the distributions of $|r|_{(n)}$ and of $|r'|_{(n)}$ are quite complicated. However, bounds on their critical values are easily obtained using Bonferroni-like inequalities. Since the bounds for the critical values of $|r'|_{(n)}$ will be easier to compute than those for $|r|_{(n)}$, we will only discuss those for $|r'|_{(n)}$.

For any events $E_1, E_2, \dots, E_n$, let $E = \bigcup_{i=1}^n E_i$. Then the first and second Bonferroni inequalities are:

$$\mathcal{L}_1 = \sum_{i=1}^n P(E_i) - \sum_{i < j} P(E_i \cap E_j) \leq P(E) \leq \sum_{i=1}^n P(E_i) = \mathcal{U}_1.$$

A sharper upper bound for $P(E)$ was derived by Hunter (1976). It is given as Theorem 2.3.

Theorem 2.3 Hunter's Inequality:

Let the events $\{E_i\}_{i=1}^n$ represent the vertices of a graph G , with edges g_{ij} if and only if $P(E_i \cap E_j) > 0$. Let T be a subgraph of G . Then,

$$P(E) = P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i) - \sum_{((i,j):g_{ij} \in T)} P(E_i \cap E_j) = \mathcal{U}_2$$

if and only if T is a tree. Furthermore, equality holds whenever G is itself a tree and $T = G$.

As is clear from the statement of Theorem 2.3, the smallest upper bound will be obtained from the maximum spanning tree which can be found using Kruskal's (1956) algorithm. Any tree, however, will give an upper bound. By comparison with Bonferroni's lower bound $\mathcal{L}_1$, Hunter's upper bound $\mathcal{U}_2$ could be quite tight.

Now, to obtain bounds on the critical values for $|r^*|_{(n)}$, let $A_i(c) = \{\omega \in \Omega: |r_i^*| > c\}$. Then,

$$P(|r^*|_{(n)} > c) = P\left(\bigcup_{i=1}^n A_i(c)\right).$$

To compute Bonferroni's lower bound $\mathcal{L}_1$ and Hunter's upper bound $\mathcal{U}_2$ in practice, we would need to evaluate

$$\alpha_{ij} = P(A_i \cap A_j) \quad \text{for } i, j = 1(1)n, i \neq j.$$

Many of the studies reviewed by Barnett and Lewis (1984), Beckman and Cook (1983) and by Hawkins (1980) are concerned with that particular aspect for a general design matrix X and for some more particular cases. A recent example is Cook and Prescott (1981). It also appears from these reviews that the Bonferroni bounds $\mathcal{L}_1$ and $\mathcal{U}_1$ (and hence $\mathcal{U}_2$) should be adequate for most regressions.

Three particular results point in that direction. The first is due to Ellenberg(1976) who showed that $\alpha_{ij} = 0$ whenever

$$1 + |\rho_{ij}| \leq 2(n-p)^{-1} t_{\alpha/2n}^2$$

where $\rho_{ij} = -h_{ij}(1-h_{ii})^{-1/2}(1-h_{jj})^{-1/2}$ is the correlation between $\hat{e}_i$ and $\hat{e}_j$, and where t_α is the upper percentage point of Student's distribution $\mathcal{T}_{n-p-1}$ with $n-p-1$ degrees of freedom, i.e. $P(\mathcal{T}_{n-p-1} > t_\alpha) = \alpha$.

It is important to remark here that when

$$1 + \max_{i,j} |\rho_{ij}| \leq 2(n-p)^{-1} t_{\alpha/2n}^2,$$

the two Bonferroni inequalities give the exact critical value, i.e. $P(|r^*|_{(n)} > t_{\alpha/2n}^2) = \alpha$.

The second result concerns the case where the correlation between any two residuals is strong. As was noted by Schwager (1984), the effectiveness of Bonferroni's inequalities may be considerably reduced in that case. This was also noted by Hawkins(1980) and by Cook and Prescott (1981). But for each $i < j$ for which $|\rho_{ij}| = 1$, Cook and Prescott(1981) suggest to decrease n by one as the maximum is effectively obtained from a reduced set of components of $\hat{e}$.

The third result is much more interesting and is stated as Theorem 2.4. It stems from a remark of Hawkins(1980, pp. 94). The theorem is stated in a more useful form and we

provide some details to its proof. This theorem possibly provides the tightest bounds on the distribution of $|r^*|_{(n)}$ obtainable without computing the α_{ij} 's, and it yields a test à la Durbin and Watson(1950) to detect at most one outlier in a standard linear model.

Theorem 2.4 (Hawkins 1980):

For a likelihood ratio test whose level is at least $1-\alpha$ of

$H_0: y_i \sim \mathcal{N}(x_i'\beta, \sigma^2)$ for all $i=1(1)n$, against

$H_1: y_i \sim \mathcal{N}(x_i'\beta + \delta, \sigma^2)$ with $\delta \neq 0$ for at most one $i=1(1)n$,

let $t_{\alpha/2n}$ be such that $P(\mathcal{T}_{n-p-1} > t_{\alpha/2n}) = \frac{\alpha}{2n}$ where $\mathcal{T}_{n-p-1}$

has Student's distribution with $n-p-1$ degrees of freedom,

and let $t_{\alpha/2(n-p)}$ be defined similarly. Then, reject H_0

whenever $|r^*|_{(n)} \geq t_{\alpha/2n}$; 'accept' H_0 whenever $|r^*|_{(n)} \leq t_{\alpha/2(n-p)}$

while for $t_{\alpha/2(n-p)} < |r^*|_{(n)} < t_{\alpha/2n}$ the test is inconclusive.

Proof: Since we reject H_0 for $|r^*|_{(n)} \geq t_{\alpha/2n}$, we have, using

the Bonferroni inequality $\mathcal{U}_1$,

$$P(|r^*|_{(n)} > t_{\alpha/2n}) \leq \sum_{i=1}^n P(|r_i^*| > t_{\alpha/2n}) = 2n \frac{\alpha}{2n} = \alpha.$$

For the lower bound on the critical value, Hawkin's argument is as follows. Since $I_n - H$ is symmetric, non-negative definite and its rank is $n-p$, there exists a real matrix B such that

$$B(I_n - H)B' = \begin{pmatrix} 0_{p \times p} & 0_{p \times p} \\ 0_{p \times p} & I_{n-p} \end{pmatrix}.$$

Note that a permutation of the columns of $I_n - H$ may be required to obtain this factorization, but this will not affect the result.

Now, the vector $\tilde{\epsilon} = B\hat{\epsilon}$ is such that

$$V[\tilde{\epsilon}] = \sigma^2 B(I_n - H)B' = \sigma^2 \begin{pmatrix} 0_{p \times p} & 0_{p \times p} \\ 0_{p \times p} & I_{n-p} \end{pmatrix}.$$

The last $n-p$ components of $\tilde{\epsilon}$ are the recursive residuals introduced by Theil(1965), which are i.i.d. $\mathcal{N}(0, \sigma^2)$. Furthermore, it can be shown that

$$|\tilde{\epsilon}|_{(n-p)} = \max_{j=p+1(1)n} \{|\tilde{\epsilon}_j|\} \leq \max_{i=1(1)n} \{|\hat{\epsilon}_i|(1-h_{ii})^{-1/2}\}.$$

Letting $\tilde{r}_j = \tilde{\epsilon}_j/\hat{\sigma}$ we have $|\tilde{r}|_{(n-p)} \leq |r|_{(n)}$. The transformation from $|r_i|$ to $|r_i^*|$ being monotonic in $|r_i|$, define for $j=p+1(1)n$

$$\tilde{r}_j^* = \tilde{r}_j \left(\frac{n-p}{n-p-1} - \frac{\tilde{r}_j^2}{n-p-1} \right)^{-1/2}.$$

Then, $|\tilde{r}^*|_{(n-p)} \leq |r^*|_{(n)}$. Since $(n-p)^{-1}\tilde{r}_j^2 \sim \mathcal{B}\left(\frac{1}{2}, \frac{n-p-1}{2}\right)$, we have $\tilde{r}_j^{*2} \sim \mathcal{T}_{n-p-1}$.

Using the Bonferroni inequality $\mathcal{U}_1$, we get $t_{\alpha/2(n-p)}$ as an upper bound for the $1-\alpha$ critical value of $|\tilde{r}^*|_{(n-p)}$, i.e.

$$P(|\tilde{r}^*|_{(n-p)} > t_{\alpha/2(n-p)}) \leq \sum_{j=p+1}^n P(|\tilde{r}_j^*| > t_{\alpha/2(n-p)}) = \alpha.$$

Let $\tilde{\alpha}_{ij} = P(|\tilde{r}_i^*| > t_{\alpha/2(n-p)}, |\tilde{r}_j^*| > t_{\alpha/2(n-p)})$. Since the recursive residuals are independent, we have

$$\tilde{\alpha}_{ij} = P(|\tilde{r}_i^*| > t_{\alpha/2(n-p)})P(|\tilde{r}_j^*| > t_{\alpha/2(n-p)}) = \frac{\alpha^2}{(n-p)^2}.$$

Now, using the Bonferroni inequality $\mathcal{L}_1$ we obtain:

$$P(|\tilde{r}^*|_{(n-p)} > t_{\alpha/2(n-p)}) \geq \alpha - \sum_{j < i} \tilde{\alpha}_{ij} = \alpha - \alpha^2 \left(1 - \frac{1}{n-p}\right).$$

Thus, for α small $P(|\tilde{r}^*|_{(n-p)} > t_{\alpha/2(n-p)}) \approx \alpha$.

Since $|\tilde{r}^*|_{(n-p)} \leq |r^*|_{(n)}$, using $t_{\alpha/2(n-p)}$ we get

$$P(|r^*|_{(n)} > t_{\alpha/2(n-p)}) \geq P(|\tilde{r}^*|_{(n-p)} > t_{\alpha/2(n-p)}) \approx \alpha.$$

Q.E.D.

Whenever the number of observations, n , is large compared to the number of parameters, p , these bounds should be tight enough to declare $|r^*|_{(n)}$ discordant, or not, unambiguously.

2.4 Measures of Influence

Many measures of influence have been developed for the linear model. The main reason for this proliferation is that many were developed with the purpose of providing a single measure of discordancy for each observation.

As evidenced by the discussion in Chatterjee and Hadi (1986), for instance, and from our perspective on discordancy, it is clear that no unique measure of discordancy could be exempt from the possible confounding in the effects of leverage, outlyingness and influence.

However, these three forms of discordancy are still confused. For instance, Chatterjee and Hadi(1986) categorize measures of influence as follows [still more categories are given in Chatterjee and Hadi(1988)]:

1. Measures based on residuals
2. Measures based on the prediction matrix
3. Measures based on the volume of confidence ellipsoids
4. Measures based on influence functions.

The first category is composed of measures of outlyingness. As can be seen in Propositions 2.1 and 2.2, an outlying observation may or may not be influential for the parameter β . But an outlying observation is influential for the residual variance σ^2 . The second category corresponds to measures of leverage. Here again an observation with high leverage need not be influential for some specific statistics.

The third category is composed of measures of change in the volume of ellipsoids incurred by the removal of an observation in their computation. These are thus measures of

influence on the volume of those ellipsoids. The fourth category is composed of measures of influence, but not all for the same statistic. It is important to specify exactly for which statistic the influence of an observation is measured.

Another example of this confusion can be seen in Cook and Weisberg(1982):

"The emphasis in this chapter is on detecting influential cases rather than on how to deal with them when they are found [...] Some of the possible actions can be given, however.[...] If an influential case cannot be traced to conditions that justify its removal and the model is known to be approximately correct, a formal outlier test might be useful, although such tests cannot be expected to be powerful."

Cook and Weisberg(1982, p.104)

We may wonder why we should test an observation for one form of discordancy (here outlyingness) when it is already viewed as discordant in another (here influence).

In section 2.4.1, we briefly introduce the influence function from which various measures of influence are defined. In section 2.4.2, we generalize Cook and Weisberg's(1982) and Critchley's(1985) approach for deriving the most frequently used measures of influence in the case of an independent and identically distributed (i.i.d.) sample. In section 2.4.3 we review and discuss the various methods used for obtaining thresholds (critical values) for identifying influential observations.

2.4.1 The Influence Function

The influence function stems from the study of the robustness properties of a statistic. Many statistics used for inference are adopted on the basis of their optimality properties with respect to a specific family of probability distributions. For instance, the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is the uniformly minimum variance unbiased estimator of μ for the family $\{\mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}$ of distributions. It is not even consistent, however, for the family of translated Cauchy distributions

$$\{\mathcal{C}(\mu): f(x) = \pi^{-1}[1+(x-\mu)^2]^{-1}, \mu \in \mathbb{R}\}.$$

In practice, the family of distributions chosen is but an approximation to the true distribution generating the observations. It is thus important that the statistics used maintain their optimality properties under slight departures from the hypothesized family of distributions. When they do, the statistics are said to be (model-)robust.

These departures from the assumed model are of two forms: a few observations in the sample come from a probability distribution that is 'far' from the assumed one; or all the observations in the sample come from a distribution that is in a small neighbourhood of the assumed one.

Note that the first form of departure is akin to the manifestation of discordant observations. Hence, robust statistics will necessarily make some accommodations to reduce the effect of discordant observations.

To study the robustness properties of a statistic, T say, it is necessary to represent it as a function of probability distributions. For many statistics this is possible since they depend on the sample only through the empirical distribution function (EDF):

$$F_n(X) = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$$

where δ_{x_i} is the degenerate distribution:

$$\delta_{x_i}(A) = \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$T(X_1, X_2, \dots, X_n) = T[F_n].$$

For example, the mean $\bar{X}$ can be written as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \int x dF_n(x).$$

When a statistic admits such a functional representation, its robustness can be easily studied. One particularly useful tool for this study is Von Mises' representation of a functional $T[F]$ which parallels the Taylor's series representation of a function. For F in a neighbourhood of the probability distribution G , the Von Mises' representation of $T[F]$ is:

$$T[F] = T[G] + \int j(x) dF(x) + \dots$$

To study the infinitesimal robustness properties of T , the kernel of the first term in Von Mises' representation, $J(x)$, is of prime importance. It is called the influence function of T at the distribution F , and is defined as

$$J(X;T,F) = \lim_{h \rightarrow 0} \frac{T[(1-h)F + h\delta_X] - T[F]}{h} .$$

Note that the influence function is defined as above even if the regularity conditions ensuring the existence of Von Mises' representation are not satisfied.

A rigorous treatment of the subject can be found in the books of Hampel, et al. (1986), Huber(1981), Rey(1983) and in the references therein.

2.4.2 Obtaining Measures of Influence

The theoretical influence function defined in section 2.4.1 is a Gâteaux directional derivative. It measures the effect of an infinitesimal change in the weight given by the distribution function F to X , on the value of the statistic T . Most finite-sample versions of the influence function retain this interpretation. To distinguish a theoretical influence function from a finite-sample one, we shall call the latter a measure of influence.

The three most frequently used measures of influence are:

1. The empirical added influence (EIC_i), obtained by evaluating the influence function at the i -th sampled point with the theoretical distribution function (F) replaced by the empirical one (F_n), i.e.

$$EIC_i = J(X_i; T, F_n).$$

2. The empirical deleted influence ($EIC^{(i)}$), obtained by evaluating the influence function at the i -th sampled point with the theoretical distribution function F replaced by the EDF $F^{(i)}$ computed without the i -th observation, i.e.

$$F^{(i)} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n \delta_{x_j},$$

and

$$EIC^{(i)} = J(X_i; T, F^{(i)}).$$

3. The sample measure of influence (SIC_i), obtained by approximating the limit defining the theoretical influence. Let

$$m(X; T, F, h) = \frac{T[(1-h)F + h\delta_x] - T[F]}{h}.$$

Then, the influence function is defined as

$$J(X; T, F) = \lim_{h \rightarrow 0} m(X; T, F, h).$$

Since the sample is finite, the influence function is approximated by:

$$\begin{aligned}
SIC_i &= m(X_i; T, F_n, -(n-1)^{-1}) \\
&= -(n-1) (T[(1 + (n-1)^{-1})F_n - (n-1)^{-1}\delta_{x_i}] - T[F_n]) \\
&= (n-1) (T[F_n] - T[F_{(i)}]).
\end{aligned}$$

These three measures of influence are clearly different. However, there is an interesting relationship linking them, which is shown to hold quite generally in Theorem 2.5 below. This relationship was proved by Cook and Weisberg(1982) and by Pregibon(1981) for the special case of $\hat{\beta}$ in the standard linear model. To obtain the relationship, Pregibon(1981) and Cook and Weisberg(1982) assumed that the i -th observation was in discordancy with the others, having an inflated variance; i.e. $y_j \sim \mathcal{N}(x_j'\beta, \sigma^2)$ for $j=1(1)n$, $j \neq i$, and $y_i \sim \mathcal{N}(x_i'\beta, \sigma^2/\omega_i)$ with $\omega_i \in (0, 1)$.

Then, writing

$$\hat{\beta}[\omega_i] = \hat{\beta} - \frac{(X'X)^{-1}x_i(1-\omega_i)\hat{\epsilon}_i}{1-(1-\omega_i)h_{ii}},$$

they obtained:

$$\begin{aligned}
n^{-1}EIC_i &= \frac{\partial \hat{\beta}}{\partial \omega_i}(1) = (X'X)^{-1}x_i\hat{\epsilon}_i, \\
(n-1)^{-1}EIC_{(i)} &= \frac{\partial \hat{\beta}}{\partial \omega_i}(0) = \frac{(X'X)^{-1}x_i\hat{\epsilon}_i}{(1-h_{ii})^2}
\end{aligned}$$

and

$$(n-1)^{-1}SIC_i = \int_0^1 \frac{\partial \hat{\beta}}{\partial \omega_i} d\omega_i = \frac{(X'X)^{-1}x_i\hat{\epsilon}_i}{1-h_{ii}}.$$

The major difficulty with their derivation is that the perturbation scheme used may not apply in other contexts, and may not be the desired one. As Theorem 2.5 below shows, the underlying perturbation scheme is, in fact, one of removing the i -th observation, an interpretation which can be carried into other situations. Both schemes, the inflated variance and the removal of an observation, are equivalent in the standard linear model.

Theorem 2.5 is not entirely new. Critchley(1985) obtained the same result but under more stringent conditions. Furthermore, the implications of this relationship are not fully stated in Critchley(1985), as he concludes that SIC_i is to be preferred because of its 'leave-one-out' interpretation, and because the other two are constructed assuming that an infinitely large sample was used in obtaining F_n and $F(i)$. But SIC_i is also constructed under the assumption that n is large enough for the limit to have been reached.

We shall now give a complete and much simpler proof of this relationship.

Theorem 2.5: Let $\{Y_i\}_{i=1}^n$ be a sample of i.i.d. random variables, and let T be a statistic expressible as a functional of the EDF F_n , i.e. for some ψ ,

$$T(Y_1, \dots, Y_n) = T[F_n] = \int \psi(x) dF_n(x).$$

For each $i=1(1)n$, let $\omega_i \in [0,1]$ and define

$$T(\omega_i) = T \left[\left(1 - \frac{\omega_i}{n} \right) F^{(i)} + \frac{\omega_i}{n} \delta_{Y_i} \right].$$

Then, for $i=1(1)n$ $\frac{\partial T}{\partial \omega_i}$ exists over $[0,1]$ and we have

$$(n-1)EIC_i = \frac{\partial T}{\partial \omega_i}(1), \quad nEIC(i) = \frac{\partial T}{\partial \omega_i}(0)$$

and

$$SIC_i = (n-1) \int_0^1 \frac{\partial T}{\partial \omega_i} d\omega_i.$$

Proof: First note that $T(0) = T[F^{(i)}]$ and that

$$T(1) = T[n^{-1}(n-1)F^{(i)} + n^{-1}\delta_{Y_i}] = T[n^{-1}(\delta_{Y_i} + \sum_{\substack{j=1 \\ j \neq i}}^n \delta_{Y_j})] = T[F_n].$$

Hence, for $\omega_i=0$, $T(\omega_i)$ is the value of the statistic obtained when the i -th observation is excluded from the sample, and for $\omega_i=1$, $T(\omega_i)$ is the value of the statistic obtained from the full sample.

Now, by definition $\frac{\partial T}{\partial \omega_i} = \lim_{h \rightarrow 0} \frac{T(\omega_i+h) - T(\omega_i)}{h}$.

For $\omega_i=0$,

$$T(\omega_i+h) = T(h) = T \left[\left(1 - \frac{h}{n} \right) F^{(i)} + \frac{h}{n} \delta_{Y_i} \right].$$

Hence,

$$\begin{aligned}
 \frac{\partial T}{\partial \omega_1}(0) &= \lim_{h \rightarrow 0} \frac{T[n^{-1}(n-h)F(i) + n^{-1}h\delta_{Y_i}] - T[F(i)]}{h} \\
 &= n^{-1} \lim_{h \rightarrow 0} \frac{T\left[\left(1 - \frac{h}{n}\right)F(i) + \frac{h}{n} \delta_{Y_i}\right] - T[F(i)]}{h/n} \\
 &= n^{-1} \lim_{\lambda \rightarrow 0} \frac{T[(1-\lambda)F(i) + \lambda\delta_{Y_i}] - T[F(i)]}{\lambda} \\
 &= n^{-1} J(Y_i; T, F(i)) = n^{-1} EIC(i) \quad \text{by definition.}
 \end{aligned}$$

For $\omega_1=1$,

$$T(\omega_1+h) = T(1+h) = T\left[\left(1 - \frac{1+h}{n}\right)F(i) + \frac{1+h}{n} \delta_{Y_i}\right].$$

Notice that $F(i) = (n-1)^{-1}(nF_n - \delta_{Y_i})$, hence

$$T(1+h) = T\left[\left(1 - \frac{h}{n-1}\right)F_n + \frac{h}{n-1} \delta_{Y_i}\right].$$

Thus,

$$\begin{aligned}
 \frac{\partial T}{\partial \omega_1}(1) &= (n-1)^{-1} \lim_{h \rightarrow 0} \frac{T\left[\left(1 - \frac{h}{n-1}\right)F_n + \frac{h}{n-1} \delta_{Y_i}\right] - T[F_n]}{h/(n-1)} \\
 &= (n-1)^{-1} J(Y_i; T, F_n) = (n-1)^{-1} EIC_i \quad \text{by definition.}
 \end{aligned}$$

The last relation is immediate, since

$$(n-1) \int_0^1 \frac{\partial T}{\partial \omega_1} d\omega_i = (n-1)[T(1)-T(0)] = SIC_i.$$

Q.E.D.

Thus we see that EIC_i is, up to a multiplicative constant, the instantaneous rate of change in the value of T as the i -th observation is smoothly added to the sample; $EIC(i)$ is, up to a multiplicative constant, the instantaneous rate of change in T as the i -th observation is smoothly removed from the sample; and SIC_i is, up to a multiplicative constant, the average rate of change in the value of the statistic T as the weight given to the i -th observation by the EDF is smoothly varied between 0 and n^{-1} .

This result is important for it provides an interpretation and a means for computing these measures of influence which can be extended to more complex situations. It also shows that the measure of sample influence (SIC_i) is to be preferred to the two others.

As Theorem 2.5 shows, SIC_i is a summary of the full curve $\frac{\partial T}{\partial \omega_i}$, being its average value, whereas EIC_i and $EIC(i)$ are only two of its values. On this ground alone, SIC_i is a more appropriate measure of influence than both $EIC(i)$ and EIC_i .

Note that Cook(1986) has suggested other summaries of the curve $\frac{\partial T}{\partial \omega_i}$. Since their properties are still to be explored, we shall not consider them in this thesis.

2.4.3 Identifying Influential Observations

Welsch(1982) identified five distinct methods to assess the potentially discordant influence of an observation:

1. By comparison with a balanced design;
2. By statistical analogy;
3. Using Atkinson's(1981,1985) envelopes;
4. Using exploratory data analysis (EDA) techniques;
5. By a formal test of significance.

Only the first four have been used to any extent in the literature that we have reviewed, the most common being EDA techniques and statistical analogies. The fifth approach is usually discarded, or not even mentioned, as in Cook and Weisberg(1982) for example.

There are three reasons for this variety of methods, which also explain why formal tests of significance are not used. The first is that most measures of influence depend on the design matrix on which no structure is imposed by the standard linear model. Hence the variability in the measures due to the random component modelled can be properly taken care of, whereas the variability due to the design cannot.

Clearly, this is a major obstacle for deriving meaningful probability statements about influence. But since this

problem is similar to that of identifying points with discordant leverage, it need not be insurmountable as we shall see later.

The second reason is that the equivalence between looking for overly influential observations and identifying an observation whose influence is discordant has not yet been fully realized. This being most likely due to the confusion between outliers, in the response space, with 'outliers' in the 'weight' space (highly influential observations).

For example, Cook(1986) wrote:

"I view perturbing the model and testing model elaborations as distinct diagnostic methodologies. In regression models, for example, deleting a single case to assess its influence is philosophically and practically different than performing an outlier test based on an elaborated model. Similarly, formal significance tests based on the null distributions of C_{max} and ℓ_{max} seem largely irrelevant for the purpose of assessing influence..."

Cook(1986 168)

This confusion is completely resolved by taking our perspective of classifying discordant observations according to the space in which they are viewed.

The third reason is related to the second. It is the failure to recognize that for every measure of influence there is an appropriate hypothesis testing framework.

For instance, both Prescott(1986) and Beckman(1986), in their discussion of Cook(1986), raised doubts about the

validity of significance testing and of 'benchmarks' proposed for the summary of the curve $\frac{\partial T}{\partial \omega}$ studied by Cook(1986), as both depend on the perturbation scheme used in deriving $\frac{\partial T}{\partial \omega}$.

Both Prescott's(1986) and Beckman's(1986) doubts can be set aside. For, in their review of tests of discordancy, Barnett and Lewis(1984) emphasized that the optimality properties of any such test cannot be assessed without a precisely defined alternative hypothesis which specify the mechanism generating the discordancy. Said differently, many optimal tests for discordancy may exist simultaneously in a given situation, each being tailored to discriminate against a specific mechanism generating discordancies.

But for any specific measure of influence, there is a model, induced by the one for the observations, which appropriately reflects the non-existence of discordancies in influence, i.e. a model for the null hypothesis of no discordantly influential observation. Also, the perturbation scheme chosen to measure influence entails a perturbation of the null model which results in an appropriate alternative hypothesis for the presence of some discordantly influential observations. This is particularly evident from Theorem 2.5.

Before illustrating the above discussion with the sample measure of the influence of an observation on the ordinary least-squares estimator of $\hat{\beta}$, we will discuss the pros and

cons of the methods that are currently used for identifying an overly influential observation.

All the measures of influence that have been considered so far for $\hat{\beta}$ are functions of the measure of leverage h_{ii} , the residuals $\hat{e}_i$ and their estimated standard deviation $\hat{\sigma}$. In the case of a balanced design, all the h_{ii} 's are constant and entirely determined by the sample size n . Thus, measures of influence computed from a balanced design depend only on the residuals and their estimated standard deviation, and so are simple transformations of the measures of outlyingness discussed in section 2.3. Hence, for balanced designs, measures of outlyingness are equivalent to measures of influence.

Therefore, comparing a measure of influence from a general design with that from a balanced design is an uninformative oversimplification.

The second method, statistical analogies, relies on finding an interpretation of the measure of influence as a test statistic for some hypothesis, and uses this test's critical values as cut-offs for overly influential observations. A good example of this method is given by Cook(1977, 1979). Two major objections to this approach are that we should, at least, test the right hypothesis instead of an unrelated one, and it still does not solve the problem

caused by assuming a fixed design as this is still assumed by the unrelated hypotheses. The sole redeemable quality of this method is that it does yield something to work with.

The third approach consists of using Atkinson's (1981, 1985) envelopes. Basically, this technique uses a resampling method to determine the $100(1-\alpha)\%$ critical values for the maximum and/or the minimum of the measure of influence considered. Graphical displays of these envelopes are an integral part of the method. The limits of this approach are that it is assumed that the design matrix is fixed, again an oversimplification, and its extension to include some variability in the design space could get bogged down in computations.

Exploratory data analysis does not provide critical values. However, some EDA techniques should always be used in practice, whether with critical values or without.

From the foregoing discussion, it is clear that the major difficulties in identifying overly influential observations in the standard linear model are the apparent lack of an appropriate hypothesis testing framework, and the lack of a probabilistic structure on the design points.

We shall now illustrate how these difficulties can be surmounted for the sample measure of influence. This will yield critical values that could be used to identify overly influential observations.

The sample measure of influence SIC_i was defined as

$$SIC_i = (n-1)(\hat{\beta} - \hat{\beta}^{(i)})$$

which simplifies to:

$$SIC_i = (n-1) \frac{(X'X)^{-1} x_i \hat{e}_i}{1-h_{ii}}.$$

Now, the i -th observation will be overly influential whenever SIC_i is discordant in the sample S formed by the SIC_i 's, i.e. $S = \{SIC_i\}_{i=1}^n$. The mechanism generating discordancies, implied by the perturbation scheme used in deriving this measure of influence in Theorem 2.5, is clearly one of slippage in the mean.

Starting with the usual hypothesis of the standard linear model, the design matrix is fixed and $\hat{e} \sim \mathcal{N}(0, \sigma^2(I_n - H))$. Therefore, under the hypothesis that the influence of the i -th observation is not discordant, the p -vector SIC_i is normally distributed, being a linear combination of normally distributed random variables, with $E[SIC_i] = 0$ and

$$\text{Var}[SIC_i] = \sigma^2(n-1)^2 \frac{(X'X)^{-1} x_i x_i' (X'X)^{-1}}{1-h_{ii}}.$$

Notice that the SIC_i 's are correlated, as

$$\text{Cov}[SIC_i, SIC_j] = -\sigma^2(n-1)^2 \frac{h_{ij}(X'X)^{-1}x_i x_j'(X'X)^{-1}}{(1-h_{ii})(1-h_{jj})}$$

where $h_{ij} = x_i'(X'X)^{-1}x_j$. Also note that since we assume that the design matrix is fixed, only the scale, σ^2 , needs to be determined to completely specify the distribution of SIC_i .

Now, let $\Delta_i = E[SIC_i]$. To test the i -th component of S for discordancy, the hypotheses are:

$$H_0: \Delta_i = 0 \quad \text{against} \quad H_1: \Delta_i \neq 0.$$

For testing that there is no discordant observation in the full sample S , against the possibility of at most one, the hypotheses are:

$$\begin{aligned} H'_0: \Delta_i &= 0 \text{ for } i=1(1)n \\ \text{against } H'_1: \Delta_i &\neq 0, \text{ for at most one } i=1(1)n. \end{aligned}$$

However, the SIC_i 's are p -vectors, and it is not easy to judge the discordancy of an observation in a p -dimensional space. Furthermore, the rank of $\text{Var}[SIC_i]$ is equal to 1. Hence, the distributions of the SIC_i 's are singular. It is thus necessary to reduce the dimensionality of SIC_i , and there are many ways to achieve this.

Three simple transformations of SIC_i are worth considering. The first leads to $DFFIT_i$, the measure of influence introduced by Belsley, Kuh and Welsch (1980):

$$x_i' SIC_i = (n-1) \frac{x_i' (X'X)^{-1} x_i \hat{\epsilon}_i}{1-h_{ii}} = (n-1) \frac{h_{ii}}{1-h_{ii}} \hat{\epsilon}_i = (n-1) DFFIT_i$$

which can also be written as

$$x_i' SIC_i = (n-1) (x_i' \hat{\beta} - x_i' \hat{\beta}_{(i)}) = (n-1) [\hat{y}_i - \hat{y}_{(i)}]$$

where $\hat{y}_i$ is the in-sample predicted response and $\hat{y}_{(i)}$ is the out-sample predicted response, i.e. the response at x_i predicted from the remaining cases.

This transformation is a direct simplification of SIC_i , since $x_i' SIC_i$ is a scalar measuring only the displacement in the i -th predicted response resulting from the use of the i -th case in the regression. Although appealing, it does not measure the displacement in the other responses resulting from the use of the i -th case in the regression.

The second transformation, mentioned by Cook and Weisberg (1982, pp. 122-123), is

$$X_{(i)}' SIC_i = (n-1) (X_{(i)}' \hat{\beta} - X_{(i)}' \hat{\beta}_{(i)}).$$

Unlike $x_i' SIC_i$, however, it does not summarize SIC_i , since it increases the dimension of the influence measure from p to $n-1$. This is not the reason for not using it, as it only measures the change in the in-sample predicted values of y_i ,

$j=1(1)n, j \neq i$, resulting from the use of the i -th case in the regression.

The third transformation is more appropriate since it combines the two previous transformations, and thus measures the effect of using the i -th case on all of the predicted values:

$$XSIC_i = (n-1)(X\hat{\beta} - X\hat{\beta}_{(i)}).$$

It is an n -vector whose i -th component is $x_i' SIC_i$ and the remaining $n-1$ components are given by $X_{(i)} SIC_i$. Again, it is not a direct summary of SIC_i , since it increases the dimension of the influence measure from p to n .

But in the context of the standard linear model, the squared length of $XSIC_i$ and that of $X_{(i)} SIC_i$ provide adequate summaries. Hence, for $XSIC_i$, we get

$$\begin{aligned} d_i &= (XSIC_i)'(XSIC_i) \\ &= \sum_{j=1}^n (n-1)^2 \frac{[x_j'(X'X)^{-1}x_i]^2}{(1-h_{ii})^2} \hat{\epsilon}_i^2 \\ &= \frac{(n-1)^2 \hat{\epsilon}_i^2}{(1-h_{ii})^2} \left(\sum_{j=1}^n h_{ji}^2 \right). \end{aligned}$$

Since $H = \{h_{ij}\}_{i,j=1}^n$ is idempotent, $\sum_{j=1}^n h_{ji}^2 = h_{ii}$. Thus,

$$d_i = \frac{(n-1)^2 h_{ii} \hat{\epsilon}_i^2}{(1-h_{ii})^2} = p(n-1)^2 \hat{\sigma}^2 C_i$$

where C_i is Cook's (1977) distance.

We further note that d_i can be written as

$$d_i = \frac{(n-1)^2 h_{ii} \hat{\sigma}^2 r_i^2}{1-h_{ii}} = \frac{(x_i' \text{SIC}_i)^2}{h_{ii}} = \frac{(X_{(i)} \text{SIC}_i)' (X_{(i)} \text{SIC}_i)}{1-h_{ii}}$$

where r_i is the internally standardized residual used for testing the outlyingness of the i -th observation.

Thus, assuming the design fixed, there is not much to choose between the three transformations.

Now, for testing H_0 against H_1 , we note that the randomness of SIC_i , assuming the design fixed, and thus that of d_i , is totally attributable to that of $\hat{\epsilon}_i$. Therefore, from our discussion of tests of outlyingness in section 2.3, the likelihood ratio test statistic for H_0 against H_1 is equivalent to

$$\tilde{d}_i^* = \frac{d_i}{\hat{\sigma}^2(i)^2} = \frac{(n-1)^2 h_{ii} r_i^2}{1-h_{ii}} = (n-1)^2 h_{ii}^* r_i^{*2} = p(n-1)^2 C_i^*$$

where r_i^* is the externally standardized residual, h_{ii}^* is the externally standardized measure of leverage, and C_i^* is the externally standardized Cook's distance.

The externally standardized Cook distance C_i^* is therefore an appropriate statistic for testing H_0 against H_1 . Similarly, the maximum externally standardized Cook distance $C_{(n)}^*$ can be seen to be the appropriate statistic for testing H_0' against H_1' .

We would now have to remove the hypothesis that the design is fixed. To do so requires a model for the observed design points. Not just any model would do, however, as we have to remain coherent with our discussion of high leverage points of section 2.2. Hence, we have to assume that the rows of the design matrix X are i.i.d. multivariate gaussian.

If such an hypothesis is made, the distribution of C_i^* is then approximatively that of the product of two independent $\mathcal{F}$ distributions. Also, using Bonferroni's upper bound $\mathcal{U}_1$, a conservative upper bound for the $100(1-\alpha)\%$ critical value for $C_{(n)}^*$ can be obtained.

It is to be remarked, however, that since the cumulative distribution function of an $\mathcal{F}$ distribution can be expressed as an incomplete Beta function, no exact expression for the distribution function of C_i^* in terms of simpler functions can be expected.

We shall not pursue this discussion any further, as the evaluation of critical values for Cook's distance is a problem of numerical analysis, and as we have achieved our goal of presenting in a unified way the theory of discordant observations in the standard linear model.

CHAPTER III

INFLUENTIAL OBSERVATIONS IN TIME SERIES

We will now apply the principles reviewed in Chapter II to the study of influential observations in time series. We will focus our study on a statistic which is central in the analysis of time series: the autocorrelation function (ACF).

In section 3.1, we describe the standard framework for the study of time series, emphasizing the importance of the theoretical and sample ACF. In section 3.2, we introduce the extension to time series of the customary influence function due to Martin and Yohai(1986). In section 3.3, we develop three measures of influence for the ACF.

Section 3.3 is divided into four parts. In the first part, section 3.3.1, we introduce the customary estimator for the k -th autocorrelation coefficient and state some of its properties, e.g. its mean and covariances, which will be required later on.

In the second part, section 3.3.2, we present expressions for the natural estimator of the k -th autocorrelation coefficient when one observation is missing at random, and derive its two first moments in relation to those of the estimator based on a full sample.

In section 3.3.3, we obtain three measures of influence for the k -th autocorrelation coefficient and compare them to those suggested by previous authors. Finally, the proofs of the propositions of section 3.3.2 are given in section 3.3.4.

3.1 Introduction to ARIMA Models

We will mostly be concerned with regular, weakly stationary time series. As for the linear model, the study of these series relies mainly on the geometry of their space to obtain a characteristic linear representation providing a natural parameterization for making statistical inferences. On certain occasions however, it will be necessary to specify a probability law, in which cases we shall state it explicitly.

A time series is an ordered collection of random variables, denoted $(X_t)_{t \in \mathbb{Z}}$, defined over the same probability space and such that X_t has a finite variance for all $t \in \mathbb{Z}$.

Three particular overlapping sub-classes of time series processes have been extensively studied.

The first class is that of regular processes, which consist of those series $(X_t)_{t \in \mathbb{Z}}$ whose whole path over $t \in \mathbb{Z}$ cannot be completely described probabilistically by a finite number of parameters. Otherwise, the process is said to be

singular, as the joint density of a finite subcollection of $(X_t)_{t \in \mathbb{Z}}$ may fail to exist. Singular processes are sometimes called purely deterministic. All the time series that we shall consider will be assumed to be regular.

The second class is that of strictly stationary processes, which are such that the joint distribution function of a finite subcollection of $(X_t)_{t \in \mathbb{Z}}$ does not depend on the time origin, i.e. for any $(t_1, t_2, \dots, t_n) \subset \mathbb{Z}$, and for all $h \in \mathbb{Z}$,

$$P(X_{t_1+h}, X_{t_2+h}, \dots, X_{t_n+h}) = P(X_{t_1}, X_{t_2}, \dots, X_{t_n}).$$

An important example of a strictly stationary process is a sequence of i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

The third class is that of weakly stationary processes, or simply stationary processes, which is made of those series $(X_t)_{t \in \mathbb{Z}}$ whose two first moments do not depend on the time origin, i.e. $E[X_t] = \mu$ a constant, and $\text{Cov}[X_t, X_{t+k}] = \gamma_X(k)$ for all $t, k \in \mathbb{Z}$.

An important example of a stationary series is a white noise process $(\epsilon_t)_{t \in \mathbb{Z}}$ which is such that $E[\epsilon_t] = 0$ and

$$\text{Cov}[\epsilon_t, \epsilon_{t+k}] = \begin{cases} \sigma^2 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0. \end{cases}$$

A strictly stationary series is weakly stationary, but the converse is not necessarily true.

The quantity $(\gamma_x(k))_{k \in \mathbb{Z}}$ is called the autocovariance function, and for $k \in \mathbb{Z}$ the quantity

$$\rho_x(k) = \frac{\gamma_x(k)}{\gamma_x(0)}$$

is called the autocorrelation function.

An important property of a regular and weakly stationary time series for modelling purposes is its infinite moving average representation in terms of a white noise process. It can be shown (see e.g. Azencott and Dacunha-Castelle 1984) that for $(X_t)_{t \in \mathbb{Z}}$ regular and stationary, there exists a white noise process $(\epsilon_t)_{t \in \mathbb{Z}}$ and constants $\theta_k \in \mathbb{R}$ such that $\sum_{k=0}^{\infty} \theta_k^2 < \infty$ and such that $X_t = \sum_{k=0}^{\infty} \theta_k \epsilon_{t-k}$.

When a process is represented as above, it is called an infinite moving average process and is denoted $MA(\infty)$.

Because of the simple properties of a white noise process, this representation is useful for making inferences. However, given a finite sample we cannot determine the infinite number of parameters in that model unless they have a special structure.

One such special structure is the moving average process of order q , denoted $MA(q)$, where $\theta_0 = 1$, $\theta_q \neq 0$ and the θ_k 's are equal to zero for $k > q$, i.e.

$$X_t = \sum_{k=0}^q \theta_k \epsilon_{t-k}.$$

Other sequences of parameters also allow a finite parameterization. Consider the following MA(∞) process

$$X_t = \sum_{k=0}^{\infty} \phi^k \epsilon_{t-k} \quad \text{where } |\phi| < 1.$$

Then,

$$X_t - \phi X_{t-1} = \sum_{k=0}^{\infty} \phi^k \epsilon_{t-k} - \sum_{k=0}^{\infty} \phi^{k+1} \epsilon_{t-k-1} = \epsilon_t.$$

Generalizing this argument yields the class of autoregressive processes of order p, denoted AR(p) where, with $\phi_0 = 1$ and $\phi_p \neq 0$,

$$\sum_{k=0}^p \phi_k X_{t-k} = \epsilon_t.$$

In the sequel, we shall make use of the following notation. The backward shift operator B is defined as $BX_t = X_{t-1}$. The difference operator of order s is defined as $\nabla_s = 1 - B^s$. For simple differences, i.e. when the order of differencing is equal to one, the difference operator will simply be denoted by ∇ .

Using the backward shift operator, the AR(p) and MA(q) models can be respectively rewritten as:

$$\epsilon_t = \sum_{k=0}^p \phi_k X_{t-k} = \sum_{k=0}^p \phi_k B^k X_t = \phi(B) X_t$$

and as

$$X_t = \sum_{k=0}^q \theta_k \epsilon_{t-k} = \sum_{k=0}^q \theta_k B^k \epsilon_t = \theta(B) \epsilon_t.$$

Henceforth, we shall always assume that all polynomials in B are never degenerate in their degree, e.g. for $\phi(B)$ $\phi_p \neq 0$, and that they have the following structure:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p.$$

We will also use the sets S_q of vectors $\theta \in \mathbb{R}^q$ for which $\theta(B)$, viewed as a polynomial in B , has all its roots outside the complex unit circle; i.e.

$$S_q = \{(\theta_1, \dots, \theta_q) \in \mathbb{R}^q: \forall \lambda \in \mathbb{C} \text{ with } 1 - \theta_1 \lambda - \dots - \theta_q \lambda^q = 0 \Rightarrow |\lambda| > 1\}.$$

Now, by mixing an $AR(p)$ process with an $MA(q)$ process we obtain a large class of regular and stationary processes admitting a finite parameterization. Let $(X_t)_{t \in \mathbb{Z}}$ be a time series such that, for some $(\phi_1, \dots, \phi_p) \in S_p$ and some $(\theta_1, \dots, \theta_q) \in S_q$, $\phi(B)X_t = \theta(B)\epsilon_t$, where $\phi(B)$ and $\theta(B)$ are coprime and $(\epsilon_t)_{t \in \mathbb{Z}}$ is a white noise process. Then $(X_t)_{t \in \mathbb{Z}}$ is said to be an autoregressive moving-average process of order p and q , denoted $ARMA(p, q)$.

Even though the class of ARMA processes is quite large, it does not cover many cases in practice. For example, the observed time series are frequently not stationary, and they may have a strong seasonal behaviour.

To transform a non-stationary time series into a stationary one, we usually do one or both of the following:

'linearize' it with respect to its own past, e.g. using a power transformation; or filter it with simple differences. These reductions are approximate. In the sequel however, we will always assume that the proper transformations have been applied when necessary.

When the observed non-stationary process $(X_t)_{t \geq 1}$ is such that there exists an integer d for which $\nabla^d X_t = (1-B)^d X_t$ is an ARMA(p, q) process, it is called an autoregressive-integrated-moving-average process of orders p , d and q , denoted ARIMA(p, d, q). Note that usually $d = 0, 1$ or 2 is adequate to model the series.

Frequently, the time series observed have a seasonal behaviour, i.e. their trajectories exhibit a regularly repetitive pattern. This pattern may simply be an addition to a simpler stochastic process, or it may interact with it. Both cases are already included in the ARIMA(p, d, q) class. For the later case, however, it is often useful to recognize that its ARIMA model can be factored into a purely seasonal component and a non-seasonal component as follows:

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d X_t = \Theta(B^s)\theta(B)\epsilon_t$$

where s , the seasonality, is the period of the repeating pattern. Such models are noted ARIMA(p, d, q)(P, D, Q), where P and Q are the orders of $\Phi(B)$ and $\Theta(B)$, respectively.

Up until now, we have implicitly assumed that ARIMA processes had a null expectation. This is an unnecessary restriction. When the mean of a time series is a non-zero constant, it is called the drift and is represented by θ_0 . Thus the most general class of time series that we will be considering is the class of $\text{ARMA}(p,q)_0$ processes, where the subscript 0 indicates a non-zero drift. That is

$(X_t)_{t \in \mathbb{Z}} \in \text{ARMA}(p,q)_0$ if

$$\phi(B)X_t = \theta_0 + \theta(B)\epsilon_t$$

with $(\phi_1, \dots, \phi_p) \in S_p$, $(\theta_1, \dots, \theta_q) \in S_q$, and $\theta_0 \in \mathbb{R}$ for some white noise process $(\epsilon_t)_{t \in \mathbb{Z}}$ of variance σ^2 .

As popularized by Box and Jenkins(1976), to model a finite realization of a regular and stationary time series as an $\text{ARMA}(p,q)_0$ process, there are at least three steps to follow. The first step, called identification, consists of the tentative specification of the degrees p and q of $\phi(B)$ and $\theta(B)$, respectively. The second step consists of the estimation of the parameters of $\phi(B)$ and $\theta(B)$, and that of θ_0 and σ^2 . The third step consists of checking the validity of the estimated model.

In implementing these three steps two quantities are of fundamental importance, the sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF). The sample ACF will be defined in section 3.3.1 and the sample

PACF is a one-to-one transformation of the ACF. The importance of these two quantities for modelling ARMA processes can be seen as follows.

Many methods have been proposed for identifying the degrees, p and q , of $\phi(B)$ and $\theta(B)$. A recent review is given by de Gooijer et al.(1985). One of the most frequently used methods is that of Box and Jenkins(1976), which consists of a study of the patterns of the sample ACF and PACF. The other methods of identification also rely on the sample ACF and PACF as these methods attempt to automate the recognition of the patterns in these functions.

There are also many methods for estimating the parameters of an $ARMA(p,q)_0$ process: least-squares, maximum likelihood and their variations. Differing in their choice of metric, they nonetheless all depend on the difference between the sample and the theoretical ACF (or PACF) which characterizes an ARMA process with a given set of parameters. This is most evident from the work of Godolphin(1984).

To check the validity of a fitted ARMA model, the most widely used statistics are the ACF and PACF of the residuals as well as the Ljung-Box(1978) statistic which is a weighted sum of the squares of the coefficients of the residuals' ACF.

Because of the importance of the ACF and of the PACF in deriving an ARMA model, the identification of observations

which unduly influence these quantities will be equally important. In the following, we will be concerned with the identification of overly influential observations for the autocorrelation function.

3.2 Influence Functions for Time Series

The customary influence function (see section 2.4.1),

$$f(X; T, F) = \lim_{\epsilon \rightarrow 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T[F]}{\epsilon}$$

cannot be used for time series, since the perturbation scheme used,

$$(1-\epsilon)F + \epsilon\delta_x \tag{3.1}$$

does not correspond to any interesting form of contamination in a time series.

Let $F: \mathbb{R}^N \rightarrow [0,1]$, where $\mathbb{R}^N = \prod_{n \in \mathbb{N}} \mathbb{R}$, denote the time series' distribution function. The perturbation scheme (3.1) corresponds to mixing F with the distribution function δ_x which gives full weight to a whole path in $\mathbb{R}^N$. In other words, the perturbation scheme (3.1) corresponds to a shift of the observed paths in their entirety by a fixed amount.

In practice, the contaminations encountered are those where only some of the coordinates of the time series are affected, or where the whole sample path is contaminated by another time series whose distribution function is not

degenerate. Thus, Martin and Yohai(1986) define the influence functions for time series as:

$$J(X;T,\eta_x,\eta_w) = \lim_{\epsilon \rightarrow 0} \frac{T[\eta_\gamma(\epsilon)] - T[\eta_x]}{\epsilon}$$

where $\eta_\gamma(\epsilon)$ is the probability measure of the contaminated process

$$Y_t(\epsilon) = [1 - Z_t(\epsilon)]X_t + Z_t(\epsilon)W_t ,$$

$(X_t)_{t \in \mathbb{Z}} \sim \eta_x$ is a realization of the nominal process,

$(W_t)_{t \in \mathbb{Z}} \sim \eta_w$ is a realization of a contaminating process, and

$(Z_t(\epsilon))_{t \in \mathbb{Z}} \sim \eta_z(\epsilon)$ is a 0-1 process where $\epsilon \in [0,1]$ and

$P(Z_t(\epsilon)=1) = g(\epsilon) = \epsilon + o(\epsilon)$ for some function g .

Note that the main difference between the definition of Martin and Yohai(1986) and the standard one is that an influence function for time series depends on the probability measure η_w of the contaminating process. Nonetheless, it keeps its interpretation as the infinitesimal change in the value of the statistic T as the weight given to the contaminating process by the probability measure of the observed process is smoothly decreased to zero. Related work on influence functions for time series appear in Künsch(1984).

Although clearly different, Martin and Yohai(1986) show that for a particular form of contamination $J(X;T,\eta_x,\eta_w)$ is truly a generalization of $J(X;T,F)$ as they coincide when $(X_t)_{t \in \mathbb{Z}}$ is a sequence of i.i.d. random variables.

3.3 Measures of Influence for the Autocorrelation Function

In Chapter II we defined three particularly important measures of influence in the i.i.d. case, the empirical added influence EIC_i , the empirical deleted influence $EIC(i)$ and the sample measure of influence SIC_i .

For time series these three measures of influence cannot be obtained from the influence function by simply 'deleting' the i -th observation from the computation of T as this may alter the structure of dependence between the observations.

This is quite clear using Martin and Yohai's (1986) definition of influence function given in section 3.2. The i -th observation must be removed without distorting the dependency structure of the remaining observations. As remarked by Peña (1986, 1987), this will often correspond to considering the i -th observation as missing at random, and by replacing it in the computation of T by a forecast based solely on the remaining observations.

Let $(Z_t)_{t=1}^n$ denote a realization from an n -dimensional random vector with joint density $g(\cdot)$. For $t=1(1)n$, define $a_t=1$ when Z_t is observed, and $a_t=0$ otherwise, with $P(a_t=0)=p_t$. Then the i -th observation is said to be missing at random, when the joint density of a_i and $(Z_t)_{t=1}^n$ can be written as

$$h(a_i, Z_1, Z_2, \dots, Z_n) = p_i g(Z_1, Z_2, \dots, Z_n).$$

Another way to see the difficulty involved is by trying to prove Theorem 2.5 giving the relationship of the three aforementioned measures of influence to the derivative of T viewed as a functional of the EDF, in the case of dependent observations. In this situation, the factorization of the EDF is not trivial.

For instance, assuming that the observed time series has zero mean, the sample estimate of the k -th autocovariance coefficient is:

$$\tilde{C}_k = n^{-1} \sum_{t=1}^{n-k} X_t X_{t+k}$$

which can be written as a functional of the EDF:

$$\tilde{C}_k = \frac{n-k}{n} \int xy dF[x, y]$$

where

$$F[x, y] = (n-k)^{-1} \sum_{t=1}^{n-k} \delta_{x_t, x_{t+k}}(x, y)$$

and

$$\delta_{x_t, x_{t+k}}(x, y) = \begin{cases} 1 & \text{if } (x, y) = (X_t, X_{t+k}) \\ 0 & \text{otherwise.} \end{cases}$$

Because of the dependency between the observations, when one observation is missing the joint probability mass $\delta_{x_t, x_{t+k}}$ cannot be factored without making special assumptions on the nature of the dependency.

For example, Li and Hui(1987) obtained a measure of influence for the k -th autocorrelation coefficient by assuming that the observed time series was an $AR(p)$ process.

Instead of attempting to prove Theorem 2.5 for dependent observations using Martin and Yohai's(1986) influence function for time series, we shall use directly the relationship of Theorem 2.5 to obtain the three main measures of influence. As we shall see in section 3.3.3, these three measures of influence, as well as others, can then be obtained for any statistic T computed from a time series.

The three measures of influence EIC_i , $EIC^{(i)}$ and SIC_i will be derived, even if the sample measure of influence SIC_i is preferred because it is a more appropriate summary of $\frac{\partial T}{\partial \omega_i}$ than the two others. This derivation will allow us to compare the three and hence gain further insight into their properties.

Because we will study the statistical properties of $SIC_{i,k} = (n-1)(r_k - r_k^{(i)})$ in Chapter IV, we will need some of the statistical properties of r_k and those of $r_k^{(i)}$. Thus, the remainder of this section is divided into four parts. In the first part, section 3.3.1, we introduce the usual estimator of the ACF, denoted $(r_k)_{k=1}^L$, and state some of its properties, in particular its mean and covariances. In the second part, section 3.3.2, we give both an exact and an approximate representation of the sample ACF when one observation is

missing at random, denoted $(r_k^{(i)})_{k=1}^L$, in terms of $(r_k)_{k=1}^L$. We also give approximations to its first two moments in terms of those for the full-sample ACF, up to terms of order $O(n^{-2})$ where n is the length of the observed time series. The results of sections 3.3.1 and 3.3.2 will be used in section 3.3.3 and in Chapter IV.

In the third part, section 3.3.3, we obtain the three measures of influence, one of which appears to be new, as well as approximations thereof, and compare them with other measures of influence that have already been suggested in the literature.

However, we will see in section 3.3.3 that the results of sections 3.3.1 and 3.3.2 are sufficient to compute the expectation of the sample measure of influence, but that they are not precise enough for computing its covariances. Thus a more useful representation of the sample measure of influence will also be developed in section 3.3.3 which will be used in Chapter IV to obtain good approximations to its expectation and covariances.

Finally, section 3.3.4 contains the proofs of the results given in section 3.3.2.

3.3.1 The Theoretical and Sample ACFs

The theoretical autocorrelation function (ACF) for a weakly stationary and regular process $(X_t)_{t \in \mathbb{Z}}$ is defined as $\rho_k = \gamma_k / \gamma_0$ for $k \in \mathbb{Z}$, where γ_k is the k -th autocovariance of $(X_t)_{t \in \mathbb{Z}}$ and μ is the mean of the process, i.e.

$$\mu = E[X_t] \quad \text{and} \quad \gamma_k = E[(X_t - \mu)(X_{t+k} - \mu)].$$

The important properties of ρ_k are that it is symmetrical about zero, $\rho_k = \rho_{-k}$, that each ρ_k is bounded, $|\rho_k| \leq 1$, and they form a square-summable sequence, $\sum_{k=1}^{\infty} \rho_k^2 < \infty$. Furthermore, the sequence $(\rho_k)_{k \in \mathbb{Z}}$ is positive-definite, i.e. for any set of real numbers $(a_j)_{j=1}^m$ whose elements are not all zero, and for any finite set of indices $(i_j)_{j=1}^m \subset \mathbb{Z}$,

$$\sum_{j=1}^m \sum_{k=1}^m a_j a_k \rho_{i_j - i_k} > 0.$$

The proof of these properties can be found in Fuller (1976). Note however that the square-summability of the ACF is a consequence of the square-summability of the coefficients in the infinite moving-average representation of a regular and weakly stationary process.

Many estimators have been proposed for the autocovariance γ_k , and for the autocorrelation ρ_k . See chapters 6 and 8 of Anderson(1971). Letting n denote the length of the observed

process, the most frequently used estimators of γ_k and ρ_k are respectively

$$c_k = n^{-1} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})$$

and

$$r_k = c_k / c_0 ,$$

where

$$\bar{X} = n^{-1} \sum_{t=1}^n X_t .$$

Although both c_k and r_k are biased and are not the minimum mean square estimators of γ_k and ρ_k respectively, their choice as estimators is based on the fact that both form a positive-definite sequence whenever the realization is not constant, i.e. whenever $X_t \neq X_{t'}$, for some $t \neq t'$. See McLeod and Jiménez(1984). They further have the following properties which correspond to those of the theoretical autocovariance and autocorrelation functions: both c_k and r_k are symmetric about $k = 0$ and r_k is bounded.

The following propositions give the first two moments of $\bar{X}$, the usual estimator of the mean μ of the process, those of $(c_k)_{k=1}^L$ and those of $(r_k)_{k=1}^L$ up to terms of order $O(n^{-2})$, where n is the length of the realization. The proofs of these results can also be found in Fuller(1976, chapter 6). They are stated here for future reference.

Proposition 3.1: Let $(X_t)_{t=1}^n$ be a finite realization of a weakly stationary and regular process with mean μ and absolutely summable autocovariance function $(\gamma_k)_{k \in \mathbb{Z}}$, i.e.

$$\sum_{k \in \mathbb{Z}} |\gamma_k| = \mathcal{M} < \infty.$$

Then $\bar{X} = n^{-1} \sum_{t=1}^n X_t$ is an unbiased estimator of μ , and

$$\text{Var}[\bar{X}] = f_0 = n^{-2} \sum_{s=1}^n \sum_{t=1}^n \gamma_{s-t} = O(n^{-1}).$$

Thus $\lim_{n \rightarrow \infty} \bar{X} = \mu$.

We remark that both f_0 and $f_{n-k} = n^{-2} \sum_{t=1}^k \sum_{s=1}^n \gamma_{t-s}$ will occur frequently in what follows. When $k = 0$, $f_{n-k} = f_n$ is defined as zero. Note also that it is shown in Lemma 3.8, section 3.3.4, that f_{n-k} is $O(n^{-2})$ only for k fixed with respect to n .

Proposition 3.2: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1 with

$$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m} \quad \text{where the } \epsilon_t \text{'s are } iid(0, \sigma^2) \text{ random}$$

variables with $E[\epsilon_t^4] = 3k\sigma^4 < \infty$. Then for k and l fixed,

$$E[c_k] = \left(1 - \frac{k}{n}\right) \gamma_k - f_0 + O(n^{-2})$$

and

$$\text{Cov}[c_k, c_l] = 3(k-1)\gamma_k\gamma_l n^{-1} + n^{-2} \sum_{s=1}^{n-l-n-k} \sum_{t=1}^{n-l-n-k} (\gamma_{t-s}\gamma_{t-s+k-l} + \gamma_{t-s+k}\gamma_{t-s+l}) + O(n^{-2}).$$

Thus $\lim_{n \rightarrow \infty} c_k = \gamma_k$.

Note that Proposition 3.2 corrects the expression for the expected value of c_k given by Fuller(1976, p. 239) as:

$$E[c_k] = \left(1 - \frac{k}{n}\right)\gamma_k - \left(1 - \frac{k}{n}\right)f_0 + o(n^{-2}).$$

The 'correct' coefficient of f_0 is $\left(1 + \frac{k}{n}\right)$. But since f_0 is of order $o(n^{-1})$, $\pm \frac{k}{n}$ is $o(n^{-2})$. Hence, up to terms of order $o(n^{-2})$, Fuller's expression for $E[c_k]$ agrees with the one above.

Proposition 3.3: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1 with

$$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m} \quad \text{where the } \epsilon_t \text{'s are } iid(0, \sigma^2) \text{ random}$$

variables with $E[\epsilon_t^4] = 3k\sigma^4 < \infty$ and $E[\epsilon_t^6] < \infty$.

Then for k and l fixed,

$$E[r_k] = \left(1 - \frac{k}{n}\right)\rho_k - (1 - \rho_k)\frac{f_0}{\gamma_0} - \frac{\text{Cov}[A_k, C_0]}{\gamma_0} + o(n^{-2})$$

and

$$\text{Cov}[r_k, r_l] = \text{Cov}[A_k, A_l] + o(n^{-2})$$

where

$$A_k = \frac{c_k - \rho_k c_0}{\gamma_0}.$$

Thus $\lim_{n \rightarrow \infty} r_k = \rho_k$.

3.3.2 The Sample ACF When One Observation is Missing

When observations are missing in a realization, i.e. the observed $(X_t)_{t \in T}$ is such that $T \neq \{1, 2, \dots, n\}$, the usual estimators of γ_k and ρ_k , namely c_k and r_k , cannot be computed as the observations are not consecutive.

From the review of Dunsmuir(1984), when the i -th observation of a weakly stationary and regular time series is missing at random the most natural estimators of μ , γ_k and ρ_k are respectively:

$$\bar{X}^{(i)} = \frac{1}{n-1} \sum_{\substack{t=1 \\ t \neq i}}^n X_t,$$

$$c_k^{(i)} = (n-1)^{-1} \sum_{t \in T_i} (X_t - \bar{X}^{(i)}) (X_{t+k} - \bar{X}^{(i)})$$

and

$$r_k^{(i)} = \frac{c_k^{(i)}}{c_0^{(i)}},$$

where

$$T_i = \begin{cases} \{t=1(1)n-k, t \neq i\} & \text{for } k=0 \text{ or } 1 \leq i \leq k \\ \{t=1(1)n-k, t \neq i, t \neq i-k\} & \text{for } k < i \leq n-k \\ \{t=1(1)n-k, t \neq i-k\} & \text{for } n-k < i \leq n. \end{cases}$$

Note that $\bar{X}^{(i)}$, as an estimator of μ , enjoys the same statistical properties as $\bar{X}$; in particular it is unbiased. Also, the properties of the process $(Y_t = X_t - \bar{X}^{(i)})_{t=1}^n$ are

similar to those of $(Z_t = X_t - \bar{X})_{t=1}^n$. The most straightforward way to see this is to note that

$$Y_t = Z_t + \frac{Z_1}{n-1},$$

and that

$$\bar{X} = \frac{(n-1)\bar{X}^{(1)} + X_1}{n}. \quad (3.2)$$

Thus the process $(Y_t)_{t \in \mathbb{Z}}$ has mean zero and becomes weakly stationary as n tends to infinity.

The estimator $c_k^{(1)}$ of γ_k is also related to the full sample quantity c_k . Lattin(1983) gave formulae relating $c_k^{(1)}$ to c_k , but they are in error. The correct relationships are:

$$nc_0 = (n-1)c_0^{(1)} + \frac{(n-1)}{n} Y_1^2 \quad (3.3)$$

$$nc_k = (n-1)c_k^{(1)} + \begin{cases} Y_1[Y_{1+k} - E_k^{(1)}] - \frac{kY_1^2}{n^2} & 1 \leq i \leq k \\ Y_1[Y_{1-k} + Y_{1+k} - E_k^{(1)}] - \frac{(n+k)Y_1^2}{n^2} & k < i < n-k \\ Y_i[Y_{i+k} - E_k^{(1)}] - \frac{kY_1^2}{n^2} & n-k < i < n \end{cases} \quad (3.4)$$

where

$$E_k^{(1)} = \frac{1}{n} \left(\sum_{t=1}^{n-k} Y_t + \sum_{t=1+k}^{n-k} Y_t \right).$$

Or again, in terms of the process $(Z_t)_{t=1}^n$:

$$nc_0 = (n-1)c_{0^{(1)}} + \frac{n}{n-1} Z_1^2 \quad (3.5)$$

$$nc_k = (n-1)c_{k^{(1)}} + \begin{cases} \frac{n}{n-1} Z_1 [Z_{1+k} - E_k] + \frac{kZ_1^2}{(n-1)^2} & 1 \leq i \leq k \\ \frac{n}{n-1} Z_1 [Z_{1+k} + Z_{1+k} - E_k] + \frac{(n+k)Z_1^2}{(n-1)^2} & k < i \leq n-k \\ \frac{n}{n-1} Z_1 [Z_{1+k} - E_k] + \frac{kZ_1^2}{(n-1)^2} & n-k < i \leq n \end{cases} \quad (3.6)$$

where

$$E_k = n^{-1} \sum_{t=1}^{n-k} (Z_t + Z_{t+k}).$$

These relationships between c_k and $c_{k^{(1)}}$ are straightforward to obtain and are exact. From these relationships we can obtain the expected value and the covariances of $c_{k^{(1)}}$ and of $r_{k^{(1)}}$ in terms of those of c_k and of r_k , respectively.

The following Proposition 3.4, which does not appear to have been published before, gives the expected value and the covariances of $(c_{k^{(1)}})_{k=1}^n$ up to terms of order $O(n^2)$. The proof of Proposition 3.4 is given in section 3.3.4.

Proposition 3.4: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1 with

$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m}$ where the ϵ_t 's are $\mathcal{N}(0, \sigma^2)$ random variables with $E[\epsilon_t^4] = 3k\sigma^4 < \infty$. Then for i, j, k and l fixed,

$$E[c_k^{(i)}] = \begin{cases} \left(\frac{n(n-2)}{(n-1)^2} - \frac{k}{n-1} \right) \gamma_k - \frac{n^2}{(n-1)^2} f_0 + o(n^{-2}) \\ \text{for } k=0, \text{ or } 1 \leq i \leq k, \text{ or } n-k < i < n, \\ \\ \left(\frac{n(n-3)}{(n-1)^2} - \frac{k}{n-1} \right) \gamma_k - \frac{n(n+1)}{(n-1)^2} f_0 + o(n^{-2}) \\ \text{for } k < i \leq n-k, \end{cases}$$

and

$$\text{Cov}[c_k^{(i)}, c_l^{(j)}] = \frac{n^2}{(n-1)^2} \text{Cov}[c_k, c_l] + o(n^{-2})$$

while

$$\text{Cov}[c_k^{(i)}, c_l] = \frac{n}{n-1} \text{Cov}[c_k, c_l] + o(n^{-2}).$$

The next proposition, Proposition 3.5, gives the expected value and the covariances of $(r_k^{(i)})_{k=1}^L$ in terms of those of $(r_k)_{k=1}^L$ up to terms of order $o(n^{-2})$. As Proposition 3.4, it does not appear to have been published before. The proof of Proposition 3.5 is given in section 3.3.4.

Proposition 3.5: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1, with

$$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m} \quad \text{where the } \epsilon_t \text{'s are } \mathcal{U}(0, \sigma^2) \text{ random}$$

variables with $E[\epsilon_t^4] = 3k\sigma^4 < \infty$ and $E[\epsilon_t^6] < \infty$.

Then for i, j, k and l fixed,

$$E[r_k^{(i)}] = \begin{cases} \left(1 - \frac{k}{n-1}\right) \rho_k - \frac{n^2(1-\rho_k)f_0}{(n-1)^2\gamma_0} - \frac{n^2\text{Cov}[A_k, C_0]}{(n-1)^2\gamma_0} + o(n^{-2}) \\ \quad \text{when } 1 \leq i \leq k \text{ or } n-k < i \leq n \\ \\ \left(1 - \frac{k+1}{n-1}\right) \rho_k - \frac{n^2(1-\rho_k)f_0}{(n-1)^2\gamma_0} - \frac{n^2\text{Cov}[A_k, C_0]}{(n-1)^2\gamma_0} + o(n^{-2}) \\ \quad \text{for } k < i \leq n-k, \end{cases}$$

where

$$A_k = \frac{C_k - \rho_k C_0}{\gamma_0}.$$

Also,

$$\text{Cov}[r_k^{(i)}, r_l^{(j)}] = \frac{n^2}{(n-1)^2} \text{Cov}[r_k, r_l] + o(n^{-2})$$

and

$$\text{Cov}[r_k^{(i)}, r_l] = \frac{n}{n-1} \text{Cov}[r_k, r_l] + o(n^{-2}).$$

Together, Propositions 3.3 and 3.5 show that it is necessary to consider terms of order $o(n^{-2})$ in the expression for the covariance of $r_k^{(i)}$ with $r_l^{(j)}$ when comparing $r_k^{(i)}$ with r_k , since $\text{Cov}[r_k^{(i)} - r_k, r_l^{(j)} - r_l] = o(n^{-2})$.

3.3.3 Measures of Influence for the Sample ACF

We will now derive the three main measures of influence for the sample autocorrelation function (ACF). In Chapter II, Theorem 2.5, we saw that these main measures of influence for the statistic T were obtained from $\frac{\partial T}{\partial \omega_i}$ in the i.i.d. case, where ω_i is the weight given to the i -th observation by the empirical distribution function (EDF).

These measures of influence are:

$$EIC_i = (n-1)^{-1} \frac{\partial T}{\partial \omega_i}(1)$$

the empirical added influence measuring the infinitesimal change in T as the i -th observation is smoothly added to the sample;

$$EIC(i) = n^{-1} \frac{\partial T}{\partial \omega_i}(0)$$

the empirical deleted influence measuring the infinitesimal change in T as the i -th observation is smoothly removed from the sample; and

$$SIC_i = (n-1)(T(i)-T)$$

the sample measure of influence which simply measures the change in T due to the i -th observation.

As was noted in Chapter II, SIC_i is to be preferred over the two others as it is a more appropriate summary of the whole curve $\frac{\partial T}{\partial \omega_i}$ than the others.

These three measures of influence can be obtained for any statistic T computed from a finite realization of a time series. But they cannot be obtained by simply 'deleting' the i -th observation from the computation of T , as this may alter the structure of dependency between the observations.

To obtain the above measures of influence in the time series context, the i -th observation must be removed without distorting the dependency structure of the remaining observations. This will be achieved by considering the i -th observation as missing at random, as defined in section 3.3, and by replacing it in the computation of T by a forecast based solely on the remaining observations.

For the sample ACF, and in our context where we principally wish to measure the influence of each observation at the identification stage of an ARMA model, one forecast that can be made of a missing observation is the mean of the remaining observations, i.e.

$$\bar{X}^{(i)} = (n-1)^{-1} \sum_{\substack{t=1 \\ t \neq i}}^n X_t.$$

Furthermore, to ensure that the 'missing' observation will not distort the sample ACF, we use $r_k^{(i)} = c_k^{(i)}/c_0^{(i)}$ as an estimator of the k -th autocorrelation. Its properties were discussed in section 3.3.2.

The following Theorem 3.6 gives the three measures of influence for the sample ACF. Note that the empirical measure of deleted influence $EIC_k(i)$ does not appear to have been derived before.

Theorem 3.6: For the sample ACF $r_k = c_k/c_0$, let $\tilde{E}_k = E_k/c_0^{1/2}$, $\tilde{E}_k(i) = E_k(i)/c_0(i)^{1/2}$, $\tilde{Z}_t = (X_t - \bar{X})/c_0^{1/2}$ for $t=1(1)n$ and 0 otherwise, and let $\tilde{Y}_t = (X_t - \bar{X}(i))/c_0(i)^{1/2}$ for $t=1(1)n$ and 0 otherwise, where $\bar{X}$, $\bar{X}(i)$, c_k , $c_k(i)$, r_k , $r_k(i)$, E_k and $E_k(i)$ are as defined in sections 3.3.1 and 3.3.2. Then

$$EIC_k(i) = \frac{\tilde{Y}_1[\tilde{Y}_{1+k} + \tilde{Y}_{1-k} - \tilde{E}_k(i)]}{n^2},$$

$$SIC_{i,k} = (n-1)(r_k - r_k(i)) = \frac{\tilde{Z}_1[\tilde{Z}_{1+k} + \tilde{Z}_{1-k} - \tilde{E}_k] - (r_k - \beta_{1,k})\tilde{Z}_1^2}{1 - \frac{\tilde{Z}_1^2}{n-1}}$$

and

$$EIC_{i,k} = \frac{n\tilde{Z}_1[\tilde{Z}_{1+k} + \tilde{Z}_{1-k} - \tilde{E}_k]}{(n-1)^3} + \left[\frac{\beta_{1,k}}{(n-1)^2} - \left(2 + \frac{1}{n-1}\right)r_k \right] \frac{\tilde{Z}_1^2}{(n-1)^2}$$

where

$$\beta_{1,k} = \frac{k}{n(n-1)} + \begin{cases} (n-1)^{-1} & \text{for } k < i < n-k \\ 0 & \text{otherwise.} \end{cases}$$

Proof: As was already noted, obtaining these measures of influence from the representation of the sample ACF as a functional of the EDF is not trivial.

However, relations (3.3) and (3.4) can be used to this effect as $r_k^{(i)}$ is an estimator of ρ_k which does not depend on the i -th observation, and which does not distort the dependency structure of the observations. Furthermore, relations (3.3) and (3.4) clearly show how the i -th observation enters into the computation of r_k .

We will thus obtain the three measures of influence from relations (3.3) and (3.4), giving the details only for the case $k < i < n-k$.

Recall that in the notation of Chapter II, $T[\omega_i]$ represents a statistic as a function of the weight given to the i -th observation by the EDF. From equations (3.3) and (3.4), for i fixed with $k < i \leq n-k$, and for $\omega_i \in [0,1]$ we can write:

$$c_0[\omega_i] = \left(1 - \frac{\omega_i}{n}\right)c_0^{(i)} + \frac{(n-1)}{n^2}Y_i^2\omega_i^2$$

and

$$c_k[\omega_i] = \left(1 - \frac{\omega_i}{n}\right)c_k^{(i)} + \frac{\omega_i Y_i}{n} [Y_{i+k} + Y_{i-k} - E_k^{(i)}] - \frac{(n+k)}{n^2}Y_i^2\omega_i^2.$$

Notice that both $c_k[\omega_i]$ and $c_0[\omega_i]$ smoothly varies from $c_k^{(i)}$, $c_0^{(i)}$ ($\omega_i=0$) to c_k and c_0 ($\omega_i=1$) when ω_i smoothly varies in $[0,1]$. Thus this representation as a function of the weight given to the i -th observation can be used directly to obtain the three measures of influence.

Since

$$\frac{\partial r_k}{\partial \omega_1}[\omega_1] = \left(\frac{\partial C_k}{\partial \omega_1}[\omega_1] - r_k[\omega_1] \frac{\partial C_0}{\partial \omega_1}[\omega_1] \right) C_0[\omega_1],$$

we get the stated results for $EIC_{1,k}$ and $EIC_k^{(1)}$ after straightforward differentiation and algebra.

To obtain a useful representation of $SIC_{1,k}$, we proceed as follows. We have

$$d_{1,k} = r_k - r_k^{(1)} = \frac{C_k C_0^{(1)} - C_0 C_k^{(1)}}{C_0 \dot{C}_0^{(1)}}.$$

Using relations (3.5) and (3.6), we get, after some simplifications,

$$d_{1,k} = \frac{n[Z_1(Z_{1+k} + Z_{1-k} - E_k) - (r_k - \beta_{1,k})Z_1^2]}{(n-1)^2 C_0^{(1)}}$$

where $\beta_{1,k}$ is as given in the statement of the theorem.

Using again relation (3.5) and dividing both the numerator and denominator by C_0 , we get

$$d_{1,k} = \frac{\tilde{Z}_1(\tilde{Z}_{1+k} + \tilde{Z}_{1-k} - \tilde{E}_k) - (r_k - \beta_{1,k})\tilde{Z}_1^2}{(n-1)\left(1 - \frac{\tilde{Z}_1^2}{n-1}\right)}$$

so that $SIC_{1,k} = (n-1)d_{1,k}$ is as stated in the theorem.

Q.E.D.

Theorem 3.6 gives exact expressions for the ACF's measures of influence. As shown in the following Corollary 3.7, keeping only the term of highest order relative to the sample size n greatly simplifies the expressions for $EIC_{i,k}$, $EIC_k^{(1)}$ and $SIC_{i,k}$, and will facilitate the comparison with previous work. The expression for $SIC_{i,k}$ in Corollary 3.7 will also serve in Chapter IV to obtain its expectation and covariances, since the approach followed to prove Propositions 3.4 and 3.5 cannot be easily carried out in this case.

Corollary 3.7: For the sample ACF r_k , where $(\tilde{Z}_t)_{t \in \mathbb{Z}}$ and $(\tilde{Y}_t)_{t \in \mathbb{Z}}$ are as defined in Theorem 3.6,

$$EIC_{i,k} = \frac{\tilde{Z}_i \tilde{Z}_{i+k} + \tilde{Z}_i \tilde{Z}_{i-k} - 2r_k \tilde{Z}_i^2}{(n-1)^2} + o_p(n^{-3})$$

$$EIC_k^{(1)} = \frac{\tilde{Y}_i \tilde{Y}_{i+k} + \tilde{Y}_i \tilde{Y}_{i-k}}{n^2} + o_p(n^{-4})$$

and

$$SIC_{i,k} = \frac{(\tilde{Z}_i \tilde{Z}_{i+k} + \tilde{Z}_i \tilde{Z}_{i-k} - r_k \tilde{Z}_i^2)}{1 - \frac{\tilde{Z}_i^2}{n-1}} + o_p(1).$$

The results of Theorem 3.6 and of Corollary 3.7 are essentially new. Lattin(1983) suggested using $SIC_{i,k}$ as a measure of influence for the sample ACF. However, he was using $r_k' = (n-k)^{-1}nr_k$, which is not the commonly used

estimator of ρ_k , as the resulting sample ACF is not positive-definite. This may lead to trouble wherever the sample ACF is needed, as positive-definiteness is a crucial property of a covariance-correlation matrix. Furthermore, the relationships between r'_k and $r'_k(i)$ stated in the paper are in error. Two other shortcomings limit the applicability of Lattin's(1983) study. We shall describe them in Chapter IV.

Polasek(1982) computed the local resistance of r_k , as defined by Mosteller and Tukey(1977):

$$\frac{\partial r_k}{\partial X_i} = n^{-1}(\tilde{Z}_{i+k} + \tilde{Z}_{i-k} - 2\tilde{Z}_i r_k - \tilde{E}_k).$$

Here again, the formulae given by Polasek(1982) are in error. Nonetheless, the relationship between $\frac{\partial r_k}{\partial X_i}$ and $EIC_{i,k}$ is immediately obtained from Corollary 3.7:

$$EIC_{i,k} = \frac{n\tilde{Z}_i}{(n-1)^2} \frac{\partial r_k}{\partial X_i} + o_p(n^{-3}).$$

It is thus seen that the local resistance of r_k is simply a component of the empirical added influence. Hence, it only measures a fraction of the change induced in r_k by smoothly adding the i -th observation.

A third suggestion for detecting influential observations is that of Chernick, Downing and Pike(1982). They used the influence function for the correlation coefficient in a

2

MicroD
DIGITAL DATA STORAGE

bivariate normal distribution to obtain the following measure of influence for the k -th coefficient of the ACF:

$$C_{i,k}^+ = \bar{z}_i \bar{z}_{i+k} - \frac{1}{2}(\bar{z}_i^2 + \bar{z}_{i+k}^2) r_k.$$

The major shortcoming of $C_{i,k}^+$ as a measure of influence is that it does not consider the interaction of z_i with z_{i-k} in the computation of r_k . It could be extended to consider this interaction by defining $C_{i,k}^-$ as:

$$C_{i,k}^- = \bar{z}_i \bar{z}_{i-k} - \frac{1}{2}(\bar{z}_i^2 + \bar{z}_{i-k}^2) r_k.$$

However, from Corollary 3.7 we have the following relationship:

$$EIC_{i,k} = \frac{C_{i,k}^+ + C_{i,k}^- + \frac{1}{2}(\bar{z}_{i+k}^2 + \bar{z}_{i-k}^2 - 2\bar{z}_i^2) r_k}{(n-1)^2} + o_p(n^{-3})$$

which shows that $C_{i,k}^+$ and $C_{i,k}^-$ would need to be combined in a non-trivial manner to simply measure the full empirical added influence of the i -th observation.

Another related work is that of Li and Hui(1987) which has already been referred to in section 3.3. They made the hypothesis that the observed time series was a realization of an AR(p) process to derive a measure of influence. The structure of the resulting measure is identical to that of $C_{i,k}^+$, since both measures are identical under certain conditions.

Now, from our review of measures of influence in the standard linear model and from the definition of the three measures of influence in Theorem 3.6, $SIC_{i,k}$ is the most appropriate summary of $\frac{\partial r_k}{\partial \omega_i}$ among the three.

Furthermore, noting that

$$\dot{z}_t = \bar{z}_t \left(1 - \frac{\bar{z}_t^2}{n-1} \right)^{-1/2} = \frac{n^{1/2}(X_t - \bar{X})}{(n-1)^{1/2} C_0(i)^{1/2}}$$

is a multiple of the externally standardized observed time series, and that $\dot{z}_{i+k} + \dot{z}_{i-k} - r_k \dot{z}_i$ is a smoothed version of the observed process, $SIC_{i,k}$ can be written as:

$$SIC_{i,k} = \dot{z}_i (\dot{z}_{i+k} + \dot{z}_{i-k} - r_k \dot{z}_i) + o_p(1).$$

$SIC_{i,k}$ may thus be interpreted as the product of a measure of 'leverage' $(\dot{z}_{i+k} + \dot{z}_{i-k} - r_k \dot{z}_i)$ with a measure of 'outlyingness' $(\dot{z}_i)$. It is thus seen that the structure of $SIC_{i,k}$ is similar to that of Cook's distance, discussed in section 2.4.3.

To use $SIC_{i,k}$ in practice, two problems remain to be resolved. The first is to obtain critical values for declaring an observation's influence discordant and the second is to obtain an appropriate summary for the influence of an observation on the full ACF with its critical values. These two problems are the subject of Chapter IV.

3.3.4 Proofs of Propositions 3.4 and 3.5

The proofs of various propositions in this section require the next two lemmas. The first, Lemma 3.8, basically states that the sample mean of $(X_t)_{t=1}^{n-k}$ and that of $(X_t)_{t=k+1}^n$ can be replaced by the mean of the whole realization. The second, Lemma 3.9, gives some properties of a realization after it has been corrected for its mean: $(Z_t = X_t - \bar{X})_{t=1}^n$. The resulting process $(Z_t)_{t=1}^n$ is not weakly stationary but for large samples it can be considered as such.

Lemma 3.8: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1. For fixed $k < n$, define

$$\bar{X}_k^+ = n^{-1} \sum_{t=1}^{n-k} X_t \quad \text{and} \quad \bar{X}_k^- = n^{-1} \sum_{t=k+1}^n X_t$$

Then

$$E[\bar{X}_k^+ - \bar{X}] = -\frac{k\mu}{n},$$

$$\text{Var}[\bar{X}_k^+ - \bar{X}] = n^{-2} \sum_{s=1}^k \sum_{t=1}^k \gamma_{t-s} = O(n^{-2})$$

and

$$\text{Cov}[\bar{X}, \bar{X} - \bar{X}_k^+] = n^{-2} \sum_{t=1}^k \sum_{s=1}^n \gamma_{t-s} = f_{n-k} = O(n^{-2}).$$

Proof: We have

$$E[\bar{X}_k^* - \bar{X}] = \frac{1}{n} \left(\sum_{t=1}^{n-k} \mu - \sum_{t=1}^n \mu \right) = -\frac{k\mu}{n}$$

which is clearly $O(n^{-1})$. The expectation $E[\bar{X}_k^* - \bar{X}]$ is obtained similarly.

Now, to derive $\text{Var}[\bar{X}_k^* - \bar{X}]$ assume that $\mu = 0$ without loss of generality. Then

$$\text{Var}[\bar{X}_k^* - \bar{X}] = E[(\bar{X}_k^* - \bar{X})^2] = n^{-2} \sum_{s=n-k+1}^n \sum_{t=n-k+1}^n \gamma_{t-s}.$$

But

$$\sum_{t=n-k+1}^n \sum_{s=n-k+1}^n \gamma_{t-s} = \sum_{t=n-k+1}^n \sum_{s=1}^k \gamma_{t-s-n+k} = \sum_{t=1}^k \sum_{s=1}^k \gamma_{t-s}$$

and is independent of n . Thus

$$\text{Var}[\bar{X}_k^* - \bar{X}] = n^{-2} \sum_{t=1}^k \sum_{s=1}^k \gamma_{t-s} = O(n^{-2}).$$

Now,

$$\text{Cov}[\bar{X}, \bar{X} - \bar{X}_k^*] = E[\bar{X}(\bar{X} - \bar{X}_k^*)] = n^{-2} \left(\sum_{t=1}^n \sum_{s=1}^n \gamma_{t-s} - \sum_{t=k+1}^n \sum_{s=1}^n \gamma_{t-s} \right).$$

By hypothesis, $\mathcal{M} = \sum_{k \in \mathbb{Z}} |\gamma_k| < \infty$. Then,

$$|\text{Cov}[\bar{X}, \bar{X} - \bar{X}_k^*]| \leq n^{-2} \sum_{t=1}^k \sum_{s=1}^n |\gamma_{t-s}| \leq \frac{k\mathcal{M}}{n^2} = O(n^{-2}).$$

To show that $\text{Cov}[\bar{X}, \bar{X} - \bar{X}_k^*] = \text{Cov}[\bar{X}, \bar{X} - \bar{X}_k]$, we only have to show that

$$\sum_{t=n-k+1}^n \sum_{s=1}^n \gamma_{t-s} = \sum_{t=1}^k \sum_{s=1}^n \gamma_{t-s}.$$

But for $t < n$,

$$\sum_{s=1}^n \gamma_{t-s} = \sum_{p=0}^{t-1} \gamma_p + \sum_{p=1}^{n-1} \gamma_p,$$

while for $t = n$

$$\sum_{s=1}^n \gamma_{n-s} = \sum_{p=0}^{n-1} \gamma_p.$$

Hence,

$$\sum_{t=n-k+1}^n \sum_{s=1}^n \gamma_{t-s} = \sum_{t=n-k+1}^n \sum_{p=0}^{t-1} \gamma_p + \sum_{t=n-k+1}^{n-1} \sum_{p=1}^{n-t} \gamma_p.$$

Now, for $t = n-k+1(1)n$, $t-1 = n-k(1)n-1$

and for $t = n-k+1(1)n-1$, $n-t = 1(1)k-1$, so that

$$\begin{aligned} \sum_{t=n-k+1}^n \sum_{s=1}^n \gamma_{t-s} &= \sum_{t=1}^k \sum_{p=0}^{n-t} \gamma_p + \sum_{t=1}^{k-1} \sum_{p=1}^t \gamma_p \\ &= \sum_{t=1}^{k-1} \left(\sum_{p=0}^{t-1} \gamma_p + \sum_{p=1}^{n-t} \gamma_p \right) + \sum_{t=1}^{k-1} \gamma_p + \sum_{p=0}^{n-k} \gamma_p \\ &= \sum_{t=1}^k \sum_{s=1}^n \gamma_{t-s}. \end{aligned}$$

Q.E.D.

Lemma 3.9: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1, with

$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m}$ where the ϵ_t 's are *iid*(0, σ^2) random

variables with $E[\epsilon_t^4] = 3k\sigma^4$. For $t=1(1)n$, define

$Z_t = X_t - \bar{X}$. Then $E[Z_t] = 0$,

$$E[Z_t Z_{t+k}] = \gamma_k + f_0 - M_t - M_{t+k}$$

and

$$E[Z_t Z_{t+u} Z_{t+v} Z_{t+w}] = \begin{cases} \gamma_{w-v} E[Z_t Z_{t+u}] + \gamma_{w-u} E[Z_t Z_{t+v}] + \gamma_{v-u} E[Z_t Z_{t+w}] \\ -\gamma_w (M_{t+v} + M_{t+u} - f_0) - \gamma_v (M_{t+w} + M_{t+u} - f_0) - \gamma_u (M_{t+w} + M_{t+v} - f_0) \\ + \frac{3(k-1)}{n} K(t, u, v, w) + O(n^{-2}) \end{cases}$$

where

$$M_t = n^{-1} \sum_{s=1}^n \gamma_{s-t} = O(n^{-1}),$$

$$K(t, u, v, w) = \sum_{s=1}^n \left\{ \begin{aligned} &\mathcal{K}(v-u, w-u, s-t-u) + \mathcal{K}(v, w, s-t) \\ &+ \mathcal{K}(u, w, s-t) + \mathcal{K}(u, v, s-t) \end{aligned} \right\}$$

with

$$\mathcal{K}(u, v, w) = \sum_{p=0}^{\infty} \theta_p \theta_{p+|u|} \theta_{p+|v|} \theta_{p+|w|}.$$

Proof: The expected value and covariances of Z_t are straightforward to obtain. Note however that since we assume that

$$\mathcal{M} = \sum_{k \in \mathbb{Z}} |\gamma_k| < \infty, \text{ then}$$

$$|M_t| \leq n^{-1} \sum_{s=1}^n |\gamma_{s-t}| \leq \frac{\mathcal{M}}{n} = O(n^{-1}).$$

Without loss of generality, assume that $\mu = E[X_t] = 0$. Under the hypothesis that $(X_t)_{t \in \mathbb{Z}}$ is a weakly stationary and regular process, there exists a white noise process $(\epsilon_t)_{t \in \mathbb{Z}}$ such that $X_t = \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m}$ with $E[\epsilon_t^2] = \sigma^2$. Assuming that $E[\epsilon_t^4] = 3k\sigma^4 < \infty$, we obtain

$$E[\epsilon_t \epsilon_u \epsilon_v \epsilon_w] = \begin{cases} 3k\sigma^4 & \text{when } t = u = v = w, \\ \sigma^4 & \text{when the subscripts are equal} \\ & \text{in pairs, but not all equal,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus we have

$$E[X_t X_{t+u} X_{t+v} X_{t+w}] = 3(k-1)\sigma^4 \mathcal{K}(u, v, w) + \gamma_u \gamma_{v-w} + \gamma_v \gamma_{w-u} + \gamma_w \gamma_{v-u}$$

where

$$\mathcal{K}(u, v, w) = \sum_{p=0}^{\infty} \theta_p \theta_{p+|u|} \theta_{p+|v|} \theta_{p+|w|}.$$

But

$$\begin{aligned} E[\bar{X} X_{t+u} X_{t+v} X_{t+w}] &= \gamma_{w-v} M_{t+u} + \gamma_{w-u} M_{t+v} + \gamma_{v-u} M_{t+w} \\ &+ \frac{3(k-1)\sigma^4}{n} \sum_{s=1}^n \mathcal{K}(v-u, w-u, s-t-u), \end{aligned}$$

$$\begin{aligned} E[\bar{X}^2 X_{t+u} X_{t+v}] &= \gamma_{v-u} f_0 + 2M_{t+u} M_{t+v} \\ &+ \frac{3(k-1)\sigma^4}{n^2} \sum_{r,s=1}^n \mathcal{K}(v-u, s-t-u, r-t-u), \end{aligned}$$

$$E[\bar{X}^3 X_{t+u}] = 3f_0 M_{t+u} + \frac{3(k-1)\sigma^4}{n^3} \sum_{q,r,s=1}^n \mathcal{K}(s-t-u, r-t-u, q-t-u)$$

and

$$E[\bar{X}^4] = 3f_0^2 + \frac{3(k-1)\sigma^4}{n^4} \sum_{p,q,r,s=1}^n \mathcal{K}(s-p, r-p, q-p).$$

However, we saw that f_0 and M_{t+k} were of order $O(n^{-1})$. Furthermore, since we assume that $\mathcal{M} < \infty$, it can be seen from the definition of $\mathcal{K}(u, v, w)$ that

$$E[\bar{X}^2 X_{t+u} X_{t+v}] = \gamma_{v-u} f_0 + O(n^{-2}),$$

and that

$$E[\bar{X}^3 X_{t+u}] = E[\bar{X}^4] = O(n^{-2}).$$

Thus, $E[Z_t Z_{t+u} Z_{t+v} Z_{t+w}] = E[(X_t - \bar{X})(X_{t+u} - \bar{X})(X_{t+v} - \bar{X})(X_{t+w} - \bar{X})]$ is as given in the statement of the lemma.

Q.E.D.

To study the large sample statistical properties of $c_k^{(i)}$ as functions of those of c_k , the terms $E_k^{(i)}$ and E_k appearing in relations (3.4) and (3.6) can be taken as zero as Proposition 3.10 below shows.

Proposition 3.10: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1, and

let $Z_t = X_t - \bar{X}$. Also, define

$$E_k^{(i)} = \frac{1}{n} \sum_{\substack{t=1 \\ t \neq i}}^{n-k} (X_t - \bar{X}^{(i)}) + \frac{1}{n} \sum_{\substack{t=1 \\ t \neq i-k}}^{n-k} (X_{t+k} - \bar{X}^{(i)})$$

and

$$E_k = \frac{1}{n} \sum_{t=1}^{n-k} (Z_t + Z_{t+k}) .$$

Then

$$E[E_k] = E[E_k^{(i)}] = 0, \quad \text{Var}[E_k] = \text{Var}[E_k^{(i)}] = O(n^{-2})$$

and

$$\text{Cov}[Z_i, E_k] = -\frac{1}{n} \left(\sum_{t=1}^k \gamma_{t-i} + \sum_{t=n-k+1}^n \gamma_{t-i} \right) + O(n^{-2}) .$$

Proof: We first show that $E[E_k]=0$ and $\text{Var}[E_k]=O(n^{-2})$. Now,

$$E[E_k] = \frac{1}{n} \sum_{t=1}^{n-k} E[X_t - \bar{X}] + \frac{1}{n} \sum_{t=1}^{n-k} E[X_{t+k} - \bar{X}] = 0$$

and note that

$$E_k = \bar{X}_k^* - \left(1 - \frac{k}{n}\right) \bar{X} + \bar{X}_k^- - \left(1 - \frac{k}{n}\right) \bar{X} = (\bar{X}_k^* - \bar{X}) + (\bar{X}_k^- - \bar{X}) + \frac{2k}{n} \bar{X}$$

where $\bar{X}_k^*$ and $\bar{X}_k^-$ were defined in Lemma 3.8. Hence,

$$\text{Var}[E_k] = \begin{cases} \text{Var}[\bar{X}_k^* - \bar{X}] + \text{Var}[\bar{X}_k^- - \bar{X}] + 2\text{Cov}[\bar{X}_k^* - \bar{X}, \bar{X}_k^- - \bar{X}] \\ + \frac{4k^2}{n^2} \text{Var}[\bar{X}] + \frac{4k}{n} \text{Cov}[\bar{X}, \bar{X}_k^* - \bar{X}] + \frac{4k}{n} \text{Cov}[\bar{X}, \bar{X}_k^- - \bar{X}] . \end{cases}$$

Now from Lemma 3.8, all the terms appearing in $\text{Var}[E_k]$ are of order $O(n^{-2})$ except $\text{Var}[\bar{X}]$ which is of order $O(n^{-1})$.

But the coefficient of $\text{Var}[\bar{X}]$ is $\frac{4k^2}{n^2}$. Therefore

$$\text{Var}[E_k] = O(n^{-2}).$$

Next, we have

$$\text{Cov}[Z_i, E_k] = E[Z_i E_k] = \frac{1}{n} \sum_{t=1}^{n-k} (E[Z_i Z_t] + E[Z_i Z_{t+k}]).$$

But from Lemma 3.9, $E[Z_i Z_t] = \gamma_{t-i} + f_0 - M_i - M_t$ where

$$M_t = \frac{1}{n} \sum_{s=1}^n \gamma_{s-t}.$$

Hence,

$$\text{Cov}[Z_i, E_k] = \begin{cases} \frac{1}{n} \sum_{t=1}^{n-k} (\gamma_{t-i} + \gamma_{t+k-i}) + \frac{2(n-k)}{n} f_0 \\ - \frac{2(n-k)}{n} M_i - \frac{1}{n} \sum_{t=1}^{n-k} M_t - \frac{1}{n} \sum_{t=1}^{n-k} M_{t+k} \end{cases}$$

Now, in Lemma 3.8 we saw that

$$\frac{1}{n} \sum_{t=1}^{n-k} M_t = \frac{1}{n} \sum_{t=1}^{n-k} M_{t+k} = f_0 - f_{n-k}.$$

Hence,

$$\text{Cov}[Z_i, E_k] = - \frac{1}{n} \sum_{t=1}^k \gamma_{t-i} - \frac{1}{n} \sum_{t=n-k+1}^n \gamma_{t-i} - \frac{2kf_0}{n} + \frac{2k}{n} M_i + 2f_{n-k}.$$

Therefore,

$$\text{Cov}[Z_i, E_k] = - \frac{1}{n} \left(\sum_{t=1}^k \gamma_{t-i} + \sum_{t=n-k+1}^n \gamma_{t-i} \right) + O(n^{-2}).$$

Finally, to obtain the expectation and the variance of $E_k(i)$, we note that

$$E_k(i) = \begin{cases} E_k + \frac{(n-2k)}{n(n-1)} (X_i - \bar{X}) & \text{for } 1 \leq i \leq k \text{ or } n-k < i \leq n, \\ E_k - \frac{2k(X_i - \bar{X})}{n(n-1)} & \text{for } k < i \leq n-k. \end{cases}$$

Thus, $E[E_k(i)] = 0$ and $\text{Var}[E_k(i)] = O(n^{-2})$ from Lemma 3.8 and the value of the k -th autocovariance of $Z_t = X_t - \bar{X}$.

Q.E.D.

We can now prove Proposition 3.4, giving the expected value and the covariances of $(c_k(i))_{k=1}^n$ up to terms of order $O(n^{-2})$.

Proposition 3.4: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1 with

$$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m} \quad \text{where the } \epsilon_t \text{'s are } iid(0, \sigma^2) \text{ random}$$

variables with $E[\epsilon_t^4] = 3k\sigma^4 < \infty$. Then for i, j, k and l fixed, we have

$$E[c_k(i)] = \begin{cases} \left(\frac{n(n-2)}{(n-1)^2} - \frac{k}{n-1} \right) \gamma_k - \frac{n^2}{(n-1)^2} f_0 + O(n^{-2}) \\ \quad \text{for } k=0, \text{ or } 1 \leq i \leq k, \text{ or } n-k < i \leq n \\ \\ \left(\frac{n(n-3)}{(n-1)^2} - \frac{k}{n-1} \right) \gamma_k - \frac{n(n+1)}{(n-1)^2} f_0 + O(n^{-2}) \\ \quad \text{for } k < i \leq n-k, \end{cases}$$

and

$$\text{Cov}[c_k^{(i)}, c_l^{(j)}] = \frac{n^2}{(n-1)^2} \text{Cov}[c_k, c_l] + O(n^{-2})$$

while

$$\text{Cov}[c_k^{(i)}, c_l] = \frac{n}{n-1} \text{Cov}[c_k, c_l] + O(n^{-2}).$$

Proof: For each case, determined by the position of the removed observation relative to the end of the series, the expectation of $c_k^{(i)}$ is straightforward to obtain using relations (3.5) and (3.6), Lemma 3.9, and Propositions 3.2 and 3.10.

To obtain the covariances in terms of $\text{Cov}[c_k, c_l]$, there are only six cases to consider by symmetry. As all cases are proved similarly, we shall only give the details for the case $k < i \leq n-k$ with $1 < j \leq n-1$.

Using Lemma 3.9, Proposition 3.2 and since

$$\left| \frac{1}{n} \sum_{s=1}^{n-1} M_s y_{w-s} \right| \leq \frac{\mathcal{M}^2}{n^2} = O(n^{-2})$$

and

$$\frac{1}{n} \sum_{s=1}^{n-1} M_s = f_0 + O(n^{-2}),$$

we get

$$E[Z_i Z_{i+k} c_l] = E[c_l] E[Z_i Z_{i+k}] + \kappa_1 + O(n^{-1}),$$

where κ_1 is a term due to the fourth-order cumulant. But

since we assume that $\mathcal{M} < \infty$ we have $\sum_{p=0}^{\infty} |\theta_p| < \infty$, so that

it can be shown that $\kappa_1 = O(n^{-2})$. Thus

$$E[Z_i Z_{i+k} c_l] = E[c_l] E[Z_i Z_{i+k}] + O(n^{-1}).$$

Also,

$$E[Z_i^2 c_l] = \frac{(n-1)}{n} \gamma_0 \gamma_l + O(n^{-1})$$

and

$$E[Z_i Z_{i+k} Z_j Z_{j+l}] = E[Z_i^2 Z_j Z_{j+l}] = E[Z_i^2 Z_j^2] = O(1).$$

Now from relations (3.5) and (3.6), simplified using Proposition 3.10, and using the above five expectations, we get, keeping only terms of order $O(n^{-1})$ and higher:

$$\text{Cov}[c_k^{(i)}, c_l^{(j)}] = \frac{n^2}{(n-1)^2} \text{Cov}[c_k, c_l] + O(n^{-2}).$$

The covariances between $c_k^{(i)}$ and c_l are obtained similarly.

Q.E.D.

The following Theorem 3.11 will be used in the proof of Proposition 3.5 to obtain the expected value and the covariances of the sample ACF with the i -th observation missing. Theorem 3.11 is a particular case of Theorem 5.4.3 of Fuller (1976, pp. 205-209) and we simply state it for the record.

Theorem 3.11 (Fuller 1976): Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of

k -dimensional random variables, and let g be a mapping from $\mathbb{R}^k$ into $\mathbb{R}$. Assume that for some $N \in \mathbb{N} - \{0\}$ and $\xi \in \mathbb{R}^k$:

$$(1) \quad E[\|X_n - \xi\|^N] = a_n^N \quad \text{where } a_n \rightarrow 0 \text{ as } n \rightarrow \infty \text{ and}$$

$\|\cdot\|$ denotes the Euclidian norm;

$$(2) \quad \text{For all } x \in \mathbb{R}^k, \quad |g(x)| < \infty;$$

$$(3) \quad \text{That } g^{(i_1, i_2, \dots, i_N)} = \frac{\partial^N g}{\partial x_{i_1} \dots \partial x_{i_N}} \text{ is continuous and bounded}$$

over a closed and bounded sphere S with $\xi \in S$;

$$(4) \quad \text{And that } |g^{(i_1, i_2, \dots, i_j)}(\xi)| < \infty \text{ for } j=1(1)N-1.$$

Then

$$E[g(X_n)] = g(\xi) + \sum_{j=1}^{N-1} \frac{E[H^{(j)}(0)]}{j!} + o(a_n^N)$$

where

$$H^{(j)}(0) = \left. \frac{d^j}{d\alpha^j} g(\xi + \alpha(X - \xi)) \right|_{\alpha=0}.$$

Note that for $N=1$ the conclusion of the theorem is simply

$$E[g(X_n)] = g(\xi) + o(a_n).$$

We are now ready to prove Proposition 3.5, giving the expected value and the covariances of $(r_k^{(i)})_{i=1}^L$ up to terms of order $O(n^{-2})$. It is worth remarking that the approximations given for the expected value and the covariances of $(r_k^{(i)})_{k=1}^L$ in Proposition 3.5 cannot be rigorously obtained

with the usual Taylor approximation used, for example, for the ratio estimator in sampling theory. As shown by David and Sukhatme(1974), the appropriateness of the resulting expressions can be justified by assuming that the random denominator is bounded from below by a non-zero constant. This assumption, however, is not tenable in the time series context.

Proposition 3.5: Let $(X_t)_{t=1}^n$ be as in Proposition 3.1, with

$$X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m} \quad \text{where the } \epsilon_t \text{'s are } iid(0, \sigma^2) \text{ random}$$

variables with $E[\epsilon_t^4] = 3k\sigma^4 < \infty$ and $E[\epsilon_t^6] < \infty$.

Then for i, j, k and l fixed,

$$E[r_k^{(i)}] = \begin{cases} \left(1 - \frac{k}{n-1}\right)\rho_k - \frac{n^2(1-\rho_k)f_0}{(n-1)^2\gamma_0} - \frac{n^2\text{Cov}[A_k, C_0]}{(n-1)^2\gamma_0} + o(n^{-2}) \\ \quad \text{when } 1 \leq i \leq k \text{ or } n-k < i \leq n \\ \\ \left(1 - \frac{k+1}{n-1}\right)\rho_k - \frac{n^2(1-\rho_k)f_0}{(n-1)^2\gamma_0} - \frac{n^2\text{Cov}[A_k, C_0]}{(n-1)^2\gamma_0} + o(n^{-2}) \\ \quad \text{for } k < i \leq n-k, \end{cases}$$

where

$$A_k = \frac{C_k - \rho_k C_0}{\gamma_0}.$$

Also,

$$\text{Cov}[r_k^{(i)}, r_l^{(j)}] = \frac{n^2}{(n-1)^2} \text{Cov}[r_k, r_l] + o(n^{-2})$$

and

$$\text{Cov}[r_k^{(i)}, r_l] = \frac{n}{n-1} \text{Cov}[r_k, r_l] + o(n^{-2}).$$

Proof: We first note that whenever $c_0 = 0$, i.e. when $X_t = x$ a constant for $t=1(1)n$, r_k is not defined. For a regular process however, the event $c_0 = 0$ has zero probability. Hence it does not affect the evaluation of $E[r_k]$ and of $\text{Cov}[r_k, r_l]$. Similarly, the event $c_{0(i)} = 0$ has zero probability of occurring. Hence it will not affect the evaluation of the moments of $r_{k(i)}$. We thus assume that $c_{0(i)}$ and c_0 are strictly greater than zero in the remainder.

To obtain the expected value of $r_{k(i)}$, we use Theorem 3.11 with $g(c_{k(i)}, c_{0(i)}) = c_{k(i)}/c_{0(i)}$, $N=3$ and $\xi = (\gamma_k, \gamma_0)$. We note that

$$|g(c_{k(i)}, c_{0(i)})| \leq 1,$$

and that

$$\frac{\partial g}{\partial c_{k(i)}}(\xi) = \gamma_0^{-1}, \quad \frac{\partial g}{\partial c_{0(i)}}(\xi) = -\frac{\rho_k}{\gamma_0}, \quad \frac{\partial^2 g}{\partial c_{k(i)}^2}(\xi) = 0,$$

$$\frac{\partial^2 g}{\partial c_{0(i)}^2}(\xi) = 2\frac{\rho_k}{\gamma_0^2} \quad \text{and} \quad \frac{\partial^2 g}{\partial c_{k(i)} \partial c_{0(i)}}(\xi) = -\gamma_0^{-2}$$

are all bounded. Also,

$$\frac{\partial^3 g}{\partial c_{k(i)}^3} = \frac{\partial^3 g}{\partial c_{k(i)}^2 \partial c_{0(i)}} = 0,$$

$$\frac{\partial^3 g}{\partial c_{k(i)} \partial c_{0(i)}^2} = 2c_{0(i)}^{-3} \quad \text{and} \quad \frac{\partial^3 g}{\partial c_{0(i)}^3} = -6 \frac{c_{k(i)}}{c_{0(i)}^4}$$

are all continuous and bounded over any closed and bounded sphere S such that $\xi \in \hat{S}$. Hence, for this particular choice of g , N and ξ , all the hypothesis of Theorem 3.11 are satisfied.

We also note that since we assume that $E[\epsilon_t^4] < \infty$, then for $q=0(1)3$, $E[(c_{k(i)} - \gamma_k)^q (c_{0(i)} - \gamma_0)^{3-q}] = O(n^{-3})$.

Now, there are three cases to consider depending on the position of the missing observation relative to the ends of the series. We shall only give the details for $k < i \leq n-k$ (and $1 < j \leq n-1$ when applicable) as they are all proved similarly. From Theorem 3.11 we get

$$E[r_k(i)] = \rho_k + \left(2 - \frac{E[c_0(i)]}{\gamma_0} \right) E[A_k(i)] - \frac{\text{Cov}[A_k(i), c_0(i)]}{\gamma_0} + O(n^{-3}).$$

From Proposition 3.4 we have

$$2 - \frac{E[c_0(i)]}{\gamma_0} = 1 + \frac{n^2 f_0}{(n-1)^2} + O(n^{-2}),$$

$$E[A_k(i)] = - \left(\frac{n}{(n-1)^2} + \frac{k}{n-1} \right) \rho_k - \frac{n^2 (1-\rho_k) f_0}{(n-1)^2 \gamma_0} + O(n^{-2})$$

and

$$\text{Cov}[A_k(i), c_0(i)] = \frac{n^2}{(n-1)^2} \text{Cov}[A_k, c_0] + O(n^{-2}).$$

We thus get the desired result.

To obtain the covariance of $r_k(i)$ with $r_l(j)$, we again use Theorem 3.11, with $h(c_k(i), c_0(i), c_l(j), c_0(j)) = \frac{c_k(i)c_l(j)}{c_0(i)c_0(j)}$, $N=3$ and $\xi = (\gamma_k, \gamma_0, \gamma_l, \gamma_0)$. Again, all the hypotheses are satisfied and we obtain

$$\text{Cov}[r_k(i), r_l(j)] = \begin{cases} \text{Cov}[A_k(i), A_l(j)] - E[A_k(i)]E[A_l(j)] \\ - (E[r_k(i)] - \rho_k)(E[r_l(j)] - \rho_l) \end{cases} + O(n^{-3})$$

which reduces to the stated result by keeping only terms of order $O(n^{-1})$ and higher. The covariance between $r_k(i)$ and r_l is obtained similarly.

Q.E.D.

CHAPTER IV

IDENTIFYING INFLUENTIAL OBSERVATIONS IN TIME SERIES

We will now discuss the identification of overly influential observations for the autocorrelation function (ACF), using the measure of sample influence introduced in section 3.3.3.

In practice, the influence of the i -th observation can be measured both on the k -th sample autocorrelation r_k , and on the whole sample ACF $(r_k)_{k=1}^L$, where L is fixed but large enough to capture most of the behaviour of $(\rho_k)_{k \in \mathbb{Z}}$, while small enough for r_L to provide an adequate estimate of ρ_L .

Letting $\mathcal{I}_{i,k}$ denote any measure of the influence of the i -th observation on the k -th sample autocorrelation, and letting ℓ_k denote the normal amount of influence that any observation has on r_k , we are interested in testing for overly influential observations in the following four contexts:

$$H_1: \mathcal{I}_{i,k} = \ell_k$$

against $H'_1: \mathcal{I}_{i,k} \neq \ell_k$, i and k fixed,

i.e. the i -th observation has a normal influence on r_k against the possibility that its influence is discordant;

$$H_2: \mathcal{J}_{i,k} = \ell_k \text{ for } k=1(1)L \text{ and } i \text{ fixed}$$

$$\text{against } H'_2: \mathcal{J}_{i,k} \neq \ell_k \text{ for some } k=1(1)L,$$

i.e. the i -th observation has a normal influence on the whole sample ACF against the possibility that its influence is discordant;

$$H_3: \mathcal{J}_{i,k} = \ell_k \text{ for } i=1(1)n \text{ and } k \text{ fixed}$$

$$\text{against } H'_3: \mathcal{J}_{i,k} \neq \ell_k \text{ for at most one } i=1(1)n,$$

i.e. no unique observation has a discordant influence on r_k ;

$$H_4: \mathcal{J}_{i,k} = \ell_k \text{ for } k=1(1)L \text{ and } i=1(1)n$$

$$\text{against } H'_4: \mathcal{J}_{i,k} \neq \ell_k \text{ for some } k \text{ and for at most one } i,$$

i.e. no unique observation has a discordant influence on the whole sample ACF.

Since we have already chosen a particular measure of influence, the sample measure of influence $SIC_{i,k}$, and since the univariate time series theory entertained herein fully specifies a model for the data, we are thus looking for discordant observations in a particularly well-defined context.

In Chapter III, section 3.3.3, we reviewed previous work on measures of influence for the ACF, in particular the work of Chernick, Downing and Pike(1982), Lattin(1983), Li and Hui

(1987) and of Polasek(1982). Of these, only Lattin(1983) suggested using the sample measure of influence. However, apart from Chernick et al.(1982), none of the others suggested a means for obtaining thresholds, or critical values, for testing the hypotheses H_i against H'_i , for $i=1,2,3$ and 4, described above.

Chernick et al.(1982) did suggest using 1.0 as a critical value for testing H_i against H'_i using their measure of influence $C_{i,k}^+$, but this choice is quite arbitrary as it does not correspond to the distribution theory of $C_{i,k}^+$, and they did not explain how this value was chosen.

Thus in this chapter, we will derive the distributional theory required for obtaining thresholds for declaring a time series observation discordant in its influence, using $SIC_{i,k}$.

Another limitation of previous work on measures of influence for the ACF is in the choice of reduction applied to the vector $(\mathcal{I}_{i,k})_{k=1}^L$ to measure the influence of the i -th observation of the whole ACF, for testing the hypothesis H_2 against H'_2 . Only one reduction was suggested, which leads to an overly complicated distributional theory, and which mixes the influence of the i -th observation with that of others. Therefore, we suggest a different reduction whose distributional properties are easily obtained, and which properly measures the influence of the i -th observation on the

whole sample ACF. We will also see that this same reduction can be applied for testing H_1 against H'_1 , thus facilitating the computation of critical values for this test.

The finite-sample statistical properties of $SIC_{i,k}$ being unmanageable in practice, in section 4.1 we obtain its asymptotic equivalent, denoted $Sic_{i,k}$, whose properties are more readily derivable.

Note that since $SIC_{i,k}$ and $Sic_{i,k}$ are asymptotically equivalent, all of our discussion on the distributional properties of $Sic_{i,k}$ will apply to the asymptotic distribution of $SIC_{i,k}$. In the following sections we restrict our study to Gaussian time series.

In section 4.2, we develop asymptotic critical values for testing H_1 against H'_1 using $SIC_{i,k}$. Also in section 4.2, we develop an appropriate transformation of $SIC_{i,k}$ for testing the influence of an observation on the whole sample ACF $(r_k)_{k=1}^L$, and obtain its asymptotic distribution. Hence, we obtain asymptotic critical values for testing H_2 against H'_2 .

Finally, in section 4.3 we obtain bounds on the asymptotic critical values for testing H_3 against H'_3 and H_4 against H'_4 using the transformed sample measure of influence.

4.1 The Asymptotic Measure of Sample Influence

From Theorem 3.6 and Corollary 3.7, it can be seen that the measure of sample influence $SIC_{i,k}$ depends on the full sample k -th autocorrelation coefficient r_k , as well as being a ratio of dependent random variables:

$$SIC_{i,k} = \frac{(\tilde{Z}_i \tilde{Z}_{i+k} + \tilde{Z}_i \tilde{Z}_{i-k} - r_k \tilde{Z}_i^2)}{1 - \frac{\tilde{Z}_i^2}{n-1}} + o_p(1)$$

where

$$\tilde{Z}_t = (X_t - \bar{X})/c_0^{1/2}.$$

The density of r_k for a Gaussian time series is known, but is quite complicated as it possesses k discontinuities; see Pan(1964). Hence, the density of $SIC_{i,k}$ can be expected to be quite complex. We thus have to use asymptotic theory to reduce the complexity.

The following Theorem 4.1 gives the asymptotic equivalent $Sic_{i,k}$ of $SIC_{i,k}$ as well as its two first moments. Notice that it is not a statistic as it depends on the unknown theoretical ACF. However, since it is asymptotically equivalent to $SIC_{i,k}$, all of our discussion on the distributional properties of $Sic_{i,k}$ will apply to the asymptotic distribution of $SIC_{i,k}$.

Theorem 4.1: Let $(X_t)_{t=1}^n$ be a finite realization of a weakly stationary and regular process with autocovariance function $(\gamma_k)_{k \in \mathbb{Z}}$, such that $X_t = \mu + \sum_{m=0}^{\infty} \theta_m \epsilon_{t-m}$ where the ϵ_t 's are $iid(0, \sigma^2)$ random variables with $E[\epsilon_t^4] = 3k\sigma^4$ and $E[\epsilon_t^6] < \infty$. Then

$$Sic_{i,k} = \lim_{n \rightarrow \infty} SIC_{i,k} = \tilde{X}_i (\tilde{X}_{i+k} + \tilde{X}_{i-k} - \rho_k \tilde{X}_i)$$

where

$$\tilde{X}_t = \begin{cases} (X_t - \mu) / \gamma_0^{1/2} & \text{for } t=1(1)n \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore,

$$E[Sic_{i,k}] = \begin{cases} \rho_k & \text{for } i=k+1(1)n-k \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\text{Var}[Sic_{i,k}] = \begin{cases} 1 - \rho_k^2 + 3(k-1)\sigma^4\tau_1/\gamma_0^2 & \text{for } i=1(1)k \\ 2(1 - \rho_k^2 + \rho_{2k}) + 6(k-1)\sigma^4\tau_2/\gamma_0^2 & \text{for } i=k+1(1)n-k \\ 1 - \rho_k^2 + 3(k-1)\sigma^4\tau_3/\gamma_0^2 & \text{for } i=n-k+1(1)n \end{cases}$$

where

$$\tau_1 = \mathcal{H}(0, k, k) - 2\rho_k \mathcal{H}(0, 0, k) + \rho_k^2 \mathcal{H}(0, 0, 0)$$

$$\tau_2 = \begin{cases} \mathcal{H}(0, k, k) + \mathcal{H}(k, k, 2k) - \rho_k \mathcal{H}(0, 0, k) \\ - \rho_k \mathcal{H}(k, k, k) + \rho_k^2 \mathcal{H}(0, 0, 0) / 2 \end{cases}$$

$$\tau_3 = \mathcal{H}(0, k, k) - 2\rho_k \mathcal{H}(k, k, k) + \rho_k^2 \mathcal{H}(0, 0, 0)$$

and

$$\mathcal{H}(u, v, w) = \sum_{p=0}^{\infty} \theta_p \theta_{p+u} \theta_{p+v} \theta_{p+w}.$$

Proof: From Corollary 3.7 we already know that

$$SIC_{i,k} = \dot{Z}_i (\dot{Z}_{i+k} + \dot{Z}_{i-k} - r_k \dot{Z}_i) + o_p(1)$$

where

$$\dot{Z}_t = \begin{cases} \left(\frac{n}{n-1}\right)^{1/2} \frac{(X_t - \bar{X})}{C_0(i)^{1/2}} & t=1(1)n \\ 0 & \text{otherwise.} \end{cases}$$

Now, under the stated hypothesis we have, using Propositions 3.1 to 3.5,

$$\text{plim}_{n \rightarrow \infty} \bar{X} = \mu, \quad \text{plim}_{n \rightarrow \infty} r_k = \rho_k \quad \text{and} \quad \text{plim}_{n \rightarrow \infty} C_0(i) = \gamma_0.$$

Thus, $\text{plim}_{n \rightarrow \infty} \dot{Z}_t = \tilde{X}_t$, so that $\text{plim}_{n \rightarrow \infty} SIC_{i,k} = Sic_{i,k}$.

To obtain the expectation and variance of $Sic_{i,k}$, we simply use the fact that, under the conditions of the theorem, we have $E[\tilde{X}_t] = 0$, $E[\tilde{X}_t \tilde{X}_{t+u}] = \rho_u$, and

$$E[\tilde{X}_t \tilde{X}_{t+u} \tilde{X}_{t+v} \tilde{X}_{t+w}] = \rho_u \rho_{v-w} + \rho_v \rho_{w-u} + \rho_w \rho_{v-u} + 3(k-1) \sigma^4 \mathcal{H}(u, v, w) / \gamma_0^2.$$

Q.E.D.

We note that, although not given in Theorem 4.1, the covariances $\text{Cov}[Sic_{i,k}, Sic_{j,l}]$ are also straightforward to obtain.

Clearly, the expression for the variance of $Sic_{i,k}$ is simplified when normality is assumed. This is stated as Corollary 4.2 below.

Corollary 4.2: Let $(X_t)_{t=1}^n$ be as in Theorem 4.1 and assume further that it is normally distributed. Then $k = 1$,

$$E[Sic_{i,k}] = \begin{cases} \rho_k & \text{for } i=k+1(1)n-k \\ 0 & \text{otherwise} \end{cases}$$

and

$$\text{Var}[Sic_{i,k}] = \begin{cases} 2(1 - \rho_k^2 + \rho_{2k}) & \text{for } i=k+1(1)n-k \\ 1 - \rho_k^2 & \text{otherwise.} \end{cases}$$

To obtain critical values for identifying overly influential observations, some distributional assumption has to be made. We will assume that the time series is Gaussian. We will also assume that the observed time series $(X_t)_{t=1}^n$ has been centered and standardized by its true mean and variance, so that $E[X_t] = 0$, $\text{Var}[X_t] = 1$, $\text{Cov}[X_t, X_{t+u}] = \rho_u$ and $E[X_t^4] = 3$. This amounts to supposing that in practice the length of the observed time series, n , is large enough for $\bar{X}$ and c_0 to have converged in probability to μ and σ^2 respectively.

We will now obtain critical values for testing H_j against H'_j , $j=1,2,3$ and 4, in the following sections.

4.2 Testing the Influence of a Specific Observation

For testing H_1 against H'_1 , i.e. that the influence of the i -th observation on the k -th autocorrelation is not discordant, assuming normality, the density of the asymptotic sample measure of influence $Sic_{i,k}$ is given in Theorem 4.3 below. The density of $Sic_{i,k}$ is that of the product of two Gaussian random variables. The density of the product of two Gaussian random variables has been studied as such by Craig (1936) and by Aroian, Taneja and Cornwell (1978) among others. A similar distributional theory also appears in the work of Abraham and Yatawara (1988) and of Chang (1982) on outliers in ARMA models.

Theorem 4.3: Let $(X_t)_{t=1}^n$ be a strictly stationary Gaussian process with $E[X_t] = 0$, $\text{Var}[X_t] = 1$ and $\text{Cov}[X_t, X_{t+u}] = \rho_u$. Then the density of $Sic_{i,k}$ is

$$f(s; \alpha, \sigma^2) = e^{\alpha s / \sigma^2} K_0(|s| / \sigma^2) / \pi \sigma \quad (4.1)$$

where $K_\nu(x)$ is the modified Bessel function of the second kind of order ν :

$$K_\nu(x) = \int_0^\infty e^{-x \cosh(t)} \cosh(\nu t) dt \quad \text{for } x > 0,$$

and where

$$\alpha = \begin{cases} \rho_k(\sigma^2 + \rho_k^2)^{-1/2} & \text{for } i=k+1(1)n-k \\ 0 & \text{otherwise} \end{cases}$$

with

$$\sigma^2 = \begin{cases} 2(1 + \rho_{2k} - 2\rho_k^2) & \text{for } i=k+1(1)n-k \\ 1 - \rho_k^2 & \text{otherwise.} \end{cases}$$

Proof: Since we assume normality, $X_i \sim \mathcal{N}(0,1)$ and

$$X_{i+k} + X_{i-k} - \rho_k X_i \sim \begin{cases} \mathcal{N}(0, \sigma^2 + \rho_k^2) & \text{for } i=k+1(1)n-k \\ \mathcal{N}(0, \sigma^2) & \text{otherwise.} \end{cases}$$

Then, since $Sic_{i,k} = X_i(X_{i+k} + X_{i-k} - \rho_k X_i)$ with

$$\text{Cov}[X_i, X_{i+k} + X_{i-k} - \rho_k X_i] = \begin{cases} \rho_k & \text{for } i=k+1(1)n-k \\ 0 & \text{otherwise,} \end{cases}$$

the result follows by applying Craig's (1936) result on the product of two correlated normal random variables.

Q.E.D.

The density (4.1) of $Sic_{i,k}$ given in Theorem 4.3 has a singularity at zero and is asymmetric unless $\alpha = 0$.

Since $SIC_{i,k}$ converges in probability to $Sic_{i,k}$ as the length of the observed time series tends to infinity, then $SIC_{i,k}$ is asymptotically distributed as $Sic_{i,k}$.

Thus, asymptotic critical values for testing H_1 against H'_1 using $SIC_{i,k}$ can be obtained. In fact, asymptotic critical values for testing one-sided hypotheses, such as $H_1: SIC_{i,k} = \rho_k$ against $H'_1: SIC_{i,k} < \rho_k$, i.e. that the influence of the i -th observation on the k -th autocorrelation coefficient is smaller than normal, can also be obtained. But such hypotheses will rarely be of particular interest compared to two-sided ones.

Notice that even if $Sic_{i,k}$ is standardized, its distribution depends on the autocorrelation function. This hinders the tabulation of critical values for testing H_1 against H'_1 .

However, the $SIC_{i,k}$'s need to be transformed for testing H_2 against H'_2 , i.e. that the i -th observation has a normal amount of influence on the whole sample ACF. When this transformation is used for testing H_1 against H'_1 , it yields a statistic whose asymptotic distribution is independent of the ACF. Thus, asymptotic critical values for both tests can be readily tabulated.

We will first obtain the appropriate reduction for the asymptotic measure $Sic_{i,k}$. To begin, we need some notation. For fixed L and n , we define $h(i) = \min\{i, n-i, L\}$ and

$$J_{h(i)} = \text{diag}(1'_{h(i)}, 0_{1 \times (L-h(i))}).$$

Further, let

$$R'_{k,l} = (\rho_k, \rho_{k+1}, \dots, \rho_l),$$

$$T_{k,l} = \begin{pmatrix} \rho_k & \rho_{k+1} & \rho_{k+2} & \cdots & \rho_l \\ \rho_{k+1} & \rho_k & \rho_{k+1} & \cdots & \rho_{l-1} \\ \rho_{k+2} & \rho_{k+1} & \rho_k & \cdots & \rho_{l-2} \\ \vdots & & & \ddots & \vdots \\ \rho_l & \rho_{l-1} & \rho_{l-2} & \cdots & \rho_k \end{pmatrix},$$

and

$$H_{k,l} = \begin{pmatrix} \rho_k & \rho_{k+1} & \rho_{k+2} & \cdots & \rho_{\frac{l}{2}} \\ \rho_{k+1} & \rho_{k+2} & \rho_{k+3} & \cdots & \rho_{\frac{l}{2}+1} \\ \rho_{k+2} & \rho_{k+3} & \rho_{k+4} & \cdots & \rho_{\frac{l}{2}+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{\frac{l}{2}} & \rho_{\frac{l}{2}+1} & \rho_{\frac{l}{2}+2} & \cdots & \rho_l \end{pmatrix}.$$

Here, $T_{k,l}$ is a Toeplitz matrix while $H_{k,l}$ is a Hankel matrix. Both matrices are symmetric. Finally, we also define

$$\tilde{T}_{k,l} = T_{k,l} - \rho_k R_{1,l-k+1} R'_{1,l-k+1}$$

and

$$\tilde{H}_{k,l} = H_{k,l} - \rho_{k-2} R_{1,\frac{l}{2}-k+1} R'_{1,\frac{l}{2}-k+1}.$$

Now, using the asymptotic measure of sample influence $Sic_{i,k}$ the hypothesis H_2 can be written as

$$H_2: Sic_i = \begin{pmatrix} Sic_{i,1} \\ Sic_{i,2} \\ \vdots \\ Sic_{i,L} \end{pmatrix} = J_{h(i)} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_L \end{pmatrix} = J_{h(i)} R_{1,L}.$$

Since we wish to test the whole vector Sic_i , for discordancy, we need to reduce its dimensionality in a meaningful way. The standard Euclidian norm is not useful here as each component of Sic_i , has a different variance; the

resulting random variable would have an unnecessarily complicated distribution.

To simplify the exposition, we will temporarily assume that $L < i \leq n-L$.

The random vector Sic_i is the product of two terms:

$$Sic_i = X_i Y_i = X_i \begin{pmatrix} X_{i+1} + X_{i-1} - \rho_1 X_i \\ X_{i+2} + X_{i-2} - \rho_2 X_i \\ \vdots \\ X_{i+L} + X_{i-L} - \rho_L X_i \end{pmatrix}$$

Any reduction of the dimensionality of Sic_i will come from a reduction of that of Y_i .

Under the hypothesis that the observed time series is Gaussian, the joint distribution of

$$(X_{i+1}, X_{i+2}, \dots, X_{i+L}, X_{i-1}, X_{i-2}, \dots, X_{i-L}, X_i)'$$

is $\mathcal{N}(0, W)$, where

$$W = \begin{pmatrix} T_{0,L-1} & H_{2,2L} & R_{1,L} \\ H_{2,2L} & T_{0,L-1} & R_{1,L} \\ R'_{1,L} & R'_{1,L} & 1 \end{pmatrix}$$

and where the matrices T and H and the vector R are as defined earlier.

A well-known result from multivariate normal theory is that for

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right),$$

the conditional distribution of Z_1 , given Z_2 is

$$Z_1|Z_2 \sim \mathcal{N}(A_{12}A_{22}^{-1}Z_2, A_{11}-A_{12}A_{22}^{-1}A_{21}).$$

Thus, the distribution of $Y_i|X_i$ is $\mathcal{N}(\mu_i, \Sigma_i)$, where

$$\mu_i = R_{1,L}X_i$$

and

$$\Sigma_i = 2(\tilde{T}_{0,L-1} + \tilde{H}_{2,2L}) = 2(T_{0,L-1} + H_{2,2L} - 2R_{1,L}R_{1,L}').$$

Note here that we attach the subscript i to Σ_i to emphasize that it varies for $i=1(1)n$, even though it is constant for $i=L+1(1)n-L$.

Hence, if we define

$$q_i = (Y_i - \mu_i)' \Sigma_i^{-1} (Y_i - \mu_i)$$

then, for fixed X_i , q_i has a Chi-square distribution with L degrees of freedom; i.e. $q_i|X_i \sim \chi_L^2$.

Now, by defining Q_i as

$$\begin{aligned} Q_i &= (Sic_i - X_i\mu_i)' \Sigma_i^{-1} (Sic_i - X_i\mu_i) \\ &= X_i^2 (Y_i - \mu_i)' \Sigma_i^{-1} (Y_i - \mu_i) \\ &= X_i^2 q_i \end{aligned}$$

we shall see that it forms an appropriate reduction of Sic_i , via the following Theorem 4.4 giving its unconditional distribution.

Theorem 4.4: Let $(X_t)_{t=1}^n$ be as in Theorem 4.3. For L fixed with respect to n , and for $n \geq 2L+1$, $i=1(1)n$ let

$$\mu_i = J_{h(i)} R_{1,L} X_i$$

and

$$\Sigma_i = \tilde{T}_{0,L-1} + J_{h(i)} \tilde{T}_{0,L-1} J_{h(i)} + \tilde{H}_{2,2L} J_{h(i)} + J_{h(i)} \tilde{H}_{2,2L}.$$

Then the density of

$$Q_i = (Sic_i - X_i \mu_i)' \Sigma_i^{-1} (Sic_i - X_i \mu_i)$$

is

$$f(q;L) = \frac{q^{\frac{L-3}{2}} K_{\frac{L-1}{2}}(q^{1/2})}{2^{\frac{L-1}{2}} \pi^{1/2} \Gamma(\frac{L}{2})} \quad (4.2)$$

where $K_\nu(x)$ is the modified Bessel function of the second kind of order ν :

$$K_\nu(x) = \int_0^\infty e^{-x \cosh(t)} \cosh(\nu t) dt \quad \text{for } x > 0.$$

Furthermore,

$$E[Q_i^t] = \frac{4^t \Gamma(\frac{L}{2}+t) \Gamma(\frac{1}{2}+t)}{\Gamma(\frac{L}{2}) \Gamma(\frac{1}{2})},$$

so that $E[Q_i] = L$ and $\text{Var}[Q_i] = 2L(L+3)$.

Proof: As was shown prior to the theorem, $Q_i = X_i^2 q_i$ was constructed such that the conditional distribution of q_i given X_i is χ_L^2 . It was also constructed for $i=L+1(1)n-L$. In the statement of the theorem, the appropriate μ_i and Σ_i are given for each $i=1(1)n$, as can be readily verified.

We thus have that $q_i | X_i \sim \chi^2_{L_i}$ and we know that $X_i^2 \sim \chi^2_{L_i}$. Hence for X_i fixed, $q_i | X_i$ has a Gamma $\mathcal{G}(\frac{L_i}{2}, 2)$ distribution, where the density of $\mathcal{G}(\alpha, \beta)$ is, for $x > 0$,

$$\frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}.$$

Now, $Q_i = X_i^2 q_i$. Hence, $Q_i | X_i \sim \mathcal{G}(\frac{L_i}{2}, 2X_i^2)$. But $X_i^2 \sim \mathcal{G}(\frac{L_i}{2}, 2)$. Therefore the density of the unconditional distribution of Q_i is, using the total probability theorem:

$$\begin{aligned} f(q; L) &= \int_0^\infty \frac{x^{L/2-1} e^{-x/2}}{2^{L/2} \Gamma(\frac{L}{2})} \frac{q^{\frac{L}{2}-1} e^{-q/2x}}{(2x)^{L/2} \Gamma(\frac{L}{2})} dx \\ &= \frac{q^{\frac{L}{2}-1}}{2^{\frac{L+1}{2}} \pi^{1/2} \Gamma(\frac{L}{2})} \int_0^\infty x^{-(L+1)/2} e^{-(x + \frac{q}{x})/2} dx. \end{aligned}$$

From Gradshteyn and Ryzhik(1980, p. 959), formula 8.432(7) we have

$$\int_0^\infty x^{-(L+1)/2} e^{-(x + \frac{q}{x})/2} dx = 2K_{\frac{L-1}{2}}(q^{1/2})/q^{\frac{L-1}{4}}.$$

Thus the result. To obtain the uncentered moments of Q_i , we use (4.2) directly:

$$E[Q_i^t] = \frac{1}{2^{\frac{L+1}{2}} \pi^{1/2} \Gamma(\frac{L}{2})} \int_0^\infty q^{\frac{L-3}{4} + t} K_{\frac{L-1}{2}}(q^{1/2}) dq.$$

Letting $\nu = \frac{L-1}{2}$ and making the change of variable $x^2 = q$, $E[Q_i^t]$ becomes

$$E[Q_i^t] = \frac{1}{2^{\nu+1} \pi^{1/2} \Gamma(\frac{L}{2})} \int_0^\infty x^{\nu+2t} K_\nu(x) dx.$$

From Gradshteyn and Ryzhik(1980, p.684), formula 6.651.16, we have

$$\int_0^{\infty} x^{\nu+2t} K_{\nu}(x) dx = 2^{\nu+2t-1} \Gamma(\nu+t+\frac{1}{2}) \Gamma(t+\frac{1}{2}).$$

Therefore,

$$E[Q_i^t] = \frac{4^t \Gamma(\frac{1}{2}+t) \Gamma(\frac{1}{2}+t)}{\Gamma(\frac{1}{2}) \Gamma(\frac{1}{2})}.$$

Q.E.D.

Note that the distribution of the product of two independent Gamma random variables has been derived before by Miller(1965) and Bhattacharya(1966) among others. Results on a related class of distributions were reviewed by Barndorff-Nielsen, Kent and Sørensen(1982). Although mentioned somewhat often in theoretical work, these random variables do not appear to have been used to any extent in practice. As a matter of notation, we shall designate these distributions by $\mathcal{W}_l$. I.e. $X \sim \mathcal{W}_l$ if and only if the density of X is (4.2).

However, the suggested use of Q_i as reduction for the asymptotic sample measure of influence of the i -th observation on the whole sample ACF, as well as the derivation of its distribution in the above Theorem 4.4 are new.

Like $Sic_{i,k}$, Q_i is not a statistic as it depends on the unknown true ACF of the observed time series. In practice,

$Sic_{i,k}$ will be replaced by $SIC_{i,k}$, while μ_i and Σ_i , entering into the definition of Q_i , must be estimated.

Since μ_i and Σ_i only depend on the true ACF $(\rho_k)_{k=1}^L$, we can either choose r_k or $r_k^{(i)}$ as estimator of ρ_k . Recall that in our review of discordant observations in the standard linear model in Chapter II, it was noted that many discussions revolved around the choice between an externally standardized statistic and its internally standardized version. There we saw that the final choice could be based on the distributional properties of the statistics.

For the sample measure of influence $SIC_{i,k}$ this choice is straightforward to make as it can be shown, in a manner similar to that of Theorem 3.6 and Corollary 3.7, that:

$$\begin{aligned} SIC_{i,k} &= \frac{\tilde{Z}_i(\tilde{Z}_{i+k} + \tilde{Z}_{i-k} - r_k \tilde{Z}_i)}{[1 - \tilde{Z}_i^2/(n-1)]} + o_p(1) \\ &= \frac{\tilde{Y}_i(\tilde{Y}_{i+k} + \tilde{Y}_{i-k} - r_k^{(i)} \tilde{Y}_i)}{[1 + \tilde{Y}_i^2/n]} + o_p(1), \end{aligned}$$

where $\tilde{Z}_t = (X_t - \bar{X})/c_0^{1/2}$ and $\tilde{Y}_t = (X_t - \bar{X}^{(i)})/c_0^{(i)1/2}$.

It is thus much simpler to use $(r_k)_{k=1}^{2L}$ for estimating μ_i and Σ_i . Hence, if we replace ρ_k by r_k in the definitions of μ_i and Σ_i , we obtain the consistent estimators $\hat{\mu}_i$ and $\hat{\Sigma}_i$ of μ_i and Σ_i , respectively. However, if the true process is $MA(q)$ with $q > L$, or if it is an $ARMA(p,q)$ process with $p \geq 1$ and $q \geq 0$, the Hankel matrix $H_{2,2L}$ entering into the

computation of Σ_i cannot be estimated consistently if it is assumed that the autocorrelation coefficients of order higher than L are zero. To circumvent this difficulty, we need to ensure that $L < \frac{n}{2}$, so that the ACF can be estimated up to its $(2L)$ -th lag, even though the estimates may not be based on a sufficient number of observations to be considered acceptable when taken singly. Thus, if we always take L as fixed but less than $\frac{n}{2}$, then $\hat{\Sigma}_i \xrightarrow{P} \Sigma_i$. And we obtain the following Corollary 4.5 to Theorem 4.4.

Corollary 4.5: Let $(X_t)_{t=1}^n$ be as in Theorem 4.3. For L fixed with respect to n , with $L < \frac{n}{2}$, and for $n \geq 2L+1$, $i=1(1)n$, let $\hat{\mu}_i$ and $\hat{\Sigma}_i$ be as defined in Theorem 4.4 but computed from $(r_k)_{k=1}^{2L}$. Then the asymptotic distribution of Q_i is $\mathcal{W}_L$, where

$$Q_i = (\text{SIC}_{i.} - X_i \hat{\mu}_i)' \hat{\Sigma}_i^{-1} (\text{SIC}_{i.} - X_i \hat{\mu}_i).$$

The advantages of Q_i as a measure of influence can be seen as follows. First, by itself Corollary 4.5 provides an argument for choosing Q_i as an appropriate reduction of $\text{SIC}_{i.}$, since the asymptotic distribution of Q_i depends only on L , the length of the ACF on which the influence of the i -th observation is measured. The asymptotic critical values for

testing for a discordant influence can be readily tabulated; see Table 1. (The computational details for Table 1 are given in section 5.4 of Chapter V.)

Another advantage of this reduction is that it is straightforwardly extended to measure the influence on any subset of autocorrelation coefficients. In particular, for testing H_1 against H'_1 , i.e. that the i -th observation has a normal influence of the k -th autocorrelation coefficient, the same reduction yields the test statistic ${}_kQ_i$, equation (4.3) in Corollary 4.6 below, which is simply the square of the 'normalized' $SIC_{i,k}$. Compared with the asymptotic density (4.1) of $SIC_{i,k}$, shown in Theorem 4.3, the density of $\mathcal{W}_1$ is simpler and independent of the ACF.

Corollary 4.6: Let $(X_t)_{t=1}^n$ be as in Theorem 4.3. For k fixed and for $n \geq 2k+1$ and $i=1(1)n$, let

$${}_kQ_i = \frac{X_i^2 [\delta_{i+k}(X_{i+k} - r_k X_i) + \delta_{i-k}(X_{i-k} - r_k X_i)]^2}{1 - r_k^2 + \delta_{i+k}\delta_{i-k}(1 + 2r_{2k} - 3r_k^2)} \quad (4.3)$$

where

$$\delta_j = \begin{cases} 1 & \text{for } 1 \leq j \leq n \\ 0 & \text{otherwise.} \end{cases}$$

Then the asymptotic distribution of ${}_kQ_i$ is $\mathcal{W}_1$, i.e. the asymptotic density of ${}_kQ_i$ is simply

$$f(q) = K_0(q^{1/2})/\pi q^{1/2}.$$

TABLE 1
Percentage Points of the $\mathcal{W}_L$ Distribution

L	.9000	.9500	.9750	.9900
1	2.544	4.761	7.764	12.990
2	5.302	8.974	13.608	21.208
3	8.011	12.984	19.046	28.695
4	10.705	16.917	24.323	35.883
5	13.394	20.812	29.517	42.911
6	16.083	24.687	34.663	49.841
7	18.773	28.550	39.777	56.707
8	21.464	32.405	44.870	63.527
9	24.156	36.256	49.949	70.314
10	26.849	40.103	55.017	77.076
11	29.544	43.949	60.076	83.818
12	32.239	47.792	65.130	90.545
16	43.030	63.159	85.305	117.351
20	53.830	78.520	105.445	144.059
24	64.636	93.881	125.569	170.714
28	75.446	109.241	145.683	197.337
32	86.258	124.602	165.791	223.940
36	97.072	139.964	185.896	250.529
40	107.888	155.326	205.999	277.108
44	118.705	170.688	226.100	303.680
48	129.523	186.051	246.200	330.246

TABLE 1 (Cont'd)**Percentage Points of the $\mathcal{W}_L$ Distribution**

L	.9910	.9920	.9930	.9940	.9950
1	13.685	14.485	15.421	16.542	17.922
2	22.189	23.313	24.620	26.173	28.072
3	29.922	31.323	32.948	34.871	37.213
4	37.339	38.997	40.915	43.182	45.932
5	44.585	46.489	48.689	51.282	54.423
6	51.728	53.870	56.342	59.252	62.772
7	58.801	61.177	63.915	67.135	71.024
8	65.825	68.431	71.431	74.956	79.208
9	72.814	75.647	78.906	82.731	87.342
10	79.776	82.833	86.348	90.471	95.437
11	86.717	89.997	93.766	98.184	103.501
12	93.641	97.142	101.164	105.875	111.541
16	121.226	125.603	130.622	136.490	143.533
20	148.705	153.946	159.950	166.961	175.361
24	176.125	182.225	189.206	197.350	207.098
28	203.510	210.465	218.419	227.690	238.778
32	230.872	238.680	247.603	257.998	270.421
36	258.220	266.877	276.768	288.284	302.038
40	285.556	295.062	305.918	318.553	333.637
44	312.884	323.238	335.059	348.811	365.221
48	340.206	351.407	364.191	379.060	396.795

TABLE 1 (Cont'd)**Percentage Points of the $\mathcal{W}_L$ Distribution**

L	.9960	.9970	.9980	.9990
1	19.691	22.105	25.760	32.703
2	30.487	33.746	38.621	47.717
3	40.176	44.155	50.064	60.980
4	49.403	54.045	60.907	73.495
5	58.377	63.651	71.418	85.592
6	67.193	73.077	81.719	97.422
7	75.903	82.382	91.876	109.069
8	84.535	91.600	101.929	120.581
9	93.111	100.752	111.905	131.991
10	101.643	109.854	121.819	143.322
11	110.141	118.916	131.686	154.590
12	118.611	127.945	141.514	165.806
16	152.300	163.841	180.553	210.304
20	185.802	199.515	219.319	254.430
24	219.197	235.063	257.927	298.337
28	252.526	270.531	296.435	342.104
32	285.811	305.946	334.876	385.775
36	319.065	341.325	373.269	429.377
40	352.298	376.676	411.628	472.929
44	385.514	412.007	449.960	516.443
48	418.717	447.322	488.272	559.926

TABLE 1 (Cont'd)**Percentage Points of the $\mathcal{W}_L$ Distribution**

L	.9991	.9992	.9993	.9994	.9995
1	33.836	35.126	36.620	38.385	40.529
2	49.184	50.850	52.772	55.035	57.774
3	62.728	64.711	66.995	69.678	72.917
4	75.502	77.776	80.391	83.459	87.156
5	87.845	90.394	93.322	96.754	100.885
6	99.911	102.725	105.956	109.739	114.286
7	111.788	114.859	118.384	122.507	127.459
8	123.525	126.849	130.660	135.117	140.465
9	135.156	138.728	142.822	147.606	153.343
10	146.705	150.522	154.894	160.000	166.121
11	158.188	162.247	166.894	172.319	178.819
12	169.617	173.915	178.835	184.576	191.450
16	214.954	220.193	226.182	233.161	241.508
20	259.903	266.063	273.100	281.294	291.081
24	304.623	311.693	319.765	329.156	340.365
28	349.195	357.168	366.266	376.845	389.462
32	393.667	402.537	412.654	424.413	438.429
36	438.067	447.830	458.962	471.894	487.303
40	482.414	493.067	505.210	519.312	536.108
44	526.720	538.261	551.412	566.681	584.860
48	570.994	583.421	597.579	614.011	633.570

TABLE 1 (Cont'd)**Percentage Points of the $\mathcal{W}_1$ Distribution**

L	.9996	.9997	.9998	.9999
1	43.238	46.866	52.241	62.142
2	61.216	65.800	72.543	84.830
3	76.976	82.365	90.256	104.547
4	91.780	97.905	106.846	122.965
5	106.044	112.863	122.796	140.640
6	119.959	127.447	138.331	157.831
7	133.631	141.768	153.576	174.680
8	147.125	155.895	168.606	191.276
9	160.482	169.875	183.471	207.676
10	173.732	183.738	198.206	223.921
11	186.896	197.506	212.835	240.041
12	199.989	211.198	227.377	256.056
16	251.856	265.412	284.924	319.368
20	303.200	319.051	341.820	381.888
24	354.231	372.344	398.321	443.925
28	405.057	425.410	454.562	505.641
32	455.743	478.320	510.623	567.131
36	506.326	531.115	566.552	628.457
40	556.834	583.826	622.383	689.658
44	607.283	636.471	678.137	750.763
48	657.686	689.065	733.831	811.791

The most sensitive aspect of the reduction used in Theorem 4.4 and Corollaries 4.5 and 4.6 is the centering by the conditional mean of $Sic_{i,}$, given X_i . For instance, for $L < i \leq n-L$,

$$Sic_{i,} - X_i \mu_i = X_i (X_{i+k} + X_{i-k} - 2\rho_k X_i)_{k=1}^L = \lim_{n \rightarrow \infty} n^2 (EIC_{i,k})_{k=1}^L$$

Hence, $Sic_{i,} - X_i \mu_i$ is equivalent to the asymptotic measure of empirical added influence. But this equivalence only occurs asymptotically, and if $Sic_{i,}$ is not centered, the resulting random variable Q_i would have an unnecessarily complicated non-central distribution.

Another important aspect of the suggested reduction is that it is quite different from the reduction suggested by Chernick et al. (1982), Lattin (1983), Li and Hui (1987) and by Polasek (1982) for their measures of influence which were described in section 3.3.3 of Chapter III.

In these papers, the reduction suggested is obtained by taking the sum of the squares of the elements of a 'clothes-pin' in the table $(SIC_{i,k})_{i,k}^n$. See figure 1. This leads to a more difficult distributional theory than that suggested herein. Moreover, this reduction mixes the influence theoretically attributable to the i -th observation on r_k , with that of the $(i-k)$ -th observation, thus hindering the identification of overly influential observations.

<u>Obs. #</u>	<u>Lags</u>								
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>...</u>	<u>L</u>
.				.					
.				.					
.				.					
i-4	.	.	.	+	.	.	.	...	.
i-3	.	.	+	.	.	.	.	...	.
i-2	.	+	.	.	.	.	.	...	.
i-1	+	.	.	.	.	.	.	...	.
i	+	+	+	+	+	+	+	...	+
i+1	.	.	.	.	.	.	.	...	.
i+2	.	.	.	.	.	.	.	...	.
.				.					
.				.					
.				.					

Figure 1. The Elements (+) of a 'Clothes-Pin'

4.3 Testing the Most Influential Observation

After having computed the influence of each observation on each autocorrelation coefficient, i.e. ${}_kQ_i$ for $i=1(1)n$ and $k=1(1)L$, we will be interested in testing H_3 against H'_3 , i.e. that no single observation has a discordant influence on the k -th autocorrelation coefficient r_k . The largest influence on r_k , ${}_kQ_{(n)}$, will clearly be an appropriate test statistic for this situation.

Similarly, after having computed the influence of each observation on the whole sample ACF, i.e. Q_i for $i=1(1)n$, we will be interested in testing H_4 against H'_4 , i.e. that no single observation has a discordant influence on the whole sample ACF. The largest influence on $(r_k)_{k=1}^L$, $Q_{(n)}$, will clearly be an appropriate test statistic for this situation.

We note here that as for testing H_1 against H'_1 , i.e. testing the influence of a single observation on a single autocorrelation coefficient, we could use for testing H_3 against H'_3 :

$$|SIC|_{(n),k} = \max_{i=1(1)n} \{ |SIC_{i,k}| \}.$$

But using the transformation ${}_kQ_{(n)}$ of $|SIC|_{(n),k}$ has the advantage of leading to a single distributional theory for testing both H_3 against H'_3 and H_4 against H'_4 . Hence, we will

first discuss the testing of H_4 against H'_4 and then specialize our discussion to the testing of H_3 against H'_3 .

To obtain the exact critical values for $Q_{(n)}$ we need the joint distribution of $(Q_i)_{i=1}^n$. But this distribution is expected to be intractable. Furthermore, except for inequalities of the Bonferroni type, all the usual multiple comparison procedures require the distribution of the range $Q_{(n)} - Q_{(1)}$, or a transformation thereof, which also requires the joint distribution of $(Q_i)_{i=1}^n$; see Miller(1981).

Thus, the best that can be achieved at the present time is to obtain bounds on the critical values for $Q_{(n)}$ using inequalities of the Bonferroni type, and use these bounds as thresholds for testing the extremeness of $Q_{(n)}$.

The most useful Bonferroni inequalities were described in section 2.3 of Chapter II. Letting A_i denote the event $\{Q_i > c\}$, these inequalities are:

$$\sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) \leq P(Q_{(n)} > c) \leq \sum_{i=1}^n P(A_i) - \sum_{(i,j) \in T} P(A_i \cap A_j) \quad (4.4)$$

where T is the maximal spanning tree obtained from the graph whose vertices are the events $(A_i)_{i=1}^n$, which are connected whenever $P(A_i \cap A_j) > 0$.

In the context of a strictly stationary process, both inequalities in (4.4) simplify in an interesting manner. This simplification may not be achievable, however, if we only

assume weak stationarity. We note in passing that this simplification does not seem to be fully exploited in the current theory for extremes of stationary stochastic processes as expounded, for example, in the books of Galambos(1978) and of Leadbetter, Lindgren and Rootzén(1983).

If $(Q_i)_{i=1}^n$ is a realization of a strictly stationary time series, then $P(A_i \cap A_j)$ depends on i and j only through $|i-j|$:

$$P(A_i \cap A_j) = P(Q_i > c, Q_j > c) = p_{|i-j|}(c), \quad \text{say.}$$

Hence the lower bound in (4.4) becomes

$$LB_{n-1}(c) = n \left[p_0(c) - \sum_{k=1}^{n-1} \left(1 - \frac{k}{n} \right) p_k(c) \right]$$

where the subscript of LB indicates the largest lag used in its computation.

To simplify Hunter's upper bound in (4.4), let $p_{(k)}(c)$ denote the ordered probabilities $p_k(c)$ for $k=1(1)n-1$, i.e.

$$p_{(n-1)}(c) \geq p_{(n-2)}(c) \geq \dots \geq p_{(1)}(c),$$

and let σ denote the induced permutation of $\{1(1)n-1\}$, i.e. σ is such that $p_{(k)}(c) = p_{\sigma(k)}(c)$.

Then, using Kruskal's(1956) algorithm, the maximum spanning tree is found to be the graph formed by taking the m largest probabilities $p_j(c)$ such that

$$n-1 = \sum_{i=1}^m [n - \sigma(n-i)],$$

i.e. m is such that

$$1 = \frac{1}{n} + \sum_{i=1}^m \left[1 - \frac{\sigma(n-i)}{n} \right].$$

Thus, Hunter's upper bound in (4.4) becomes:

$$UB_{n-1}(c) = n \left[p_0(c) - \sum_{k=1}^m \left(1 - \frac{\sigma(n-k)}{n} \right) p_{(n-k)}(c) \right].$$

Therefore, for a strictly stationary process, the inequalities (4.4) reduce to:

$$LB_{n-1}(c) \leq P(Q_{(n)} > c) \leq UB_{n-1}(c). \quad (4.5)$$

There are however three difficulties in computing $UB_{n-1}(c)$ and $LB_{n-1}(c)$ in our particular context:

1. The stochastic process $(Q_i)_{i=1}^n$ is not necessarily strictly stationary.
2. We cannot compute $p_k(c)$ up to $k = n-1$ without an additional hypothesis.
3. There is no easily computable expression yet for $p_k(c)$.

These difficulties are not insurmountable, however. We attempted to resolve them as follows.

First, notice that the segment $(Q_i)_{i=L+1}^{n-L}$ is strictly stationary, as each Q_i in that segment is obtained from the same measurable transformation $F: \mathbb{R}^{2L+1} \rightarrow [0, \infty)$, of a finite segment of the strictly stationary time series $(X_t)_{t=1}^n$, i.e.

$$Q_i = F(X_{i+L}, X_{i+L-1}, \dots, X_{i+1}, X_i, X_{i-1}, \dots, X_{i-L+1}, X_{i-L}).$$

But for $i=1(1)L$ and $i=n-L+1(1)n$, the mapping changes from one observation to another: e.g., for Q_{L-2} , $F_{L-2}: \mathbb{R}^{2L-2} \rightarrow [0, \infty)$, while for Q_1 , $F_1: \mathbb{R}^{L+1} \rightarrow [0, \infty)$. Hence, any finite realization of $(Q_i)_{i=1}^n$ may not be stationary. From the expression for the covariances τ_k of the process $(Q_i)_{i=k+1}^{n-L}$ given in Theorem 4.7 below, it can be deduced that $\text{Cov}[Q_i, Q_{i+k}]$ may depend on i for $i=1(1)L$ or $i=n-L+1(1)n$. Thus the process $(Q_i)_{i=1}^n$ may not even be weakly stationary.

However, since the effect of the non-stationary terms appearing in the bounds (4.4) will certainly decrease as the length of the time series increases, the bounds in (4.5) should nonetheless be adequate.

Theorem 4.7: Let $(X_t)_{t=1}^n$ be as in Theorem 4.3. For L fixed with respect to n , and for $i=L+1(1)n-L$ and $k=0(1)n-2L+1$, let $\Sigma = 2(\tilde{T}_{0,L-1} + \tilde{H}_{2,2L})$ and let D be such that $DD' = \Sigma^{-1}$. Also let

$$V_k = R_{k+1,k+L} + R_{k-1,k-L} - 2\rho_k R_{1,L},$$

and

$$B_k = \tilde{T}_{k,k+L-1} + \tilde{T}_{k,k-L+1} + \tilde{H}_{k+2,k+2L} + \tilde{H}_{k-2,k-2L} - 2V_k R_{1,L}' - 2R_{1,L} V_k'.$$

Finally, let $W_k = D'V_k$ and $A_k = D'B_kD$. Then,

$$\lim_{n \rightarrow \infty} \text{Cov}[Q_i, Q_{i+k}] = \tau_k,$$

where

$$\tau_k = 2(1+2\rho_k^2) \text{tr}(A_k' A_k) + 16\rho_k W_k' A_k W_k + 2\rho_k^2 L^2 + 4L W_k' W_k + 4(W_k' W_k)^2.$$

Proof: Let $Z'_{i.} = Y'_{i.} - \mu'_{i.}$. We have

$$(Q_i, Q_{i+k}) = (X_i^2 q_i, X_{i+k}^2 q_{i+k}),$$

where

$$q_i = \begin{pmatrix} Z_{i.} \\ Z_{i+k.} \end{pmatrix}' \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Z_{i.} \\ Z_{i+k.} \end{pmatrix}$$

and

$$q_{i+k} = \begin{pmatrix} Z_{i.} \\ Z_{i+k.} \end{pmatrix}' \begin{pmatrix} 0 & 0 \\ 0 & \Sigma^{-1} \end{pmatrix} \begin{pmatrix} Z_{i.} \\ Z_{i+k.} \end{pmatrix}.$$

Now, as is readily verified, the joint distribution of $(Z'_{i.}, Z'_{i+k.}, X_i, X_{i+k})$ is $\mathcal{N}(0, \Omega)$, where

$$\Omega = \begin{pmatrix} \Sigma & B_k & 0 & V_k \\ B'_k & \Sigma & V_k & 0 \\ 0 & V'_k & 1 & \rho_k \\ V'_k & 0 & \rho_k & 1 \end{pmatrix}.$$

Defining now $\tilde{Z}_{i.} = D'Z_{i.}$ and $\tilde{Z}_{i+k.} = D'Z_{i+k.}$ where $DD' = \Sigma^{-1}$, the above reduce to $q_i = \tilde{Z}'_{i.} \tilde{Z}_{i.}$ and $q_{i+k} = \tilde{Z}'_{i+k.} \tilde{Z}_{i+k.}$, where the joint distribution of $(\tilde{Z}'_{i.}, \tilde{Z}'_{i+k.}, X_i, X_{i+k})$ is $\mathcal{N}(0, \tilde{\Omega})$, with

$$\tilde{\Omega} = \begin{pmatrix} I_L & A_k & 0 & W_k \\ A'_k & I_L & W_k & 0 \\ 0 & W'_k & 1 & \rho_k \\ W'_k & 0 & \rho_k & 1 \end{pmatrix}.$$

Then, for X_i and X_{i+k} fixed, the joint conditional distribution of $(\tilde{Z}'_i, \tilde{Z}'_{i+k})'$ is:

$$\mathcal{N} \left[\begin{pmatrix} (X_{i+k} - \rho_k X_i) \frac{W_k}{1-\rho_k^2} \\ (X_i - \rho_k X_{i+k}) \frac{W_k}{1-\rho_k^2} \end{pmatrix}, \begin{pmatrix} I_L - \frac{W_k W'_k}{1-\rho_k^2} & A_k + \frac{\rho_k}{1-\rho_k^2} W_k W'_k \\ A'_k + \frac{\rho_k}{1-\rho_k^2} W_k W'_k & I_L - \frac{W_k W'_k}{1-\rho_k^2} \end{pmatrix} \right].$$

We will temporarily drop the subscript k from the elements of $\tilde{\Omega}$ as k is fixed and no confusion can arise.

From multivariate normal theory, we know that for $\theta_1 = Z' \theta_1 Z$ and $\theta_2 = Z' \theta_2 Z$ with $Z \sim N(\xi, \Lambda)$ we have

$$E[\theta_j] = \text{tr}(\theta_j \Lambda) + \xi' \theta_j \xi \quad \text{for } j=1,2,$$

and

$$\text{Cov}[\theta_1, \theta_2] = 2 \text{tr}(\theta_1 \Lambda \theta_2 \Lambda) + 4 \xi' \theta_1 \Lambda \theta_2 \xi.$$

Now, $\tau_k = \tau_k^{(1)} + \tau_k^{(2)}$ where

$$\tau_k^{(1)} = \text{Cov}[E[Q_i | X_i, X_{i+k}], E[Q_{i+k} | X_i, X_{i+k}]]$$

and

$$\tau_k^{(2)} = E[\text{Cov}[Q_i, Q_{i+k} | X_i, X_{i+k}]].$$

We now evaluate the first term $\tau_k^{(1)}$.

$$\tau_k^{(1)} = \text{Cov}[X_i^2 E[q_i | X_i, X_{i+k}], X_{i+k}^2 E[q_{i+k} | X_i, X_{i+k}]].$$

Using the above properties of quadratic forms, we get

$$E[q_i | X_i, X_{i+k}] = L - \frac{W'W}{1-\rho^2} + \frac{(X_{i+k} - \rho X_i)^2 W'W}{(1-\rho^2)^2}$$

and

$$E[q_{i+k} | X_i, X_{i+k}] = L - \frac{W'W}{1-\rho^2} + \frac{(X_i - \rho X_{i+k})^2 W'W}{(1-\rho^2)^2}.$$

Since X_i and X_{i+k} are jointly normal with correlation ρ ,

$$E[X_i^2 X_{i+k}^2] = 1 + 2\rho^2,$$

$$E[X_i^2 X_{i+k}^2 (X_{i+k} - \rho X_i)^2] = E[X_i^2 X_{i+k}^2 (X_i - \rho X_{i+k})^2] = 3(1-\rho^2)$$

and

$$E[X_i^2 X_{i+k}^2 (X_{i+k} - \rho X_i)^2 (X_i - \rho X_{i+k})^2] = 3(1-\rho^2)^2 (3-8\rho^2+8\rho^4).$$

Gathering all these terms, simplifying and since $E[Q_i] = E[Q_{i+k}] = L$, we get the first term of τ_k as:

$$\tau_k^{(1)} = 2\rho^2 L^2 + 4LW'W + \frac{2(2-11\rho^2+12\rho^4)(W'W)^2}{(1-\rho^2)^2}.$$

We now evaluate the second term $\tau_k^{(2)}$.

$$\tau_k^{(2)} = E[X_i^2 X_{i+k}^2 \text{Cov}[q_i, q_{i+k} | X_i, X_{i+k}]].$$

But

$$E[X_i^2 X_{i+k}^2 (X_{i+k} - \rho X_i)(X_i - \rho X_{i+k})] = 3\rho(1-2\rho^2)(1-\rho^2),$$

and using the expression for the covariance of normal quadratic forms, we get the second term of τ_k as:

$$\tau_k^{(2)} = (1+2\rho^2)2tr(A'A) + 16\rho W'AW + \frac{2\rho^2(7-10\rho^2)(W'W)^2}{(1-\rho^2)^2}.$$

Adding $\tau_k^{(1)}$ to $\tau_k^{(2)}$ and simplifying we get the stated value of τ_k . Now, since $Q_i \xrightarrow{P} Q_i$, for $i=1(1)n$, we get the stated result.

Q.E.D.

The following Corollary 4.8 gives the asymptotic covariances of ${}_jQ_i$, the measure of influence of the i -th observation on the j -th autocorrelation coefficient.

Corollary 4.8: Let $(X_t)_{t=1}^n$ be as in Theorem 4.3. For fixed

j , and for $i=j+1(1)n-j$ and $k=0(1)n-2j+1$, let

$$\sigma_j^2 = 2(1 + \rho_{2j} - 2\rho_j^2),$$

$${}_jv_k = \rho_{k-j} + \rho_{k+j} - 2\rho_j\rho_k$$

and

$${}_jb_k = 2\rho_k(1 - 2\rho_j^2) + \rho_{k-2j} + \rho_{k+2j} - 4\rho_j{}_jv_k.$$

Then,

$$\lim_{n \rightarrow \infty} \text{Cov}[_jQ_i, {}_jQ_{i+k}] = {}_j\tau_k$$

where

$${}_j\tau_k = 2\rho_k^2 + \frac{4{}_jv_k^2}{\sigma_j^2} + \frac{2(1 + 2\rho_k^2){}_jb_k^2}{\sigma_j^2} + \frac{16\rho_k{}_jv_k^2{}_jb_k}{\sigma_j^4} + \frac{4{}_jv_k^4}{\sigma_j^4}.$$

The second difficulty in computing the bounds in (4.5) stems from the finite length, n , of the observed times series, which does not permit the computation of $p_k(c)$ for k up to $n-1$. Indeed, as can be seen from Theorem 4.7, an estimate of ρ_{k+2L} is required simply for computing the covariance τ_k .

A similar difficulty was encountered in obtaining a consistent estimate of Σ_i . There, it was suggested to ensure that $L < \frac{n}{2}$, so that the ACF could be estimated up to its

(2L)-th coefficient, thus preserving the consistency of the estimator of Σ_i . For the bounds in (4.5), consistency is of lesser importance.

Since L is usually chosen such that the sample ACF $(r_k)_{k=1}^L$ contains all the salient features of the true ACF, and since this is equivalent to supposing that the observed time series is a realization of an MA(L) process, the bounds in (4.5) can be computed with the hypothesis that the observed time series is an MA(M) process, where $M \geq L$.

If it is supposed that $(X_t)_{t=1}^n$ is an MA(M) process, then the asymptotic measures of influence Q_i and Q_{i+k} are independent whenever $k > M + 2L$, while the measures ${}_j Q_i$ and ${}_j Q_{i+k}$ become independent whenever $k > M + 2j$. These results are immediate upon inspecting the proof of Theorem 4.7.

Thus in practice, we would estimate the first $2L$ coefficients of the ACF and assume that the remaining ones are all zero. We would hence only need to compute the joint probabilities $p_k(c)$ for $k=1(1)4L$, and assume that the remaining ones are given by $p_k(c) = p_0(c)^2$ for $k=4L+1(1)n-1$.

The last difficulty is the most important one of the three. We do not yet have an easily computable expression for $p_k(c)$, $k=1(1)4L$. To see this, we can use the argument of Theorem 4.7 to obtain the following expression for $p_k(c)$:

$$p_k(c) = \int_{\mathbb{R}^2} \int_{c/x^2}^{\infty} \int_{c/y^2}^{\infty} \dot{p}(q_1, q_2 | x, y) \phi_{\rho_k}(x, y) dx dy dq_1 dq_2,$$

where $\dot{p}$ is the joint density of two dependent non-central quadratic forms in normal variables, and ϕ_ρ is the density of a standard bivariate normal random variable of correlation ρ .

Expressions for the density $\dot{p}$ are known; see e.g. Jensen (1970a,b) and Khatri, Khrishnaiah and Sen(1977). These expressions, however, are in terms of Laguerre and zonal polynomials and are thus inappropriate for our purpose. Furthermore, although adequate approximations to the distribution of a single quadratic form in normal variables are available, see e.g. Jensen and Solomon(1972) and Solomon and Stephens(1977), none are readily available for $\dot{p}$, and their integrability properties are not well known.

Hence, we need to approximate $p_k(c)$. The easiest way is by finding a transformation f such that $(f(Q_i), f(Q_{i+k}))'$ is approximately jointly normal. The following considerations lead to choosing the logarithmic transformation.

First, recall that each of Q_i and Q_{i+k} is the product of two independent chi-square random variables, one with one degree of freedom, the other with L . Wilson and Hilferty (1931) showed that in the family of power transformations, $(\chi^2/L)^{1/3}$ could be approximated by a $\mathcal{N}(1 - \frac{2}{9L}, \frac{2}{9L})$ distribution. Thus the distribution of $(Q_i/L)^{1/3}$ could be approximated by that of the product of two normal random variables. But we

saw in section 4.2 that this is a more complicated distribution than that of Q_i .

If we examine the density (4.2), which is the marginal density of Q_i and of Q_{i+k} , we can see that any power transformation will lead us back to the same modified Bessel function kernel.

The logarithmic transformation however, which belongs to the family of power transformations, will transform each of Q_i and Q_{i+k} into a sum of independent 'log- χ^2 ' random variables. The 'log- χ^2 ' distribution can be approximated by a normal distribution, and has been studied by Olshen(1938) and Bartlett and Kendall(1946). See also Johnson and Kotz(1970).

A second consideration is the similarity between the shape of the density of a $\mathcal{W}_L$ random variable and that of a log-normal random variable, hereafter denoted by $\mathcal{LN}(\mu, \sigma^2)$. Both random variables are strictly positive and both densities have a singularity at zero. Recall that $Y \sim \mathcal{LN}(\mu, \sigma^2)$ is such that $\ln Y \sim \mathcal{N}(\mu, \sigma^2)$.

A third consideration in the choice of a transformation is the behaviour of the moment generating function. In Theorem 4.4, we showed that $E[\mathcal{W}_L^t]$ is a polynomial in L of degree t . Taking

$$\mu_L = 0.5 \ln(L^3/3(L+2))$$

and

$$\sigma_L^2 = \ln\{3(L+2)/L\}$$

then $Y \sim \mathcal{LN}(\mu_L, \sigma_L^2)$ is such that $E[Y] = E[\mathcal{W}_L] = L$ and $E[Y^2] = E[\mathcal{W}_L^2] = 3L(L+2)$.

Furthermore, for $t \geq 0$,

$$E[Y^t] = 3^{t^2-t} L^{(3t-t^2)/2} (L+2)^{(t^2-t)/2}$$

i.e. the moments of $Y \sim \mathcal{LN}(\mu_L, \sigma_L^2)$ are also polynomials of the same degree in L .

Finally, as shown in the following Table 2, the upper percentage points of the $\mathcal{LN}(\mu_L, \sigma_L^2)$ are not too distant from those of the $\mathcal{W}_L$ distribution for various values of L .

Thus as a reasonable first approximation to the distribution of $\mathcal{W}_L$ we can use a $\mathcal{LN}(\mu_L, \sigma_L^2)$ distribution. This suggests to use a bivariate log-normal distribution to approximate $p_k(c)$.

Let $\begin{pmatrix} X \\ Y \end{pmatrix}$ be bivariate normal with

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \eta\sigma_x\sigma_y \\ \eta\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Then we say that $(Z_x, Z_y) = (e^x, e^y)$ has a bivariate log-normal distribution with mean vector μ and covariance matrix Σ , which we denote by $\mathcal{LN}(\mu, \Sigma)$.

Hence,

$$E[Z_x^s Z_y^t] = E[e^{sx+ty}] = \exp\{s\mu_x + t\mu_y + 0.5(s^2\sigma_x^2 + 2st\eta\sigma_x\sigma_y + t^2\sigma_y^2)\}.$$

TABLE 2

Comparison of the Upper Percentage Points of $\mathcal{W}_L$
With Those From a Log-Normal Approximation

L	α				
	.1	.05	.01	.001	.0005
1	.0852	.0364	.0067	.0010	.0006
2	.0811	.0367	.0075	.0012	.0007
3	.0795	.0368	.0079	.0013	.0008
4	.0784	.0368	.0081	.0014	.0009
5	.0776	.0368	.0083	.0015	.0009
6	.0769	.0367	.0085	.0016	.0010
7	.0764	.0367	.0086	.0016	.0010
8	.0759	.0366	.0087	.0017	.0011
9	.0755	.0365	.0088	.0017	.0011
10	.0751	.0365	.0089	.0018	.0011
12	.0746	.0364	.0090	.0018	.0012
24	.0728	.0360	.0094	.0021	.0014
36	.0720	.0358	.0095	.0022	.0015
48	.0716	.0357	.0096	.0023	.0016

By equating the first and second moments of the joint distribution of (Q_i, Q_{i+k}) with those of the bivariate log-normal distribution, we obtain

$$\mu_L = \mu_x = \mu_y = 0.5 \ln\{L^3/3(L+2)\},$$

$$\sigma_L^2 = \sigma_x^2 = \sigma_y^2 = \ln\{3(L+2)/L\}$$

and

$$\eta_k = \ln\{1 + \tau_k/L^2\}/\ln\{3(L+2)/L\}.$$

By equating the first and second moments of the joint distribution of (Q_i, Q_{i+k}) with those of the bivariate log-normal distribution, we obtain

$$\mu_1 = \mu_x = \mu_y = -\ln 3 \approx -1.0986,$$

$$\sigma_1^2 = \sigma_x^2 = \sigma_y^2 = \ln 9 \approx 2.1972$$

and

$${}_j\eta_k = \ln\{1 + {}_j\tau_k\}/\ln 9 \approx \ln\{1 + {}_j\tau_k\}/2.1972 \approx \ln\{1 + {}_j\tau_k\}^{1/2}.$$

We thus propose to approximate $p_k(c)$ with $\bar{p}_k(c)$ where

$$\bar{p}_k(c) = \Phi_{\eta_k}\left(\frac{\ln c - \mu_L}{\sigma_L}, \frac{\ln c - \mu_L}{\sigma_L}\right),$$

$$\Phi_{\eta}(x, y) = \int_x^{\infty} \int_y^{\infty} \phi_{\eta}(x, y) dx dy$$

and $\phi_{\eta}(x, y)$ is the standard bivariate normal density with correlation η .

Clearly, the approximation of $p_k(c)$ could be studied further on its own. But since the approximation enters into

the computation of the bounds in (4.5), it would be preferable to study further the approximate bounds.

Summarizing, from a time series of length n , to test at level $1-\alpha$ the most influential observation on the j -th sample autocorrelation coefficient, we compute ${}_jC_\alpha^-$ and ${}_jC_\alpha^+$ such that

$$\tilde{p}_0({}_jC_\alpha^+) - \sum_{k=1}^m [1 - \sigma(n-k)/n] \tilde{p}_{(n-k)}({}_jC_\alpha^+) = \alpha/n$$

and

(4.6)

$$\tilde{p}_0({}_jC_\alpha^-) - \sum_{k=1}^{n-1} (1 - k/n) \tilde{p}_k({}_jC_\alpha^-) = \alpha/n$$

where m is such that

$$1 = \frac{1}{n} + \sum_{k=1}^m [1 - \sigma(n-k)/n]$$

and

$$\tilde{p}_k(c) = \begin{cases} P(\mathcal{W}_1 > c) & \text{for } k = 0 \\ \Phi_{1, \eta_k} \left(\frac{\ln 3c}{\ln 9}, \frac{\ln 3c}{\ln 9} \right) & \text{for } k=1(1)(2L+2j) \\ P(\mathcal{W}_1 > c)^2 & \text{for } k=(2L+2j+1)(1)(n-1). \end{cases}$$

Then if ${}_jQ_{(n)} \geq {}_jC_\alpha^+$ we would reject the hypothesis H_3 , that no single observation has a discordant influence on r_j ; if ${}_jQ_{(n)} \leq {}_jC_\alpha^-$ we would 'accept' hypothesis H_3 , while for ${}_jC_\alpha^- < {}_jQ_{(n)} < {}_jC_\alpha^+$ the test would be inconclusive.

To test at level $1-\alpha$ the most influential observation on the whole sample ACF, we compute C_α^- and C_α^+ as in (4.6) above, where now

$$\tilde{p}_k(c) = \begin{cases} P(\mathcal{W}_L > c) & \text{for } k = 0 \\ \Phi_{\eta_k} \left(\frac{\ln c - \mu_L}{\sigma_L}, \frac{\ln c - \mu_L}{\sigma_L} \right) & \text{for } k=1(1)4L \\ P(\mathcal{W}_L > c)^2 & \text{for } k=(4L+1)(1)(n-1). \end{cases}$$

Then if $Q_{(n)} \geq C_a^+$ we would reject the hypothesis H_0 , that no single observation has a discordant influence on the whole sample (2); if $Q_{(n)} \leq C_a^-$ we would 'accept' hypothesis H_0 , while for $C_a^- < Q_{(n)} < C_a^+$ the test would be inconclusive.

The tighter these bounds are, the more useful they will be. Their tightness may, however, depend on the particular time series studied. In the following Chapter V, three examples of their usefulness for detecting overly influential observations in time series are given.

CHAPTER V

EXAMPLES

In this chapter, we give three examples of the use of the sample measure of influence with its bounds to detect overly influential observations on the sample autocorrelation function (ACF) when modelling time series as ARIMA processes.

The three times series chosen were used in the literature for demonstrating methods to detect outlying observations. Fox(1972) described two types of outliers in ARMA models: additive outliers (AO's) which are discordant observations in the observed time series, and innovation outliers (IO's) which are discordant observations in the innovation process underlying the observed time series.

Since outlying observations may overly influence the ACFs that are examined in the identification of ARIMA models, these outliers could also be identified using the sample measure of influence. For instance, Chang and Tiao(1983) and Burn and Ord(1984) have shown that AO's are more likely to distort the ACF of the observed time series than IO's.

In the following three sections we will only illustrate the identification of overly influential observations in the model-building process without regards for outlying observations and for explanations of their causes. We will however compare the results of our analyses with those obtained from the identification of outliers.

We do not claim, however, that the measures of influence developed in the previous chapters can distinguish between various forms of outliers in time series, or even that they can detect outliers, since they were not develop for this purpose. But the examples in this chapter are provided to show how useful these measures of influence can be for detecting at an early stage of the analysis the overly influential observations, which may also happen to be outliers.

The procedure for detecting overly influential observations while modelling a time series as an ARIMA process will generally be as follows.

1. We first determine the degrees of regular and seasonal differencing, and the power transformation required to stationarize the time series.
2. Then the sample ACF and PACF are computed, along with the sample measure of influence and its critical bounds to

detect overly influential observations on the ACF. If any observation is overly influential, it is temporarily set aside, and the influences with the critical bounds are recomputed without this observation. This is iterated until all overly influential observations on the ACF are detected. The ACF (and PACF) used for model identification is obtained by considering all the overly influential observations as missing.

3. Once some ARMA models are tentatively identified using the clean ACF and PACF, these models are estimated using standard techniques, but excluding the overly influential observations identified in step 2. The ACF of the residuals of these models is then analyzed, and if any residual is found to be overly influential, it too is declared missing. At this stage, however, measures of outlyingness should also be used. This step is repeated until an adequate ARMA model is found.
4. At the end of step 3 we will have a model for the bulk of the series and a list of discordant observations that are not modelled. However, because the discordant observations were detected one at a time, it is then preferable to formally test whether each of them could have been generated by the model or not, as a preventive measure against masking and swamping (see Chapter II).

Those observations that could have been generated by the model are reintroduced in the analysis. For those that could not, it is the analyst's duty to figure out why these observations are discordant and how the model should be modified to encompass them.

In section 5.1 we use a time series reported by Barnett and Lewis(1978) which was analyzed by Chedzoy(1973) using a very simple procedure for detecting outliers. This procedure is seen to be too conservative, as too many observations are declared discordant.

In section 5.2 and 5.3, the time series used were analyzed for outliers by Chang(1982) and by Chang, Tiao and Chen(1988), respectively, both using a more appropriate method than that of Chedzoy. In both cases it is seen that outliers which are also overly influential can be detected quite straightforwardly using the sample measure of influence.

We close this chapter with section 5.4 giving some of the details on the algorithms used in the computation of the bounds on the critical values for the sample measures of influence.

5.1 Proportions of Fatal Road Accidents in the British Isles.

Our first example is taken from Barnett and Lewis(1978). The time series consists of the Monthly Proportions of Road Accidents Resulting in Death in the British Isles. There are 120 observations from January 1961 (1961/1) through December 1970 (1970/12). The series is plotted in Figure 2.

According to Barnett and Lewis(1978, p. 13), this series was analyzed by Chedzoy(1973) for discordant observations, by comparing each observation to the average value of the same month for the ten years. For example, the value for March 1961 was compared with the average of the ten observed March values. This method of detection is much too simple, but it is related to the method used in the seasonal adjustment procedure X-11-ARIMA. See Dagum(1980).

With this simple approach, Chedzoy indicated that the following 12 observations might be discordant: January and February 1963, March 1961, April 1965, May and June 1961, July 1969, August 1961, September and October 1969, November and December 1966. Note that for the observed ten years, all twelve months are found discordant once, and only once. From Barnett and Lewis'(1978) discussion, it is not clear if Chedzoy considered all of these as truly discordant.

However, using our measure of influence for the sample autocorrelation function (ACF), we can determine whether any of the above observations are discordant in their influence.

The first 24 lags of the ACF and of the PACF of the original series are shown in Figure 3, where the dotted lines delimit the 95% confidence intervals for the ACF and PACF of a Gaussian white noise process. Since the ACF in Figure 3 fails to decay to zero rapidly, and since the first PACF coefficient is truly significant, the data needs to be differenced once for identifying an ARMA model. The ACF and PACF of the differenced data are shown in Figure 4, where again the dotted lines mark the 95% confidence intervals for the ACF and PACF of a Gaussian white noise process.

It is usually at this stage that the ACF and the PACF are analyzed further to specify some ARMA models. It is thus at this stage that we should determine first whether the ACF has been distorted by overly influential observations. Thus, the influence of each observation on the first 24 coefficients of the ACF is computed, and is plotted in Figure 5 along with the bounds on the 95% critical value for the largest influence. These bounds are 286.5 and 347.9.

From Figure 5 it is seen that none of the observations indicated by Chedzoy are discordant in their influence on the sample ACF. The observation having the largest influence,

January 1965 with a value of 154.1, is well below the lower bound. Therefore the sample ACF computed is reliable for determining and estimating an ARMA model.

Letting X_t denote the Monthly Proportion of Fatal Road Accidents, we obtained model (5.1) for this series, where the residuals e_t form a white noise process with $\hat{\sigma}_e = .1304$. (The number in square brackets below a parameter's estimate is its t-statistic.)

$$(1 - \underset{[3.3]}{.31}B)(1 - \underset{[4.5]}{.42}B^{12} - \underset{[3.7]}{.34}B^{24})X_t = \underset{[24.4]}{2.096} + e_t. \quad (5.1)$$

Notice that model (5.1) is stationary. In theory, we should have been able to compute the bounds on the critical value of the largest observed influence without having to take the first difference of the data. However, since the ACF is slowly decaying to zero, the correlation between the measures of influence is large enough to render the lower Bonferroni bound useless, a phenomenon discussed by Schwager(1984). We shall discuss this further in section 5.4.

The second autocorrelation coefficient of the residuals e_t of model (5.1) is $r_e(2) = .23$, whereas the 95% confidence interval assuming white noise is ± 0.2 . This indicates that we may add a moving average parameter of order 2 to improve model (5.1). But was the value of $r_e(2)$ distorted by some residual?

The influence of each residual on their second autocorrelation coefficient, together with the bounds on the 95% critical value for the maximum value, are shown in Figure 6. These bounds are 38.1 and 38.7. Since there are no overly influential residuals on the second autocorrelation coefficient, we can add the moving-average parameter of order 2 obtaining model (5.2), where $\hat{\sigma}_e = .1275$.

$$(1 - \underset{[3.1]}{.28B})(1 - \underset{[4.2]}{.38B^{12}} - \underset{[4.0]}{.36B^{24}})X_t = \underset{[26.3]}{2.097} + (1 + \underset{[2.3]}{.25B^2})e_t. \quad (5.2)$$

As a final check, the maximum influence of the residuals of model (5.2) was also verified to be non-significant at the 95% level.

Therefore, the ARMA model (5.2) is appropriate for the Monthly Proportions of Fatal Road Accidents, and we are assured that no observations, nor any residuals are unduly influencing the ACF's. Hence, we could now apply any outlier detection technique that relies on a prespecified model, as the method of Chang(1982) discussed in the next sections.

Figure 2. Proportions of Fatal Road Accidents in the British Isles

Figure 3. Sample ACF and PACF of Fatal Road Accidents

Figure 4. Sample ACF and PACF of First Differences of Fatal Road Accidents

Figure 5. Influences on the whole sample ACF (24 Lags) of Fatal Road Accidents

Figure 6. Influences on the Second Autocorrelation Coefficient of the Residuals from Model (5.2)

5.2 Monthly Sales of Variety Stores in the U.S.

The second example is taken from Chang(1982). The time series is the Monthly Retail Sales of Variety Stores in the United States. There are 153 observations, from January 1967 (1967/1) through September 1979 (1979/9). The series was adjusted for trading days by Chang(1982) and is logarithmically transformed to stationarize the variance of the observations. The series is shown in Figure 7.

The series of Variety Stores Sales was used by Chang (1982) to exemplify her proposed method for identifying outlying observations in time series.

Chang's(1982) approach is as follows. First we identify and estimate an ARIMA model for the series. Then, two statistics are computed from the residuals of the fitted model. One statistic is tailored for detecting Additive Outliers (AOs), the other is tailored for detecting Innovation Outliers (IOs). If any outliers are found using these statistics, the model is modified for their inclusion and the new model is estimated. The procedure stops when no outliers are found and the model is acceptable. Details on this method can also be found in Chang, Tiao and Chen(1988).

Since outliers can only be defined with respect to a model, and since they may also overly influence the estimated

parameters of the initial model, there always is the possibility that outliers that are also overly influential would go undetected. It is therefore important to determine first the overly influential observations.

Chang(1982) detected 9 outliers in the Variety Stores Sales series. They are listed in Table 3.

Letting Y_t denote the logarithm of Monthly Retail Sales of Variety Stores in the U.S., the initial ARIMA model used by Chang to identify the nine outliers was

$$(1 + .4B + .27B^2)\nabla\nabla_{12}Y_t = (1 - .81B^{12})e_t, \quad (5.3)$$

where $\hat{\sigma}_e = .0319$. The final model obtained by Chang is:

$$(1 + .66B + .34B^2)\nabla\nabla_{12}(Y_t + A_t) = (1 - .9B^{12})(e_t + I_t), \quad (5.4)$$

where $\hat{\sigma}_e = .0219$, A_t is a correction for the AOs and I_t is a correction for the IOs.

We will now reanalyze the series Y_t , to include the detection of overly influential observations, and compare our results with those of Chang. Note however that we could not include Chang's outlier analysis as we did not have the necessary software at our disposal. This is not a problem

here since our goal is to compare the results of both methods when they are used at an early stage of the modelling process.

In Figure 8 are shown the first 24 lags of the ACF and PACF of the series Y_t , where the dotted lines delimit the 95% confidence intervals for the ACF and PACF of a Gaussian white noise process. Since the ACF fails to decay rapidly to zero both on its non-seasonal lags (i.e. $k=1,2,3,\dots$) and on its seasonal lags (i.e. $k=12,24,36,\dots$), and since the first, twelfth and thirteenth coefficients of the PACF are truly significant, these functions clearly show the need for one regular and one seasonal differences to stationarize the series. The first 24 lags of the ACF and PACF of the differenced data, $\nabla\nabla_{12}Y_t$, are shown in Figure 9, where again the dotted lines mark the 95% confidence intervals for the ACF and PACF of a Gaussian white noise process.

Now, before attempting to interpret the ACF and PACF of Figure 9 for identifying an ARMA model, we compute the influence of each observation on the whole ACF to determine whether any observation is overly influential. In Figure 10 is shown the influence of each observation on the first 24 lags of the ACF, together with the bounds on the 95% critical value for the largest influence. These bounds are 303.3 and 359.9.

As can be seen in Figure 10, the April 1976 and April 1977 observations are overly influential. Their measures of influence are 309.6 and 295.7, respectively. In theory, we should only declare 1976/4 as overly influential since we are testing for at most one discordant observation. But if we remove 1976/4 from the computations, by temporarily declaring it as missing, and recompute the influence of all the remaining observations, we find that the influence of the 1977/4 observation jumps to 487.3, while the bounds remain essentially the same. Thus, both observations are overly influential.

These two observations mostly influence the first autocorrelation coefficient, but they also greatly influence other coefficients. This is seen in Figure 11, showing the ACF and PACF of $VV_{12}Y_t$ computed without the 1976/4 and 1977/4 observations, where the dotted lines mark the 95% confidence intervals for the ACF and PACF of a Gaussian white noise process. The PACF now clearly shows that the non-seasonal component of the process is AR(2).

The observations for 1976/4 and 1977/4 are the only two that can be declared overly influential at this stage of the analysis, and thus are temporarily set aside.

We recomputed the influence of all the remaining observations and found no other overly influential

observations. This is shown in Figure 12, where the solid lines are the bounds on the 95% critical value for the largest influence. These bounds are 301.2 and 358.1.

Interestingly, however, the largest influential observation at this stage is 1970/9, because of its impact on the third autocorrelation coefficient. If taken singly, the influence of 1970/9 on the first 24 lags of the ACF is significantly important. It measures 167.5, while the 95% critical value is 93.9. This critical value is shown as a dotted line in Figure 12. But since we were testing for the largest influence on the first 24 four lags of the ACF, this is not a significant largest influence. Therefore, we cannot declare this observation discordant at this stage.

Figure 12 also shows that there are a total of 5 observations whose influence is significant when they are taken singly: 1970/9, 1973/2, 1975/7, 1975/8 and 1976/1.

Continuing with the ARMA analysis, we obtain model (5.5) for the corrected series of Variety Stores Sales $\tilde{Y}_t$, i.e. the original series Y_t with the 1976/4 and 1977/4 observations replaced by the mean of the remaining observations, where $\hat{\sigma}_e = .0269$.

$$(1 + \underset{[8.1]}{.54}B + \underset{[5.4]}{.35}B^2)\nabla\nabla_{12}\tilde{Y}_t = (1 - \underset{[7.0]}{.61}B^{12} - \underset{[3.5]}{.31}B^{24} + \underset{[2.3]}{.22}B^{36})e_t. \quad (5.5)$$

Then, if we analyse the influence of the residuals of model (5.5) on their ACF, we observe that observations 1970/9, 1972/4, 1973/1, 1974/12, 1975/7 would be significantly influential on the first 24 lags of the residuals' ACF if each of these observations had been tested singly, but that the largest influence measured, for 1974/12, is not within the bounds for the 95% critical value.

In practice, it is at this point that we would start analyzing the residuals for outliers, using for example Chang's method. But we can now compare our results with those obtained by Chang(1982).

First, we note that Chang's models (5.3) and (5.4) are different from model (5.5) in that the latter contains moving-average parameters of order 24 and 36, which are absent in models (5.3) and (5.4). Furthermore, both models (5.3) and (5.4) are peculiar in that the seasonal moving-average term of order 12 nearly cancels the seasonal difference applied to Y_t , while none of the three seasonal roots of model (5.5) are near one.

Model (5.5) differs from Chang's model (5.3) because her model was obtained from the original series Y_t including the overly influential observations 1976/4 and 1977/4, while model (5.5) was found from the corrected series $\tilde{Y}_t$. If model (5.5) is fitted to the original series Y_t , the moving-average

parameter of order 36 becomes non-significant, and the resulting moving-average operator $(1 - \phi_{12}B^{12} - \phi_{24}B^{24})$ has one seasonal root almost equal to 1. Thus, this is one effect of the overly influential observations 1976/4 and 1977/4.

As for the identified outlying and overly influential observations, Table 3 summarizes our findings and those of Chang(1982). In Table 3 it is seen that apart from 1972/4, which was recognized as a potentially influential residual, there are essentially no differences between Chang's analysis and ours for the data prior to 1976.

There is however a major difference for the period 1976/1 to 1979/9. We declared the observations 1976/4 and 1977/4 overly influential. While 1977/4 was not recognized as an outlier in Chang's analysis, the three contiguous observations 1976/4, 1976/5 and 1976/7 were declared IO's. But she recognized the existence of a change in level starting in 1976/4, that caused these three consecutive IO's.

Now, because of this change in the level of the original series Y_t in 1976/4, when the series is differenced once regularly and once seasonally, the change is 'spread' on adjacent observations.

Let $\delta_t^{(1)}$ represent a change in level of α at time T , i.e. $\delta_t^{(1)} = \alpha > 0$ for $t \geq T$, and is equal to zero otherwise. Then

$$\nabla \delta_t^{(T)} = \begin{cases} \alpha & \text{at time } T \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\nabla \nabla_{12} \delta_t^{(T)} = \begin{cases} \alpha & \text{at time } T \\ -\alpha & \text{at time } T+12 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, our measure of influence correctly detected this change in level.

Chang modelled this effect by a correction to the original series, assuming that the three observations 1976/4, 1976/5 and 1976/6 are IO's. Since we suggest to analyze the series with the observations 1976/4 and 1977/4 considered as missing, which simply consists of replacing the observations 1976/4 and 1977/4 in the stationarized series $\nabla \nabla_{12} Y_t$ by the average of the remaining observations, we may compare Chang's correction with our correction to the level of the series.

Figure 13 shows the seasonally differenced series of Variety Stores, i.e. $\nabla_{12} Y_t$, as a solid line together with the two corrected series. The first correction, the dotted line in Figure 13, is that of Chang, and the second correction, the dashed line, is that used in our analysis. Figure 13 clearly shows that the two corrections are essentially the same.

The other differences with regards to the status of observations 1976/1, 1977/1 and 1977/9, may have been caused by the difference in the treatment of the observations going from 1976/4 to 1977/4, and in the ARMA models used.

We do not claim that our measure of influence can distinguish between various forms of outliers in time series, or even that it can detect outliers, since it was not developed for this purpose. This example however shows how useful it can be for detecting at an early stage of the analysis the overly influential observations, which may also be outliers.

TABLE 3

Comparison of Chang's(1982) Outlier Analysis
with the Influence Analysis of Variety Stores Sales

<u>Outlying Observations</u>		<u>Influential Observations</u>	
<u>Time</u>	<u>Type</u>	<u>Time</u>	<u>Type</u>
1970/9	IO	1970/9	pI
		1972/4	pI
1973/1	IO	1973/1	pI
		1973/2	pI
1974/12	AO	1974/12	pI
1975/7	IO	1975/7	pI
		1975/8	pI
		1976/1	pI
1976/4	IO	1976/4	I
1976/5	IO		
1976/6	IO		
1977/1	AO		
		1977/4	I
1977/9	IO		

AO: Additive Outlier
IO: Innovation Outlier
pI: Potentially Influential
I : Overly Influential

Figure 7. Logarithm of Retail Sales of Variety Stores in the U.S.

Figure 8. Sample ACF and PACF of Variety Stores Sales

Figure 9. Sample ACF and PACF of the Differenced Variety Stores Sales

Figure 10. Influences on the whole sample ACF (24 Lags) of Variety Stores Sales

Sample ACF

Sample PACF

Figure 11. Sample ACF and PACF of the Differenced Variety Stores Sales
Excluding Observations 1976/4 and 1977/4

Figure 12. Influences on the Whole Sample ACF (24 Lags) of Variety Stores Sales,
Excluding Observations 1976/4 and 1977/4

Figure 13. Comparison of the Seasonally Differenced Variety Stores Sales Series with Chang's Correction and the Correction in Level

5.3 Chemical Process Concentration Readings

Our final example is taken from Box and Jenkins(1976), where it is referred to as Series "A". It is a series of concentration readings from a chemical process, taken every 2 hours; there are a total of 197 observations. The series is shown in Figure 14.

Series "A" was used by Chang, Tiao and Chen(1988) to demonstrate Chang's(1982) method for the identification of outliers in time series. We briefly described this method in section 5.2. We will analyze this series for overly influential observations and compare the results with those of the outlier analysis performed by Chang, Tiao and Chen(1988).

Since the ACF of the original series fails to decay rapidly to zero, one regular difference is required for identifying an ARMA model. The first 20 lags of the ACF and PACF of the differenced data are shown in Figure 15, where the dotted lines mark the 95% confidence intervals for the ACF and PACF of a Gaussian white noise process. Again, before attempting to identify an ARMA model for Series "A", we compute the influence of each observation on the first twenty lags of the ACF. The influences are shown in Figure 16, together with the bounds on the 95% critical value for the largest influence. These bounds are 290.3 and 324.5.

From Figure 16 it can be immediately deduced that observations #44 and #64 are overly influential. Hence they are temporarily set aside, and the influences of all the remaining observations are recomputed, as well as new bounds. These new bounds are 295.0 and 326.0. Upon inspecting these new measures of influence, it is revealed that the observation having the largest influence, observation #43, is almost significantly too large, its influence measuring 249.0. We thus also removed observation #43 from the analysis, and no other overly influential observations were found at this stage.

Next, we obtain model (5.7) for Series "A", where $\hat{\sigma}_e = .2938$.

$$(1 - \underset{[10.6]}{.85B}) A_t = \underset{[119.3]}{17.07} + (1 - \underset{[3.6]}{.38B}) e_t. \quad (5.7)$$

We examined the influence of the residuals of model (5.7) on the twenty first lags of their ACF, and none were found to be overly influential. There are however eight residuals whose influence would have been found significant if they had been tested singly. The two most influential residuals correspond to observations #172 and #191, and their influences were measured at 136.0 and 209.1 respectively, while the 95% critical value is 78.5.

Chang, Tiao and Chen(1988) found only two outliers in Series "A": observations #43 and #64. However, observation #43 was categorized as an additive outlier (AO), while observation #64 was categorized as an innovation outlier (IO).

We found that observations #43, #44 and #64 were overly influential for the first difference of the data. Hence if observation #43 is an AO, the first difference will 'spread' part of the discordancy to the next observation, observation #44. Therefore our results are consistent with those of Chang, Tiao and Chen(1988), and this examples shows once more the usefulness of testing for overly influential observation.

Hence, these three examples clearly demonstrate the utility of the proposed measures of influence for the whole sample ACF and of their approximate bounds obtained in section 4.3 of Chapter IV.

In the first example, section 5.1, no overly influential observations were found. Since our method was not designed to detect outliers, we could apply an appropriate outlier detection technique, such as Chang's(1982), to the series of Fatal Road Accidents. Because no overly influential observations were found, we are assured of the power of these outlier detection techniques that rely on a prespecified ARIMA model.

In the second example, section 5.2, the usefulness of an early detection of overly influential observations was demonstrated by showing that two observations had indeed distorted the model required for detecting outliers. Although these two observations were nonetheless declared outliers, the distortion in the model may have biased the identification of the remaining outliers in the series.

In the third example in this section, the influence analysis was sufficient to detect all the known discordant observations in the time series.

Figure 14. Concentration Readings of a Chemical Process (Series A of Box and Jenkins)

Figure 15. Sample ACF and PACF of the First Differences of Series A

Figure 16. Influences on the Whole Sample ACF (20 Lags) of the Differenced Series A

3

OF/DE

3

MicroD

5.4 Computational Details

All the computations pertaining to the measures of influence and their bounds were performed with version 7.1 of APL*PLUS from STSC Inc.. All the regressions were performed with version 6.0 of Micro-TSP from QMS Inc..

To compute the bounds on the influence measures and the distribution function of the $\mathcal{W}_L$ random variable, we need to compute the modified Bessel function of the second kind $K_\nu(x)$ for integer orders $\nu = 0, 1, 2, \dots$, and for fractional orders $\nu = 0.5, 1.5, 2.5$, etc. We also need to compute integrals of the form

$$J_\nu(x) = \int_0^x y^\nu K_\nu(y) dy.$$

For the functions $K_0(x)$, $K_1(x)$, $J_0(x)$ and $J_1(x)$ we used the Chebychev polynomial approximations given by Luke(1969). For the functions $K_{0.5}(x)$, $K_{1.5}(x)$ we used their definitions:

$$K_{0.5}(x) = \left(\frac{\pi}{2x} \right)^{1/2} e^{-x}$$

and

$$K_{1.5}(x) = \left(\frac{\pi}{2x} \right)^{1/2} e^{-x} (1 + x^{-1}),$$

while $J_{0.5}(x)$ and $J_{1.5}(x)$ were obtained by direct integration of the above.

For the functions of higher order, we used the following recurrence formula for $K_\nu(x)$ which is straightforward to modify to obtain that for $J_\nu(x)$:

$$xK_{\nu-1}(x) + xK_{\nu+1}(x) = 2xK_\nu(x).$$

Finally to compute the distribution function of a standard bivariate normal distribution with non-zero correlation, we used Donnelly's(1973) algorithm. This algorithm requires the cumulative distribution function of a standard univariate normal distribution, for which we used the continued fractions of Hill(1973).

All inversions of these functions, e.g. when calculating the critical values of the $\mathcal{W}_l$ distribution for a fixed probability as given in Table 1, were performed using the Newton-Raphson algorithm. The precision of the resulting values is always of at least one decimal more than what is shown herein.

Now, two difficulties were encountered when computing the bounds on the critical value for the largest influence. The first is directly related to the use of Donnelly's(1973) algorithm for computing the cumulative distribution function of a bivariate normal distribution. This algorithm is very slow. In future implementations of these computations, a much faster algorithm should be used, e.g. that of Divgi(1979).

The second difficulty is that on certain occasions, the lower Bonferroni bound on the upper tail probability of the largest influence becomes negative. This phenomenon was reported by Schwager(1984), and it usually occurs when the correlations between the events entering into the computation of the Bonferroni bounds is strong. When this happens the lower bound on the critical value cannot be computed.

However, we saw in Chapter II that when we are testing for outliers in the standard linear model, we could reduce the number of events entering into the computation of the Bonferroni bounds, under certain conditions. It is quite likely that similar conditions could be found to reduce the number of events used to compute the bounds on the critical value of the largest influence.

Nonetheless, Hunter's(1976) upper bound is always computable and often constitutes a definite improvement over the usual Bonferroni upper bound.

CHAPTER VI

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

In this thesis we developed tools for identifying overly influential observations on a fundamental quantity in time series analysis, the autocorrelation function (ACF).

As was shown with three examples in Chapter V, these tools usefully complement the outlier detection techniques for time series relying on a prespecified ARIMA model, such as the methods by Abraham and Box(1979), Abraham and Yatawara(1988), Chang(1982), Chang, Tiao and Chen(1988), Fox(1972) and Tsay (1988). The power of these methods may be severely reduced when outliers are also overly influencing the initial ARIMA model. Hence, if the overly influential observations are detected prior to the ARIMA modelling, the power of the outlier detection methods will be increased.

By comparison with previous work on the identification of overly influential observations for the ACF, particularly the papers by Chernick, Downing and Pike(1982), Lattin(1983), Li and Hui(1987) and Polasek(1982), the most commonly used measures of influence were obtained in Chapter III from a more appropriate framework for the analysis of influence measures, and, more importantly, in Chapter IV we obtained thresholds for declaring an observation overly influential.

To describe the proper theoretical framework for the study of influential observations in time series, we needed to review the current theory behind the identification of discordant observations in the standard linear regression model, a situation for which the theory is almost complete. In reviewing this theory, however, we found that most presentations were failing to clearly show the common thread in case analysis.

Thus, in Chapter II we briefly presented all the major aspects of this theory in a way that should prove useful both in averting any confusion in the relevant concepts of case analysis, and for judging the intrinsic qualities of many, if not all, measures of discordancy that have been suggested for the standard linear model.

This theory is not however complete, as we noted that the currently proposed critical values for determining whether an observation is a discordant leverage point, and those for determining if an observation is overly influential for the least-squares estimator of the coefficients in the model, had not been derived within the theoretical framework for analyzing discordant observations. Clearly, the fact that the probabilistic structure of the design points is not considered relevant in the standard linear model is the major obstacle for deriving the required critical values.

By imposing some structure on the design points, e.g. by assuming that they form a random sample from a multivariate Gaussian distribution, this problem is resolved, and meaningful probability statements about these potential discordancies can be made. As Chatterjee and Hadi(1988), we suggested that Bonferroni's upper bound could be used for identifying design points of discordant leverage, but we did not tackle the problem of determining the critical values for identifying overly influential observations using Cook's(1977) distance.

In both cases, bounds such as those that we developed for overly influential observations on the ACF could be obtained using Hunter's(1976) upper bound and Bonferroni's lower bound described in section 2.3 of Chapter II.

In Chapter II, Theorem 2.5, we obtained a really useful characterization of the three most frequently used measures of influence, as summaries of the derivative of the statistic with respect to the weight given to an observation by the empirical distribution function (EDF). Theorem 2.5 generalizes the results of Pregibon(1981) and Cook and Weisberg(1982) for the linear model, and is shown to hold under less stringent conditions than those used by Critchley(1985).

The characterization of Theorem 2.5 was used in Chapter III to obtain the three main measures of influence for the ACF. However, the proof of Theorem 2.5 requires the assumption of independent and identically distributed (i.i.d.) observations. It also uses the standard definition for the theoretical influence function, which is inadequate for time series, as shown by Martin and Yohai(1986). Furthermore, there currently are at least two definitions for the EDF of an observed time series; one definition is used by Martin and Yohai(1986), another is used by Künsch(1984).

Thus, there remains open the question as to whether a theorem similar to Theorem 2.5 could be proved using Martin and Yohai's(1986) definition of influence function for time series, without too stringent hypotheses on the time series structure. Nonetheless, as was seen in Chapter III the characterization of Theorem 2.5 provides an effective framework for obtaining and comparing measures of influence for time series.

Our principal contribution is given in Chapter IV, where the statistical properties of the sample measure of influence were derived in order to obtain an appropriate summary of the influence of one observation on the whole sample ACF. In previous work, only one summary had been suggested, and we saw that it had the defects of mixing the influence of various

observations, and of having a possibly very complicated distributional theory.

Furthermore, we derived some of the distributional properties of these measures of influence, from which asymptotic critical values for declaring a specific observation as overly influential either for a specific autocorrelation coefficient, or for the whole sample ACF can be obtained.

We also obtained bounds for the critical values required to test the extremeness of the most influential observation on either a specific autocorrelation coefficient, or on the whole sample ACF. Leadbetter, Lindgren and Rootzén(1983) have shown that the distributions of the extremes of a stationary Gaussian process varies according to the speed with which the true ACF of the process decays to zero. Since our measures of influence are not normally distributed, and since we wish to measure influence at an early stage of the ARIMA model-building process, we cannot expect to obtain exact critical values.

The bounds that we developed depend on a distribution whose density is the integral, with respect to a correlated standard bivariate normal distribution, of the density of two dependent non-central quadratic forms in normal variables.

Current expressions for the joint density of these

quadratic forms are in terms of Laguerre and zonal polynomials, see e.g. Khatri, Krishnaiah and Sen(1977). Hence, they are inappropriate for our purpose as the integration with respect to the bivariate normal distribution cannot be carried out easily, either numerically or theoretically. Furthermore, the integrability properties for the usual approximations to a single quadratic forms are not well known in the non-central case, and it is not clear if these approximations could be extended to the dependent non-central bivariate case. Therefore, a good integrable approximation for this density should be developed.

Meanwhile, we used a bivariate log-normal approximation to obtain the desired probabilities. This log-normal approximation should be viewed as the first-term of an Edgeworth-type approximation. The quality of this approximation, as well as the the coverage probability of the resulting bounds remain to be determined either theoretically or through simulation studies.

Also, when the sample ACF is not decaying rapidly enough to zero, the lower Bonferroni bound may be useless. This phenomenon was described by Schwager(1984). It is most likely that when this occurs, we would only have to reduce the effective number of observations used in computing this bound to obtain a reasonable value. However, there remains to

determine the conditions under which this can be done, and the amount by which to reduce the number of observations.

Finally, using the framework described in Chapters II and III, it would be possible to develop the equivalent measures of influence for the partial auto-correlation function (PACF), and to study their relationship to those for the ACF, since the PACF is a non-linear one-to-one transformation of the ACF.

BIBLIOGRAPHY

- Abraham, B., Box, G.E.P.(1979) Bayesian Analysis of Some Outliers Problems in Time Series, Biometrika, **66**, 2, pp. 229-236
- Abraham, B., Yatawara, N.(1988) A Score Test for Detection of Time Series Outliers, J. Time Series Anal., **9**, 2, pp. 109-119
- Anderson, T.W.(1971) The Statistical Analysis of Time Series, J. Wiley, New York
- Aroian, L.A., Taneja, V.S., Cornwell, L.W.(1978) Mathematical Forms of the Distribution of the Product of Two Normal Variables, Commun. Statist.-Theor. Meth., **A7**, 2, pp. 165-172
- Atkinson, A.C.(1981) Two Graphical Displays for Outlying and Influential Observations in Regression, Biometrika, **68**, 1, pp. 13-20
- Atkinson, A.C.(1985) Plots, Transformations, and Regression, Oxford University Press, Oxford, England
- Azencott, R., Dacunha-Castelle, D.(1984) Séries d'observations irrégulières. Modélisation et prévision, Masson, Paris, France
- Barndorff-Nielsen, O., Kent, J., Sørensen, M.(1982) Normal Variance-Mean Mixtures and z Distributions, Internat. Statist. Rev., **50**, pp. 145-159
- Barnett, V., Lewis, T.(1978) Outliers in Statistical Data, First Edition, J. Wiley, New York
- Barnett, V., Lewis, T.(1984) Outliers in Statistical Data, Second Edition, J. Wiley, New York
- Bartlett, M.S., Kendall, D.G.(1946) The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation, J. Roy. Statist. Soc. Ser. B, **8**, pp. 128-138
- Beckman, R.J.(1986) Discussion of Cook(1986), J. Roy. Statist. Soc. Ser. B., **48**, 2, pp. 161-162

- Beckman, R.J., Cook, R.D.(1983) Outlier.....s (with discussion), Technometrics, 25, 2, pp. 119-163
[Correction: 25, 4, p. 390]
- Belsley, D.A., Kuh, E., Welsch, R.E.(1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, J. Wiley, New York
- Bhattacharya, S.K.(1966) A Modified Bessel Function Model in Life Testing, Metrika, 11, pp. 133-144
- Box, G.E.P., Jenkins, G.M.(1976) Time Series Analysis Forecasting and Control, Holden-Day, San-Francisco
- Burn, D.A., Ord, J.K.(1984) The Effect of Time Series Outliers on the Sample Autocorrelation Coefficient, Amer. Statist. Assoc. Proc. Bus. Econom. Statist. Section, pp. 581-585
- Chang, I.(1982) Outliers in Time Series, Ph. D. Thesis, University of Wisconsin-Madison
- Chang, I., Tiao, G.C.(1983) Effect of Exogenous Interventions on the Estimation of Time Series Parameters, Amer. Statist. Assoc. Proc. Bus. Econom. Statist. Section, pp. 532-537
- Chang, I, Tiao, G.C., Chen, C.(1988) Estimation of Time Series Parameters in the Presence of Outliers, Technometrics, 30, 2, pp. 193-204
- Chatterjee, S., Hadi, A.S.(1986) Influential Observations, High Leverage Points, and Outliers in Linear Regression (with discussion), Statist. Sci., 1, 3, pp. 379-416
- Chatterjee, S., Hadi, A.S.(1988) Sensitivity Analysis in Linear Regression, J. Wiley, New York
- Chedzoy, O.B.(1973) Paper read at the annual conference of the Royal Statistical Society, Newcastle upon Tyne
- Chernick, M.R., Downing, D.J., Pike, D.H.(1982) Detecting Outliers in Time Series Data, J. Amer. Statist. Assoc., 77, 380, pp. 743-747

- Cook, R.D.(1977) Detection of Influential Observations in Linear Regression (with discussion), Technometrics, 19, 1, pp. 15-18, 19, 3, pp. 348-350
- Cook, R.D.(1979) Influential Observations in Linear Regression, J. Amer. Statist. Assoc., 74, 365, pp. 169-174
- Cook, R.D.(1986) Assessment of Local Influence (with discussion), J. Roy. Statist. Soc. Ser. B, 48, 2, pp. 133-169
- Cook, R.D., Holschuh, N., Weisberg, S.(1982) A Note on an Alternative Outlier Model, J. Roy. Statist. Soc. Ser. B, 44, 3, pp. 370-376
- Cook, R.D., Prescott, P.(1981) On the Accuracy of Bonferroni Significance Levels for Detecting Outliers in Linear Models, Technometrics, 23, 1, pp. 59-63
- Cook, R.D., Weisberg, S.(1982) Residuals and Influence in Regression, Chapman and Hall, London, England
- Craig, C.C.(1936) On the Frequency Function of xy , Ann. Math. Stat., 7, pp. 1-15
- Critchley, F.(1985) Influence in Principal Components Analysis, Biometrika, 72, 3, pp. 627-636
- Dagum, E.B.(1980) The X-11-ARIMA Seasonal Adjustment Method, Statistics Canada, Ottawa, Cat. No. 12-564E
- David, I.P., Sukhatme, B.V.(1974) On the Bias and Mean Square Error of the Ratio Estimator, J. Amer. Statist. Assoc. 59, 346, pp. 464-466
- Divgi, D.R.(1979) Calculation of Univariate and Bivariate Normal Probability Functions, Ann. Statist., 7, 4, pp. 903-910
- Donnelly, T.G.(1973) Algorithm 462. Bivariate Normal Distribution, Commun. Assoc. Comput. Mach., 16, p. 638
- Dunsmuir, W.(1984) Large Sample Properties of Estimation in Time Series Observed at Unequally Spaced Times in Time Series Analysis of Irregularly Observed Data, Parzen, E.(Ed.), Lect. Notes in Statistics #25, Springer-Verlag, Berlin

- Durbin, J., Watson, G.S.(1950) Testing for Serial Correlation in Least Squares Regression I, Biometrika, 37, pp. 409-428
- Ellenberg, J.H.(1976) Testing for a Single Outlier from a General Linear Regression Model, Biometrics, 32, pp. 637-645
- Fox, A.J.(1972) Outliers in Time Series, J. Roy. Statist. Soc. Ser. B, 34, 3, pp. 350-363
- Fuller, W.A.(1976) Introduction to Statistical Time Series, J. Wiley, New York
- Galambos, J.(1978) The Asymptotic Theory of Extreme Order Statistics, J. Wiley, New York
- Godolphin, E.J.(1984) A Direct Representation for the Large-Sample Maximum Likelihood Estimator of a Gaussian Autoregressive-Moving Average Process, Biometrika, 71, 2, pp. 281-289
- de Gooijer, J.G., Abraham, B., Gould, A., Robinson, L.(1985) Methods for Determining the Order of an Autoregressive-Moving Average Process: A Survey, Internat. Statist. Rev., 53, pp. 301-329
- Gradshteyn, I.S., Ryzhik, I.M.(1980) Table of Integrals, Series, and Products, Academic Press, New York
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.(1986) Robust Statistics. The Approach Based on Influence Functions, J. Wiley, New York
- Hawkins, D.M.(1980) Identification of Outliers, Chapman and Hall, London, England
- Hill, I.D.(1973) Algorithm AS66. The Normal Integral, Appl. Statist., 22, pp. 424-427
- Huber, P.J.(1981) Robust Statistics, J. Wiley, New York
- Hunter, D.(1976) An Upper Bound for the Probability of a Union, J. Appl. Prob., 13, pp. 597-603
- Jensen, D.R.(1970a) The Joint Distribution of Quadratic Forms and Related Distributions, Austral. J. Statist., 12, 1, pp. 13-22

- Jensen, D.R.(1970) The Joint Distribution of Traces of Wishart Matrices and Some Applications, Ann. Math. Statist., **41**, 1, pp. 133-145, [Correction: **41**, 6, p. 2186]
- Jensen, D.R., Solomon, H.(1972) A Gaussian Approximation to the Distribution of a Definite Quadratic Form, J. Amer. Statist. Assoc., **67**, 340, pp. 898-902, [Correction: **68**, p. 757]
- Johnson, N.L., Kotz, S.(1970) Continuous Univariate Distributions-1, J. Wiley, New York
- Khatri, C.G., Krishnaiah, P.R., Sen, P.K.(1977) A Note on the Joint Distribution of Correlated Quadratic Forms, J. Statist. Plann. Inference, **1**, pp. 299-307
- Kruskal, J.B.(1956) On the Shortest Spanning Subtree of a Graph and the Travelling Salesman Problem, Proc. Amer. Math. Soc., **7**, pp. 48-50
- Künsch, H.(1984) Infinitesimal Robustness for Autoregressive Processes, Ann. Statist., **12**, 3, pp. 843-863
- Lattin, J.M.(1983) Identifying Influential Observations in Time Series Data, in Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface, Heiner, K.W. et al. (Eds.), Springer-Verlag, Berlin, pp. 283-286
- Leadbetter, M.R., Lindgren, G., Rootzén, H.(1983) Extremes and Related Properties of Random Sequences and Processes, Springer-Verlag, Berlin
- Li W.K., Hui, Y.V.(1987) On the Empirical Influence Functions of Residual Autocorrelations in Time Series, Amer. Statist. Assoc. Proc. Bus. Econom. Statist. Section, pp. 465-468
- Ljung, G.M., Box, G.E.P.(1978) On a Measure of Lack of Fit in Time Series Models, Biometrika, **65**, 2, pp. 297-303
- Luke, Y.L.(1969) The Special Functions and Their Applications Vol. 2, Academic Press, New York
- Martin, R.D., Raftery, A.E.(1987) Comment: Robustness, Computation, and Non-Euclidian Models, J. Amer. Statist. Assoc., **82**, 400, pp. 1044-1050

- Martin, R.D., Yohai, V.J.(1986) Influence Functionals for Time Series (with discussion), Ann. Statist., 14, pp. 781-855
- McLeod, A.I., Jiménez, C.(1984) Nonnegative Definiteness of the Sample Autocovariance Function (with discussion), Amer. Statist., 38, 4, pp. 297-298, 39, 3, pp. 237-238
- Miller, K.S.(1965) Some Multivariate Density Functions of Products of Gaussian Random Variates, Biometrika, 52, pp. 645-646
- Miller, R.G.(1981) Simultaneous Statistical Inference, Second Edition, Springer-Verlag, Berlin
- Mosteller, F., Tukey, J.W.(1977) Data Analysis and Regression Addison-Wesley, Reading, Mass.
- Olshen, A.C.(1938) Transformations of the Pearson Type III Distribution, Ann. Math. Statist., 9, pp. 176-200
- Pan, J.J.(1964) Distributions of the Noncircular Serial Correlation Coefficients, Shuxue Jinzhan, 7, pp. 328-337 (in Chinese), Also in Selected Translations in Mathematical Statistics and Probability, 7, 1968, American Mathematical Society, pp. 281-292
- Peña, D.(1986) Discussion of Cook(1986), J. Roy. Statist. Soc. Ser. B, 48, 2, pp. 164-165
- Peña, D.(1987) Measuring the Importance of Outliers in ARIMA Models, in New Perspectives in Theoretical and Applied Statistics, Puri, M.L.(Ed.), J. Wiley, New York, pp. 109-118
- Polasek, W.(1982) Robust Estimation and Resistance Analysis for the Auto-Correlation Function, Preprint #47, Institute of Statistics and Computer Science, University of Vienna, Austria.
- Pregibon, D.(1981) Logistic Regression Diagnostics, Ann. Statist., 9, 4, pp. 705-724
- Prescott, P.(1986) Discussion of Cook(1986), J. Roy. Statist. Soc. Ser. B, 48, 2, p. 161

- Rey, W.J.J.(1983) Introduction to Robust and Quasi-Robust Statistical Methods, Springer-Verlag, Berlin
- Rousseeuw, P.J., Leroy, A.M.(1987) Robust Regression and Outlier Detection, J. Wiley, New York
- Schwager, S.J.(1984) Bonferroni Sometimes Loses (with discussion), Amer. Statist., **38**, 3, pp. 192-197, **39**, 3, pp. 235-236
- Solomon, H., Stephens, M.A.(1977) Distribution of a Sum of Weighted Chi-Square Variables, J. Amer. Statist. Assoc., **72**, 360, pp. 881-885
- Theil, H.(1965) The Analysis of Disturbances in Regression Analysis, J. Amer. Statist. Assoc., **60**, 312, pp. 1067-1079
- Tsay, R.S.(1988) Outliers, Level Shifts, and Variance Changes in Time Series, J. Forecasting, **7**, 1, pp. 1-20
- Welsch, R.E.(1982) Influence Functions and Regression Diagnostics, *in* Modern Data Analysis, Launer, R.L., Siegel, A.F. (Eds.), Academic Press, New York, pp. 149-169
- Wilson, E.P., Hilferty, M.M.(1931) The Distribution of Chi-Square, Proc. Nat. Acad. Sci., **17**, pp. 684-688

END

21/09/90

FIN