Acknowledgement

I would like to thank my supervisor, Professor Stephan Morgenthaler, for having given me the opportunity to pursue my doctoral studies at EPFL and then being my mentor in academia. In these past four years, he has not only been my thesis advisor but also part of my support system that has made my stay here in Switzerland so comfortable and fruitful. While his guidance and encouragement have made this thesis possible in the first place, his warm and friendly outlook has made this long journey so much more pleasant and memorable. From Stephan I have learnt more about, not only Statistics, but also patience, compassion and a myriad of other aspects of life. While this knowledge of Statistics will be part of my professional life forever and I am extremely grateful to Stephan for it, it is the better understanding of various aspects of life that have enriched my personal life for which I will be indebted to Stephan forever.

I would also like to thank Professor Anthony Davison, Professor Olivier Renaud and Professor Stefan Van Aelst for agreeing to be members of the defense jury and Professor John Maddocks for presiding over as the jury president.

Coming back to that support system, I would like to thank Anne-Lise Courvoisier for her support and immense help right from my very first day in Switzerland. I'd like to thank my past and present colleagues and friends from EPFL : Andreas, Sandro, Baptiste, Sahar, Ismail, Vahid, Mehdi, Nicolas, Jean-Marc, Sonja, Maya, Laurence and many many more, for my pleasant and memorable time at STAP. I'd also like to thank my other friends for making my stay in Switzerland, a pleasant one.

Last but definitely not the least, I'd like to thank my family for their encouragement and steadfast support during these past four years. In particular, I'd like to thank my father and mother for their belief in me and constant encouragement all through my life. But for them, I'd probably have taken another path in life. Finally, I'd like to immensely thank my wife, Krupa, for her unconditional support all through. For the past four years, she has stood by me and been a solid pillar of strength through smooth and rough times. Her presence and support have filled this long journey of mine with many pleasant surprises and events that I could only have dreamt of otherwise.

Abstract

Time series modeling and analysis is central to most financial and econometric data modeling. With increased globalization in trade, commerce and finance, national variables like gross domestic productivity (GDP) and unemployment rate, market variables like indices and stock prices and global variables like commodity prices are more tightly coupled than ever before. This translates to the use of multivariate or vector time series models and algorithms in analyzing and understanding the relationships that these variables share with each other.

Autocorrelation is one of the fundamental aspects of time series modeling. However, traditional linear models, that arise from a strong observed autocorrelation in many financial and econometric time series data, are at times unable to capture the rather nonlinear relationship that characterizes many time series data. This necessitates the study of nonlinear models in analyzing such time series. The class of bilinear models is one of the simplest nonlinear models. These models are able to capture temporary erratic fluctuations that are common in many financial returns series and thus, are of tremendous interest in financial time series analysis.

Another aspect of time series analysis is homoscedasticity versus heteroscedasticity. Many time series data, even after differencing, exhibit heteroscedasticity. Thus, it becomes important to incorporate this feature in the associated models. The class of conditional heteroscedastic autoregressive (ARCH) models and its variants form the primary backbone of conditional heteroscedastic time series models.

Robustness is a highly underrated feature of most time series applications and models that are presently in use in the industry. With an ever increasing amount of information available for modeling, it is not uncommon for the data to have some abberations within itself in terms of level shifts and the occasional large fluctuations. Conventional methods like the maximum likelihood and least squares are well known to be highly sensitive to such contaminations. Hence, it becomes important to use robust methods, especially in this age with high amounts of computing power readily available, to take into account such abberations.

While robustness and time series modeling have been vastly researched individually in the past, application of robust methods to estimate time series models is still quite open. The central goal of this thesis is the study of robust parameter estimation of some simple vector and nonlinear time series models.

More precisely, we will briefly study some prominent linear and nonlinear models in the time series literature and apply the robust S-estimator in estimating parameters of some simple models like the vector autoregressive (VAR) model, the (0, 0, 1, 1) bilinear model and a simple conditional heteroscedastic bilinear model. In each case, we will look at the important aspect of stationarity of the model and analyze the asymptotic behavior of the S-estimator.

Keywords : Vector models, multivariate time series, robust estimation, outlier propagation, stationarity, vector autoregression, bilinear series, conditional heteroscedasticity, S-estimator, Fast-S.

Résumé

La modélisation et l'analyse de séries temporelles est un sujet fondamental des mathématiques financieres et de la modélisation de données économiques. Avec la mondialisation accrue des échanges, du commerce et de la finance, les variables nationales telles que le produit intérieur brut (PIB), le taux de chômage, des variables telles que les indices et les cours boursiers ainsi que des variables globales telles que les prix des produits de base sont de plus en plus étroitement liées. Cela se traduit par l'utilisation de plusieurs variables ou des modèles de séries temporelles vectorielles dans l'analyse et la modélisation.

L'autocorrélation est un des aspects fondamentaux de la modélisation des séries temporelles. Toutefois, les modèles traditionnels linéaires, qui sont inspirés de la forte autocorrélation des données financières et économétriques, ne sont parfois pas en mesure de saisir la relation, plutôt nonlinéaire qui caractérise de nombreuses séries temporelles. La classe de modèles bilinéaires est l'un des modèles non linéaire le plus simple. Ces modèles sont capables de capturer des fluctuations erratiques qui sont courantes dans de nombreuses séries de rendements financiers et, pourtant, sont d'un grand intérêt dans l'analyse financière.

Un autre aspect de l'analyse des séries temporelles est le contraste entre la variation constante (homoscédasticité) et la variation elle-même variable dans le temps (hétéroscédasticité). Beaucoup de séries temporelles, même après la dérivation, montrent de telles sous-structures de la variabilité. Ainsi, il devient important d'intégrer cette fonctionnalité dans les modèles associés. La classe d'hétéroscédasticité condi-

tionnelle autorégressive (Les modèles ARCH) et ses variantes constituent l'ossature primaire de tels modèles.

La robustesse est une caractéristique sous-estimée dans la plupart des applications des séries temporelles et de modèles qui sont actuellement en usage dans l'industrie. Avec le volume croissant d'informations disponibles pour la modélisation, il n'est pas rare pour les données de contenir des aberrations en termes de changements de niveau et les fluctuations occasionnelles large. Les méthodes conventionnelles comme le maximum de vraisemblance et des moindres carrés sont bien connues pour être très sensibles à de telles contaminations. Il devient donc important d'utiliser des méthodes robustes, qui limitent l'influence de telles données.

Alors que la robustesse et la modélisation des séries temporelles ont été largement étudiées individuellement dans le passé, l'application de méthodes robustes pour estimer les séries temporelles est encore ouvert. L'objectif central de cette thèse est l'étude de l'estimation des paramètres robustes de certains vecteurs simples et non linéaire des modèles de séries temporelles.

Plus précisément, nous allons brièvement étudier quelques-uns des modèles linéaires et non linéaires de premier plan dans la littérature des séries temporelles et d'appliquer le S-estimateur robuste l'estimation des paramètres de certains modèles simples comme le vecteur autorégressif (VAR) modèle, le modèle (0, 0, 1, 1) bilinéaire et un modèle bilinéaire simple hétéroscédastiques conditionnel. Dans chaque cas, nous étudions l'aspect important de la stationnarité du modèle et analysons le comportement asymptotique du S-estimateur.

Mots-clés : Modèles vectoriels, séries multivaries, estimation robuste, outlier propagation, stationnarité, autorégression vectorielle, séries bilinéaire, hétéroscédasticité conditionnelle, S-estimateur, Fast-S.

Contents

CHAPTER 1

Introduction

1.1 Time series - Motivation and definition

Time series analysis is central to many financial and econometric applications. Simply put, a time series refers to any indexed dataset $\{\mathbf{x_t} : t = 1, .., T\}$ where $\mathbf{x_t}$ is a d-dimensional vector that represents the value of some variable at time t. Daily closing prices of the IBM stock in the year 2008 is an example of a time series. The annual gross domestic output (GDP) of Switzerland for the years between 1960 and 2008 is another example of a time series. When $d = 1$, the series is termed univariate while in the case of $d > 1$, it is termed multivariate. Figure 1.1 shows the daily log returns (based on the daily closing price) of the IBM stock listed on the New York stock exchange (NYSE). For further discussion, we will restrict ourselves to the univariate case.

Statistical analysis of time series generally involves studying the evolution of data over time. The aim of such a study could be to forecast future movements, discover any underlying driving factors or variables, or simply better understand the dynamics of the series in terms of its variance and other characteristics. Considering the examples cited before, one could be interested in forecasting the annual gross domestic product (GDP) of Switzerland for 2009 or perhaps understanding how the IBM and NASDAQ composite daily closing affect each other's present and future movements. One could also be interested in quantifying the volatility of the daily

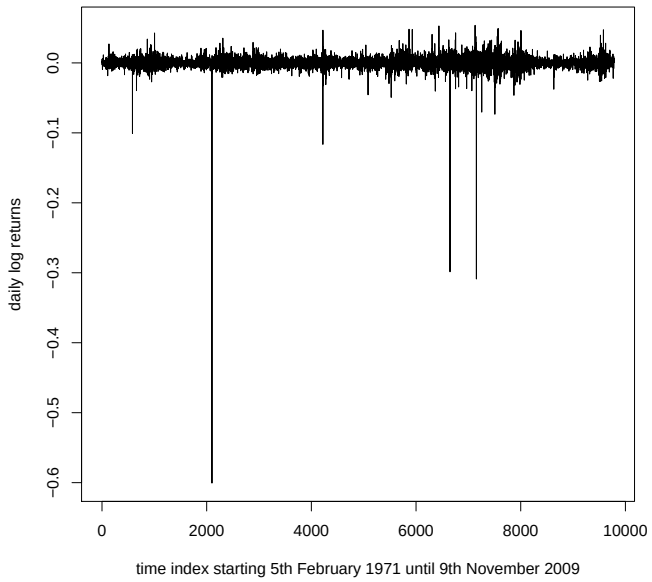

Figure 1.1: The daily log returns of the IBM stock on the NYSE. The return of a stock is defined as $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ where P_t is the closing price on day t.

closing price of IBM shares and use that in pricing some derivative.

1.2 Modeling

Studying a time series typically involves assuming a dependence of future values on past values by way of some stochastic functions. In the finance parlance, this is known as technical analysis. Consider a stochastic process $\{X_t\}_{t\in\mathcal{Z}}$. Let $\{x_t\}_{t\in\mathcal{Z}}$ be a realization of this process. Then, a simple mathematical model for this stochastic process is :

$$X_t = f_t(X_{t-1}, X_{t-2}, ...) + g_t(X_{t-1}, X_{t-2}, ...)\epsilon_t \tag{1.1}$$

where

1. $\{\epsilon_t : t = 1, ..., T\}$ are independent and identically distributed (i.i.d) random variables having some distribution F_ϵ with zero mean and unit variance,

2. f_t and g_t are measurable functions that govern the conditional mean and variance respectively, of x_t. That is, the conditional mean and standard deviation of X_t given $(X_{t-1}, X_{t-2}, ...)$ are $f_t(x_{t-1}, x_{t-2} \ldots,)$ and $g_t(x_{t-1}, x_{t-2} \ldots,)$ respectively.

ϵ_t is the new information at time t and is often referred to as the *innovation* or *shock* at time t. The series $\{\epsilon_t\}$ itself is referred to as a white noise series. In simple models, f_t and g_t are linear combinations of past shocks. However, as will be seen later in the chapter on nonlinear time series modeling, they could also contain nonlinear combinations of past shocks. Stochastic processes are not usually written in this form. If one attempts to do it, regularity conditions are required to assure existence of a corresponding stochastic process.

1.3 Homoscedasticity and heteroscedasticity

The model given in Equation (1.1), in its present form, is of little use since it is too general. Hence, most time series models assume f_t and g_t to be independent of t as in $f_t \equiv f$ and $g_t \equiv g$. This is the first level of tractability since we now have two rather than $2(T-1)$ functions to deal with.

Further tractability is achieved by assuming some forms for f and g. In the most trivial case, we can assume $f(.) \equiv c_f$, a constant, and $g(.) \equiv c_g$, another constant.

In this case, we have a sequence of i.i.d random variable realizations and we can use the empirical data characteristics to summarize the time series. In particular, when $c_f = 0$ and $c_g = \sigma$, the series $\{x_t\}$ is called a *white noise* series with variance σ^2. If g is not constant, then the time series is termed *conditional heteroscedastic*. In other words, when the conditional variance of x_t is not time-invariant, the series is called conditional heteroscedastic. Sometimes, this definition is simply referred to as heteroscedasticity. In general, heteroscedasticity refers to the time varying nature of the marginal variance of x_t. In contrast, when $g \equiv c_g$, a constant, the series is termed *homoscedastic*. In this thesis, we will use the terms heteroscedasticity as well as conditional heteroscedasticity to refer to the time variant nature of the conditional variance of a series. Figure 1.2 illustrates the heteroscedastic nature of the daily NASDAQ index returns.

1.4 Multivariate analysis

In many econometric applications, it is not unusual to study many time series together. This is where considering $\mathbf{x_t}$ as a vector comes in handy and thus multivariate time series analysis comes into play. Examples for multivariate time series in econometrics include population growth, GDP and unemployment rate since these are highly interrelated. In financial applications, one might be interested in studying many indices like the SMI, NASDAQ and BSE, together. Figure 1.3 shows the simultaneous evolutions of the daily log returns of IBM and NASDAQ. Even though the overall trends of the two series do not show any relationship, they seem to tend to move together locally. This suggests that the two series may have some dependence. This is valuable information that can be utilized in studying the series by considering them jointly as opposed to in isolation. Hence vector time series modeling is an important tool in financial applications.

1.5 Robustness

We saw that in time series modeling, we assume some kind of dependence structure in the data. The next natural step is to try and find this structure. This is where some of the major hurdles arise. The main reasons for this can be listed as follows :

1. The dependence structure we assume may not be the right one in the first place. For example, we may assume a linear dependence, i.e., autoregression, while the actual dependence may be quadratic.

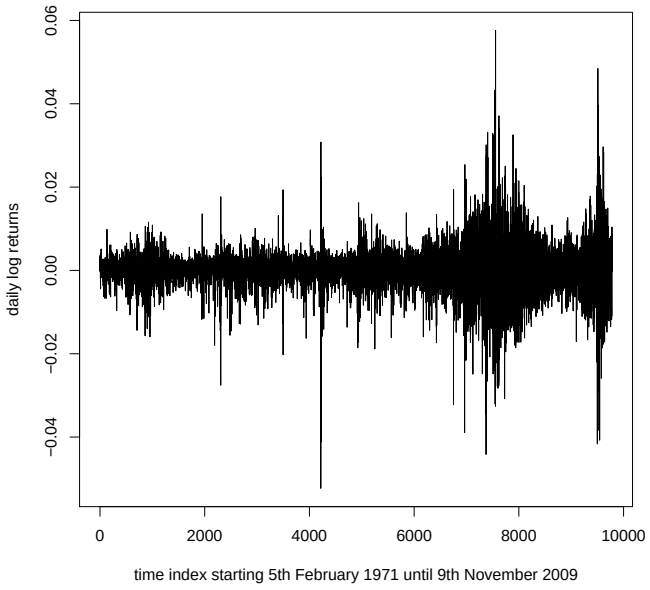

Figure 1.2: The daily NASDAQ log returns.

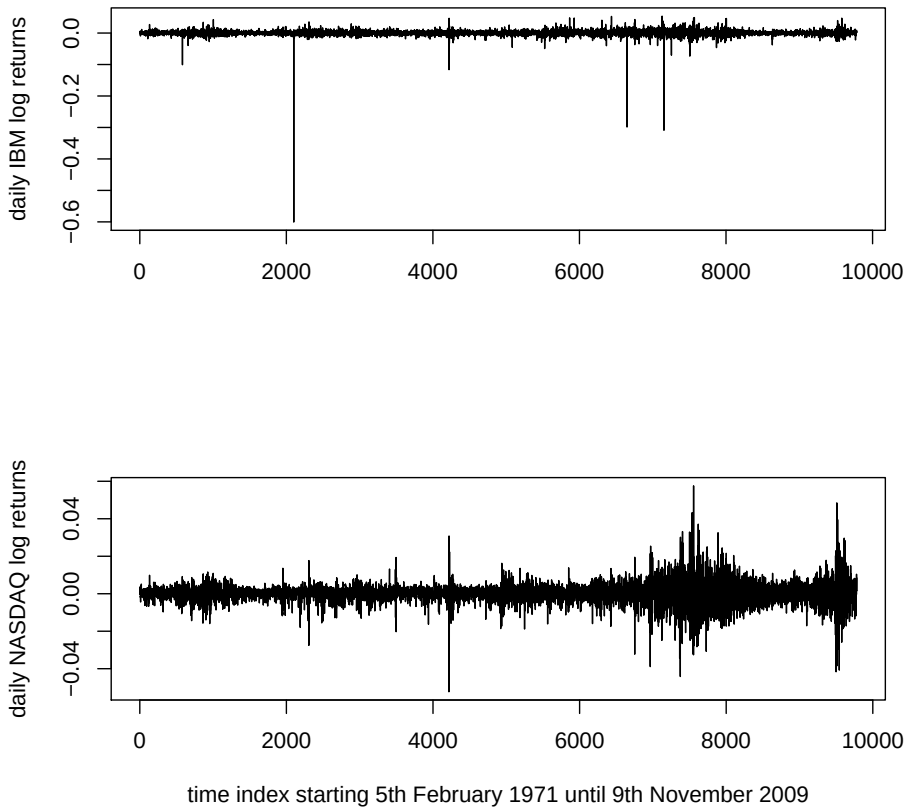

Figure 1.3: The daily log returns for IBM and NASDAQ.

2. The assumption of the nature of F_ϵ, which is also important in modeling the series, may not be accurate.

3. There may be aberrant data in the series that may not behave according to the model that is adequate for the majority of the data. Such data are called **outliers** in the statistics parlance.

The first problem mentioned can be tackled to a large extent by running simple tests to see if the data fits into the proposed structure. For example, to check for linear dependence, one can look at empirical correlations and significance tests to see their strengths. One can fit the data to the model by estimating the parameters and then test the residuals to see how they behave.

The second problem mentioned can be tackled to some extent by plotting the residual series (the series of the residuals) and analyzing its behavior.

The third problem is where the actual method employed to find the dependence structure, comes into play. This is because depending on the sensitivity of the method to outliers, one could get misled to a completely different structure. This can be seen in Figure 1.4 where the method of least squares, which is known to be highly sensitive to even a single outlier, estimates the single parameter (the slope of the line, in red) of a linear regression problem, with a large error. The correct parameter is represented by the blue line. This is where robustness comes into the picture. The way to tackle the problem with outliers is to have the methods robust enough so as to withstand some amount of contamination in the data. Ideally, one would like to have methods that withstand up to 50% contamination since any further contamination renders the data without the structure that we are interested in finding in the first place.

The theory of robust statistics is as old as that of statistics itself. However, robust methods have not been very popular in actual applications until very recently. This is because the insensitivity of these methods to outliers comes at the cost of efficiency. Robust methods typically involve complex calculations that are difficult and time consuming. Also, there are no closed form expressions for most robust estimators and in almost all cases, one depends on some heuristics to compute the estimator. However again, with recent advents in computing technologies, complex computations is no longer an issue and hence robust statistics is gaining more importance now.

While time series analysis and robust statistics have both been studied extensively, applications of robust methods in time series analysis still has some open areas. In

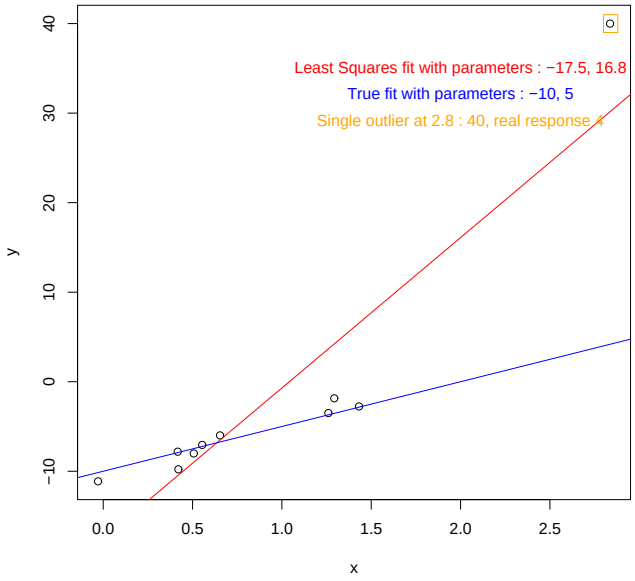

Figure 1.4: Effect of a single outlier on the Least Squares Fit.

particular, application of robust methods in multivariate time series models' estimation has been somewhat limited. The work by Li and Hui (1989) [33] looks at robust multivariate linear time series modeling. In particular, they study the application of the residual autocovariance (RA) method to the problem of parameter estimation in vector autoregressive (VAR) models. Ben et al. (1998) [5] further refined this method to make it affine equivariant (independent of the choice of co-ordinate axes of the data). Further, Ben et al. (2001) [6] proposed the τ-estimator for a VAR model which is also affine equivariant. More recently, Croux and Joossens (2008) [18] applied a multivariate least trimmed squares (MLTS) method to estimate parameters of VAR models. The central aim of this thesis is to study the VAR model and examine the application of a special τ-estimator, the S-estimator, in estimating parameters of this model. The reason for choosing the S-estimator is its good robustness and computational properties. It has a breakdown point of 50%, an asymptotic convergence rate of $O(\sqrt{n})$ and its objective function can be computed in O(n) time. In addition, the Fast-S method of Yohai and Salibian-Barrera (2006) [76] is an iterative algorithm for computing an approximation of the S-estimator in a reasonable time. This algorithm will be adapted to the multivariate regression scenario as is the case in VAR models.

1.6 Nonlinear modeling and conditional heteroscedasticity

The form of the conditional mean equation, $f(.)$, of a time series has been studied primarily within the linear framework. However, there is a need for inclusion of nonlinear terms that is driven by observed nonlinearity in many financial time series data. By considering stochastic parameters in a traditional linear model, one can think of modeling conditional heteroscedastic time series data in a simple yet more effective manner than simple linear modeling. Such a consideration leads one to the well known bilinear model. The bilinear model of Granger and Andersen (1978) [24] is one of the few nonlinear models that considers a quadratic form for f. Another aim of this thesis is to study the bilinear model. In particular, application of the S-estimator in parameter estimation will be examined for the univariate bilinear model. In addition, we will briefly examine the vector bilinear model specification.

Conditional heteroscedasticity in time series is a widely observed phenomenon and the literature in this area is vast. The autoregressive conditional heteroscedastic (ARCH) and generalized ARCH (GARCH) are some of the famous models that are

central to conditional heteroscedastic time series modeling. The main idea of these models is to fit an autoregressive model to the squared residual series. However, estimating parameters of these models jointly with those of the conditional mean equation is not straightforward. The final aim of this thesis is to study a particular kind of conditional heteroscedastic model that considers a simple linear conditional mean equation and is easy to deal with in estimating and interpreting the parameters. In particular, the stationarity conditions and robust parameter estimation using the S-estimator will be considered.

The analysis of common robust estimators in nonlinear regression problems has been studied to a large extent. Stromberg and Ruppert(1992) [64] have analyzed the breakdown points of the least squares and least median of squares estimators in some nonlinear regression problems. Sakata and White (2001) [63] further studied the properties of the S-estimator in the nonlinear regression context and found it to be resistent and consistent. Hence our persistence with the S-estimator for the robust estimation of parameters of the bilinear and conditional heteroscedastic models as well.

1.7 Outline

To summarize, there are three aims of this thesis which are as follows :

1. To study the VAR model and apply the S-estimator in robustly estimating its parameters.

2. To study the class of univariate bilinear models and apply the S-estimator in robustly estimating the parameters of a simple model in this class.

3. To study conditional heteroscedasticity in time series by analyzing a conditional heteroscedastic bilinear model that integrates linear conditional mean functions and a simple ARCH type conditional variance function.

Chapter 2 gives a short introduction to time series analysis that includes multivariate time series as well. In particular, it talks about, among other topics, concepts of stationarity and autocorrelation that are central to time series analysis. It also briefly describes the need for robust models in time series analysis and in particular, discusses the M-estimator. In Chapter 3 we will see how to compute an S-estimator for a VAR model using the Fast-S method of Yohai and Salibian-Barrera (2006) [76]. Chapter 4 talks about bilinear models in time series applications. In chapter

5, we will look at a conditional heteroscedastic diagonal bilinear model to handle conditional heteroscedasticity in time series. Finally, we will summarize the analysis in the concluding chapter.

CHAPTER 2

A short introduction to robust linear time series analysis

2.1 Introduction

The data used in the examples of this thesis are mainly financial. It is therefore good to briefly look at some of the basic terminology and methods used when dealing with financial data.

Financial data mostly refers to indices, prices and returns (of assets as well as commodities) and also macroeconomic variables like GDP and interest rates. Most applications in this domain deal with returns rather than actual prices. Amongst the reasons for this are two important ones. Firstly, returns are dimensionless and therefore easily comparable across industries and sectors. Secondly, returns exhibit interesting statistical properties like stationarity (which will be discussed later) that prices do not. A further refinement involves dealing with log returns rather than returns itself since properties of log returns are more tractable. For example, the variance of log returns is smaller than that of returns. Also, a multi-period log return is simply the sum of simple one-period log returns whereas in the case of returns, it is the product.

Mathematically, consider the price series $\{p_t : t = 1, \ldots, T\}$, where t denotes the time index. A time index refers to a particular time in a series of equally spaced times. A series may be hourly, weekly, monthly, yearly, etc.. For example, we could have $T = 52$ where each index represents the week in the year 2008. Then, the

one-period simple return, q_t, is given by

$$q_t = \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1.$$

The simple one-period log return, x_t, is given by

$$x_t = \log(1 + q_t) = \log\big(\frac{p_t}{p_{t-1}}\big).$$

Note that when $p_t = 0$ for some t, then all the following prices, $\{p_s : s > t\}$, are also all equal to zero, which is a property of typical financial data like prices. Such a scenario would occur when a company goes bankrupt, for example.

Log returns are typically intra-correlated. That is, there is a correlation between x_t and x_{t-1}. This can be seen in typical bull and bear runs in a stock market. This is termed as autocorrelation which is described is the next section. Then, one can also observe trends and seasonal variation in many return series. For example, a typical index like NASDAQ tends to have a long term positive trend. Commodity prices always have a long term positive trend due to inflation. Sales data typically exhibit seasonal trends. Sales generally go up just around festival times which is followed by a period of lull.

Time series analysis thus involves studying various characteristics of data like trends, seasonality and autocorrelation. When empirical evidence suggests a trend or seasonal component, it is usually removed before analyzing the data further for the presence of autocorrelation.

In this chapter, we will briefly examine the fundamentals of time series modeling and look at some simple linear models. In addition, we will also look at robust analysis of time series problems. In the following sections, we will look at some of these concepts, namely, autocorrelation, stationarity and causality.

2.2 Autocorrelation

Time series models are generally concerned with the conditional mean function, f, and the conditional variance function, g. Since $f \equiv c_f$, a constant, is mathematically too simple and empirically very rare if not impossible, we look at the next most simple function - the linear function :

$$f(X_{t-1}, X_{t-2}, ...) = \mu + \sum_{i=1}^{t-1} \phi_i X_{t-i} \tag{2.1}$$

where μ and ϕ_i are constants. This form of f appeals to the statistician because it captures dependence in a time series in the most simple way - linear dependence. There is ample empirical evidence of linear dependence within many time series data and hence to study f in this form seems like a good idea. Restricting f to the linear family is referred to as linear time series analysis.

Linear dependence is synonymous with correlation. When correlations exist in a time series, it is termed as *autocorrelation* or *serial correlation*. For a given time series $\{x_t : t = 1, .., T\}$, $corr(x_t, x_s) = \rho_{ts}$ is the correlation between x_t and x_s. When $\mathbf{x_t}$ is a vector, $corr(x_{ti}, x_{sj})$ is termed as *cross-correlation*, while $corr(x_{ti}, x_{tj})$ is termed as *concurrent-correlation*.

2.3 Stationarity, causality and invertibility

The concept of stationarity is central to time series analysis. A time series $\{x_t\}$ is said to be *strictly stationary* if the joint distribution of $(x_{t_1}, \ldots, x_{t_k})$ is identical to that of $(x_{t_1+s}, \ldots, x_{t_k+s})$ for all s where k is an arbitrary positive integer and $(t_1, \ldots, t_k)$ is a collection of k positive integers. Put succinctly, strict stationarity requires the joint distribution of $(x_{t_1}, \ldots, x_{t_k})$ to be invariant under time shift. This is a rather strong condition that is difficult to verify empirically.

$\{x_t\}$ is said to be *weakly stationary* if both the mean of x_t and the covariance between x_t and x_{t+l} are finite and time invariant, where l is an arbitrary integer. More precisely, weak stationarity requires that $\mathrm{E}[x_t] = \mu$, a constant, and $Cov(x_t, x_{t+l}) = \gamma_l$, which only depends on l, the lag. In applications, weak stationarity enables one to construct forecast models for time series data. Henceforth, we will refer to a weakly stationary series as a stationary series.

Another form of stationarity is the *periodic stationarity*. $\{x_t\}$ is said to be periodically stationary with period d, a positive integer, if $\mathrm{E}[x_t] = \mu_t$ and $Cov(x_t, x_{t+k}) = \gamma_t(k)$ for all integers k, are both bounded and periodic functions of t with the same period d. Examples of periodically stationary time series data are sales data which typically have a monthly and/or yearly period since most people tend to spend more during the beginning of the month (hence monthly) when they have the money as

well as during festive seasons (hence yearly).

Causality is another important concept in the domain of time series analysis. A time series $\{x_t\}$ having an associated model

$$X_t = f(X_{t-1}, X_{t-2}, ...) + \epsilon_t,$$

where f is the conditional mean function and $\{\epsilon_t\}$ is a white noise process, is said to be causal if there exists a sequence, (ϕ_i), such that

$$\sum_{i=0}^{\infty} |\phi_i| < \infty$$

and

$$x_t = \sum_{i=0}^{\infty} \phi_i \epsilon_{t-i}. \tag{2.2}$$

Invertibility is yet another aspect of time series analysis. A time series $\{x_t\}$ with an associated model f given by

$$x_t = f(x_{t-1}, ..., x_{t-p}) + \epsilon_t$$

where $\{\epsilon_t\}$ is a white noise process with zero mean, is termed invertible if there exist constants $\pi_0, \pi_1, \ldots$ such that

$$\epsilon_t = \pi(B) x_t$$

and

$$\sum_{j=0}^{\infty} |\pi_i| < \infty$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_i B^i$ and B is the back-shift operator, i.e., $B(x_t) = x_{t-1}$.

2.4 The linear time series model

A time series $\{x_t\}$ is said to be linear if it can be written as

$$x_t = \mu + \sum_{i=0}^{\infty} \phi_i \epsilon_{t-i}, \tag{2.3}$$

where $\phi_0 = 1$,

$$\sum_{i=0}^{\infty} |\phi_i| < \infty$$

and $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with mean zero and a well-defined distribution, i.e., $\{\epsilon_t\}$ is a white-noise series. It is easy to see that a linear time series is causal, since the mean can be adjusted with the white noise process in order for Equation (2.2) to be satisfied.

If $\{x_t\}$ as defined in Equation (2.3) is stationary, then its mean and variance can be obtained as

$$E(x_t) = \mu$$

and

$$Var(x_t) = \sigma_\epsilon^2 \sum_{i=0}^{\infty} \phi_i^2$$

where σ_ϵ^2 is the variance of ϵ_t. Since $\{x_t\}$ is stationary, $\text{Var}(x_t) < \infty$ and so we must have that $\{\phi_i^2\}$ is a convergent series converging to zero, that is, $\phi_i^2 \to 0$. Hence, for a stationary series, the impact of the remote shock ϵ_{t-i} on the current value x_t diminishes with i.

After some simple calculations, it can be seen that the lag-l autocovariance is given by

$$\gamma_l = Cov(x_t, x_{t+l}) = \sigma_\epsilon^2 \sum_{i=0}^{\infty} \phi_i \phi_{i+l}. \tag{2.4}$$

Consequently, the lag-l autocorrelation is given by

$$\rho_l = Corr(x_t, x_{t+l}) = \frac{\gamma_l}{\gamma_0} = \frac{\sum_{i=0}^{\infty} \phi_i \phi_{i+l}}{1 + \sum_{i=1}^{\infty} \phi_i^2}. \tag{2.5}$$

Since, under stationarity, $\phi_i^2 \to 0$, it's clear from the above equation that $\rho_l \to 0$ as $l \to \infty$.

2.5 The Autoregressive (AR) Model

The fact that the monthly return x_t of the Bombay Stock Exchange (BSE) index has a statistically significant lag-1 autocorrelation indicates that the lagged return x_{t-1} might be useful in predicting x_t. A simple linear model that makes use of such predictive power is the AR(1) model defined as :

$$x_t = \phi_0 + \phi_1 x_{t-1} + \epsilon_t, \tag{2.6}$$

where $\{\epsilon_t\}$ is a white noise series with zero mean and variance σ_ϵ^2. This model is in the form of the simple linear regression model with x_t being the dependent variable and x_{t-1} the explanatory variable.

The AR(1) model described has some interesting properties. From Equation (2.6), the conditional distribution of the return x_t given x_{t-1} is given by

$$E(x_t|x_{t-1}) = \phi_0 + \phi_1 x_{t-1}, \quad \text{and} \quad Var(x_t|x_{t-1}) = Var(\epsilon_t) = \sigma_\epsilon^2.$$

That is, given the past return x_{t-1}, the present return is centered around $\phi_0 + \phi_1 x_{t-1}$ with standard deviation σ_ϵ. This is a Markov property since conditional on x_{t-1}, the return x_t does not depend on x_{t-i} for $i > 1$.

The AR(1) model considers the immediate past, i.e., the lag-1, return to determine the present return. A natural generalization of this is the AR(p) model that takes into account the last p returns to determine the distribution of the present return as follows :

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t, \tag{2.7}$$

where p is a positive integer and $\{\epsilon_t\}$ is as defined in Equation (2.6). Again, this model is in the form of a multiple linear regression model with the p lagged values acting as the explanatory variables and the current return as the response variable.

2.5.1 Properties of AR models

AR(1) model

In this section, we will look at some basic properties of AR models. We start with the most simple AR(1) model. We begin by looking at the stationarity properties of

this model. First, we look at the mean of the series. Suppose the model is causal and stationary. Then, we must have

$$\mu = E(x_t) = \phi_0 + \phi_1 E(x_{t-1}) = \phi_0 + \phi_1 \mu.$$

Hence, $\mu = \frac{\phi_0}{1-\phi_1}$. For this to make sense, we must have $\phi_1 \neq 1$. Next, we look at the variance of the series. Under stationarity, we must have

$$\sigma_x^2 = Var(x_t) = \phi_1^2 \text{var}(x_{t-1}) + \sigma_\epsilon^2 = \phi_1^2 \sigma_x^2 + \sigma_\epsilon^2$$

since $Cov(\epsilon_t, x_{t-1}) = 0$ because ϵ_t is independent of the past (recall that ϵ_t is the innovation at time t and the process is causal). Hence, we have that $\sigma_x^2 = \frac{\sigma_\epsilon^2}{1-\phi_1^2}$. For this to make sense, we must have $|\phi_1| < 1$. The stationarity as well as mean and variance results can also be obtained in another way.

Considering $\phi_0 = (1 - \phi_1)\mu$ and $|\phi_1| < 1$, the AR(1) process can be rewritten as

$$s_t = x_t - \mu = \phi_1(x_{t-1} - \mu) + \epsilon_t = \phi_1 s_{t-1} + \epsilon_t.$$

By repeated substitutions, we have that

$$s_t = x_t - \mu = \sum_{i=0}^{\infty} \phi_1^i a_{t-i}.$$

From the above equation, it is clear that the unconditional mean and variance of this AR(1) process is exactly as derived earlier. In addition, from Equation (2.5), it is clear that the lag-l autocovariance does not depend on t and is finite. Hence, the necessary and sufficient condition for stationarity of an AR(1) process is that $|\phi_1| < 1$.

We now look at the autocovariance in an AR(1) model. From Equation (2.6), it can be deduced that

$$\gamma_l = \phi_1 \gamma_{l-1} \quad \text{for} \quad l > 0. \tag{2.8}$$

From the above, it is clear that

$$\rho_l = \phi_1^l \quad \text{for} \quad l \geq 0.$$

That is, the autocorrelation of a stationary AR(1) process decays exponentially with rate ϕ_1.

AR(2) model

We will now examine an AR(2) model to see how the addition of a single lag affects the properties of the model. An AR(2) model can be written as

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t, \tag{2.9}$$

where ϵ_t is as defined in Equation (2.6). Using the same technique we used in the AR(1) case, we obtain, under the assumption of stationarity,

$$\mu = E(x_t) = \frac{\phi_0}{1 - \phi_1 - \phi_2},$$

for $\phi_1 + \phi_2 \neq 1$. For the variance and autocovariance, we can use the well known Yule-Walker equations. From the model definition, the following sets of equations can be arrived at :

$$\sigma_x^2 = \mathrm{Var}(x_t) = (\phi_1^2 + \phi_2^2)\sigma_x^2 + \sigma_\epsilon^2 + 2\phi_1\phi_2\gamma_1$$

and

$$\gamma_1 + \mu^2 = \phi_0\mu + \phi_1(\sigma_x^2 + \mu^2) + \phi_2(\gamma_1 + \mu^2) \quad \text{since} \quad \gamma_l = \gamma_{-l}.$$

From the above two equations, γ_1 and σ_x can be found as

$$\sigma_x^2 = \frac{\sigma_\epsilon^2}{1 - (\phi_1^2 + \phi_2^2 + 2\theta\phi_1\phi_2)}$$

and

$$\gamma_1 = \theta\sigma_x^2,$$

where $\theta = \frac{\phi_1}{1-\phi_2}$. From the autocovariance equation, we get the recurrence relation for the autocorrelation as

$$\rho_l = \phi_1\rho_{l-1} + \phi_2\rho_{l-2} \quad \text{for} \quad l > 0. \tag{2.10}$$

Denoting by B, the backshift operator, i.e., $B(\rho_l) = \rho_{l-1}$, we can rewrite the above equation as

$$(1 - \phi_1 B - \phi_2 B^2)\rho_l = 0.$$

This is known as a difference equation. This equation determines the properties of the autocorrelation function (ACF) of a stationary AR(2) process. Corresponding to this

difference equation is the second order polynomial equation called the characteristic equation given by

$$1 - \phi_1 x - \phi_2 x^2 = 0,$$

solutions of which are

$$x = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}.$$

In time series analysis, inverse of these two solutions are known as the *characteristic roots* of the AR(2) model. Denote the two solutions of the characteristic equation as ω_1 and ω_2.

If both ω_i are real valued, then the second order difference equation of the model can be factored as

$$(1 - \omega_1 B)(1 - \omega_2 B)$$

and the AR(2) model can be thought of as an AR(1) model operating on top of another AR(1) model. The ACF of x_t is then a mixture of two exponential decays. In the other case, when the roots are complex, a plot of the ACF of x_t would show a picture of damping sine and cosine waves, which represent business cycles in many econometric and financial time series. A business cycle can be thought of as a sequence of expansions and contractions in a time series data.

2.5.2 Stationarity

A necessary condition for stationarity of AR models is that the its characteristic roots be less than one in modulus. Under such a condition, for an AR(2) model, it is easy to see that the ACF given in equation (2.10) converges to zero with the lag l. Applying this condition to an AR(1) model gives us the characteristic root of the difference equation $1 - \phi_1 x = 0$, which is ϕ_1 (recall that the characteristic root of the difference equation is the inverse of the root of the difference equation). The condition on the characteristic root to be less than one in modulus is the same condition we obtained before.

The results of the AR(2) model can be generalized to AR(p) models in the context of stationarity. That is, a necessary condition for the stationarity of an AR(p) model is

that all its characteristic roots lie within the unit space. In this case, the difference equation is given by

$$1 - \sum_{i=1}^{p} \phi_i x^i = 0.$$

Unit root non-stationarity

When the stochastic process governing the evolution of a time series has a root of its associated difference equation on the unit space, the time series is non-stationary and is called unit root non-stationary.

Unit root non-stationary process are useful in modeling non-stationary processes. A consequence of the presence of unit roots is the non-decaying effect the shock at time t has on all future realizations. Unit root non-stationarity can be overcome by differencing a series. Differencing a series $\{x_t\}_{t\in T}$ once yields the series $\{y_t\}_{t\in T}$ where $y_t = x_t - x_{t-1}$. When a series has a unit root of order greater than one, the series can be differenced multiple times to achieve a stationary series.

2.5.3 Model identification, estimation and checking

The first step in fitting an AR(p) model to a time series is to determine the order, p. This is typically done in either of two ways. The first method is to use the partial ACF plot to determine the order while the second is to use some information criteria like the Akaike information criteria (AIC). Once identified, the next step involves estimating the autocorrelation parameters, ϕ_i. This can be done by using either the least squares method or the maximum likelihood method. In both the methods, for a given time series $\{x_t : t = 1, \ldots, T\}$, the sample size becomes $T - p$ since the first $p - 1$ data give information regarding only the impulse variable. Finally, the fitted model is checked for the goodness of fit. This is done by looking at the behavior of the residual series as well as the statistical significance of the parameter estimates. If the residual series does not behave like a white noise series and shows some evidence of autocorrelation or some of the estimated parameters are statistically insignificant, then the model is refined, re-estimated and checked again. The R-square (R^2) statistic is also employed as a measure of the goodness of the fitted model.

2.6 Linear multivariate time series analysis

Multivariate time series models are natural in econometric applications because one is often interested in modeling many variables jointly. There are strong underlying relationships between and within market variables and macro-econometric indicators like indices, GDP, unemployment rate, etc.. Thus, it becomes important to study these quantities together rather than in isolation. In this section, we will look at linear time series analysis from a multivariate perspective. The ideas of the preceding sections can be extended to the multivariate context.

2.7 Cross-correlation and the vector autoregressive (VAR) model

Before we get to multivariate models, it is important to understand cross covariance. Earlier, we saw the concept of autocorrelation. In a multivariate model, however, in addition to a time series being correlated with its own past, it could also be correlated with another time series. This is termed as cross-correlation and just like autocorrelation, will be central in the definition of stationarity.

More precisely, consider a d-dimensional vector time series $\{\mathbf{x_t} : t = 1, \ldots, T\}$ where $\mathbf{x_t} = (x_{t1}, \ldots, x_{td})^T$. Then, weak stationarity means that the mean, $\boldsymbol{\mu}$, of $\mathbf{x_t}$, and the lag-l autocovariance, $\boldsymbol{\Sigma_l}$, between $\mathbf{x_t}$ and $\mathbf{x_{t-l}}$, are both time-invariant and finite. Mathematically,

$$\boldsymbol{\Sigma_l} = \begin{pmatrix} \rho_{11l}\sigma_1^2 & \cdots & \rho_{1dl}\sigma_1\sigma_d \\ \vdots & \ddots & \cdots \\ \rho_{d11}\sigma_d\sigma_1 & \cdots & \rho_{dd1}\sigma_d^2 \end{pmatrix},$$

where σ_i^2 is the stationary variance of x_{ti} and ρ_{ijl} is the correlation between x_{ti} and $x_{(t-l),j}$. While ρ_{iil} represents the autocorrelation within the univariate series $\{x_{ti}\}$, ρ_{ijl} for $i \neq j$ represents the *cross correlation* between x_{ti} and x_{tj}. When $l = 0$, ρ_{ij0} represents the *concurrent correlation* within the vector $\mathbf{x_t}$.

With the above definitions, we are now ready to define and analyze a vector autoregressive (VAR) model. For the vector time series, $\{\mathbf{x_t}\}$ as defined above, a VAR model or order p, VAR(p), is defined as

$$\mathbf{x_t} = \mathbf{A_0} + \sum_{i=1}^{p} \mathbf{A_i} \mathbf{x_{t-i}} + \boldsymbol{\epsilon_t}, \tag{2.11}$$

where $\{\boldsymbol{\epsilon_t}\}$ is a d-dimensional vector white noise process with zero mean and variance $\boldsymbol{\Sigma_\epsilon}$, $\mathbf{A_0}$ is a d-dimensional constant vector and $\mathbf{A_i}$ are $d \times d$ matrix parameters governing the model behavior [69].

2.7.1 Properties of the VAR model

We start with a zero mean stationary VAR(1) process to understand its properties. Consider the d-dimensional VAR(1) model defined as

$$\mathbf{x_t} = \mathbf{A_1} \mathbf{x_{t-1}} + \boldsymbol{\epsilon_t}, \tag{2.12}$$

where $\{\boldsymbol{\epsilon_t}\}$ is a white noise process with covariance matrix $\boldsymbol{\Sigma_\epsilon}$ and $\mathbf{A_1}$ is a $d \times d$ constant matrix. Note that we have omitted the mean vector $\boldsymbol{\mu}$ or $\mathbf{A}_0$ from the equation without loss of generality since any weakly stationary AR process can be written in a mean-corrected form by a simple linear transformation.

If all eigenvalues of $\mathbf{A_1}$ are less than one in modulus, then by repeated substitutions, the above equation can be rewritten as

$$\mathbf{x_t} = \sum_{i=0}^{\infty} \mathbf{A}_1^i \boldsymbol{\epsilon_t}.$$

The above equation is well-defined since we have assumed all eigenvalues of $\mathbf{A_1}$ to be less than one in modulus which makes the infinite sum in the above equation well-defined. We can now use this equation to compute the form of the lag-l auto-covariance matrices as

$$\boldsymbol{\Gamma_l} = \sum_{i=0}^{\infty} \mathbf{A}_1^{l+i} \boldsymbol{\Sigma_\epsilon} (\mathbf{A}_1^i)^T.$$

Such a VAR(1) process, with modulus of all eigenvalues of the sole parameter matrix, $\mathbf{A_1}$, less than 1, is called a stable VAR(1) process. This result can be extended to the general VAR(p) model as well by considering it as a VAR(1) model; see Lütkepohl (2007) [38] for more details. A VAR(p) model is said to be stable if

$$\det(\mathbf{I_{dp}} - \mathbf{A}x) \neq 0 \quad \text{for} \quad |x| \leq 1,$$

where $\mathbf{I_{dp}}$ is the $dp \times dp$ identity matrix and $\mathbf{A}$ is the $dp \times dp$ matrix given by :

$$\mathbf{A} = \begin{pmatrix} \mathbf{A_1} & \mathbf{A_2} & \dots & \mathbf{A_{p-1}} & \mathbf{A_p} \\ \mathbf{I_d} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_d} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I_d} & \mathbf{0} \end{pmatrix}. \tag{2.13}$$

The above condition can be simplified to the familiar characteristic equation that we saw in the univariate case as follows. Consequently, a VAR(p) process as defined in Equation (2.11) is stable if

$$\det(\mathbf{I_d} - \mathbf{A_1}x - \ldots - \mathbf{A_p}x^p) \neq 0 \quad \text{for} \quad |x| \leq 1.$$

Stability implies stationarity and unstable process are of little interest in time series analysis. However, stationarity does not imply stability.

Just like we considered the VAR(p) process as a VAR(1) process with a modified matrix parameter, to arrive at the stability condition, we will arrive at the form of the lag-l autocovariance matrix of a stationary VAR(p) process as follows. Consider the $dp \times 1$ vector

$$\mathbf{X_t} = \begin{pmatrix} \mathbf{x_t} \\ \mathbf{x_{t-1}} \\ \vdots \\ \mathbf{x_{t-p+1}} \end{pmatrix}.$$

Then, for the modified $pd \times 1$ process $\{\mathbf{X_t}\}$, we get the following expression for the lag-0 autocovariance :

$$\mathbf{\Gamma_X(0)} = \mathbf{A\Gamma_X(0)A'} + \mathbf{\Sigma_\epsilon},$$

where $\mathbf{A}$ is as defined in Equation (2.13) and $\mathbf{\Sigma_\epsilon}$ is the covariance matrix of the VAR associated white noise process. From the above equation, it is easy to see that

$$\text{vec}(\mathbf{\Gamma_X(0)}) = (\mathbf{I_{(dp)^2}} - \mathbf{A} \otimes \mathbf{A})^{-1}\text{vec}(\mathbf{\Sigma_\epsilon}).$$

Computing directly the lag-0 autocovariance matrix for $\{\mathbf{X_t}\}$, we get

$$\mathbf{\Gamma_X(0)} = E(\mathbf{X_t X_t^T}) = \begin{pmatrix} \mathbf{\Gamma_x}(0) & \mathbf{\Gamma_x}(1) & \dots & \mathbf{\Gamma_x}(p-1) \\ \mathbf{\Gamma_x}(-1) & \mathbf{\Gamma_x}(0) & \dots & \mathbf{\Gamma_x}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Gamma_x}(-p+1) & \mathbf{\Gamma_x}(-p+2) & \dots & \mathbf{\Gamma_x}(0) \end{pmatrix},$$

where $\mathbf{\Gamma_x(i)}$ is the lag-i autocovariance matrix for the original process, $\{\mathbf{x_t}\}$, of interest. From the equations defining the lag-i autocovariances we can deduce $\mathbf{\Gamma_x}(i)$. Parameters of a VAR model can be estimated by the least squares method or the maximum likelihood (MLE) method. Both methods yield estimates that are asymptotically normal.

2.8 Parameter estimation in AR and VAR models

In this section, we will briefly look at the least squares method to estimate parameters in AR and VAR models. This will include the computation of the estimator as well as its asymptotic properties.

2.8.1 The AR model

Recall the definition of an AR(p) model given in Equation (2.7). This model is in the form of a multivariate linear regression model with the explanatory vector variable $(1, x_{t-1}, \dots, x_{t-p})^T$ and the response variable x_t. Define the following :

$$\mathbf{Y} = \begin{pmatrix} x_T \\ \vdots \\ x_{p+1} \end{pmatrix}, \quad \mathbf{\Phi} = \begin{pmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{T-1} & \dots & x_{T-p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_p & \dots & x_1 \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} \epsilon_T \\ \vdots \\ \epsilon_{p+1} \end{pmatrix}.$$

Then, equation (2.7) can be written in a concise matrix form as :

$$\mathbf{Y} = \mathbf{X\Phi} + \mathbf{A}.$$

This is the familiar multiple linear regression model and the least squares parameter estimate is given by :

$$\hat{\mathbf{\Phi}} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} = \mathbf{\Phi} + (\mathbf{X^T X})^{-1}\mathbf{X^T A}.$$

If $\mathbf{\Sigma_X} = E[\mathbf{X^T X}/(T-p)]$ exists and is nonsingular, then, from the law of large numbers, the above equation implies that

$$\frac{1}{\sqrt{T}}(\hat{\mathbf{\Phi}} - \mathbf{\Phi}) \xrightarrow{d} N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{\Sigma_X^{-1}}).$$

From the definition of $\mathbf{X}$, we can compute $\mathbf{\Sigma_X}$ as

$$\mathbf{\Sigma_X} = \begin{pmatrix} 1 & \mu & \mu & \mu & \dots & \mu \\ \mu & \gamma_0 & \gamma_1 & \gamma_2 & \dots & \gamma_{p-1} \\ \mu & \gamma_1 & \gamma_0 & \gamma_1 & \dots & \gamma_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mu & \gamma_{p-2} & \gamma_{p-3} & \dots & \gamma_0 & \gamma_1 \\ \mu & \gamma_{p-1} & \gamma_{p-2} & \dots & \gamma_1 & \gamma_0 \end{pmatrix},$$

where γ_i is the lag-i autocovariance of the stationary AR process, $\{x_t\}$, and μ is its mean.

2.8.2 The VAR model

The VAR model can also be written in the form of a multiple linear regression model. Consider the general VAR(p) model given in Equation (2.11). To be clear, let

$$\mathbf{x_t} = (x_{t1}, \dots, x_{td}),$$

$$\boldsymbol{\epsilon_t} = (\epsilon_{t1}, \dots, \epsilon_{td}),$$

$$\mathbf{u}_t = (1, \mathbf{x_{t-1}}, \dots, \mathbf{x_{t-p}}),$$

$$\mathbf{\Phi_0} = \mathbf{A_0} = (A_{01}, \dots, A_{0d}) \quad \text{and}$$

$$\mathbf{\Phi_i} = \mathbf{A_i} = \begin{pmatrix} \phi_{i11}, \dots, \phi_{i1d} \\ \vdots \ddots \vdots \\ \phi_{id1}, \dots, \phi_{idd} \end{pmatrix}, \quad i = 1, \dots, p.$$

Now define the $(dp+1) \times d$ matrix $\mathbf{\Phi}$ as

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi_0} \\ \mathbf{\Phi_1} \\ \vdots \\ \mathbf{\Phi_p} \end{pmatrix}.$$

Then, we can write equation (2.11) succinctly as

$$\mathbf{x_t} = \mathbf{u_t}\mathbf{\Phi} + \boldsymbol{\epsilon_t}.$$

Writing

$$\mathbf{Y} = \begin{pmatrix} \mathbf{x_t} \\ \vdots \\ \mathbf{x_{p+1}} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{u_t} \\ \vdots \\ \mathbf{u_{p+1}} \end{pmatrix}, \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} \boldsymbol{\epsilon_t} \\ \vdots \\ \boldsymbol{\epsilon_{p+1}} \end{pmatrix},$$

Equation (2.11) can be written in a concise matrix form as :

$$\mathbf{Y} = \mathbf{X}\mathbf{\Phi} + \mathbf{A}.$$

This is the familiar multiple linear regression model and the least squares parameter estimate is given by :

$$\hat{\mathbf{\Phi}} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} = \mathbf{\Phi} + (\mathbf{X^T X})^{-1}\mathbf{X^T A}.$$

As in the AR case, if $\mathbf{\Sigma_X} = E[\mathbf{X^T X}/(T-p)]$ exists and is nonsingular, then, from the law of large numbers, the above equation implies that

$$\frac{1}{\sqrt{T}}(\text{vec}(\hat{\mathbf{\Phi}}) - \text{vec}(\mathbf{\Phi})) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_\epsilon \otimes \mathbf{\Sigma_X^{-1}}).$$

From the definition of $\mathbf{X}$, we can compute $\mathbf{\Sigma_X}$ as

$$\mathbf{\Sigma_X} = \begin{pmatrix} 1 & \boldsymbol{\mu} & \boldsymbol{\mu} & \boldsymbol{\mu} & \dots & \boldsymbol{\mu} \\ \boldsymbol{\mu^T} & \mathbf{\Gamma_0} & \mathbf{\Gamma_1} & \mathbf{\Gamma_2} & \dots & \mathbf{\Gamma_{p-1}} \\ \boldsymbol{\mu^T} & \mathbf{\Gamma_1} & \mathbf{\Gamma_0} & \mathbf{\Gamma_1} & \dots & \mathbf{\Gamma_{p-2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\mu^T} & \mathbf{\Gamma_{p-2}} & \mathbf{\Gamma_{p-3}} & \dots & \mathbf{\Gamma_0} & \mathbf{\Gamma_1} \\ \boldsymbol{\mu^T} & \mathbf{\Gamma_{p-1}} & \mathbf{\Gamma_{p-2}} & \dots & \mathbf{\Gamma_1} & \mathbf{\Gamma_0} \end{pmatrix},$$

where $\mathbf{\Gamma_i}$ is the lag-i $d \times d$ autocovariance matrix of the stationary VAR process $\{\mathbf{x_t}\}$, and $\boldsymbol{\mu}$ is its $1 \times d$ stationary mean vector.

2.9 Robustness in time series models : The need

Sudden and unexpected transient yet large movements in financial and econometric time series data are not uncommon due to external factors like political and regulatory changes. In Figure 2.1, the sharp one day rise of the Bombay Stock Exchange (BSE) Sensex index, highlighted by the box, from 12173.42 on 15^{th} May 2009 (time index 2930) to 14284.21 on the 18^{th} May 2009 (time index 2931) of 2110.79 points (17.33%), can be attributed to the fact that the 2009 general assembly election results in India were announced on the 18^{th}. The incoming government was perceived by the investor community to be highly investor friendly which in turn boosted investor confidence. Minor corrections took place post the 18^{th} but the general positive trend continued. This can also be seen in Figure 2.2. In addition, in spite of the digital age, faulty observations and corrupt data exist due to the various data warehousing processes involved.

When fitting a model to data, such large and/or erratic movements tend to have a sizeable impact on the fitted model, which is not desirable. For example, in fitting a trend to time series data, one could be interested in, not the precise data but rather the general path the series follows over time. Smoothing by taking the moving average over some lag gives one an idea about this trend. However, any malicious values in the time series data could potentially alter the trend curve.

Thus robust time series models are important. In the context of multivariate models, cross-correlations need to be taken into account in categorizing outliers. For example, in the 2-dimensional standard Normal distribution with a significant positive correlation coefficient, the value (0.5, 1.5) could possibly be considered as an outlier even though the first component in itself is not one.

2.10 Robust methods : definition and properties

Robust statistics is somewhat underrated in financial applications. This is probably because of two reasons. One is the simplicity of standard non-robust methods in comparison to robust methods. Secondly, robust methods typically involve complex computations that are time and resource intensive. Both these reasons are, however, countered by the recent advances in computing technologies. Hence, robust methodologies are now starting to gain some attention.

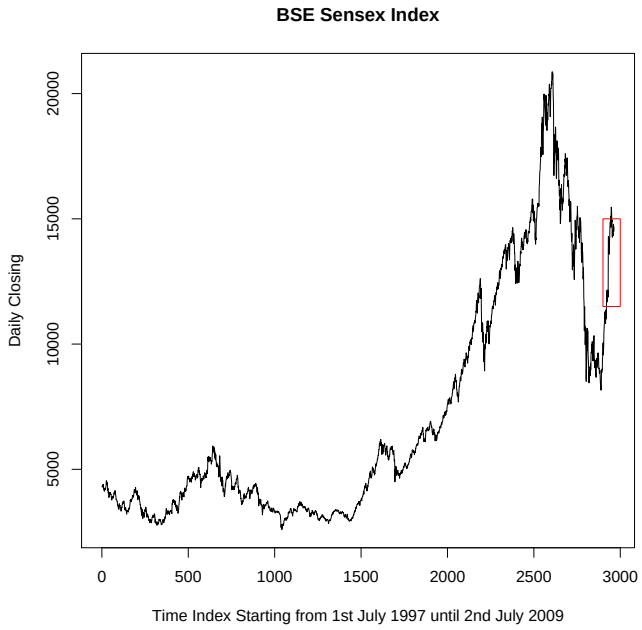

Figure 2.1: Daily closing of the Bombay Stock Exchange (BSE) Sensex index.

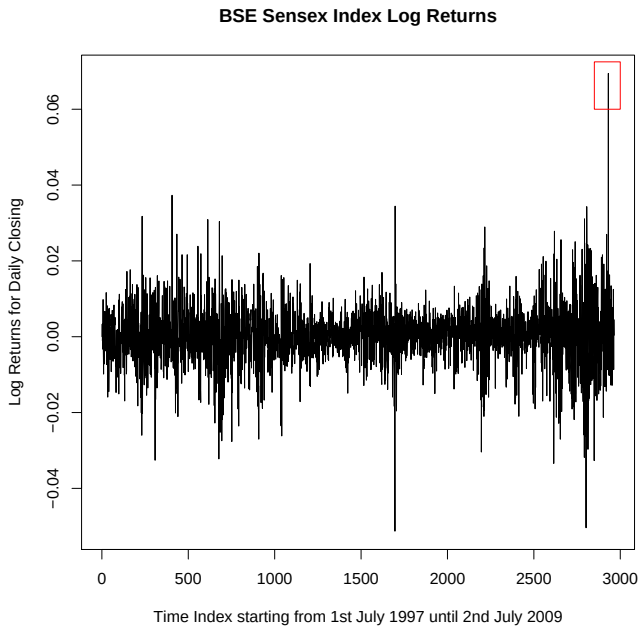

Figure 2.2: Log returns for daily closing of the Bombay Stock Exchange (BSE) Sensex index.

2.10.1 Robustness and outliers

A resistant method or model is one that, in analyzing any underlying structures of a given dataset, is resistant to reasonable amounts of contaminations in the dataset. Ideally, one would like the model to be resistant to any contaminations less than 50% since any further contamination renders the data void of any structure. Any data point that is distinct from the majority of the data points in terms of the structure we are interested in analyzing, is termed as an *outlier*. Hence, the definition of an outlier depends on the properties the data we are interested in studying. For instance, in wanting to study the mean characteristics of a given data, one may encounter location outliers that tend to be far away from the majority of the data. In linear regression or quadratic regression, an outlier could be thought of as a data point lying far away from the line or second order curve respectively, that characterizes the majority of the data. Let us look in detail at an example.

Suppose we are given a two dimensional stationary time series, $\{\mathbf{x_t} = (x_{t1}, x_{t2}) : t = 1, \ldots, T\}$, with a strong positive correlation of 0.9 and individual variance of 1.0 each. First, suppose we are interested in estimating its stationary mean $\boldsymbol{\mu} = (0, 0)$. Suppose $(x_{s1}, x_{s2}) = (0.5, 0.5)$ for some $1 < s < T$. The sample mean $\overline{\mathbf{x}}_T = \sum_{t=1}^{T} \mathbf{x_t}/T$ is a consistent estimate of $\boldsymbol{\mu}$. Now suppose now we corrupt the data by letting $(x_{s1}, x_{s2}) = (0.5, -0.5)$. Then, as far as estimating the mean of the bivariate series is concerned, the contaminated data is not an outlier since both the series individually maintain their respective mean structure.

Second, suppose now that we are interested in estimating the correlation between x_{t1} and x_{t2}. We now see that, the contaminated point shows an exact opposite structure (perfectly negative correlation) to that of the general data which is highly positively correlated (correlation of 0.9). Hence, in this application of estimating the correlation of the bivariate series, the contaminated data acts as an outlier.

Hence, the definition of an outlier is dependant on the context in which it is being defined. A corollary of the above example is when $\mathbf{x_s}$ is contaminated as $(x_{s1}, x_{s2}) = (2, 2)$. In this case, this point acts as an outlier when estimating the mean of the series while in the estimation of the correlation, it does not act as an outlier since it preserves the strong positive correlation.

2.10.2 Properties : Breakdown point (BP), influence function and sensitivity curve

Robustness in a model or method is measured by certain properties. These are the breakdown point (BP) and sensitivity curve which we will briefly discuss here.

As the name suggests, the breakdown point (BP) of a robust estimator of parameters in a model is the point at which the estimator breaks down. The point here refers to a level of contamination and breaking down means deviating from the true parameter by an arbitrarily large amount. Mathematically, assume a given dataset $\mathbf{X} = \{\mathbf{x_t} : t = 1, \ldots, T\}$ of size T. Assume that this data follows some model $H_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$. Consider another dataset, $\mathbf{X'}$, which is a copy of $\mathbf{X}$ but with a fraction ϵ of the data contaminated arbitrarily. Contamination refers to a data point being modified by an arbitrarily large amount in any direction (positive or negative). Let $\hat{\boldsymbol{\theta}}(\mathbf{X})$ be an estimator of $\boldsymbol{\theta}$. Then the *maximum bias* of this estimator for this contamination is defined as

$$bias(\epsilon, \mathbf{X}, \hat{\boldsymbol{\theta}}) = sup_{\mathbf{X'}} |\hat{\boldsymbol{\theta}}(\mathbf{X'}) - \hat{\boldsymbol{\theta}}(\mathbf{X})|.$$

The *finite sample BP* of an estimator $\hat{\boldsymbol{\theta}}$ for a data sample $\mathbf{X}$ is defined as

$$\epsilon^*(\hat{\boldsymbol{\theta}}, \mathbf{X}) = inf\{\epsilon : bias(\epsilon, \mathbf{X}, \hat{\boldsymbol{\theta}}) = \infty\}.$$

2.10.3 Influence function

Simply put, the influence function (IF) measures the influence of the addition of a data point to the sample data on the model parameter estimator.

The *empirical IF* or *sensitivity curve* measures the IF by actually adding the additional data point and comparing the new estimate with the estimate obtained with the original sample data. Mathematically, for a given data sample $\mathbf{X_T} = \{\mathbf{x_t} : t = 1, \ldots, T\}$ and a parameter estimator $\hat{\boldsymbol{\theta}}(\mathbf{X_T})$, the sensitivity curve is defined pointwise as

$$SC_T(\mathbf{x}) = \frac{\hat{\boldsymbol{\theta}}(\mathbf{X_T}(\mathbf{X_T}) \bigcup \{x\}) - \hat{\boldsymbol{\theta}}(\mathbf{X_T})}{1/T}.$$

While the sensitivity curve measures the influence of a data point on the estimator $\hat{\boldsymbol{\theta}}$ by simply adding that point to the sample data, the influence curve measures this influence by altering the underlying data distribution. Specifically, suppose the

distribution of the sample data is F and $\hat{\boldsymbol{\theta}}(F)$ is now the estimator for a given data distribution F. Let Δ_x denote the point mass distribution concentrated at $\mathbf{x}$. Then, the IF is defined point-wise as

$$\mathrm{IF}(\mathbf{x}, F, \hat{\boldsymbol{\theta}}) = \lim_{t \to 0} \frac{\hat{\boldsymbol{\theta}}((1-t)F + t\Delta_x) - \hat{\boldsymbol{\theta}}(F)}{t}, \tag{2.14}$$

if this limit exists for every $\mathbf{x}$.

2.11 Robust linear regression : The case of parameter estimation in AR and VAR models

We saw earlier that one can employ the method of least squares in estimating the parameters of AR and VAR models by considering them as a linear regression problem. However, it is known that the method of least squares is not robust to even small levels of contamination in the data. In this regard, we will look at some robust alternatives. In particular, we will study briefly the M-estimator and look at some of the more sophisticated robust estimators. All these estimators can be used as an alternative to the method of least squares and, hence, in analyzing AR and VAR models.

Recall the linear regression model for a data given by $\{\mathbf{u_t}, \mathbf{v_t}\}_{t=1,\ldots,T}$, where $\mathbf{u_t}$ is the p-dimensional impulse variable and $\mathbf{v_t}$ is the q-dimensional response variable. The linear regression model is given by

$$\mathbf{v_t} = \mathbf{a}\mathbf{u_t} + \mathbf{b} + \boldsymbol{\epsilon_t}$$

where $\{\boldsymbol{\epsilon_t}\}$ is a q-dimensional white noise process with zero mean, $\mathbf{a}$ is . The method of least squares estimates $(\mathbf{a}, \mathbf{b})$ by minimizing the sum of square of residuals given by

$$\sum_{i=1}^{T} \mathbf{r_t}(\tilde{\mathbf{a}}, \tilde{\mathbf{b}})$$

where $\mathbf{r_t}(\tilde{\mathbf{a}}, \tilde{\mathbf{b}}) = \mathbf{v_t} - \tilde{\mathbf{a}}\mathbf{u_t} - \tilde{\mathbf{b}}$ are the residuals.

Consider the zero mean d-dimensional stationary time series $\{\mathbf{x_t} : t = 1, \ldots, T\}$. Suppose we wish to fit a zero mean VAR(p) model to this series, denoted by

$$\mathbf{x_t} = \sum_{i=1}^{p} \mathbf{\Phi_i x_{t-i}} + \boldsymbol{\epsilon_t}$$

where $\boldsymbol{\epsilon_t}$ is a vector white noise series with covariance $\mathbf{\Sigma}$ and $\mathbf{\Phi_i}$ are $d \times d$ matrix parameters. This can be looked at as a regression problem by considering

$$\mathbf{u_t} = (\mathbf{x_{t-1}^T}, \ldots, \mathbf{x_{t-p}^T})^T \in \mathbb{R}^{dp},$$

$$\mathbf{v_t} = \mathbf{x_t}$$

and

$$\boldsymbol{\beta} = [\mathbf{\Phi_1} | \ldots | \mathbf{\Phi_p}] \in \mathbb{R}^{d \times dp}.$$

Then, the VAR(p) equation can be rewritten in the regression notation as

$$\mathbf{v_t} = \boldsymbol{\beta} \mathbf{u_t} + \boldsymbol{\epsilon_t}.$$

By putting a VAR model in this form, we can use the method of least squares to estimate the parameters of the VAR model. Note that the maximum likelihood estimator (MLE) reduces to the method of least squares when the noise distribution is assumed to be Gaussian. We will limit the scope of our discussion on robust methods applied to time series models to the Gaussian white noise case. This is because we are primarily concerned with dealing with outliers in the data rather than faulty white noise distribution assumptions.

2.11.1 The M-estimator and other estimators in brief

The M-estimator minimizes

$$\sum_i \rho(r_i(\hat{\beta})/s)$$

where ρ is an M-function or ρ-function (symmetric, passing through the origin, monotone non-decreasing for positive arguments and continuously differentiable on all but a finite number of exceptional points), r_i are the residuals (for a given parameter estimate $\hat{\beta}$) and s is a robust estimate of the scale of these residuals. When $\rho(r) = r^2$, the M-estimator reduces to the least squares estimator. A particular difficulty is the estimation of the scale of the residuals, which has to be done off-line.

The reason for introducing the scale in the optimization equation is to distinguish outliers and extreme values from the general data. Suppose we momentarily ignore s (i.e. assume s = 1) in the M-function, Then, when the variance of the associated white noise is large, the M-estimator will tend to try and fit the model in a way that maximizes the number of residuals close to zero. However, given the large variance, this number will be low and not representative of the underlying distribution. On the other hand, when the variance of the associated noise is small, the estimator will tend to try and bring the larger residuals closer to zero since all the residuals will already be close to zero due to the small underlying variance. Again, this will put less weights on the majority of the residuals. Thus, in both cases, the estimator will tend to overlook the majority of the residuals and therefore incorrectly estimate the model parameters. Put succinctly, dividing the residuals by s standardizes them thus giving a more accurate idea about the general distribution of the noise as well as potential outliers. This is necessary because outlyingness depends on the relative size of the residual rather than on the absolute size. This in turn leads to the estimator putting more weight on the majority of the residuals that are actually representative of the true noise distribution and thus makes the estimator less sensitive to the potentially outlying minority residuals.

The generalized M-estimator (GM Estimator) was introduced with the idea of bounding the influence of the outlying x_i (the explanatory variables in the regression model) as described below.

Define

$$\psi(r) = \frac{d}{dr}\rho(r)$$

Then, minimizing $\sum_i \rho(r_i(\hat{\beta})/s)$ is done by solving the following equations for $\hat{\beta}$:

$$\sum_i \psi(r_i(\hat{\beta})/s)\frac{\partial r_i(\hat{\beta})}{\partial \hat{\beta}} = \sum_i \psi(r_i(\hat{\beta})/s)x_i = 0. \tag{2.15}$$

From the above equation, it is clear that an outlier in the impulse variable x_i will adversely affect the estimate. Such an outlier is termed as a bad leverage point. The GM Estimators bound this influence of the outlying x_i using some weight function w. The Mallows type proposes to replace Equation (2.15) by

$$\sum_i w(x_i)\psi(r_i(\hat{\beta})/s)x_i = 0$$

while the Schweppe type suggests to use

$$\sum_i w(x_i)\psi(r_i(\hat{\beta})/(w(x_i)s))x_i$$

in place of equation (2.15). However, all GM Estimators have a breakdown point of at most $1/(1+p)$ where p is the dimension of the explanatory variable.

As seen above, the main idea of the M-estimators is the use of a penalty function (the M-function) that is more slowly increasing than the quadratic function used in the least squares procedure.

Here, it is worthwhile mentioning about the redescending M-estimators. An M-estimator is termed redescending if the corresponding ψ-function, $\psi(.) = \rho'(.)$, is redescending. By this, we mean that

$$\lim_{x\to\infty} \psi(x) = 0. \tag{2.16}$$

Often, a redescending ψ-function is also defined as that for which a constant c exists such that $\psi(x) = 0 \quad \forall \quad x \geq c$. The main purpose of using redescending estimators is to straightaway reject gross outliers, especially those resulting from bad leverage points. The cost of this insensitivity is the added complexity of the optimization function that defines the estimator. However, advanced computer power means that this drawback can be overcome in most cases.

By using redescending estimators, outliers will not get a lot more weight than the good points as the method will try to fit a model that brings data points having small to medium sized residuals closer to the fit rather than trying to minimize the size of the largest residuals. This is because, after a certain size, determined by the auxiliary scale parameter s, the residuals have a much smaller effect on the objective function to be minimized. In other words, the typical objective function is most sensitive to residual values in the range $[-s, s]$ and thereafter the sensitivity tends to decrease. This can be seen in Figure 2.3 where the ρ-function for Tukey's bi-weight M-function is seen, which is equal to

$$\rho_c(r) = \begin{cases} \frac{c^2}{6}[1 - \{1 - (r/c)^2\}^3] & \text{if } |r| \leq c \\ \frac{c^2}{6} & \text{otherwise} \end{cases}$$

The corresponding ψ-function is given by

$$\psi_c(r) = \begin{cases} r[1-(r/c)^2]^2 & \text{if } |r| \leq c \\ 0 & \text{otherwise} \end{cases}$$

The LMS (least median of squares, which minimizes the median of the residual squares) and LTS (least trimmed squares which minimizes the sum of the h smallest residuals squares) methods have a maximum breakdown point of 50%. However, the LMS converges at the rate of $n^{-1/3}$. Even though the LTS has a favorable convergence rate of $n^{-1/2}$, the computation of its objective function (for fixed p) takes O(n log n) steps (because of the ordering) compared to only O(n) for the LMS. The S-estimator, which will be discussed in the next chapter, was proposed with the intention of having an affine equivariant (independent of the choice of the coordinate axes of the x_i), 50% breakdown estimator with a convergence rate of $n^{-1/2}$, an O(n) objective function and a higher asymptotic efficiency than that of the LTS.

2.12 Summary

In this chapter, we saw some basic concepts related to robust time series analysis. We looked at stationarity, causality, linearity, autocovariance and autocorrelation in both, univariate, as well as multivariate, setups. We then examined the most simple linear time series model - the autoregressive (AR) model and looked at the properties of some simple AR models. We also saw the vector AR (VAR) model and its properties. We then saw the need for robust models in time series modeling. In the end, we saw briefly the M-estimator and some other estimators that are all robust alternatives to the least squares estimator. With this short introduction to simple robust linear time series models, we are now ready to move forward to see how to apply the S-estimator to the VAR model.

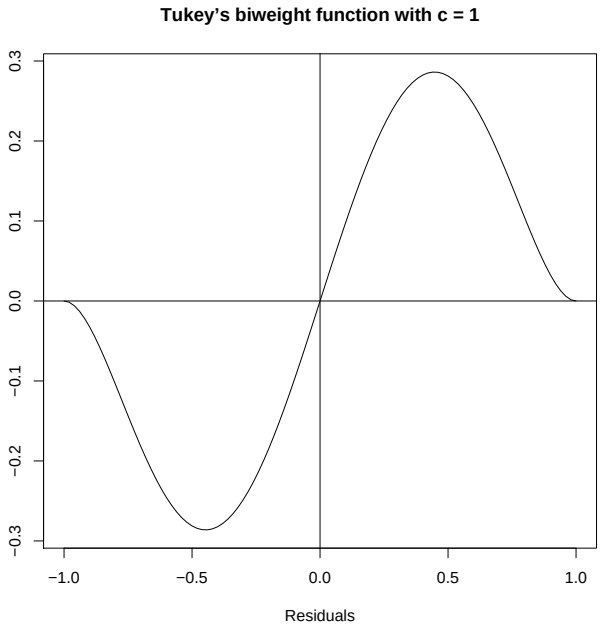

Figure 2.3: Tukey's ψ-function : $\psi_c(r) = \begin{cases} r[1-(r/c)^2]^2 & \text{if } |r| \leq c \\ 0 & \text{otherwise} \end{cases}$

CHAPTER 3

Robust estimators for VAR models : The S-estimator

3.1 Introduction

We saw in the last two chapters the concepts of autocorrelation, linear models and stationarity as well as the need for multivariate and robust models. We saw the simple VAR model and some of its important properties. Further, we saw some of the standard robust estimators and their properties. In this chapter, we will see how we can combine the concepts of robustness and multivariate time series models to come up with a robust estimator for the VAR model.

Recall the definition of a VAR model. Given a d-dimensional stationary time series $\mathbf{x_t} = (x_{1t}, \ldots, x_{dt})^T$, $t = 1, \ldots, T$, a VAR(p) model for $\mathbf{x_t}$ is given by

$$(\mathbf{I_m} - \boldsymbol{\phi}_1 B - \ldots - \boldsymbol{\phi}_p B^p)(\mathbf{x_t} - \boldsymbol{\mu}) = \boldsymbol{\epsilon_t} \tag{3.1}$$

where B is the backward shift operator ($B(\mathbf{x_t}) = \mathbf{x_{t-1}}$), $\mathbf{I_m}$ is the $d \times d$ identity matrix, $\boldsymbol{\phi}_i$ are the $d \times d$ matrix autoregressive parameters, $\boldsymbol{\mu}$ is a $d \times 1$ vector of constants and $\boldsymbol{\epsilon_t}$ are independent d-dimensional white noise with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

As seen in the last chapter, the presence of outliers coupled with strong cross-correlations necessitate the study of robust models for multivariate time series. Robust estimation of time series models is a widely researched area. The problem of

fitting VAR models robustly has been looked at already, but open questions remain. In particular, the application of S-estimators to estimate the VAR model parameters has yet to be studied.

In the next section, we will briefly describe the various kinds of outliers encountered in time series data and how they propagate. After that, we will briefly describe the residual autocovariance (RA) and the multivariate least trimmed squares (MLTS) methods which are the state of the art in robust VAR modeling. Then, we will give the motivation, definition and properties of the S-estimator for the univariate as well as multivariate cases. We will then present the Fast-S methods to compute these S-estimates. In the following section, we will compute the S-estimator for a simple 2-dimensional VAR(1) model using the multivariate Fast-S method. We will then give some comparative statistics of the S-estimator, the RA estimator and the least squares estimator. Finally, we will summarize the results in the last section.

3.2 Outlier classifications and propagation in time series data

3.2.1 Classification

There are two primary kinds of outliers in time series data. These are the *Innovative Outliers* and the *Additive Outliers*.

Innovative outliers (IO) occur when a data value is corrupted and the corruption gets propagated further according to the underlying mechanism. In other words, when the innovation or noise at any time index gets corrupted, it is an instance of an innovative outlier. This could happen, for example, as an effect of some political or regulatory development as can be seen in the spike (highlighted in the box) at index 2931 in Figure 2.2. Innovation outliers thus, by definition, are model dependent in that an outlier at any index is propagated further according to the model.

Additive outliers (AO) occur when a data value is corrupted in isolation and hence there is no propagation of that contamination. This could happen, for example, during the data warehousing processes. For a univariate series $\{x_t | t = 1, \ldots, T\}$, replacing x_s by $x_s + \alpha$ leads to an additive outlier contamination of the observation at time s. Again, this shows that additive outliers are model independent. That is, one can artificially contaminate a time series data with additive outliers without

having any knowledge of the underlying model that governs the evolution of the time series.

3.2.2 Propagation

Innovative outlier propagation

Since innovative outliers propagate further, even a single contamination leads to the contamination of all further data values. However, the effect of that single contamination on latter values diminishes with distance when the series is causal.

More precisely, consider a univariate time series model $\{x_t : t = 1, \ldots, T\}$ that evolves according to the model

$$x_t = f(x_{t-1}, \ldots, x_1) + \epsilon_t$$

where $f(x_{t-1}, \ldots, x_1)$ is a measurable function with respect to the σ-field of all information prior to time t and ϵ_t are independent and identically distributed (i.i.d) Gaussian white noise with variance σ_ϵ^2.

Suppose x_s for some $1 < s < T$ is an IO. Then x_{s+1} also gets contaminated since $x_{s+1} = f(x_s, \ldots, x_1) + \epsilon_{s+1}$. The same holds for all $x_t : t > s$. Thus, IO contamination propagates. The outlier itself is called innovative because it is most easily thought of as an isolated outlier in ϵ_s.

Suppose we are dealing with a simple stationary and causal zero mean AR(1) process. That is,

$$x_t = f(x_{t-1}, \ldots, x_1) + \epsilon_t = \phi x_{t-1} + \epsilon_t \quad \text{where} \quad |\phi| < 1, \tag{3.2}$$

where $1 \leq t \leq T$ and $x_1 = \epsilon_1$. Suppose, like before, x_s for some $1 < s < T$ is an IO. We leave it to the reader to check that from Equation (3.2) we obtain

$$x_{s+k} = \sum_{i=0}^{s+k-1} \phi^i \epsilon_{s+k-i}.$$

The above result can be verified by the principle of mathematical induction as well. The direct contribution of ϵ_s to x_{s+k} is the term $\phi^k \epsilon_s$ which diminishes with k, since $|\phi| < 1$. Thus, an IO creates a ripple effect which dies down with time. This is often

termed as a level shift.

Fortunately, x_{s+1} is not an outlier in the context of the model and parameter estimation methods. This is because it still maintains the relationship with x_s that is governed by the model. Again, the same holds for all $x_t : t > s$. Thus, if $\alpha\%$ of the time series data is contaminated by IO, any method that tries to estimate the parameters of the model that the time series follows, will encounter the same $\alpha\%$ contamination.

In the sense of parameter fitting, being model dependent, innovative outliers in time series preserve the contamination percentage in the process of fitting autoregressive time series data.

Additive outlier propagation

Consider once more the causal and stationary univariate zero mean AR(1) model

$$x_t = \phi x_{t-1} + \epsilon_t.$$

This can be viewed as a regression model fitted to the doublets $\{(x_1, x_2), \ldots, (x_{T-1}, x_T)\}$. Being model independent, AO do not propagate like IO. However, this is a drawback since an AO will show up as, not only a vertical outlier, that is an outlier in the second value of a doublet, but also a leverage point in the first value of a doublet. To be more precise, suppose that x_s is an AO. That is, we replace x_s by $\tilde{x}_s = x_s + A$ for some arbitrarily large value A. We will now have two outliers in this dataset in $(x_{s-1}, \tilde{x}_s)$ (where $\tilde{x}_s$ is a vertical outlier) and $(\tilde{x}_s, x_{s+1})$ (where $\tilde{x}_s$ will act as a bad leverage point).

More generally, in any autoregressive model of order p, a single additive outlier in the time series data shows up as $p + 1$ outliers in the regression model as p leverage points and one vertical outlier. Thus, estimators that have a breakdown point of α % could potentially breakdown when the time series data is corrupted through additive outliers by only $\alpha/(p+1)$ %. This phenomenon can be seen later in the penultimate section when a simulated time series data is corrupted by 20% which leads to a 40% corruption in the VAR(1) model data. Thus, additive outliers are the more difficult to deal with.

In the latter section that deals with simulations, we thus limit ourselves to additive outliers.

3.3 A short review of existing procedures

3.3.1 The residual autocovariance (RA) method

A widely cited method for univariate linear time series models is the one introduced by Bustos and Yohai (1986) [9], which is based on the so called residual autocovariance (RA). Li and Hui (1989) generalize the RA method in order to propose a robust estimation procedure for a vector time series. The method is based on formulating the estimating equations in terms of the residuals and then using *appropriately scaled* values of these residuals to obtain the estimates. More precisely, given a d-dimensional stationary time series $\mathbf{x_t} = (x_{1t}, \ldots, x_{dt})$, $t = 1, \ldots, T$, that evolves according to a zero mean VAR(p) model with Gaussian white noise $\boldsymbol{\epsilon_t}$ with a given covariance matrix $\boldsymbol{\Sigma}$, and parameters $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p$ the maximum likelihood equation (MLE) for

$$\boldsymbol{\beta} = \text{vec}(\boldsymbol{\phi}) = \text{vec}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p)$$

is obtained by minimizing

$$\frac{1}{2} \sum_{t=p+1}^{T} \mathbf{r_t^T}(\hat{\boldsymbol{\beta}}) \boldsymbol{\Sigma}^{-1} \mathbf{r_t}(\hat{\boldsymbol{\beta}}).$$

where $r_t(\hat{\boldsymbol{\beta}})$ are the residuals for a given parameter $\hat{\boldsymbol{\beta}}$, that is,

$$\mathbf{r}_t(\hat{\boldsymbol{\beta}}) = \mathbf{x_t} - \sum_{j=1}^{p} \hat{\boldsymbol{\Phi}}_j \mathbf{x}_{t-j}.$$

Note that the determinant of $\boldsymbol{\Sigma}$ does not appear in the MLE equation since it is assumed known. In that respect, this can be termed as the conditional MLE. Let

$$(\mathbf{I_d} - \boldsymbol{\phi}_1 B - \ldots - \boldsymbol{\phi}_p B^p)^{-1} = \sum_i \boldsymbol{\pi}_i(\boldsymbol{\beta}) B^i.$$

where $\mathbf{I_d}$ is the d-dimensional identity matrix. Then, the estimating equations reduce to

$$\sum_t \sum_i \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}) r_{t,h}(\hat{\boldsymbol{\beta}}) \mathbf{r}_{t-j-i}(\hat{\boldsymbol{\beta}}) = \mathbf{0}, \quad j = 1, \ldots, p; \quad h = 1, \ldots, d \tag{3.3}$$

where $r_{t,h}(\hat{\boldsymbol{\beta}})$ is the h^{th} component of $\mathbf{r_t}(\hat{\boldsymbol{\beta}})$ and $\mathbf{r_t}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ for $t < p + 1$. To robustify the method, the products $r_{t,h}(\hat{\boldsymbol{\beta}}) r_{t-j-i,k}(\hat{\boldsymbol{\beta}})$ are scaled appropriately by

an odd, bounded and continuous function, $\eta(u, v)$. Two possible choices for this function are $\eta(u, v) = \psi(u)\psi(v)$ and $\eta(u, v) = \psi(uv)$, where $\psi(.)$ is an odd, bounded and continuous function. The former choice is said to be of Mallows type and the latter of Hampel type. The function $\psi(.)$ may be in the Huber family

$$\psi_{H,c}(u) = \text{sgn}(u)\min(|u|, c)$$

or the biweight family

$$\psi_{B,c}(u) = u(1 - u^2/c^2)^2 \qquad (0 \leq |u| \leq c).$$

Ben et al. (1998) [5] further refined the RA method to make it affine equivariant. In this version, the residual vectors are first robustified by an odd and bounded function, $\psi(.)$, as follows. A weight function is defined as follows :

$$w(x) = \frac{\psi(x)}{x}.$$

The modified residuals are then defined as

$$\mathbf{r'}_t(\hat{\boldsymbol{\beta}}) = \mathbf{r}_t(\hat{\boldsymbol{\beta}})w\Big(\sqrt{\mathbf{r}_t^T(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}\mathbf{r}_t(\hat{\boldsymbol{\beta}})}\Big),$$

where $\boldsymbol{\Sigma}$ is assumed known or estimated robustly offline. The RA estimating Equation (3.3) can be rewritten in vector notation as

$$\sum_t \sum_i \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}})\mathbf{r}_t(\hat{\boldsymbol{\beta}})\mathbf{r'}_{t-j-i}(\hat{\boldsymbol{\beta}}) = \mathbf{0}, \quad j = 1, \ldots, p$$

From the above equation it can be seen that the difference between the standard RA and the affine equivariant RA method lies in the robustifying method of the residuals. While Li and Hui (1989) robustify the product of residuals component-wise, Ben et al. (1998) robustify the entire residuals individually and then consider the products. This refinement makes the RA method affine equivariant as now a residual that is not an outlier component-wise yet is an outlier (by way of its covariance structure for example) is downweighted accordingly as opposed to in the non affine equivariant RA method where the product of residuals gets downweighted only if there is an outlier in at least one component of any of the two residuals.

3.3.2 The multivariate least trimmed squares (MLTS) estimator

Recently, a multivariate least trimmed squares (MLTS) method was described for the VAR model by Croux and Joossens (2008) [18]. This method is based on the idea of the minimum covariance determinant (MCD) estimator of Rousseeuw (1985) [57]. The MLTS selects the subset of h observations having the property that if we performed a least squares fit to these observations, then the determinant of the covariance matrix of the corresponding residuals, is minimal.

More formally, consider the zero mean d-dimensional stationary time series $\{\mathbf{x_t} : t = 1, \ldots, T\}$. Suppose we wish to fit a zero mean VAR(p) model to this series, denoted by

$$\mathbf{x_t} = \sum_{i=1}^{p} \boldsymbol{\Phi_i} \mathbf{x_{t-i}} + \boldsymbol{\epsilon_t}$$

where $\boldsymbol{\epsilon_t}$ is a vector white noise series with covariance Σ and $\boldsymbol{\Phi_i}$ are $d \times d$ matrix parameters. This can be looked at as a regression problem by considering

$$\mathbf{u_t} = (\mathbf{x_{t-1}^T}, \ldots, \mathbf{x_{t-p}^T})^T \in \mathbb{R}^{dp},$$

$$\mathbf{v_t} = \mathbf{x_t}$$

and

$$\boldsymbol{\beta} = [\boldsymbol{\Phi_1} | \ldots | \boldsymbol{\Phi_p}] \in \mathbb{R}^{d \times dp}.$$

Then, the VAR(p) equation can be rewritten in the regression notation as

$$\mathbf{v_t} = \boldsymbol{\beta} \mathbf{u_t} + \boldsymbol{\epsilon_t}.$$

Now, given the series $\{\mathbf{x_t} : t = 1, \ldots, T\}$, to use the MLTS to estimate $\boldsymbol{\beta}$, consider the dataset $\mathbf{X} = \{(\mathbf{u_t}, \mathbf{v_t}) : t = p + 1, \ldots, T\}$. Denote by $\mathbf{U}$, the matrix consisting of the rows of the explanatory variable $\mathbf{u_t}$. That is, $\mathbf{U} = (\mathbf{u_{p+1}}, \ldots, \mathbf{u_T})^T$. Similarly, define the matrix $\mathbf{V} = (\mathbf{v_{p+1}}, \ldots, \mathbf{v_T})^T$. Let $\mathbb{H} = \{H \subset \{p+1, \ldots, T\} \quad | \quad \#H = h\}$ be the collection of all subsets of size h. For any subset $H \subset \mathbb{H}$, let $\hat{\boldsymbol{\beta}}_{OLS}(H)$ be the classical least squares fit based on the observations of the subset. This is given by :

$$\hat{\boldsymbol{\beta}}_{OLS}(H) = (\mathbf{U_H^T U_H})^{-1} \mathbf{U_H^T V_H},$$

where $\mathbf{U_H}$ and $\mathbf{V_H}$ are sub-matrices of $\mathbf{U}$ and $\mathbf{V}$, consisting of the rows of $\mathbf{U}$ and $\mathbf{V}$, respectively, having an index in H. The corresponding scatter matrix estimator computed from this subset is then :

$$\hat{\mathbf{\Sigma}}_{OLS}(H) = \frac{1}{h-p}(\mathbf{V_H} - \mathbf{U_H}\hat{\boldsymbol{\beta}}_{OLS}(H))^T(\mathbf{V_H} - \mathbf{U_H}\hat{\boldsymbol{\beta}}_{OLS}(H)).$$

The MLTS estimator is then defined as :

$$\hat{\boldsymbol{\beta}}_{MLTS}(\mathbf{X}) = \hat{\boldsymbol{\beta}}_{OLS}(\hat{H}) \quad \text{where} \quad \hat{H} = \text{argmin}_{H\subset\mathbb{H}}\{\det(\hat{\mathbf{\Sigma}}_{OLS}(H))\}.$$

3.4 The S-estimator

3.4.1 The univariate version

The motivation for the S-estimator comes from a desire to have a high breakdown point yet highly efficient estimator that has an $O(\sqrt{n})$ asymptotic convergence rate and is also not hard to compute. We will now briefly discuss its definition in the linear regression case. The S-estimator of Rousseeuw and Yohai (1984) [75] is a robust estimator with high breakdown point. It is defined as an estimate of the regression parameters with the smallest robust scale. Formally, the estimating procedure is defined as follows :

$$\text{Minimize}_{\hat{\beta}} \quad s = s(r_1(\hat{\beta}), \ldots, r_T(\hat{\beta})) \tag{3.4}$$

subject to

$$\frac{1}{T}\sum_{t=1}^{T}\rho(r_t(\hat{\beta})/s) = b. \tag{3.5}$$

Here, $s(r_1(\hat{\beta}), \ldots, r_T(\hat{\beta}))$ is a robust scale estimate, $\hat{\beta}$ is the parameter of interest, b is a constant, typically equal to $E_\Phi[\rho(X)]$ where X is a random variable having the same distribution G as that of the white noise in the model (to ensure consistency) and ρ is an M-Function. Existence of the solutions to the above optimization is briefly studied in Lopuhaä (1989) [37].

Properties

The least squares estimator is a special case of the S-estimator when ρ is chosen to be $\rho(r/s) = (r/s)^2$ and $b = 1$. The following assertions about the S-estimator are in

the context of the standard linear regression problem. The following theorem is due to Maronna and Yohai (1981) [39].

Theorem 1. *Suppose ρ satisfies the following properties :*

1. There exists a $c > 0$ such that ρ is strictly increasing on [0, c] and constant on [c, ∞). The corresponding ψ function is then called redescending.

2. $E_g[\rho(X)]/\rho(c) = 1/2$, (X being a white noise random variable with density function g).

It follows that the breakdown point of the corresponding S-estimator is

$$\epsilon_n^* = ([n/2] - p + 2)/n \tag{3.6}$$

which converges to the favorable 50% as $n \to \infty$. In condition 2, if $E_g[\rho(X)]/\rho(c) = \lambda$ where $0 < \lambda < 1/2$, then the corresponding S-estimator has a breakdown point converging to λ as $n \to \infty$. When g is the Standard Normal Density, $c = 1.547$ satisfies condition 2. Values of $c > 1.547$ yield better asymptotic efficiencies at a Gaussian model, but smaller breakdown points.

Consistency of the S-estimators follows from Theorems 2.2 and 3.1 of Maronna and Yohai (1981) because S-estimators satisfy the same first order necessary conditions as do M-estimators. The other necessary conditions for consistency are :

[C1]. $\psi(u)/u$ is non-decreasing for $u > 0$ and

[C2]. $E_H[||x||] < \infty$ where x is the explanatory variable and H is its distribution.

Asymptotic normality of the S-estimators follows from Theorem 4.1 of Maronna and Yohai (1981). The other necessary conditions for the asymptotic normality are :

[N1]. ψ is differentiable in all but a finite number of points, $|\psi'|$ is bounded and $\int \psi' d\phi > 0$ and

[N2]. $E_H[xx^T]$ is non-singular and $E_H[||x||^3] < \infty$.

When the above conditions are met,

$$n^{1/2}(\hat{\beta}-\beta) \xrightarrow{\mathbb{L}} N(0, E_H[xx^T]^{-1}\frac{\int \psi^2 d\phi}{(\int \psi' d\phi)^2})$$

and

$$n^{-1/2}(\hat{\sigma}-\sigma) \xrightarrow{\mathbb{L}} N(0, \frac{\int (\rho(y)-b)^2 d\phi(y)}{(\int y\psi(y) d\phi(y))^2})$$

An implicit assumption made in the above assertions is an i.i.d form of the regression data $\{(x_i, y_i) : i = 1, \ldots, n\}$.

3.4.2 The multivariate version

In multivariate regression, which is needed for fitting VAR models, the S-estimator is defined as in the univariate case except that we minimize a measure of the size of the covariance matrix estimate $\mathbf{S} = \mathbf{S}(r_1, \ldots, r_n)$ instead of the variance s^2. Possible size criteria include the determinant, the trace and the maximal eigenvalue. The argument of the M-function also has to be appropriately adapted to $\sqrt{\mathbf{r}(\hat{\boldsymbol{\beta}})\mathbf{S}^{-1}\mathbf{r}(\hat{\boldsymbol{\beta}})^T}$ instead of r/s.

In the following, we discuss the estimation problem :

Minimize $\mathrm{Det}(\mathbf{S}(\mathbf{r_1}, \ldots, \mathbf{r_N}))$ under the constraint

$$1/N \sum_{i=1}^{N} \rho(\sqrt{\mathbf{r_i}(\hat{\boldsymbol{\beta}})\mathbf{S}^{-1}\mathbf{r_i}(\hat{\boldsymbol{\beta}})^T}) = b \tag{3.7}$$

The multivariate S-estimator enjoys all the asymptotic and robustness properties of the univariate S-estimator. It has a favorable asymptotic breakdown point of 50% and is resistant to leverage points when $\psi(.)$ is redescending. For a detailed analysis of the regression case, the reader is referred to Van Aelst and Willems (2005) [72].

It is clear from Equation (3.7) that the S-estimator penalizes a data point even if only a few of its components are contaminated. If we were to, on the other hand, compute d independent S-estimators corresponding to the d components of $\mathbf{y_t}$, a data point may be penalized by some S-estimators but not by the others depending on the component values. However, given the cross-correlation in the data, it is perhaps better to down-weight the entire data point which is what the multivariate

S-estimator does. Hence, a VAR S-estimator is more appropriate than separate S-estimators in that it takes into account the cross-correlation in the data.

3.4.3 S-estimator for linear time series models

Given that in the linear autoregression version of a time series model, the explanatory and response variables are both random and come from the same time series, contamination in the data leads to leverage points. The use of high breakdown estimators such as the S-estimator is called for in estimating the parameters of the model. The breakdown characteristics and asymptotic behavior of the S-estimator in multivariate time series models remain essentially intact.

Breakdown point

The maximum breakdown point of 50% continues to be achievable since this depends only on the choice of a suitable $\psi(.)$ function and not on any assumptions on the distribution of the data.

Consistency

Rousseeuw and Yohai (1984) [60], in their introduction to S-estimators for the regression scenario, showed consistency and asymptotic normality under an i.i.d assumption on the carrier variables (the impulse variables). However, Davies (1990) [19] showed the consistency and asymptotic normality of the S-estimator under some milder conditions on the carrier variables. In linear time series models, the impulse variables are lagged versions of the response variables and hence the i.i.d assumption of the corresponding regression data is not valid anymore. However, simulations conducted by us seem to confirm consistency of the S-estimator in estimating parameters of AR and VAR time series models. Denote the regression data in a multivariate linear regression problem as $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, n\}$ where $\mathbf{x}_i$ are of dimension d_x. The conditions required of the data for consistency of the S-estimator for linear regression models as given by Davies are :

1. There exist positive numbers η_1, η_2 and n_0 such that

$$\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T 1_{\|\mathbf{x}_i\| < \eta_1} - n\eta_2 \mathbf{I}_{d_x}$$

is positive definite for all $n \geq n_0$, where $\mathbf{I}_{d_x}$ is the d_x dimensional identity matrix.

2.
$$\lim_{n\to\infty} \frac{1}{n} \sum \|\mathbf{x}_i\|^2 1_{\|\mathbf{x}_i\|>\delta\sqrt{n}} = 0 \quad \forall \quad \delta > 0 \tag{3.8}$$

Note that the other required conditions concern the $\rho(.)$ function which are fulfilled by most ρ-functions used in practice. Looking at the above two conditions, one observes that these conditions are expected to be fulfilled by a stationary VAR time series since in this case, stationarity implies a finite and constant covariance matrix. Thus, one could expect the S-estimator to be consistent for stationary VAR time series models which is in line with what our simulations demonstrate.

Affine equivariance

In the general regression model, the S-estimator is affine-equivariant. By this, we mean that if we transform the data affinely, then the parameter estimate with the modified data will be a related affine transformation of the parameter estimate with the old data.

In the time series context, as opposed to the general regression context, the impulse and response variables are dependent and come from the same time series. Thus, an affine transformation of a time series data transforms the impulse as well as response variables in the corresponding regression scenario of the associated model. More formally, given a d-dimensional multivariate time series $\{\mathbf{x}_t\}$, an affine transformation, $T_{(\mathbf{A},\mathbf{B})}$, transforms the series into a new series $\{\mathbf{y}_t\}$ where $\mathbf{y}_t = \mathbf{A}\mathbf{x_t} + \mathbf{B}$ where $\mathbf{A}$ is a $d \times d$ invertible matrix and $\mathbf{B}$ is a d-dimensional vector.

Now consider a stationary VAR(p) model for $\{\mathbf{x}_t\}$ given by

$$\mathbf{x}_t = \boldsymbol{\mu} + \sum_{i=1}^{p} \boldsymbol{\phi}_i \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t,$$

where $\{\boldsymbol{\epsilon_t}\}$ is a vector white noise with covariance $\boldsymbol{\Sigma}$. Now consider the transformed series $\{\mathbf{y}_t\}$ as defined earlier which is an affine transformation of $\{\mathbf{x}_t\}$. It is easy to see that this series also follows a stationary VAR(p) model given by

$$\mathbf{y}_t = \boldsymbol{\mu}' + \sum_{i=1}^{p} \boldsymbol{\phi}_i' \mathbf{y}_{t-i} + \boldsymbol{\delta}_t,$$

where

$$\boldsymbol{\mu}' = \mathbf{A}\boldsymbol{\mu} + (\mathbf{I} - \sum_{i=1}^{p} \mathbf{A}\boldsymbol{\phi}_i\mathbf{A}^{-1})\mathbf{B}$$

$$\boldsymbol{\phi}_i' = \mathbf{A}\boldsymbol{\phi}_i\mathbf{A}^{-1}$$

and

$$\boldsymbol{\delta}_t = \mathbf{A}\boldsymbol{\epsilon}_t.$$

Definition 1. *Consider the two series,* $\{\mathbf{x_t}\}$ *and* $\{\mathbf{y_t}\}$*, as defined above. Consider an estimator* $\mathbf{T}$ *of the parameters of a VAR(p) model. Denote the estimator corresponding to the original series by* $\mathbf{T}_x = (\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\phi}}_{x1}, \ldots, \hat{\boldsymbol{\phi}}_{xp}, \hat{\boldsymbol{\Sigma}}_x)$*. Denote the same estimator but now corresponding to the transformed series by* $\mathbf{T}_y = (\hat{\boldsymbol{\mu}}_y, \hat{\boldsymbol{\phi}}_{y1}, \ldots, \hat{\boldsymbol{\phi}}_{yp}, \hat{\boldsymbol{\Sigma}}_y)$*. Then,* $\mathbf{T}$ *is called affine-equivariant if*

$$\hat{\boldsymbol{\mu}}_y = \mathbf{A}\hat{\boldsymbol{\mu}}_x + (\mathbf{I} - \sum_{i=1}^{p} \mathbf{A}\hat{\boldsymbol{\phi}}_{xi}\mathbf{A}^{-1})\mathbf{B},$$

$$\hat{\boldsymbol{\phi}}_{yi} = \mathbf{A}\hat{\boldsymbol{\phi}}_{xi}\mathbf{A}^{-1},$$

and

$$\hat{\boldsymbol{\Sigma}}_y = \mathbf{A}\hat{\boldsymbol{\Sigma}}_x\mathbf{A}^T.$$

Theorem 2. *The S-estimator of the parameters of a stationary VAR(p) model is affine equivariant.*

Proof. Let $\{\mathbf{x}_t\}$ be a realization of a stationary d-dimensional VAR(p) model. Let $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \mathbf{B}$ where $\mathbf{A}$ is a $d \times d$ invertible matrix and $\mathbf{B}$ is a d-dimensional vector. Consider the S-estimators of the series $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$. As in the definition above, the indices x and y of these parameter estimates indicate the data used. Consider the auxiliary parameters defined by

$$\hat{\boldsymbol{\mu}} = \mathbf{A}^{-1}(\hat{\boldsymbol{\mu}}_y - (\mathbf{I} - \sum_{i=1}^{p} \hat{\boldsymbol{\phi}}_{yi})\mathbf{B}),$$

$$\hat{\boldsymbol{\phi}}_i = \mathbf{A}^{-1}\hat{\boldsymbol{\phi}}_{yi}\mathbf{A},$$

and

$$\hat{\mathbf{\Sigma}} = \mathbf{A}^{-1}\hat{\mathbf{\Sigma}}_y(\mathbf{A}^T)^{-1}.$$

These are well defined since $\mathbf{A}$ is invertible. Let $\mathbf{T} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\phi}}, \hat{\mathbf{\Sigma}})$. Then,

$$\hat{\boldsymbol{\mu}}_y = \mathbf{A}\hat{\boldsymbol{\mu}} + (\mathbf{I} - \sum_{i=1}^{p} \mathbf{A}\hat{\boldsymbol{\phi}}_i\mathbf{A}^{-1})\mathbf{B},$$

$$\hat{\boldsymbol{\phi}}_{yi} = \mathbf{A}\hat{\boldsymbol{\phi}}_i\mathbf{A}^{-1},$$

and

$$\hat{\mathbf{\Sigma}}_y = \mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^T. \tag{3.9}$$

Then, our aim is to show that

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_x,$$

$$\hat{\boldsymbol{\phi}}_i = \hat{\boldsymbol{\phi}}_{xi},$$

and

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_x.$$

Denote by $\mathbf{r}_{xt}(\mathbf{T}_x)$ the t^{th} residual for the S-estimator $\mathbf{T}_x$ for the original series and by $\mathbf{r}_{yt}(\mathbf{T}_y)$ the t^{th} residual for the S-estimator $\mathbf{T}_y$ for the transformed series. It is easy to see that

$$\mathbf{r}_{yt}(\mathbf{T}_y) = \mathbf{A}\mathbf{r}_{xt}(\mathbf{T}_x) \tag{3.10}$$

By definition, the S-estimator of the original data is given by

Min Det($\hat{\mathbf{\Sigma}}_x$) subject to $\sum_t \rho(\sqrt{\mathbf{r}_{xt}^T(\mathbf{T}_x)\hat{\mathbf{\Sigma}}_x^{-1}\mathbf{r}_{xt}(\mathbf{T}_x)}) = c$.

Now consider the definition of the S-estimator of the transformed series which is given by

Min Det($\hat{\mathbf{\Sigma}}_y$) subject to $\sum_t \rho(\sqrt{\mathbf{r}_{yt}^T(\mathbf{T}_y)\hat{\mathbf{\Sigma}}_y^{-1}\mathbf{r}_{yt}(\mathbf{T}_y)}) = c$.

Substituting for the parameters by their parameterized definitions (using the auxiliary parameters) and using Equations (3.9) and (3.10), the above reduces to

Min $\text{Det}(\mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^T)$ subject to $\sum_t \rho(\sqrt{\mathbf{r}_{xt}^T(\mathbf{T})\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_{xt}(\mathbf{T})}) = c$.

But minimizing $\text{Det}(\mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^T)$ is the same as minimizing $\text{Det}(\hat{\mathbf{\Sigma}})$ since $\mathbf{A}$ being a $d \times d$ square matrix, we have that

$$\text{Det}(\mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^T) = \text{Det}(\hat{\mathbf{\Sigma}}) * \text{Det}(\mathbf{A})^2 \tag{3.11}$$

and $\mathbf{A}$ is a constant matrix. Combining the reduced minimization equation and constraint we get that the parameterized S-estimator of the transformed data is defined by

Min $\text{Det}(\hat{\mathbf{\Sigma}})$ subject to $\sum_t \rho\Big(\sqrt{\mathbf{r}_{xt}^T(\mathbf{T})\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_{xt}(\mathbf{T})}\Big) = c$.

But this optimization problem is the same as the one that defines the S-estimator of the original data. Hence, we must have that

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_x,$$

$$\hat{\boldsymbol{\phi}}_i = \hat{\boldsymbol{\phi}}_{xi},$$

and

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_x.$$

Thus, the S-estimator for a VAR time series model is affine equivariant. □

Asymptotic analysis

An important tool in the asymptotic analysis of the S-estimator is the fact that it satisfies the same first order conditions as the M-estimator. This can be easily seen by using the Lagrange multipliers in the constrained optimization equation that characterizes the S-estimator. See Lopuhaä (1989) for details. As a consequence, the asymptotic behavior of the S-estimator is similar to that of the M-estimator.

In the following, we will compute the asymptotic covariance matrix of the S-estimator of a zero mean VAR model. For the following theorem, we will need some context

and definitions.

Definition 2. *Assume a given dataset* $\{\mathbf{z_t} = (\mathbf{x_t}, \mathbf{y_t}) : t = 1, \ldots, T\}$ *for which we wish to fit the linear regression model* $\mathbf{y_t} = \boldsymbol{\theta}\mathbf{x_t} + \boldsymbol{\epsilon_t}$ *where* $\mathbf{x_t}$ *is the p-dimensional explanatory variable,* $\mathbf{y_t}$ *is the q-dimensional response variable,* $\{\boldsymbol{\epsilon_t}\}$ *is a q-dimensional multivariate white noise process with covariance matrix* $\boldsymbol{\Sigma_\epsilon}$ *and* $\boldsymbol{\theta}$ *is the* $q \times p$ *matrix parameter.*

Suppose now that we estimate the parameters of this model using the S-estimator. Assume the data is contaminated with contamination percentage $r < 50\%$*. We saw earlier, in Equation* (3.6)*, that the S-estimator is bounded for contaminations less than* 50%*. Hence, the S-estimator is bounded for this data as well. Thus, the S-estimator,* $\hat{\boldsymbol{\theta}}_{\boldsymbol{s}}$*, stays in some fixed compact subset of* $\boldsymbol{\Theta}$*, the parameter space.*

Denote by $\mathbb{F}$*, the class of all distributions on* $\mathbb{R}^{p+q}$*, the data space. Let* $\mathbf{S}(.)$ *be a vector-valued mapping from a subset of* $\mathbb{F}$ *into* $\boldsymbol{\Theta}$*. If* Δ_z *denotes the atomic probability distribution concentrated in* $\mathbf{z} \in \mathbb{R}^{p+q}$*, then the influence function of* $\mathbf{S}(.)$ *at* $F \in Domain(\mathbf{S}(.))$ *is defined point-wise as*

$$IF(\mathbf{z}; \mathbf{S}, F) = \lim_{h \to 0} \frac{\mathbf{S}((1-h)F + h\Delta_z) - \mathbf{S}(F)}{h}, \tag{3.12}$$

if this limit exists for every $\mathbf{z} \in \mathbb{R}^{p+q}$*.*

Theorem 3. *Consider a zero mean d-dimensional stationary VAR(p) time series of size T given by* $\{\mathbf{z}_t : t = 1, \ldots, T\}$*. Consider the S-estimator,* $\hat{\boldsymbol{\beta}}_{\boldsymbol{z}}$*, of the parameter matrices, considering them as a vector of size* pd^2*, of this model,* $\boldsymbol{\beta}$*, with the associated* ρ*-function,* $\rho(.)$*. Denote the distribution of the associated white noise process by* $\mathbf{F}_\epsilon$*. Assume the following conditions to hold. These correspond to the conditions concerning the* ρ*-function as well as the data, given in theorem 8 of Davies (1990).*

1. *This condition concerns* F_ϵ*.*
   - (a) F_ϵ *has a bounded density* f_ϵ
   - (b) $f_\epsilon(r) = f_\epsilon(-r)$*, i.e.,* f_ϵ *is symmetric around the origin*
   - (c) f_ϵ *is non-increasing on the positive side of the origin.*
2. *This condition concerns the* ρ*-function.*

(a) $\exists \quad c$ *such that* $\rho(.)$ *is strictly increasing on* $[0, c]$ *and constant on* $[c, \infty]$, *i.e., the corresponding* ψ*-function is redescending.*

(b) $\psi(.) = \rho'(.)$ *is bounded.*

(c)

$$\lim_{(v,s)\to(0,0)} \frac{R(v, 1+s) - R(0, 1+s)}{v^2} < 0 \tag{3.13}$$

where

$$R(v, s) = \int \rho(\frac{u-v}{s}) f_\epsilon(u), \quad s > 0$$

3. ρ *and* f_ϵ *have a common point of decrease on the positive side of the origin.*

4. *This condition concerns the data.*

(a) *There exist positive numbers* η_1, η_2 *and* n_0 *such that*

$$\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t^T 1_{\|\mathbf{z}_t\| < \eta_1} - n\eta_2 \mathbf{I}_d$$

is positive definite for all $T \geq n_0$, *where* $\mathbf{I}_d$ *is the* d *dimensional identity matrix.*

(b)

$$\lim_{T\to\infty} \frac{1}{T} \sum \|\mathbf{z}_t\|^2 1_{\|\mathbf{z}_t\| > \delta\sqrt{T}} = 0 \quad \forall \quad \delta > 0 \tag{3.14}$$

Further, assume $\hat{\boldsymbol{\beta}}_{\boldsymbol{z}}$ *to be asymptotically normal with mean* $\boldsymbol{\beta}$. *This is expected to hold under the conditions given in theorem 8 of Davies (1990).*

Then the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{z}}$ *is given by*

$$\gamma \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$$

where $\boldsymbol{\Sigma}$ *is the covariance matrix of the white noise and*

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \dots & \boldsymbol{\Gamma}_{p-1} \\ \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \dots & \boldsymbol{\Gamma}_{p-2} \\ \boldsymbol{\Gamma}_2 & \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_0 & \dots & \boldsymbol{\Gamma}_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}_{p-1} & \boldsymbol{\Gamma}_{p-2} & \boldsymbol{\Gamma}_{p-3} & \dots & \boldsymbol{\Gamma}_0 \end{pmatrix}$$

where $\mathbf{\Gamma}_i$ is the lag i autocovariance of the series and γ is a constant which depends on F_ϵ, the ρ-function (chosen in the S-estimator) and d. The asymptotic norming factor is $\sqrt{T}$.

Proof. We can write the VAR model as follows :

$$\mathbf{z_t} = \begin{pmatrix} z_{1t} \\ \vdots \\ z_{dt} \end{pmatrix} = \begin{pmatrix} \phi_{111} & \dots & \phi_{11d} & \dots & \phi_{p11} & \dots & \phi_{p1d} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{1d1} & \dots & \phi_{1dd} & \dots & \phi_{pd1} & \dots & \phi_{pdd} \end{pmatrix} \begin{pmatrix} z_{1,t-1} \\ \vdots \\ z_{d,t-1} \\ z_{1,t-2} \\ \vdots \\ z_{d,t-1} \\ \vdots \\ z_{1,t-p} \\ \vdots \\ z_{d,t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \vdots \\ \epsilon_{dt} \end{pmatrix},$$

where $\boldsymbol{\epsilon_t} = (\epsilon_{1t}, \dots, \epsilon_{dt})^T$ is a d-dimensional white noise series with distribution F_ϵ.

Denote $\mathbf{x_t} = (z_{1,t-1}, \dots, z_{d,t-1}, z_{1,t-2}, \dots, z_{d,t-2}, \dots, z_{1,t-p}, \dots, z_{d,t-p})^T$, $\mathbf{y_t} = \mathbf{z_t}$ and

$$\boldsymbol{\phi}_c = \begin{pmatrix} \phi_{111} & \dots & \phi_{11d} & \dots & \phi_{p11} & \dots & \phi_{p1d} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{1d1} & \dots & \phi_{1dd} & \dots & \phi_{pd1} & \dots & \phi_{pdd} \end{pmatrix}.$$

Then, we can write the multivariate linear regression model above succinctly as

$$\mathbf{y_t} = \boldsymbol{\phi}_c \mathbf{x_t} + \boldsymbol{\epsilon_t},$$

where now, $\mathbf{y_t}$ is the d-dimensional response, $\mathbf{x_t}$ is the pd-dimensional impulse, $\boldsymbol{\phi}_c$ is the $pd \times pd$ matrix parameter and $\boldsymbol{\epsilon_t}$ is the d-dimensional white noise. Denote $\mathbf{s_t} = (\mathbf{x_t}, \mathbf{y_t})$ and call this model H. Then, Van Aelst and Willems (2005) showed that the influence function of the S-estimator, $\hat{\boldsymbol{\phi}}_{\boldsymbol{cs}}$, at a point $\mathbf{s} = (\mathbf{x}, \mathbf{y})$, for the above model is given by

$$\text{IF}(\mathbf{s}; \hat{\boldsymbol{\phi}}_{\boldsymbol{cs}}, H) = E_H[\mathbf{x}\mathbf{x}^\mathbf{T}]^{-1}\mathbf{x}\text{IF}(\mathbf{y}; \mathbf{M_d}, F_\epsilon)^T, \tag{3.15}$$

where $\mathbf{M_d}$ is the d-dimensional S-estimator of the location of $\mathbf{y}$. Lopuhaä (1989) had shown that

$$\mathrm{IF}(\mathbf{y}; \mathbf{M_d}, F_\epsilon) = \frac{1}{\beta}\psi(\|\mathbf{y}\|)\frac{\mathbf{y}}{\|\mathbf{y}\|}$$

where

$$\beta = E_{F_\epsilon}\Big[\Big(1 - \frac{1}{d}\Big)\frac{\psi(\|\mathbf{a_0}\|)}{\|\mathbf{a_0}\|} + \frac{1}{d}\psi'(\|\mathbf{a_0}\|)\Big]$$

where $\mathbf{a_0}$ is a sample random variable from the distribution F_ϵ. The asymptotic covariance matrix of $\hat{\boldsymbol{\phi}}_{\boldsymbol{cs}}$ can be computed by means of the influence function as

$$ASC(\hat{\boldsymbol{\phi}}_{cs}) = E_H[\mathrm{IF}(\mathbf{s}; \hat{\boldsymbol{\beta}}_{\boldsymbol{z}}, H) \otimes \mathrm{IF}(\mathbf{s}; \hat{\boldsymbol{\beta}}_{\boldsymbol{z}}, H)^T].$$

Denoting $\boldsymbol{\Sigma}_x = E_H[\mathbf{x}\mathbf{x}^{\mathbf{T}}]$, it follows from Equation (3.15) that

$$ASC(\hat{\boldsymbol{\phi}}_{\boldsymbol{cs}}) = ASV(\mathbf{M_d}, F_\epsilon) \otimes \boldsymbol{\Sigma}_x^{-1}, \tag{3.16}$$

where $ASV(\mathbf{M_d}, F_\epsilon)$ is the asymptotic covariance of the S-estimator of the location of $\mathbf{y}$, $\mathbf{M_d}$, under the distribution F_ϵ. Lopuhaä (1989) showed that

$$ASV(\mathbf{M_d}, F_\epsilon) = \frac{\alpha}{\beta^2}\boldsymbol{\Sigma}, \tag{3.17}$$

where

$$\alpha = \frac{1}{d}E_{F_\epsilon}[\psi^2(\|\mathbf{a_0}\|)],$$

where $\mathbf{a_0}$, as defined earlier, is a sample random variable from the distribution F_ϵ and β is as defined earlier. From the definition of $\mathbf{x}$, we compute $\boldsymbol{\Sigma}_x$ as

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \dots & \boldsymbol{\Gamma}_{p-1} \\ \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \dots & \boldsymbol{\Gamma}_{p-2} \\ \boldsymbol{\Gamma}_2 & \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_0 & \dots & \boldsymbol{\Gamma}_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}_{p-1} & \boldsymbol{\Gamma}_{p-2} & \boldsymbol{\Gamma}_{p-3} & \dots & \boldsymbol{\Gamma}_0 \end{pmatrix} \tag{3.18}$$

From Equations (3.16), (3.17) and (3.18), we get

$$ASC(\hat{\boldsymbol{\phi}}_{cs}) = \frac{\alpha}{\beta^2}\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}}^{\boldsymbol{-1}},$$

where α and β are as defined earlier. This completes the proof for the asymptotic covariance matrix form for the S-estimator for a zero mean VAR(p) model. □

Ben et al. (2001) [6] had studied the application of the τ-estimator to the VAR model. They gave a heuristic proof of the consistency of the estimator. The form of the asymptotic covariance matrix for the S-estimator for the VAR model agrees with that of the τ-estimator with appropriately tuned parameters. Ben et al. used an iterative algorithm based on weighted least squares to compute the τ-estimator for a VAR model. In the special case of the S-estimator, we will make use of the Fast-S method by extending its use in the multivariate regression case.

3.5 The Fast-S method to compute S-estimators

The computation of S-estimators is not easy. The Fast-S algorithm, proposed by Yohai and Salibian-Barrera (2006) [76], is an algorithm for the S-estimator. We will propose an extension of the Fast-S algorithm for the multiple regression scenario with particular applications to VAR Models.

3.5.1 The Fast-S algorithm for the univariate scenario

The univariate Fast-S algorithm is a recursive method, analogous to the fast-LTS algorithm (which in turn is based on the fast-MCD or fast minimum covariance determinant algorithm), to compute S-estimates. Given a starting parameter estimate β, the improvement step (I-step, which is the core of the algorithm) is as follows.

1. Compute the residuals $\hat{r}(\beta) = (\hat{r}_1(\beta), \ldots, \hat{r}_T(\beta))$.

2. Compute an approximate scale $\hat{s}$ of $\hat{r}(\beta)$ by applying to Equation (3.5), one step of any iterative algorithm starting from the MAD (median absolute deviation). Here, the iterative algorithm is chosen to be the Newton-Raphson method.

This ensures that we are close to, if not within, the feasible region. By starting from the MAD, we try to avoid outliers' influence on the scale estimate and by applying the iterative algorithm to Equation (3.5) starting from this value, we try to get closer to the feasible region. We perform just one step of this iterative algorithm to be efficient in terms of speed. Since the overall I-Step process is iterative, under suitable conditions, we will converge to a point in the feasible region.

3. Compute the weights

$$w_t = \frac{\psi(\hat{r}_t(\beta)/s)}{\hat{r}_t(\beta)/s} \tag{3.19}$$

where $\psi = \rho'$.

4. The improved candidate β^* is obtained by a weighted least squares with weights defined by (3.19). Steps 3 and 4 ensure that the new estimate has a smaller scale value than that of the previous estimate.

The S-estimator satisfies the same first order condition as the M-estimator does and steps 3 and 4 are part of the iteratively re-weighted least squares (IRLS) method to satisfy this condition. For more details regarding the algorithm, refer to Yohai and Salibian-Barrera (2006) [76].

3.5.2 The Fast-S algorithm for the multivariate scenario

In the multiple regression scenario, with the scale being replaced by a covariance matrix $\mathbf{\Sigma}$, one needs to define how to perform a step of the iterative method (in the I-Step) of step 2. We introduce a novel way of applying the Newton-Raphson method to move to the next point from a staring point.

Consider the general $\mathbb{R}^d \to \mathbb{R}$ function $y = f(x_1, \ldots, x_d)$ whose roots we are interested in finding using the Newton-Raphson method. Then, given that we are at step n of the iteration at the point $\mathbf{r_n} = (\mathbf{x_n}, y_n) = (x_{n1}, \ldots, x_{nd}, y_n)$, we are interested in improving to a better point $\mathbf{x_{n+1}}$.

The tangent at $\mathbf{r_n}$ is a hyperplane that intersects the domain in a hyperplane (which we will refer to as the cutting hyperplane henceforth). Thus, for the next point, one has infinitely many points to choose from, in this cutting hyperplane.

One way of moving forward in a consistent way is to consider the point on the cutting hyperplane that is in the direction of the gradient to the function $f(\mathbf{x_n})$ at $\mathbf{x_n}$.

This is equivalent to working with a projection of the function onto a plane that is perpendicular to the domain plane and contains the gradient vector at $\mathbf{x_n}$. Then, this is the same as restricting the domain to the projection of the gradient at $\mathbf{x_n}$. We thus reduce the problem to the familiar 2-dimensional case.

The convergence of the above mentioned method is guaranteed under appropriate conditions as is the case in the univariate Newton-Raphson method. Sufficient smoothness and starting close to a root of the equation are two key conditions.

As an example, consider the simplest non-trivial case of a 3-dimensional space. We are interested in finding the roots of the equation $z = x^2 + y^2 - 50$. Starting from $\mathbf{x_n} = (7, 7)$, we find that the gradient vector is given by $(\frac{\partial z}{\partial x}|_{(7,7)}, \frac{\partial z}{\partial y}|_{(7,7)})$ which in 3-dimensions, corresponds to the plane $x = y$. Restricting our function z to this plane reduces to the trivial case of considering $z' = 2x'^2 - 50$. Starting from $x' = 7$, we see that the next approximation is $x'' = 5.28$. Thus, by constraining ourselves to remain on the line $y = x$, we make an improvement from $(7, 7)$ to $(5.28, 5.28)$ which is already close to a solution $(5, 5)$. This is illustrated in Figure 3.1.

Formally, the improvement step is as follows.

1. First, we restrict the domain to the line

$$\frac{x_1 - x_{n1}}{\partial y/\partial x_1 \,|_{x_n}} = \ldots = \frac{x_d - x_{nd}}{\partial y/\partial x_d \,|_{x_n}} = t$$

   This is the line in the domain that is the projection of the gradient vector at the current point.

2. Next, we compute the cutting hyperplane

$$\sum_{i=1}^{d} (x_i - x_{ni}) \Big(\frac{\partial y}{\partial x_i}\Big|_{x_n} \Big) = -y_n$$

3. We now find the intersection of the cutting hyperplane and the restricted domain space to be

$$x_{n+1,i} = x_{ni} - \frac{y_n}{\sum_{i=1}^{d} (\partial y/\partial x_i \,|_{x_n})^2} \Big(\frac{\partial y}{\partial x_i}\Big|_{x_n} \Big) \qquad i = 1, \ldots, d$$

   which is the next point we are interested in. It is clear that when $d = 1$, the next point is the same as that found by the Newton-Raphson method for the univariate case. This is because in the univariate case, the restricted domain

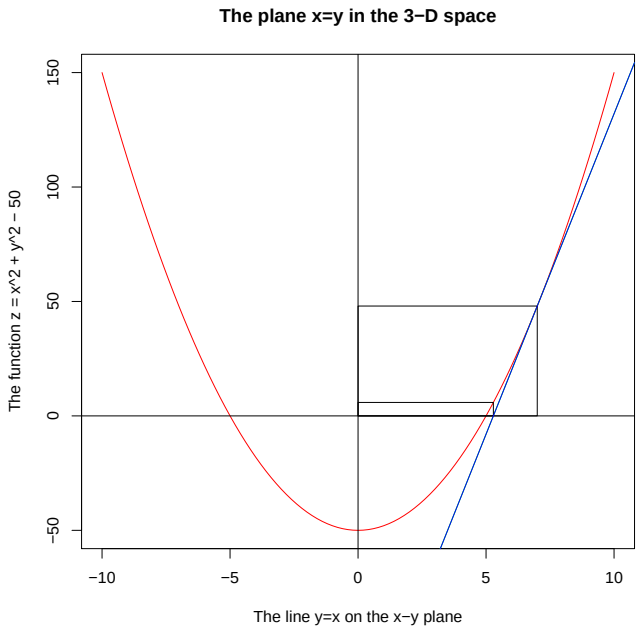

Figure 3.1: An example of performing the Newton-Raphson method in a 3-D space.

(given in (1)) is the whole domain itself which is $y = 0$ while the cutting hyperplane is just a point which is the next point $x_{n+1,1}$.

3.6 S-estimator for a VAR time series

In this section, we will see the application of the Fast-S method for obtaining the S-estimator for a d-dimensional VAR(p) time series model using the Fast-S method described in the earlier section. Given $\mathbf{x_t} = (x_{t1}, \ldots, x_{td})^T$, the model is as follows :

$$\begin{pmatrix} x_{t1} \\ \vdots \\ x_{td} \end{pmatrix} = \sum_{j=1}^{p} \begin{pmatrix} \phi_{j11} & \ldots & \phi_{j1d} \\ \vdots & \ddots & \vdots \\ \phi_{jd1} & \ldots & \phi_{jdd} \end{pmatrix} \begin{pmatrix} x_{t-j,1} \\ \vdots \\ x_{t-j,d} \end{pmatrix} + \begin{pmatrix} \epsilon_{t1} \\ \vdots \\ \epsilon_{td} \end{pmatrix}. \tag{3.20}$$

To simplify the formulae, we will use the notations

$$\mathbf{\Phi_j} = \begin{pmatrix} \phi_{j11} & \ldots & \phi_{j1d} \\ \vdots & \ddots & \vdots \\ \phi_{jd1} & \ldots & \phi_{jdd} \end{pmatrix}$$

and

$$\epsilon_{\mathbf{t}} = (\epsilon_{t1}, \ldots, \epsilon_{td})^T \sim N(\mathbf{0}, \mathbf{\Sigma})$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \ldots & \rho_{1d}\sigma_1\sigma_d \\ \vdots & \ddots & \vdots \\ \rho_{d1}\sigma_d\sigma_1 & \ldots & \sigma_d^2 \end{pmatrix}$$

where $\rho_{ij} = \rho_{ji}$ is the correlation between ϵ_{ti} and ϵ_{tj}. Let $\mathbf{\Phi} = (\mathbf{\Phi_1}, \ldots, \mathbf{\Phi_p})$. The S-estimator for this model is defined as that $\hat{\mathbf{\Theta}} = (\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Phi}})$ which satisfies the following optimization equation.

Minimize Det($\hat{\mathbf{\Sigma}}$)

subject to

$$\frac{1}{(T-p)} \sum_{t=p+1}^{T} \rho(\mathbf{r^T}(\hat{\mathbf{\Phi}})\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}(\hat{\mathbf{\Phi}})) = b \tag{3.21}$$

where $\mathbf{r}_t(\hat{\mathbf{\Phi}})$ is the t^{th} residual, ρ is the M-function and b is a constant, typically equal to $\mathrm{E}[\rho(\mathbf{X})]$ where $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$ is a d-dimensional standard Normal random variable.

The Fast-S being based on sub-sampling, we start the process by randomly choosing a predetermined number of points in the pd^2 dimensional parameter space. Due to stationarity, this space is a subset of the unit pd^2 dimensional space and we select the starting points in this unit space. The way we choose a random point in this unit parameter space is by randomly choosing pd points in the regression space corresponding to the time series data space and then computing the parameter that corresponds to the hyperplane that contains these pd data points. Since each regression data point gives us d equations, pd regression data points give us pd^2 equations and thus, we can compute the parameter vector (consisting of pd^2 scalar components) that satisfies these pd^2 equations exactly. In case the system of equations corresponding to this choice of regression data points does not have a unique solution, we can ignore these data points and repeat the process of randomly choosing pd regression data points until the system of equations corresponding to these regression data points has a unique solution.

In their introduction to the Fast-S, the authors suggest starting with a number, N, of sub-samples such that

$$N \geq \frac{\log(\alpha)}{\log(1-(1-\epsilon)^q)} \approx \frac{-\log(\alpha)}{(1-\epsilon)^q}, \tag{3.22}$$

where q is the dimension of the explanatory regression data and $\epsilon < \epsilon_0, 0 < \alpha < 1$ where ϵ_0 is the breakdown point of the estimator and $1-\alpha$ is the probability that the breakdown point of the resulting algorithm is at least ϵ. This number, N, grows exponentially with the dimension, q, of the explanatory regression variable. In the d-dimensional VAR(p) case, $q = pd^2$ and hence both, the dimension as well as the lag parameter p, play a role in the choice of the number of sub-samples to start the Fast-S process from. We now look at the core of the Fast-S algorithm applied to the d-dimensional VAR(p) model, the I-step.

In the I-Step of the Fast-S, let the parameter estimate at the n^{th} step be called $(\hat{\mathbf{\Phi}}_{\boldsymbol{n}}, \hat{\mathbf{\Sigma}}_{\boldsymbol{n}})$. Then the $(n+1)^{th}$ step involves the following.

1. Compute the residuals $\mathbf{r}(\hat{\mathbf{\Phi}}_{\boldsymbol{n}}) = (\mathbf{r}_{p+1}(\hat{\mathbf{\Phi}}_{\boldsymbol{n}}), \ldots, \mathbf{r}_T(\hat{\mathbf{\Phi}}_{\boldsymbol{n}})) = (\hat{\mathbf{r}}_{p+1}, \ldots, \hat{\mathbf{r}}_T)$.

2. Compute an approximate covariance estimate $\hat{\mathbf{\Sigma}}$ of $\mathbf{r}(\hat{\mathbf{\Phi}}_{\boldsymbol{n}})$ by applying to equation (3.21), one step of the Newton-Raphson algorithm (as described in the earlier section) starting from any robust covariance estimate like the minimum covariance determinant (MCD) estimate.

For simplicity and more importantly speed, one can start from the covariance estimate whose components are computed robustly as

$$\hat{\gamma}_{ij} = \text{Median}_t\{\hat{r}_{ti}\hat{r}_{tj}\} - \text{Median}_t\{\hat{r}_{ti}\}\text{Median}_t\{\hat{r}_{tj}\} \quad \text{for} \quad i,j = 1,\ldots,d.$$

2.1 Call this starting covariance estimate, $\hat{\mathbf{\Sigma}}'_{\boldsymbol{n}}$ and

$$f_n(\mathbf{\Sigma}) = \frac{1}{(T-p)} \sum_{t=p+1}^{T} \rho(\mathbf{r_t}(\hat{\mathbf{\Phi}}_{\boldsymbol{n}})^T \mathbf{\Sigma}^{-1} \mathbf{r_t}(\hat{\mathbf{\Phi}}_{\boldsymbol{n}})) - b \tag{3.23}$$

Then the $\hat{\mathbf{\Sigma}}_{\boldsymbol{n+1}}$ estimate turns out to be as follows.

$$\hat{\sigma}_{i,n+1} = \sigma'_{in} - \frac{f_n(\mathbf{\Sigma}_{\boldsymbol{n}}) f_{n_{\sigma_i}}(\mathbf{\Sigma}_{\boldsymbol{n}})}{\sum_{j=1}^{d} [f^2_{n_{\sigma_j}}(\mathbf{\Sigma}_{\boldsymbol{n}}) + \sum_{0<k<j}^{d} f^2_{n_{\rho_{kj}}}(\mathbf{\Sigma}_{\boldsymbol{n}})]}, \quad i = 1,\ldots,d$$

and

$$\hat{\rho}_{ij,n+1} = \rho'_{ijn} - \frac{f_n(\mathbf{\Sigma}_{\boldsymbol{n}}) f_{n_{\rho_{ij}}}(\mathbf{\Sigma}_{\boldsymbol{n}})}{\sum_{l=1}^{d} [f^2_{n_{\sigma_l}}(\mathbf{\Sigma}_{\boldsymbol{n}}) + \sum_{0<k<l}^{d} f^2_{n_{\rho_{kl}}}(\mathbf{\Sigma}_{\boldsymbol{n}})]}, \quad i,j = 1,\ldots,d, \quad i<j$$

where

$$f_{n_*}(.) = \frac{\partial f_n(.)}{\partial *}$$

Note that we need $-1 \leq \hat{\rho}_{ij,n+1} \leq 1$ and hence, when this condition is not met, we reset $\hat{\rho}_{ij,n+1}$ to zero.
3. Compute the weights

$$w_{tn} = \frac{\psi\left(\sqrt{\mathbf{r_t}^T(\hat{\mathbf{\Phi}}_{\boldsymbol{n}})\mathbf{\Sigma}^{-1}_{n+1}\mathbf{r_t}}\right)}{\sqrt{\mathbf{r_t}^T(\hat{\mathbf{\Phi}}_{\boldsymbol{n}})\mathbf{\Sigma}^{-1}_{n+1}\mathbf{r_t}}} \tag{3.24}$$

where $\psi = \rho'$

4. The improved candidate $\hat{\mathbf{\Phi}}_{\boldsymbol{n+1}}$ is obtained by a weighted least squares with weights defined by (3.24).

3.7 Examples

Example 1. *In this example, we look at the IBM and NASDAQ daily log returns based on the daily closing price. The plots of the daily log returns starting from* 5^th *February 1971 until* 9^{th} *November 2009 are given in Figure* 1.3. *The autocorrelation functions (ACFs) of the two series are also given in figures* 3.2 *and* 3.3. *The IBM and NASDAQ composite show a significant autocorrelation at lag 1. The sample correlation between the two returns is 0.037 which, though small, is not insignificant. Based on this, a VAR(1) model is fit for the bivariate series assuming a Normal noise component. The sharp fall in the IBM price around time index 2000 (from 304 on the* 31^{st} *of May 1979 to 76.25 on the next day,* 1^st *of June 1979; a log return of -0.6) acts as a strong additive outlier making parameter estimation difficult. The parameters are estimated using the multivariate least squares (MLS), the RA estimator and the S estimator.*

To gauge the effectiveness of the robust methods (RA and S) versus the conventional non-robust MLS method, we use 70% *of the data to estimate the parameters and the remainder* 30% *to determine the forecast quality. We use the mean residual quadratic norm (MSE) as a measure of the forecast quality. In addition, we contaminate a single point in the NASDAQ returns data at time index 3000 to 10.0 in one case and a single point in the IBM returns data at index 3000 to -100.0 in another case and re-estimate the parameters and their corresponding forecast qualities. The results are given in the following table. The zero mean VAR(1) model has a single* 2×2 *matrix parameter* $\mathbf{\Phi}$.

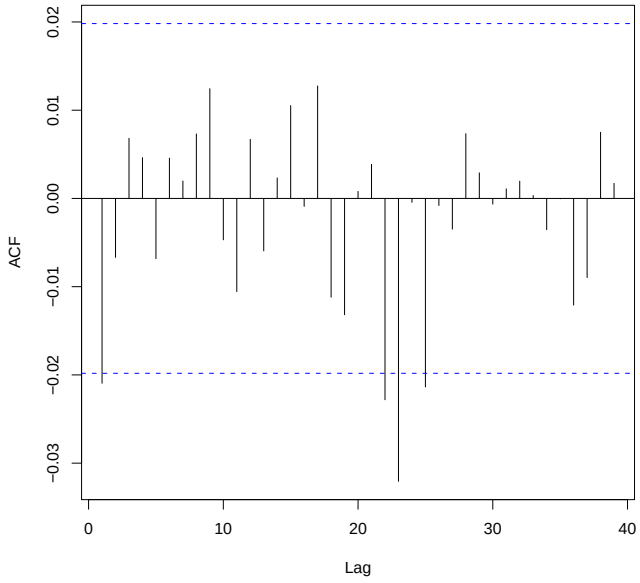

Figure 3.2: The ACF of the daily log returns of IBM closing.

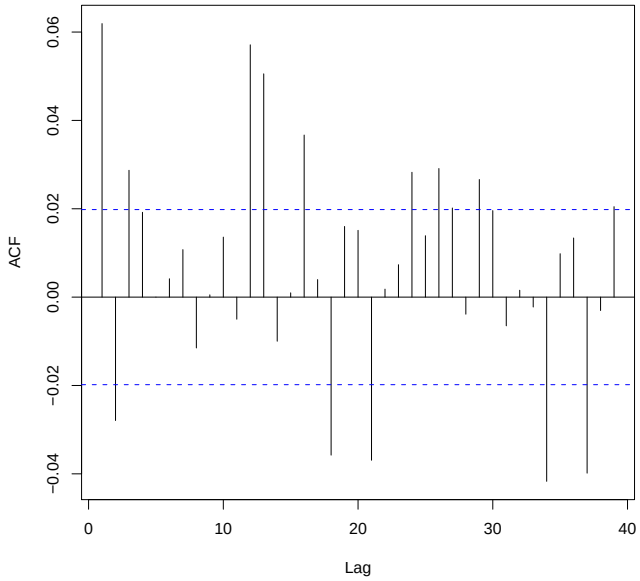

Figure 3.3: The ACF of the daily log returns of NASDAQ closing.

No artificial contamination

Estimator	*Estimate of* $\mathbf{\Phi}$ *(standard error)*		*MSE* ($\times 10^{-5}$)
MLS	*-0.008 (0.0125)*	*-0.002 (0.0127)*	*9.16*
	0.068 (0.0131)	*0.263 (0.0115)*	
S	*-0.0006 (0.0116)*	*-0.002 (0.0127)*	*9.24*
	0.139 (0.0125)	*0.273 (0.0127)*	
RA	*-0.018 (0.0115)*	*-0.0002 (0.0125)*	*9.28*
	0.113 (0.0135)	*0.303 (0.0125)*	

NASDAQ[3000] replaced by 10.0

Estimator	*Estimate of* $\mathbf{\Phi}$ *(standard error)*		*MSE* ($\times 10^{-5}$)
MLS	*-0.007 (0.0121)*	*0.195 (0.0135)*	*9.01*
	0.0001 (0.0117)	*0.0005 (0.0145)*	
S	*-0.0003 (0.0115)*	*-0.002 (0.0145)*	*9.24*
	0.140 (0.0115)	*0.273 (0.0155)*	
RA	*-0.017 (0.0121)*	*-0.0009 (0.0129)*	*9.23*
	0.109 (0.0125)	*0.286 (0.0122)*	

IBM[3000] replaced by -100.0

Estimator	*Estimate of* $\mathbf{\Phi}$ *(standard error)*		*MSE* ($\times 10^{-5}$)
MLS	*-0.0001 (0.0115)*	*0.000006 (0.0166)*	*19.07*
	-1.705 (0.0111)	*0.263 (0.0125)*	
S	*-0.0006 (0.0113)*	*-0.002 (0.0119)*	*9.24*
	0.140 (0.0111)	*0.273 (0.0123)*	
RA	*0.010 (0.0127)*	*-0.0006 (0.0125)*	*9.27*
	0.063 (0.0122)	*0.313 (0.0155)*	

The results show all three estimators doing about equally well in terms of forecasting when there is no artificial contamination. The lag-1 estimates for the IBM series in all the three cases are not significant. The NASDAQ lag 1 data seems to be a leading indicator for both the IBM and the NASDAQ series. In the case of artificial contaminations, the MLS estimator clearly breaks down giving different estimates each time. The RA and S estimators, however, show resistance and their estimates remain consistent.

3.8 Simulations

In this section, we will present some performance metrics after running some simulations.

3.8.1 Scenario 1

For the simulations, for simplicity, we generated 100 samples of size 1000 each, of two-dimensional zero mean VAR(1) data with normal white noise that was then contaminated with additive outliers to varying degrees and the Fast-S method was applied to obtain the S-estimator of the parameters. Parameters were also estimated using the multivariate least squares (MLS) method which is equivalent to the conditional maximum likelihood estimator. Finally, the mean squared error (MSE) was calculated for each of the four parameters of the two estimators (S-estimator and the Least Squares) using all the 100 samples. Given a two-dimensional zero mean VAR(1) model, there are four regression parameters and hence four MSE are reported respectively. The following table gives details of the results.

Contamination %	MLS MSE		S-estimator MSE	
0	0.0005	0.0006	0.0005	0.003
	0.0009	0.0007	0.001	0.0005
1	0.01	0.02	0.0005	0.0002
	0.01	0.01	0.003	0.0007
10	0.05	0.05	0.01	0.02
	0.06	0.08	0.02	0.02
20	0.07	0.06	0.06	0.05
	0.09	0.07	0.09	0.06

As can be seen, the S-estimator and the least squares estimator have similar MSE under no contamination. The marked increase in the MSE under even a small one percent contamination is visible for the least squares estimator whereas the S-estimator remains unaffected at this level of contamination. At 10% and 20% contamination levels, that transform to 20% and 40% contaminations in the corresponding regression data respectively, we see the S-estimator also beginning to get affected.

3.8.2 Scenario 2

In this simulation, we compare the bias and variance of the MLS, RA and S estimators under various rates of contamination and noise covariances. Like in the previous

scenario, we start with no contamination and then progressively increase it. We use samples of size 100 and run 100 simulations to compute the bias and root mean squared error (RMSE) statistics. We will also vary the correlation of the bivariate white noise to see how the results are affected. The variance of the noise components remain the same at 0.0001 and 0.0009 respectively. The additive contaminations happen by contaminating components of individual data points with a Normal noise of variance 0.01. Specifically, we will look at the following scenarios.

No contamination case with noise correlation of 0.7

$$\mathbf{\Phi} = \begin{pmatrix} 0.05 & -0.23 \\ 0.39 & 0.12 \end{pmatrix}$$

Estimator	Bias		RMSE	
MLS	0.001	0.005	0.065	0.063
	-0.005	-0.018	0.035	0.113
RA	0.003	0.025	0.07	0.21
	-0.005	-0.022	0.03	0.11
S	-0.001	0.0008	0.07	0.21
	-0.006	-0.022	0.03	0.011

No contamination case with noise correlation of -0.7

$$\mathbf{\Phi} = \begin{pmatrix} 0.36 & -0.42 \\ -0.23 & 0.21 \end{pmatrix}$$

Estimator	Bias		RMSE	
MLS	-0.018	0.040	0.09	0.25
	-0.005	0.009	0.03	0.11
RA	-0.013	0.039	0.09	0.23
	-0.006	0.004	0.03	0.12
S	-0.010	0.024	0.09	0.95
	-0.001	-0.001	0.03	0.11

No contamination case with noise correlation of 0.0

$$\mathbf{\Phi} = \begin{pmatrix} 0.40 & -0.12 \\ 0.11 & -0.46 \end{pmatrix}$$

Estimator	Bias		RMSE	
MLS	-0.004	0.014	0.09	0.31
	0.001	-0.002	0.03	0.08
RA	-0.005	0.033	0.09	0.032
	0.002	-5.31×10^{-6}	0.03	0.09
S	-0.003	0.003	0.09	0.033
	0.001	-0.003	0.03	0.09

1% contamination case with noise correlation of 0.7

$$\mathbf{\Phi} = \begin{pmatrix} -0.23 & 0.38 \\ 0.21 & -0.29 \end{pmatrix}$$

Estimator	Bias		RMSE	
MLS	0.012	-0.013	3.21	3.22
	-0.017	0.007	3.35	3.34
RA	-0.027	-0.052	0.35	0.45
	0.017	0.037	0.31	0.35
S	-0.001	-0.052	0.07	0.29
	0.002	0.037	0.03	0.11

5% contamination case with noise correlation of 0.7

$$\mathbf{\Phi} = \begin{pmatrix} 0.03 & 0.46 \\ 0.46 & 0.34 \end{pmatrix}$$

Estimator	Bias		RMSE	
MLS	0.032	0.039	6.67	6.78
	-0.002	-0.010	6.88	6.34
RA	-0.195	-0.198	1.01	1.11
	0.155	0.148	0.29	0.95
S	-0.009	-0.016	0.19	0.38
	0.001	-0.0007	0.08	0.11

10% contamination case with noise correlation of 0.7

$$\mathbf{\Phi} = \begin{pmatrix} -0.41 & 0.24 \\ 0.47 & -0.41 \end{pmatrix}$$

Estimator	Bias		RMSE	
MLS	0.804	0.803	4.89	4.45
	-0.715	-0.714	4.67	4.56
RA	0.104	0.099	0.75	0.79
	-0.234	-0.233	0.75	0.81
S	0.007	0.007	0.15	0.11
	-0.001	0.005	0.19	0.19

Like in scenario one, the S-estimator and the least squares estimator have similar bias and MSE under no contamination. The marked increase in the MSE under even a small one percent contamination is visible for the least squares estimator whereas the S-estimator is affected relatively slightly at this level of contamination. At 5% and 10% contamination levels, that transform to 10% and 20% contaminations in the corresponding regression data respectively, we see the S-estimator also beginning to get affected, albeit to a much lower level compared to the least squares estimator.

3.9 Summary

As seen in the last section, the S-estimator is highly robust and useful in estimating parameters for VAR(p) models, especially under additive contamination. The Fast-S is an efficient method to compute the S-estimator in a reasonable time period. In this chapter, we demonstrated the use of a multivariate version of the Fast-S to compute the S-estimator for a vector autoregressive model. Using the S-estimate as a starting value, one can then obtain an M-estimate leading to the well known MM-estimator which uses a robust starting estimate to compute the an M-estimate.

We also saw the propagation of outliers in time series data and how that can blow up the outlier proportion in the associated model data. This is a serious drawback of all methods that base the estimation techniques on optimizing some function of the residuals. This problem is briefly discussed in the next chapter on bilinear models where the problem's seriousness increases manifold.

CHAPTER 4

Nonlinear time series analysis : The bilinear model

4.1 Introduction

While linear time series models are relatively easy to analyze, interpret and use in many data analysis applications, the very simplicity of these models is also a weakness. This is because many real-life data are complex enough that linear models are unable to capture their features. For example, it is well known that the conditional variance of a financial time series is not constant but in fact, volatile. This can be seen when comparing the scenarios of a positive versus a negative trend. There is comparatively lesser volatility during a bull market run versus a bear market run. This can be attributed to the nervousness of the investors, especially the retail ones, when the market is going down. This translates to a pressure to sell which contributes to the already increased volatility. Thus, models like the autoregressive conditional heteroscedastic (ARCH) and the generalized ARCH (GARCH) come into play that incorporate nonlinearity in the conditional variance equation of the time series model.

Many nonlinear models have been proposed in the time series literature. The idea of using simulation and data driven methods is central to these models. Most recently, nonparametric and semi-parametric methods like kernel regression and neural networks have also been studied in the context of time series modeling.

In this chapter we will examine the class of bilinear models and look at the application

of the S-estimator to robustly estimate parameters of a simple model of this class. In the next section, we will briefly look at some of the commonly used nonlinear models in time series analysis.

4.2 State of the art

Traditional nonlinear models for time series include the bilinear model of Granger and Andersen (1978) [24], the threshold autoregressive (TAR) model of Tong (1978) [67], the state-dependent model of Priestley (1980) [50] and the Markov switching model of Hamilton (1989) [25]. The primary idea behind these models is to model the conditional mean using some parametric nonlinear function.

More recently, taking advantage of the advances in computing technologies, a number of nonlinear models have been proposed. These include the nonlinear state-space modeling of Carlin, Polson and Stoffer (1992) [10], the functional coefficient autoregressive (FCAR) model of Chen and Tsay (1993) [13], the nonlinear additive autoregressive model of Chen and Tsay (1993) [14] and the multivariate adaptive regression spline of Lewis and Stevens (1991) [32]. The ideas in these models is to use either simulation methods to describe the evolution of the conditional distribution of x_t or data-driven methods to explore the nonlinear characteristics of x_t.

Finally, most recently, nonparametric and semi-parametric methods such as kernel regression and artificial neural networks have also been applied to time series analysis to study nonlinear properties of data. We next briefly discuss some of the above mentioned models.

4.2.1 Threshold Autoregressive (TAR) Model

This model came about to describe the asymmetry in declining and rising patterns of a process, as discussed before. It uses piecewise linear models to obtain better approximations of the conditional mean process. However, as opposed to periodic processes modeling that use piecewise linear models incorporating periodicity in the time space, TAR models use the data space by incorporating thresholds to switch between piecewise linear models. The thresholds are commonly referred to as regimes.

Consider the following simple 2-regime AR(1) TAR model.

$$x_t = \begin{cases} -1.5x_{t-1} + \epsilon_t & \text{if } x_{t-1} < 0, \\ 0.5x_{t-1} + \epsilon_t & \text{if } x_{t-1} \geq 0, \end{cases} \tag{4.1}$$

where ϵ_t is a white noise process with zero mean and variance σ_ϵ^2. This model illustrates some characteristics of TAR models.

Firstly, in spite of the coefficient -1.5 in the first regime, the process x_t is geometrically ergodic and stationary. The *ergodic theorem* refers to the property that shows that the sample mean of a mean stationary time series, $\{x_t\}$, given by $\overline{x}_T = (\sum_{t=1}^{T} x_t)/T$, converges to μ, the stationary expectation of x_t as $T \to \infty$. In fact, one can go a step further and look at a general 2-regime AR(1) TAR process given as

$$x_t = \begin{cases} \alpha x_{t-1} + \epsilon_t & \text{if } x_{t-1} < 0, \\ \beta x_{t-1} + \epsilon_t & \text{if } x_{t-1} \geq 0, \end{cases} \tag{4.2}$$

For this model to be geometrically ergodic, a set of necessary and sufficient conditions are

$$\alpha < 1, \quad \beta < 1 \quad \text{and} \quad \alpha\beta < 1.$$

For more details on this model and the derivation of these conditions, the reader is referred to Petrucelli and Woolford (1984) [47] and Chen and Tsay (1991) [12].

Secondly, the series exhibits an asymmetric increasing and decreasing pattern. If x_{t-1} is negative, then x_t tends to switch to a positive value due to the negative and explosive coefficient -1.5. At the same time, however, when x_{t-1} is positive, it takes x_t multiple time indices to reduce to a negative value. As a consequence, the time plot of x_t shows that regime 2 has more observations than regime 1. In addition, the series contains large upward jumps when it becomes negative.

Finally, the model does not contain any constant terms yet $E(x_t) \neq 0$. In general, $E(x_t)$ is a weighted average of the conditional means of the two regimes, which are non-zero.

The two regime TAR AR(1) model described before can be generalized to a k-regime TAR AR(p) model with threshold variable x_{t-d} as follows :

$$x_t = \mu_i + \sum_{j=1}^{p} \phi_{ij} x_{t-j} + \epsilon_{it}, \quad \text{if} \quad \gamma_{i-1} < x_{t-d} < \gamma_i,$$

where

1. k and d are positive integers,
2. $i = 1, \ldots, k$,
3. γ_i are real numbers such that $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_{k-1} < \gamma_k = \infty$ and
4. ϵ_{it} is a white noise process for each i and they are mutually independent.

The parameter d is referred to as the *delay parameter* while the γ_i are referred to as the thresholds. Such TAR models where the thresholds are compared to some series values are referred to as *self-exciting* TAR (SETAR) models. The SETAR model is linear for $k = 1$.

4.2.2 The smooth transition AR (STAR) model

A possible drawback of the SETAR model is the discontinuity of the conditional mean equation. The thresholds, $\{\gamma_i\}$, are the discontinuity points of the conditional mean function μ_t. To overcome this, smooth TAR models have been proposed in Chan and Tong (1986) [11] and Teräsvirta (1994) [66]. A series $\{x_t\}$ is said to follow a 2-regime STAR(p) process if it satisfies the following :

$$x_t = c_0 + \sum_{i=1}^{p} \phi_{0i} x_{t-i} + F\Big(\frac{x_{t-d} - \delta}{s}\Big)\Big(c_1 + \sum_{i=1}^{p} \phi_{1i} x_{t-i}\Big) + \epsilon_t, \tag{4.3}$$

where d is the usual delay parameter, δ and s are parameters representing the location and scale of the model transition process, and $F(.)$ is a smooth transition function. In practice, $F(.)$ commonly assumes one of the following three forms, namely, logistic, exponential or a cumulative distribution function. From Equation (4.3), the conditional mean of a STAR model is a weighted linear combination of the following two equations :

$$\mu_{t1} = c_0 + \sum_{i=1}^{p} \phi_{0i} x_{t-i}$$

and

$$\mu_{t2} = (c_0 + c_1) + \sum_{i=1}^{p} (\phi_{0i} + \phi_{1i}) x_{t-i}.$$

The weights are determined in a continuous manner by $F((x_{t-d} - \delta)/s)$. The above two equations also determine some properties of a STAR model. For example, a necessary condition for the stationarity of a STAR model is that all zeros of both the AR polynomials be outside the unit circle. A possible advantage of the STAR model over the TAR model is that the conditional mean function is differentiable. However, empirical studies have shown that the estimation of the transition parameters δ and s is a difficult problem. In particular, the standard errors of these estimates are often quite large, resulting in t-ratios of about 1.0; see Teräsvirta (1994). This uncertainty leads to complications in interpreting an estimated STAR model.

4.2.3 The Markov switching model

The idea of using probability switching in nonlinear time series analysis is discussed in Tong (1983) [68]. Using a similar idea, but emphasizing aperiodic transitions between various states of an economy, Hamilton (1989) [25] considered the Markov switching autoregressive (MSA) model. Here the transition is driven by a hidden two state Markov chain. A time series x_t is said to follow a two state MSA model if

$$x_t = \begin{cases} c_0 + \sum_{i=1}^{p} \phi_{0i} x_{t-1} + \epsilon_{t0} & \text{if} \quad s_t = 0, \\ c_1 + \sum_{i=1}^{p} \phi_{1i} x_{t-1} + \epsilon_{t1} & \text{if} \quad s_t = 1, \end{cases} \tag{4.4}$$

where s_t, taking values in $\{0, 1\}$, is a first order Markov chain with transition probabilities

$$P(s_t = 1 | s_{t-1} = 0) = w_0 \quad \text{and} \quad P(s_t = 0 | s_{t-1} = 1) = w_1.$$

The series $\{\epsilon_{t0}\}$ and $\{\epsilon_{t1}\}$ are mutually independent white noise series. $1/w_i$ is the expected duration of the process to stay in state i. An MSA model thus uses a hidden Markov model chain to govern the transition from one conditional mean function to the other. This is a differentiating factor from the SETAR model where the transition is determined by the value(s) of lagged variable(s). As a consequence, a SETAR model uses a deterministic scheme to govern the model transition whereas an MSA model uses a stochastic scheme.

In practice, the stochastic nature of the state in an MSA model implies that one is never certain about which state x_t belongs to at any given time instant t. When the sample size is large, one could possibly use some filtering techniques to draw inferences on the state of x_t. In contrast, as long as x_{t-d} is observed, the regime of x_t is known in a SETAR model. This difference has important practical consequences in forecasting. For example, forecasts in an MSA model are always a linear

combination of forecasts produced by sub-models of individual states. But those of a SETAR model come from only a single regime provided x_{t-d} is observed.

Forecasts of a SETAR model also become a linear combination of those produced by sub-models of individual regimes when the forecast horizon exceeds d, the delay parameter. It is much harder to estimate parameters in an MSA model than in other models since the states are not directly observable. Hamilton (1990) [26] used the EM algorithm, a statistical method involving iterations between taking expectations and maximizations, while McCulloch and Tsay (1994) [42] considered a Markov chain Monte Carlo (MCMC) method to estimate a general MSA model.

McCulloch and Tsay (1993) [41] generalized the MSA model in Equation (4.4) by considering the transition probabilities w_0 and w_1 to be logistic or probit functions of some explanatory variables available at time $t-1$. The MSA model can be generalized to the case involving more than two states. However, the computational intensity involved in estimating the model increases rather rapidly. For more on Markov switching models in econometrics, see Hamilton (1994, Chapter 22) [27].

4.2.4 The functional coefficient autoregressive (FCAR) model

The phenomenon of differences in characteristics of increasing and decreasing patterns of a typical financial time series motivates one to consider models with time-variant parameters. The simple linear model can be extended to incorporate such time varying parameters. The functional coefficient autoregressive (FCAR) model is such a model. It is an AR model with variable parameters that are in turn driven by past data. More precisely, the model is as follows :

$$x_t = \sum_{i=1}^{p} \phi_{ti} x_{t-i} + \epsilon_t$$

where $\{\epsilon_t\}$ is a white noise process and ϕ_{ti} represent the dynamic AR coefficients. These parameters are assumed to be functionally driven by past data as $\phi_{ti} = f_i(x_{t-1}, \ldots, x_{t-q})$ for some lag q. Kernel regression and local linear regression, as described in the last section, are typically used to estimate these functional parameters. While the functions, $f_i(.)$, are assumed to have some properties like continuity and twice differentiability, research into stability and stationarity properties of the FCAR model has been somewhat limited.

4.2.5 The nonlinear additive AR model

A major difficulty in applying nonparametric methods to nonlinear time series modeling is the *curse of dimensionality*. Considering a general nonlinear autoregressive model of lag p as

$$x_t = f(x_{t-1}, \ldots, x_{t-p}) + \epsilon_t,$$

we see that estimating $f(.)$ without assuming any form of f would require p-dimensional smoothing. This is hard to do when p is large, particularly when the available data size is small. One way of going around this problem, without assuming any parametric models, is to assume an additive form of $f(.)$. That is, assume

$$f(x_{t-1}, \ldots, x_{t-p}) = \sum_{i=1}^{p} f_i(x_{t-i}).$$

With such a model, only 1-dimensional smoothing is required to estimate the function in a nonparametric way. The nonlinear additive AR (NAAR) model is thus defined as

$$x_t = f_0(t) + \sum_{i=1}^{p} f_i(x_{t-i}) + \epsilon_t$$

where $\{\epsilon_t\}$ is a white noise process, $f_i(.)$ are nonparametric functional coefficients with $f_0(t)$ representing the time dependent trend component.

4.2.6 The nonlinear state-space model and neural networks

A simple state-space model for time series modeling is as follows :

$$s_t = f_t(s_{t-1}) + u_t, \qquad x_t = g_t(s_t) + v_t,$$

where s_t is the unobservable state vector, x_t is the observed time series, $f_t(.)$ and $g_t(.)$, the transition functions, are known functions of some unknown parameters and $\{u_t\}$ and $\{v_t\}$ are mutually independent white noise series. Monte Carlo methods are typically used to estimate the transition functions while the use of smoothing and Monte Carlo Markov chain (MCMC) methods is somewhat limited.

With advances in computing technologies, neural networks have been used in analyzing nonlinear time series data. Section 10 of Ripley (1993) [54] gives some remarks concerning applications of neural networks in financial applications.

4.2.7 Nonparametric models

There exist financial and econometric applications where one may not have sufficient knowledge about the dependence structure within the data. Yet, one is interested in analyzing the functional relationship that drives the data. In such situations, one could turn to nonparametric methods. However, these methods have a cost associated with them. They are highly data driven and can therefore easily result in over-fitting.

Smoothing is central to non-parametric methods and models. Consider the following dependence structure.

$$y_t = f(x_t) + \epsilon_t,$$

where $\{(x_t, y_t)\}$ is a two dimensional time series with x_t driving y_t by the relationship function f, and $\{\epsilon_t\}$ is a white noise series. Suppose we are interested in estimating the dependence function f. To start with, suppose further that we are interested in estimating f at a particular value x of x_t. That is, we are interested in estimating $f(x)$. Suppose further that for $x_t = x$, we have s repeated independent observations in y as $\{y_{x1}, \ldots, y_{xs}\}$. Then, we have that

$$y_{xi} = f(x) + a_{xi}, \quad i = 1, \ldots, s.$$

Taking the sample mean for this sample, we have

$$\overline{y}_x = \frac{\sum_{i=1}^{s} y_{xi}}{s} = f(x) + \frac{\sum_{i=1}^{s} a_{xi}}{s}.$$

By the law of large numbers, the average of the innovations converges to zero as $s \to \infty$. Therefore, the sample average, $\overline{y}_x$, is a consistent estimator of $f(x)$. This demonstrates the method of smoothing.

In most real-life data oriented applications, we do not have the luxury of repeated observations like in the above example. But if the dependency function f is sufficiently smooth, then $y_t (= f(x_t))$, for which $x_t \approx x$, still provides a good approximation for $f(x)$. Similarly, y_t for which x_t is far away from x provides lesser reliable information regarding $f(x)$. This leads to the idea of using a weighted average of y_t with the weights being inversely proportional to some distance metric between x_t and x. More precisely, given a time series data $\{z_t\} = \{x_t, y_t : t = 1, \ldots, T\}$ with a dependence function f, we can write a smoothed estimate of $f(x)$ as

$$\hat{f}(x) = \sum_{i=1}^{T} w(x, x_t) y_t,$$

where w is the standardized weight function such that $\sum_{i=1}^{T} w(x, x_t) = 1$. It is clear from the above equation that the estimate is simply a locally weighted average with weights determined by the distance metric used and the weight function w. Kernel regression and local linear regression are two examples of smoothing techniques to estimate dependence functions.

4.3 The bilinear model

4.3.1 The univariate bilinear model

A natural way to move from linear to nonlinear models is to introduce quadratic terms in the model equation. One can thus formulate a general second order model for a time series as follows :

$$\begin{aligned} x_t = \alpha &+ \sum_{i=1}^{p} \beta_i x_{t-i} + \sum_{j=1}^{q} \gamma_j \epsilon_{t-j} \\ &+ \sum_{i=1}^{p_{xx}} \sum_{j=1}^{q_{xx}} \eta_{xxij} x_{t-i} x_{t-j} \\ &+ \sum_{i=1}^{p_{ee}} \sum_{j=1}^{q_{ee}} \eta_{eeij} \epsilon_{t-i} \epsilon_{t-j} \\ &+ \sum_{i=1}^{p_{xe}} \sum_{j=1}^{q_{xe}} \eta_{xeij} x_{t-i} \epsilon_{t-j} \\ &+ \epsilon_t \end{aligned} \tag{4.5}$$

The first line of the above equation, consisting of the first three terms, is simply an autoregressive moving average (ARMA) model. The other three terms are the quadratic terms, of which, the last one considers cross products between innovations and the series itself.

The bilinear model focuses on this last heterogenous term and does not consider the homogenous terms involving only the innovations and series terms respectively. The reasons for avoiding the homogeneous terms is because these terms make the model

less tractable and less stable. More precisely, the quadratic terms involving only the innovations makes the model non-invertible. This is not desirable since from an application point of view, invertibility is essential. For example, forecasting is an important application in time series modeling and lack of invertibility makes prediction a hard problem. On the other hand, homogeneous terms involving only the series' terms renders the time series hard to analyze, which, from an analysis point of view, makes studying the properties of the series difficult.

A bilinear model of order (p, q, r, s) is defined as

$$\begin{aligned} x_t = \alpha &+ \sum_{i=1}^{p} \beta_i x_{t-i} + \sum_{j=1}^{q} \gamma_j \epsilon_{t-j} \\ &+ \sum_{i=1}^{r} \sum_{j=1}^{s} \eta_{ij} x_{t-i} \epsilon_{t-j} \\ &+ \epsilon_t \end{aligned} \tag{4.6}$$

Special extensions of the bilinear model incorporate conditional heteroscedasticity as follows :

$$\begin{aligned} x_t = \alpha &+ \sum_{i=1}^{p} \beta_i x_{t-i} + \sum_{j=1}^{q} \gamma_j \epsilon_{t-j} \\ &+ \sum_{i=1}^{r} \sum_{j=0}^{s} \eta_{ij} x_{t-i} \epsilon_{t-j} \\ &+ \epsilon_t \end{aligned} \tag{4.7}$$

The subtle difference in the conditional heteroscedastic bilinear model is the inclusion of cross-product terms involving the current innovation and past series values in the definition of the current series value. In this chapter, we will focus on univariate bilinear models and robustly estimating their parameters.

Bilinear model categories

Bilinear models are typically classified into three categories. These are the diagonal, sub-diagonal and super-diagonal bilinear models. While the sub-diagonal model's cross-product terms look like

$$\sum_{i=1}^{r}\sum_{\substack{j=1\\ j<i}}^{s} x_{t-i}\epsilon_{t-j},$$

the super-diagonal model's cross-product terms look like

$$\sum_{i=1}^{r}\sum_{\substack{j=1\\ j>i}}^{s} x_{t-i}\epsilon_{t-j}.$$

Finally, the diagonal bilinear model's cross-product terms are given by

$$\sum_{i=1}^{r} x_{t-i}\epsilon_{t-i}.$$

Analysis

Bilinear models are known to be able to model occasional large fluctuations in time series. Figure 4.1 depicts an example. One can see the intermittent large fluctuations in the series that characterizes the occasional large variance. In such situations, bilinear models are useful in modeling time series where the conditional variance is stochastic.

These models have been studied extensively in the literature. An early paper is Granger and Anderson (1978). The bilinear model is so called because it has linear components in x_t and ϵ_t separately. In this section, we will briefly study some properties of simple bilinear models.

In order to demonstrate the properties of bilinear models, they are often put in a matrix form. That is, Equation (4.6) is written as follows. Define

$$\mathbf{y_t} = \begin{pmatrix} x_t \\ \vdots \\ x_{t-l+1} \end{pmatrix}$$

and

$$\mathbf{z_t} = \begin{pmatrix} \epsilon_t \\ \vdots \\ \epsilon_{t-q} \end{pmatrix},$$

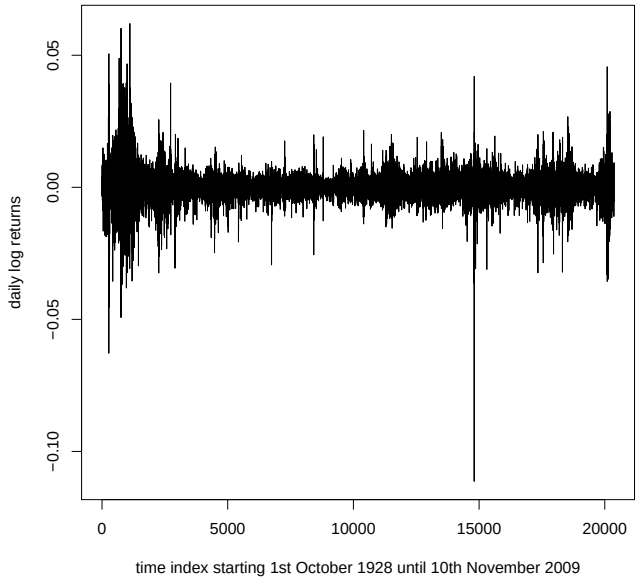

Figure 4.1: The daily log returns of the DOW Jones Industrial Average.

where $l = \max(p, r)$. Now define the $l \times l$ matrices

$$\mathbf{A} = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_p & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \ddots & & & & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & \dots & \dots & 0 & 1 & 0 \end{pmatrix}$$

and

$$\mathbf{B} = \begin{pmatrix} \eta_{1j} & \eta_{2j} & \dots & \eta_{rj} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}$$

and the $l \times (q+1)$ matrix

$$\mathbf{C} = \begin{pmatrix} 1 & \gamma_1 & \dots & \gamma_q \\ 0 & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

With the above notations, Equation (4.6) can be written as

$$\mathbf{y_t} = \mathbf{A}\mathbf{y_{t-1}} + \mathbf{C}\mathbf{z_t} + \sum_{j=1}^{s} \mathbf{B_j}\mathbf{y_{t-1}}\epsilon_{t-j}, \tag{4.8}$$

which can further be simplified as

$$\mathbf{y_t} = (\mathbf{A} + \sum_{j=1}^{s} \mathbf{B_j}\epsilon_{t-j})\mathbf{y_{t-1}} + \mathbf{C}\mathbf{z_t}. \tag{4.9}$$

In the case where $s = 1$ and $q = 0$, the above equation reduces to

$$\mathbf{y_t} = (\mathbf{A} + \mathbf{B}\epsilon_{t-1})\mathbf{y_{t-1}} + \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.10}$$

Akamanam et al. (1983) [1] showed that under the conditions

$$E(\epsilon_t) = 0 \quad \text{and} \quad E(\epsilon_t^2) = \sigma_\epsilon^2 < \infty, \tag{4.11}$$

Equation (4.10) has a strictly stationary solution provided the maximum eigenvalue of the matrix

$$\mathbf{A} \otimes \mathbf{A} + \sigma_\epsilon^2 \mathbf{B} \otimes \mathbf{B}$$

is strictly less than unity. Here, we can see that the eigenvalues of the matrix $\mathbf{B}$ need not necessarily be less than unity. This can be explained by looking at the dimension of the components of $\mathbf{B}$. If we call the dimension of the returns series d_y, then the dimension of the components of $\mathbf{B}$ is d_y^{-1} and hence these components are expected to be of order $(10\sigma_y)^{-1}$ which can take values larger than unity. The reason for dividing by 10 is that otherwise, the product $bx_{t-i}\epsilon_{t-j}$ (b being a component of $\mathbf{B}$) could explode. In other words, looking at the product, we could expect $|bx_{t-i}| < 1$ for the product to not explode.

Marginal Distribution

While the conditional distribution of x_t in a bilinear time series model is the same, F_ϵ, as that of the associated white noise process $\{\epsilon_t\}_{t\in T}$, its marginal distribution is not easy to compute. This can be demonstrated by looking at the (0, 0, 1, 1) model.

$$x_t = ax_{t-1}\epsilon_{t-1} + \epsilon_t = \epsilon_t + \sum_{i=1}^{t-1} a^i \epsilon_{t-i} \Pi_{j=1}^{i} \epsilon_{t-j}$$

This equation shows that x_t is a sum of random variables that are in the form of a product of i.i.d random variables. Distributions of products of up to three Gaussian variables have been computed. However, as is seen, the products involved in the summation are O(t).

However, it is possible to compute the first and second moment of this marginal distribution. These are given below.

$$E[x_t] = a\sigma_\epsilon^2$$

and

$$\mathrm{Var}(x_t) = \sigma_\epsilon^2 + a^2\tau_\epsilon^4 \frac{1-(a^2\sigma_\epsilon^2)^t}{1-a^2\sigma_\epsilon^2} - a^2\sigma_\epsilon^4,$$

where τ_ϵ^4 is the fourth moment of ϵ_t. We will now study two simple bilinear models, the (0, 0, 1, 1) and (1, 0, 1, 1) diagonal bilinear models and understand their behavior.

The (1,0,1,1) bilinear model

We consider a bilinear (1,0,1,1) model given by :

$$x_t = bx_{t-1} + ax_{t-1}\epsilon_{t-1} + \epsilon_t \tag{4.12}$$

where $\{\epsilon_t\}$ is a white noise process with distribution F_ϵ and variance σ_ϵ^2. We start the analysis of this model by looking at the conditional distribution. Given the past information until time $t-1$, it is easy to see that

$$x_t|x_{t-1}, x_{t-2}, ... \sim F_\epsilon(bx_{t-1} + ax_{t-1}\epsilon_{t-1}, \sigma_\epsilon^2).$$

That is, the conditional distribution of the present X_t value given the past information until time $t-1$ is centered around $bx_{t-1} + ax_{t-1}\epsilon_{t-1}$ with standard deviation σ_ϵ.

Next, we look at stationarity and causality in the context of this simple bilinear model. Similar to the definition in the linear time series case, the definition of causality for a general bilinear model as defined in Equation (4.6) is as follows. A bilinear process $\{x_t\}$ of order (p, q, r, s) is said to be causal if there exists a measurable function $f : \mathbb{R}^\infty \to \mathbb{R}$, such that $x_t = f(\epsilon_t, \epsilon_{t-1}, \ldots)$ a.s. for all $t \in (0, \pm 1, \ldots)$.

For the simple bilinear model defined by Equation (4.12), Pham and Tran (1981) [48] have shown that the condition

$$b^2 + a^2\sigma_\epsilon^2 < 1 \tag{4.13}$$

is necessary and sufficient for the existence of a causal stationary solution. Similar results for higher order bilinear models can be found in Liu and Brockwell (1988) [36]. From Equation (4.12), it can be seen that the variance and autocovariance of a stationary solution will involve higher order moments. This is because of the presence of cross-product terms in the model equation. By repeated substitutions in Equation (4.12), we get the following :

$$x_t = \sum_{i=0}^{\infty} \epsilon_{t-i}\Pi_{j=1}^{i}(b + a\epsilon_{t-j}).$$

Thus, when condition (4.13) is met, the above expression shows that that this bilinear process is causal. Under stationarity and a Gaussian white noise, we can compute the

stationary mean, variance and autocovariance functions for the (1, 0, 1, 1) bilinear model. These computations give us the following.

$$\mu_x = \frac{a\sigma_\epsilon^2}{1-b},$$

$$\sigma_x^2 = \gamma_0 = \sigma_\epsilon^2 \frac{1 + a^4\sigma_\epsilon^4 + a^2\{(1-b)^2\tau_\epsilon^4/\sigma_\epsilon^2 - 2\sigma_\epsilon^2(1-b^2)\} + 4b(1-b)a^3\sigma_\epsilon^2}{(1-b)^2(1-a^2\sigma_\epsilon^2)},$$

$$\gamma_1 = b\gamma_0 + \frac{a^2\sigma_\epsilon^4}{1-b}$$

and

$$\gamma_k = b\gamma_{k-1}, \quad k > 1.$$

We can see that the autoregressive parameter a drives the autocovariance function. The stationarity condition given in Equation (4.13) ensures that $|b| < 1$ which ensures that the autocovariance function diminishes with the lag. The lag-1 autocorrelation is positive irrespective of the model parameters. The sign of the stationary mean, μ_x, is determined by the sign of the bilinear parameter, a. The stationarity condition given by Equation (4.13) ensures that $b \neq 1$ which in turn ensures that the above mentioned moments are well defined.

The (0, 0, 1, 1) Bilinear Model

In this section, we will study the most simple diagonal bilinear process, the (0, 0, 1, 1) bilinear model. This is given by

$$x_t = ax_{t-1}\epsilon_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process with variance σ_ϵ^2. Simple arithmetic shows that, under stationarity, the stationary mean and variance of the process are given by

$$\mu_x = a\sigma_\epsilon^2, \quad \sigma_x^2 = \gamma_0 = \sigma_\epsilon^2 \frac{1 + a^4\sigma_\epsilon^4 + a^2(\tau_\epsilon^4/\sigma_\epsilon^2 - 2\sigma_\epsilon^2)}{1 - a^2\sigma_\epsilon^2}.$$

where τ_ϵ^4 is the fourth moment of the white noise ϵ_t. The stationarity condition in this case reduces to $a^2\sigma_\epsilon^2 < 1$. Further, simple calculations yield the following.

$$\gamma_k = \begin{cases} a^2\sigma_\epsilon^4 & k = 1 \\ 0 & k > 1 \end{cases} \tag{4.14}$$

The mean corrected version of this model is as follows :

$$x_t = b + ax_{t-1}\epsilon_{t-1} + \epsilon_t,$$

where b is the new term in the model that, when equal to $-a\sigma_\epsilon^2$, renders the stationary mean of the series zero. The general stationary mean and variance of this model is

$$\mu_x = b + a\sigma_\epsilon^2$$

and

$$\sigma_x^2 = \sigma_\epsilon^2 \frac{1 + a^4\sigma_\epsilon^4 + a^2(\tau_\epsilon^4/\sigma_\epsilon^2 - 2\sigma_\epsilon^2) + a^2 b(b + a\sigma_\epsilon^2)}{1 - a^2\sigma_\epsilon^2}.$$

The stationary covariance in this case is given by

$$\gamma_k = \begin{cases} a\sigma_\epsilon^2(b + a\sigma_\epsilon^2) & k = 1 \\ 0 & k > 1 \end{cases} \tag{4.15}$$

4.3.2 Outlier propagation in bilinear models

In a VAR(p) model, we saw how an additive outlier at time index t transforms into (p+1) outliers in the associated regression model as p outliers in the regressor and one outlier in the response variable. A similar but much more serious effect exists in bilinear models. This is because, a single additive outlier at index t corrupts the entire dataset in the associated regression model starting from that index. This is demonstrated in a (0, 0, 1, 1) bilinear model next :

$$x_t = ax_{t-1}\epsilon_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process with variance σ_ϵ^2. By repeated substitution, this equation can be rewritten as :

$$x_t = -\sum_{i=1}^{t-1} (-a)^i x_{t-i}^2 \Pi_{j=1}^{i-1} x_{t-j} + \epsilon_t = \sum_{i=1}^{t-1} a^i \epsilon_{t-i}^2 \Pi_{j=1}^{i-1} \epsilon_{t-j} + \epsilon_t. \tag{4.16}$$

From the above formulation, it is clear that a single additive contamination at index s shows up as an outlier, not only in the response variable at index s, but also in the impulse variables for all further data points. Again, as in the VAR case, the effect of this outlier on further regression data points diminishes according to the parameter a and the distance $t - s$. However, recalling the stationarity condition for a bilinear model, it is seen that the parameter a can take values greater than unity

in modulus for a sufficiently small σ_ϵ^2. Hence, the dampening effect of the outlier in the regression data indirectly depends on the variance of the white noise process.

Thus, for a given bilinear series $\{x_t : t = 1, \ldots, 2T+1\}$, any additive contamination at or before the time index T, contaminates more than 50% of the associated regression data and thus any estimator will break down in such a scenario. This is a serious drawback of estimators based on residuals like the least squares, Generalized M, least trimmed median of squares and S estimators. The drawback can, however, be alleviated to some extent by using robust filters to compute the residuals as introduced by Masreliez (1975) [40]. Further research in this direction can be found in the work of Muler et al. (2009) [45] who introduces the Bounded Innovation Propagation ARMA (BIP-ARMA) model to limit the affect of a single outlier to a single data point within the associated regression model.

4.3.3 Parameter estimation for a simple univariate bilinear model

We now look at parameter estimation techniques for the (0, 0, 1, 1) diagonal bilinear model discussed before. The conditional maximum likelihood (MLE) method (conditional on the first $\max(p, q, r, s)$ data points) is commonly applied to estimate parameters in bilinear models. When the white noise is assumed to be Normal, the conditional MLE reduces to the method of nonlinear least squares.

It is easy to see that the optimization problem does not have a nice closed form solution as in the usual linear regression problem. The Newton-Raphson method can be applied to compute the nonlinear least squares estimate of the autoregressive parameters. Alternately, an iterative least squares procedure can also be used to compute the estimate.

Like in the case of vector autoregressive models, the application of robust methods in analyzing bilinear models has been somewhat limited. Gabr (1998) [21] studied the application of the generalized M-estimator in the estimation of parameters in bilinear models. However, little is known about the asymptotic properties of this estimator in the bilinear model context. As a second aim of this thesis, we will study the application of the S-estimator to the univariate bilinear model parameter estimation problem. In particular, we will compute and analyze the S-estimator for the (0, 0, 1, 1) bilinear model.

The (0, 0, 1, 1) bilinear model estimation

Recall the definition of the (0, 0, 1, 1) diagonal bilinear model. This is given by :

$$x_t = ax_{t-1}\epsilon_{t-1} + \epsilon_t, \tag{4.17}$$

where $\{x_t : t = 1, .., T\}$ is the time series in question and $\{\epsilon_t : t = 1, ..T\}$ is the associated white noise process with variance σ_ϵ^2.

The least squares estimator

In this case, the optimization equation of the least squares method for estimating the sole parameter a is given by :

Minimize $\sum_{i=1}^{n} r_t^2(\hat{a})$,

where $r_t(a) = x_t - ar_{t-1}x_{t-1}$ are the residuals.

Pham and Tran (1981) showed the consistency of the nonlinear least squares estimator for a first order bilinear time series which is the (1, 0, 1, 1) bilinear model. Liu (1990) [34] further analyzed the asymptotic distribution of this estimator under some conditions. Next, we will briefly look at how to compute this estimator and its asymptotic properties.

The diagonal bilinear model as given in equation (4.17) can be written, as in equation (4.16) as :

$$x_t = -\sum_{i=1}^{t-1} (-a)^i x_{t-i}\Pi_{j=1}^{i} x_{t-j} + \epsilon_t.$$

Hence, the residuals are given by :

$$r_t(a) = \sum_{i=1}^{t-1} (-a)^i x_{t-i}\Pi_{j=1}^{i} x_{t-j} + x_t.$$

The least squares estimator is the solution to :

$$\sum_{t=2}^{T} r_t(a)r_t'(a) = 0$$

From the definition of the residuals, we can simplify the above to :

$$\sum_{t=2}^{T}(x_t - ax_{t-1}r_{t-1})r'_t(a) = 0$$

One can now expand $r_t(a)$ and $r'_t(a)$ in the above equation using the definition of the residuals and perform a Newton Raphson procedure to compute the solution.

Given weights $\{w_t : t = 2, \ldots, T\}$, the weighted version of the nonlinear least squares can be done by solving for the weighted version of the above equation given by

$$\sum_{t=2}^{T}(x_t - ax_{t-1}r_{t-1})r'_t(a)w_t = 0$$

Liu (1990) showed that when

$$E[r'^2_t(a_0)] < \infty, \quad \text{and} \quad E[|r''_t(a_0)|] < \infty,$$

where a_0 is the true parameter value, i.e., $r_t(a_0) = \epsilon_t$,

$$\sqrt{T}(\hat{a} - a_0) \xrightarrow{d} N(0, \sigma^2_\epsilon / E[r'^2_t(a_0)]). \tag{4.18}$$

In the case of the diagonal (0, 0, 1, 1) model, the necessary conditions reduce to :

$$E[(\sum_{i=1}^{t-1} i(-a)^{i-1}x_{t-i}\Pi^i_{j=1}x_{t-j})^2] < \infty$$

and

$$E[|\sum_{i=1}^{t-1} i(i-1)(-a)^{i-1}x_{t-i}\Pi^i_{j=1}x_{t-j}|] < \infty.$$

Thus, when the above conditions are met, the least squares estimator of the diagonal (0, 0, 1, 1) model is asymptotically normal as given in equation (4.18).

Pham and Tran (1981) had earlier, in their analysis of the first order bilinear time series model, left the analysis of the asymptotic distribution open. The reason was their doubt regarding whether the above mentioned conditions could be met at all. This is because, as is clearly seen, the terms involved in the sums representing $r'_t(a)$ and $r''_t(a)$ are of the type $\Pi^i_{j=1}x_{t-j}$ and hence higher order moments come into play.

Rao (1981) [52] studied the general $(p, 0, r, s)$ bilinear model and showed that the maximum likelihood estimator had an asymptotic multivariate Gaussian distribution. Kim et al. (1990) [30] considered the method of moments estimator for the (1, 0, 1, 1) bilinear model and showed that it was asymptotically normal.

Brunner and Hess (1995) [8] provide a succinct critique of the bilinear model. They list some undesirable properties of this model with particular emphasis on parameter estimation. The main critique has been the bimodal nature of the maximizing objective function and the narrow spike that characterizes the true optimum. This further reduces the power of the associated t-tests for hypothesis testing.

The above mentioned problem can be tackled to a large extent by using advanced optimization tools like artificial neural networks and advanced computing power to search a large parameter subspace.

We now look at estimating parameters of a bilinear time series model when the data is contaminated. In particular, we will look at the S-estimator and analyze its asymptotic distribution.

The S-estimator

Recall the definition of the S-estimator from the last chapter. For the bilinear (0, 0, 1, 1) model, the S-estimator is defined as

Minimize s^2 subject to

$$\frac{1}{(T-1)} \sum_{t=2}^{T} \rho(r_t(a)/s) = \alpha \tag{4.19}$$

where s represents a scale estimate, $r_t(a)$ are residuals, n is the sample data size, α is a constant, typically equal to $E_F(\rho(X))$ (where F is the white noise distribution and X is a random variable having this distribution but with unit variance) and ρ is an M function like the Tukey's bi-weight function. Like before, we use the Fast-S method to compute the S-estimate.

Given a starting parameter estimate $\boldsymbol{\theta_n} = (a_n, s_n)$ at step n, the improvement step (I-step, which is the core of the algorithm) is as follows.

1. Compute the residuals $\hat{r}(\boldsymbol{\theta_n}) = (\hat{r}_2(\boldsymbol{\theta_n}), \ldots, \hat{r}_T(\boldsymbol{\theta_n}))$.

2. Compute an approximate scale $\hat{s}$ of $\hat{r}(\boldsymbol{\theta_n})$ by applying to Equation (4.19), one step of any iterative algorithm starting from the MAD (median absolute deviation). Here, we use the Newton-Raphson method.

Call the median of the computed residuals $\hat{s}_n$. Then, the one step Newton-Raphson improvement is given by

$$\hat{s}_{n+1} = \hat{s}_n + \frac{\sum_{t=1}^{T} \rho(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_n) - \alpha(T-1)}{\sum_{t=1}^{T} \psi(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_n)(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_n^2)}$$

3. Compute the weights

$$w_t = \frac{\psi(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_{n+1})}{\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_{n+1}} \tag{4.20}$$

where $\psi = \rho'$.

4. The improved candidate $\boldsymbol{\theta_{n+1}}$ is obtained by a weighted nonlinear least squares with weights defined by (4.20).

In the case of nonlinear models too, the S-estimator tends to retain its favorable characteristic of a 50% breakdown point. For more on this, the reader is referred to Sakata and White (2001) [63].

Asymptotic distribution

Like in Liu (1990), we establish the asymptotic normality of the S-estimator under some regulatory conditions. However, we will assume consistency of the S-estimator since simulations support this assumption.

Theorem 4. *Consider a (0, 0, 1, 1) stationary bilinear time series,* $\{x_t : t = 1, .., T\}$, *given by*

$$x_t = a x_{t-1} \epsilon_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ *is a white noise process with variance* σ_ϵ^2. *Consider the S-estimator of* a, $\hat{a}$, *with the associated* ρ *function given by* $\rho(.)$. *Denote by* $r_t(\tilde{a})$ *the residual of the* t^{th} *observation for a given parameter estimate* $\tilde{a}$. *Assume that* $\rho'(r_t(a)/s) = \psi(r_t(a)/s) = r_t(a)\alpha_t$ *where* $r_t^{-1}(a)$ *is not a factor of* α_t *and*

1). $E[\epsilon_t \alpha_t] = 0$,
2). α_t *is independent of* $\{\epsilon_{t-i} : i > 0\}$,
3). $E[|r_t''(a)|] < \infty$
4). $E[(\epsilon_t \alpha_t r_t'(a))^2] = \sigma_\epsilon^2 E[\alpha_t^2 r_t'^2(a)] = \sigma_\epsilon^2 \gamma_1 < \infty$,
5). $E[r_t'^2(a)\alpha_t] = \gamma_2 < \infty$,
6). $\psi(.)$ *is redescending and bounded and*
7). The S-estimator is consistent.

Then,

$$\sqrt{T}(\hat{a} - a) \xrightarrow{d} N(0, \sigma_\epsilon^2 \gamma_1 / \gamma_2^2).$$

Proof. The proof follows from the standard Taylor expansion of the minimizing equation. We can do this since the estimator is assumed to be consistent and hence for a large sample size, we can assume the estimator to be close enough to the true parameter. The optimization equation of the S-estimator is :

Min s^2 subject to

$$1/(T-1) \sum_{t=2}^{T} \rho(r_t(a)/s) = b.$$

Using Lagrange multipliers, we get the unconstrained minimization problem as :

$$Min[s^2 + \lambda\{\sum_{t=2}^{T} \rho(r_t(a)/s) - b(T-1)\}].$$

Differentiating and solving for $\hat{a}$ gives the same conditions as those for the classical robust M-estimator, namely :

$$\sum_{t=2}^{T} \psi(r_t(\hat{a})/s) r_t'(\hat{a}) = 0.$$

From the primary assumption, namely $\psi(r_t(a)/s) = r_t(a)\alpha_t$, the above equation simplifies to

$$\sum_{t=2}^{T} r_t(\hat{a}) \alpha_t r_t'(\hat{a}) = 0.$$

Denote

$$f(a) = \sum_{t=2}^{T} r_t(a)\alpha_t r_t'(a).$$

Then, standard Taylor expansion of $f(.)$ at $\hat{a}$ gives

$$0 = f(\hat{a}) = f(a) + (\hat{a} - a)f'(a) + O_p((\hat{a} - a)^2).$$

Under consistency, ignoring second order terms, we thus get,

$$\hat{a} - a = -\frac{f(a)}{f'(a)}$$

Since $r_t(a) = \epsilon_t$,

$$f(a) = \sum_{t=2}^{T} r_t(a)\alpha_t r_t'(a) = \sum_{t=2}^{T} \epsilon_t \alpha_t r_t'(a) = \sum_{t=2}^{T} \beta_t.$$

From assumptions 1 and 2, we get

$$E[\beta_t] = 0 \quad \text{since } r_t'(a) \text{ does not depend on } \epsilon_t$$

and from assumption 4, we get

$$E(\beta_t^2) = \sigma_\epsilon^2 \gamma_1.$$

From the central limit theorem then, we conclude that

$$f(a)/(T-1) = \sum_{t=2}^{T} \epsilon_t \alpha_t r_t'(a)/(T-1) \xrightarrow{d} N(0, \sigma_\epsilon^2 \gamma_1).$$

We now look at $f'(a)$. We have that

$$f'(a) = \sum_{t=2}^{T} [r_t'^2(a)\alpha_t + r_t(a)\alpha_t' r_t'(a) + r_t(a)\alpha_t r_t''(a)].$$

Once again noting that $r_t(a) = \epsilon_t$, we can simplify the above to

$$f'(a) = \sum_{t=2}^{T} [r_t'^2(a)\alpha_t + \epsilon_t \alpha_t' r_t'(a) + \epsilon_t \alpha_t r_t''(a)].$$

From assumptions 1 and 2 and noting that $r'_t(a)$ and $r''_t(a)$ do not depend on ϵ_t, the expectations of the second and third term in the summand vanish. From assumption 5, we then get

$$\lim_{T\to\infty} f'(a)/(T-1) = \gamma_2.$$

Combining the equations governing $f(a)$ and $f'(a)$, we arrive at the desired result which is

$$\sqrt{T}(\hat{a} - a) \xrightarrow{d} N(0, \sigma_\epsilon^2 \gamma_1/\gamma_2^2).$$

□

Note how $\alpha_t = 1$ reduces the estimator to the least squares estimator and conditions 3 and 5 then are exactly as in the least squares case as was given by Liu (1990), described earlier. Conditions 1 and 2 are met by any standard ρ function like the Tukey's biweight function. Conditions 3, 4 and 5 involve higher order moments and therefore are similar in nature to the least squares estimator case.

Simulations and Examples

In this section, we will look at an example time series data and a simulated series.

Example 2. *The daily log returns of the NASDAQ composite is given in Figure* 1.2. *The ACF was plotted and is given in Figure* 3.3 *and a string lag-1 autocorrelation was observed. To account for the observed conditional heteroscedasticity in the data plot, a bilinear (0, 0, 1, 1) model was fitted for a sample of the first 714 data points. This was done because the iterative methods used to compute the least squares and S-estimator are of $O(n^3)$. A mean correction was not done to maintain simplicity. To gauge the quality of the model and the fit, the sample data was divided into two parts. The first part, consisting of* 70% *(500) of the data, was used to fit the bilinear model. The second part, consisting of the remainder data (200), was used to calculate the median absolute forecast error (MAFE). Next, the NASDAQ data was contaminated with a single additive outlier (AO). This was done by replacing the data point at index 495 by 10.0 first and then by 100.0. The parameters were then estimated under each contamination respectively. The estimated parameters are given in the following table.*

AO	*Estimator*	*Estimate (standard error)*	*MAFE*
0.0	*Least Squares*	*16.00 (0.0003)*	*0.0014*
0.0	*S*	*16.88 (0.0023)*	*0.0014*
10.0	*Least Squares*	-2.32×10^{-5} *(0.0039)*	*0.0015*
10.0	*S*	*20.16 (0.0002)*	*0.0014*
100.0	*Least Squares*	-2.32×10^{-5} *(0.0041)*	*0.0015*
100.0	*S*	*20.16 (0.0003)*	*0.0014*

As discussed before in the section about outlier propagation in bilinear time series data, the single additive outlier at time index 495 results in 5 outliers in the regression model which corresponds to a 1% *contamination level in the regression model. Even such a low level results in drastically altering the least squares estimate. The S-estimator, however, appears to be reasonably resistant at this level of contamination. The degree of contamination (10 versus 100) does not seem to be affect either estimator in any substantial way. This suggests that the contaminant level of 10 itself is perhaps higher than the threshold above which the estimators are affected in any reasonable way.*

Another point to note is the value of the estimators when the data is not contaminated. The values of 16 *and* 20.16 *are clearly greater than unity which is atypical of most stationary time series models. However, as discussed before, the stationarity condition of a bilinear model allows the parameters to be greater than unity in modulus depending on the variance of the associated white noise process. The estimated values are of* $O(10)$ *which corroborates with* $\frac{1}{10\sigma_y}$ *since* $\sigma_y \sim O(0.01)$.

Simulation scenario 1

We will now present some performance metrics after running some simulations. For the simulations, for simplicity, we generated 100 samples of size 100 each, of bilinear (0, 0, 1, 1) data with normal white noise with variance 0.01, and parameter $a = 9.0$, that was then contaminated with a single additive outlier by replacing the value at time index 97, by 10.0 first, 100.0 next, and 1000.0 finally, in three separate contamination scenarios. This translates to a 3% contamination in the associated regression model. The Fast-S method was applied to obtain the S-estimator of the parameter. The sole parameter was also estimated using the least squares (LS) method which is equivalent to the conditional maximum likelihood estimator. Finally, the bias and root mean squared error (RMSE) were calculated for the two estimators (S-estimator and the Least Squares) and the following table gives details of the results.

Additive outlier	LS (Bias and RMSE)	S-estimator (Bias and RMSE)
N/A	7.78×10^{-6}, 0.03	-1.089×10^{-4}, 0.031
10	-2.211, 26.34	0.005, 0.032
100	-5.59, 11.28	-0.002, 0.031
1000	-5.69, 8.51	-5.115×10^{-5}, 0.031

The simulations show how a single additive outlier affects the least squares estimator while the S-estimator shows some resistance to the contamination.

Simulation scenario 2

In this next simulation exercise, we generated 100 samples of size 100 each, of bilinear (0, 0, 1, 1) data with normal white noise with variance 0.01, and parameter $a = 9.0$, that was then contaminated with additive outliers. Bias and root mean squared errors (RMSE) were calculated for the S-estimator and the least squares estimator of the forecast parameter a. The contamination was done by replacing the value at time index 95 by 100.0 which translates to a 5% contamination in the associated regression model. The following table gives details of the results.

Additive outlier	LS (Bias and RMSE)	S-estimator (Bias and RMSE)
100	-6.03, 7.93	6.503×10^{-4}, 0.03

4.3.4 The multivariate bilinear model

As discussed in earlier chapters, multivariate models are important in analyzing econometric and financial time series data because of the presence of correlation within various time series data. In this section, we will briefly look at the multivariate bilinear model.

We start with the definition and some properties of the multivariate bilinear model. This model, of order (p, q, r, s) and dimension d, is defined as

$$\mathbf{x_t} = \boldsymbol{\mu} + \sum_{i=1}^{p} \boldsymbol{A_i}\mathbf{x_{t-i}} + \sum_{j=1}^{q} \boldsymbol{B_j}\epsilon_{\mathbf{t-j}} + \sum_{i=1}^{r}\sum_{j=1}^{s} \mathbf{C_{ij}} Vec(\mathbf{x_{t-i}}\boldsymbol{\epsilon_{t-j}^{T}}) + \boldsymbol{\epsilon_t}, \tag{4.21}$$

where $\{\mathbf{x_t}\}$ is the d-dimensional time series in question, $\boldsymbol{\mu}$ is a d-dimensional constant vector, A_i and B_i are $d \times d$ square matrix parameters, C_{ij} are $d \times d^2$ matrix parameters that dictate how the second order interactions between the innovations

and series contribute to the conditional mean equation and $\{\boldsymbol{\epsilon_t}\}$ is a d-dimensional vector white noise with covariance matrix $\boldsymbol{\Sigma_\epsilon}$.

Like in the univariate case, multivariate bilinear models are also analyzed by putting them in a matrix form. Assume without loss of generality, that $p \geq r$ and $q \geq s$ and define the following :

$$\mathbf{z_t} = \begin{pmatrix} \mathbf{x_t} \\ \vdots \\ \mathbf{x_{t-p+1}} \\ \boldsymbol{\epsilon_t} \\ \vdots \\ \boldsymbol{\epsilon_{t-q+1}} \end{pmatrix},$$

$$\mathbf{B} = \left(\begin{array}{cccc|cccc} \mathbf{A_1} & \dots & \mathbf{A_{p-1}} & \mathbf{A_p} & \mathbf{B_1} & \dots & \mathbf{B_{q-1}} & \mathbf{B_q} \\ \mathbf{I} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ -- & -- & -- & -- & -- & -- & -- & -- \\ & & & & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ & & \mathbf{0} & & \mathbf{I} & \dots & \mathbf{0} & \mathbf{0} \\ & & & & \vdots & \ddots & \vdots & \vdots \\ & & & & \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} \end{array}\right),$$

$$\mathbf{w_t} = \begin{pmatrix} \boldsymbol{\epsilon_{t+1}} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \boldsymbol{\epsilon_{t+1}} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

and a matrix $\mathbf{C}$ which contains the elements of the $\mathbf{C_{ij}}$ matrices in a suitable arrangement such that

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \mathbf{C_{ij}}\text{vec}(\mathbf{y_{t-i}}\boldsymbol{\epsilon}'_{\boldsymbol{t-j}}) = \mathbf{C}\text{vec}(\mathbf{z_t}\mathbf{z}'_{\mathbf{t}}).$$

Then, we can write the bilinear model in a state space form as

$$\mathbf{z_{t+1}} = \mathbf{B}\mathbf{z_t} + \mathbf{w_t} + \mathbf{C}\text{vec}(\mathbf{z_t}\mathbf{z_t'}),$$

$$\mathbf{y_t} = [\mathbf{I_d} \quad \mathbf{0} \quad \dots \quad \mathbf{0}]\mathbf{z_t}$$

where $\mathbf{I_d}$ is the d-dimensional identity matrix, $\mathbf{z}_t$ represents the unobservable state of the system and $\mathbf{y}_t$ represents the observable output of the system. Since this equation involves nonlinear terms, namely $\text{vec}(\mathbf{z_t}\mathbf{z_t'})$, this is an example of a nonlinear state space system.

The vector bilinear models are estimated using standard nonlinear least squares and the M-estimator and S-estimator can be used to estimate parameters robustly as in the VAR case. Having already demonstrated the use of robust procedures like the S-estimator in the univariate bilinear model, we leave the application of the S-estimator in the multivariate case open. Apart from other reasons, an important reason for this is the present lack of use of multivariate bilinear models in real life applications due to the rather strict stationarity conditions that are hard to verify empirically.

4.4 Summary

Many financial and econometric time series data exhibit nonlinear characteristics. In this chapter, we saw some prominent nonlinear models in time series modeling. We took a brief look at various nonlinear models proposed in literature like the threshold based, semi-parametric and nonparametric models.

We analyzed the bilinear model in particular since it is the most natural way to move from linear to nonlinear models. We saw the stationarity conditions associated with this model and looked at other properties like causality. We also saw the multivariate version of this model and how it model could be specified in a state-space framework.

We saw briefly the drawback of estimators based on residuals, in estimating bilinear models. This drawback, due to the fast propagation of additive outliers in a bilinear time series model, can be overcome by using robust filters and models based on bounded innovation propagation (BIP). Such models seem to be best suited to time series modeling because of the phenomenon of outlier propagation. Hence, application of BIP based models in modeling bilinear time series would give us good

alternatives to the M and S estimators.

Finally, we applied the multivariate Newton-Raphson method to compute the nonlinear least squares (LS) estimate of the parameters of the bilinear (0, 0, 1, 1) process and could immediately see its lack of resistance to outliers. As a result, we computed the S-estimator, using the Fast-S method, which is a robust estimator, and compared it with the LS estimator. The simulation results demonstrate the advantage of using robust estimators when there is even little contamination in the data.

CHAPTER 5

Conditional heteroscedasticity in time series

5.1 Introduction

As discussed in the earlier chapters, financial and econometric data often exhibit conditional heteroscedasticity which means that the conditional variance of the data series varies with time.

The study of the evolution of this conditional variance is important in many financial and econometric applications. For example, derivatives pricing is highly dependent on the volatility of the underlying asset's returns. The well known Black-Scholes options pricing formula consists of the variance term. Another application is quantifying the Value-at-Risk or VaR of an asset or a portfolio of assets. Finally, the volatility index of a market has recently become a financial instrument. The VIX volatility index from by the Chicago Board of Option Exchange (CBOE) is an example.

Many statistical models have been proposed in the statistical and econometric literature to study this aspect of time series. Most of these models deal with specifying an equation for the varying conditional variance. We saw some simple linear and nonlinear models in earlier chapters that dealt with the conditional mean equation governing the evolution of a time series. Combining the two aspects of conditional mean and variance is typically done by

1. specifying a mean equation first by testing for serial dependence,
2. using the residuals from the mean equation to test for serial dependence within the squared residual series,
3. if any serial dependence is found in the squared residual series, then specifying a volatility model for it and performing a joint estimation of the mean and volatility equations and finally
4. checking the fitted model for statistical significance of the parameters and refining if necessary.

The first step is done by specifying any linear or nonlinear model for the mean equation by looking at the partial autocorrelation function (PACF) and having some understanding of the data context. The second step involves looking at the square of the residuals and checking for any serial dependence in that series. This can again be done by looking at the PACF of the squared residual series. If any such serial correlation is found, then one can specify some volatility equation for the noise ϵ_t. Given a time series data $\{x_t\}$ and an associated model it follows, given by

$$x_t = f(x_{t-1}, ..., x_{t-p}) + \epsilon_t$$

where f is a measurable function, a volatility equation can be specified as

$$\epsilon_t = \sigma_t \delta_t, \quad \sigma_t^2 = f(x_{t-1}, x_{t-2}, ...).$$

where ϵ_t is the shock at time t, σ_t^2 is its variance and $\{\delta_t\}$ is a white noise series with unit variance. Then, one can estimate parameters of the mean and volatility equation jointly and finally refine the model if any parameters are found to be statistically insignificant.

The process of joint estimation is not straightforward as one needs to use an iterative procedure to solve for the estimating equations. This is because the estimating equation, even in the simplest cases, is highly nonlinear in nature. A further aim of this thesis is to consider a simple model that incorporates a linear conditional mean equation and a linear conditional variance equation in a single model and study its properties. This is the first order diagonal conditional heteroscedastic bilinear model. This will be the focus of this chapter. Before that, we will briefly discuss some of the primary models for conditional heteroscedasticity that are in existence today. In the following section, we briefly look at some volatility characteristics and basics of volatility model building.

5.2 Volatility characteristics and testing for conditional heteroscedasticity

A special feature of volatility of a financial or econometric series is that it is not directly observable. That is, as opposed to returns, volatility is not realized. More precisely, given R_t represents the return at time t and V_t^2 represents its conditional variance (both $\{R_t\}_{t\in T}$ and $\{V_t^2\}_{t\in T}$ being stochastic processes), one knows x_t, the realization of R_t, precisely at time t while one does not know σ_t^2, the realization of V_t^2, at any time. Hence, one uses $s_t^2 = (x_t - \hat{f}_t(x_{t-1}, x_{t-2}, ...))^2$ as an approximate realization of V_t^2 (where $f_t(.)$ is the conditional mean equation of R_t and $\hat{f}(.)$ is an estimate of $f_t(.)$). This is analogous to the lack of knowledge of the true variance of the white noise process in most standard linear regression problems. One uses an approximation (sample median absolute deviation (MAD) or standard deviation) to infer properties like asymptotic variance of the least squares estimator.

An an example, the daily volatility is not observable for the daily log return series of the NASDAQ index. This is because there is only one observation per day for this return series. However, intra-day data such as fifteen minute returns or the daily high-low range can possibly be used as an estimate of the daily volatility. This said, the precision of such estimates warrants some study and analysis.

Then there is the concept of *implied volatility*. As the name suggests, the volatility here is derived from some model that implicitly implies the volatility. For example, if one accepts the hypothesis that asset prices follow a geometric Brownian motion and the corresponding options are priced according to the Black-Scholes model, then the Black-Scholes options pricing formula could be used to deduce the volatility of the asset price series. As seen, the underlying hypothesis is rather strong and in cases where it is not met, using implied volatility could lead to incorrect modeling and analysis.

Testing for ARCH (autoregressive conditional heteroscedasticity) effects is the starting point for constructing volatility models. This is done by looking at the squared residual series and applying the Ljung-Box statistics or Lagrange multipliers test of Engle (1982) [20], to this series. Conditional heteroscedastic models can be classified into two general categories. Those in the first category use an exact function to describe the evolution of the conditional variance, σ_t^2, while those in the second category use a stochastic equation for the same. We will now look at some of the well known conditional heteroscedastic models currently in use.

For the remainder of this chapter, we will refer to the conditional mean as μ_t, conditional variance as σ_t^2 and innovation or shock as ϵ_t.

5.3 State of the art

The autoregressive conditional heteroscedastic (ARCH) model of Engle (1982) uses autoregression to describe the evolution of σ_t^2. The primary idea governing this model is that ϵ_t, the white noise component of the associated model, is serially uncorrelated but dependent and that this dependence can be quantified by a simple quadratic function of lagged values. Mathematically, an ARCH(m) model is given by

$$\epsilon_t = \sigma_t \delta_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i \sigma_{t-i}^2, \tag{5.1}$$

where $\{\delta_t\}$ is a white noise series with unit variance and $\alpha_i \geq 0 \quad \forall \quad i \geq 0$. Like in the AR($p$) model, the coefficients must satisfy some regularity conditions for the ARCH model to be stable, i.e., for σ_t^2 to be finite. It can be seen from the above equation that large shocks (in modulus) are followed by large shocks. This is quite often observed in financial time series and is referred to as volatility clustering.

ARCH models have some weaknesses. Most important among them is the assumption that positive and negative shocks have the same effect on volatility. In practice, it is observed that prices respond differently to positive and negative shocks. More precisely, negative shocks are typically followed by higher volatility than volatility that follows positive shocks.

The order of an ARCH model can be determined using the partial autocorrelation function (PACF). The parameters can be estimated using the conditional least squares or conditional maximum likelihood methods. The estimating equations here are conditioned on the first m shocks which are usually assumed known and hence dropped from the equation. Post estimation, the model is checked for adequacy by looking at the standardized residuals

$$\tilde{\epsilon}_t = \frac{\hat{\epsilon}_t}{\hat{\sigma}_t}.$$

Bollerslev (1986) [7] proposed an extension of the ARCH model called the generalized ARCH (GARCH) model. The idea here was to use past conditional variances, in

addition to past shocks, to describe the evolution of σ_t. Mathematically, a GARCH (m, s) model is given by

$$\epsilon_t = \sigma_t \delta_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{i=1}^{m} \beta_i \sigma_{t-i}^2, \tag{5.2}$$

where, as before, $\{\delta_t\}$ is a unit variance white noise process and all parameters are nonnegative. For stability of the process, the sum of squares of all parameters is assumed less than unity. The estimation of parameters in a GARCH model is more involved than in the ARCH case. A two pass estimation method, whereby one estimates the conditional mean equation first ignoring any ARCH effects and then uses the residual series as an observed series to estimate the conditional volatility model, works well with a large sample size. However, the statistical properties of this method has not been rigorously investigated.

Just like the autoregressive integrated moving average (ARIMA) model that has a unit root in its AR polynomial, the integrated GARCH (IGARCH) model has a unit root in its AR polynomial. More precisely, an IGARCH(1, 1) model is defined as

$$\epsilon_t = \sigma_t \delta_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + (1 - \alpha_1)\epsilon_{t-1}^2,$$

where $\{\delta_t\}$ is the usual unit variance white noise process and $0 < \alpha_1 < 1$. As mentioned in Chapter 1, such processes with unit roots in the characteristic equations are useful in modeling non-stationary processes. The resulting non-stationarity can be removed by differencing the series.

In financial return series, the return of a derivative may depend on its volatility. This is typically the case with speculative derivatives. To model such a phenomenon, one could use the MGARCH model, where the "M" stands for GARCH in the mean. In this case, the conditional volatility equation is the same as in the GARCH model but the conditional mean equation now becomes

$$\mu_t' = \mu_t + c\sigma_t^2,$$

where the parameter c is called the risk premium parameter. The above conditional mean equation implies serial correlations in the return series x_t, now given by

$$x_t = \mu_t' + \epsilon_t.$$

These correlations are introduced by those in the volatility process $\{\sigma_t^2\}$. The existence of a risk premium may thus be another reason that many historical asset

returns have serial correlations.

We saw earlier that a weakness of the ARCH model is that positive and negative shocks impact the volatility of further shocks the same way. This weakness exists in the GARCH, IGARCH and MGARCH models as well. To overcome this, Nelson (1991) [46] proposed the exponential GARCH (EGARCH) model. In particular, to allow for asymmetric effects between positive and negative shocks, he considered the EGARCH(m, s) model given below.

$$\epsilon_t = \sigma_t \delta_t, \quad ln(\sigma_t^2) = \alpha_0 + \frac{1 + \sum_{i=1}^{s-1} \beta_i B^i}{1 - \sum_{j=1}^{m} \alpha_j B^j} g(\delta_{t-1}),$$

where α_0 is a constant, B is the back-shift operator ($B(g(\delta_t)) = g(\delta_{t-1})$), all roots of the numerator and denominator polynomials have roots outside the unit circle (meaning that the absolute values of the roots are greater than 1) and finally $g(.)$, the weighted innovation, is defined as

$$g(\delta_t) = \theta \delta_t + \gamma\{|\delta_t| - E[|\delta_t|]\},$$

where θ and γ are real constants. This weighted innovation function serves to induce the asymmetric effects between positive and negative returns that is observed in real financial time series data.

Another model used to model asymmetric effects between positive and negative returns is the threshold GARCH or TGARCH model. A TGARCH(m, s) model is defined as

$$\epsilon_t = \sigma_t \delta_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{s} (\alpha_i + \gamma_i N_{t-i}) \alpha_{t-i}^2 + \sum_{j=1}^{m} \beta_j \sigma_{t-j}^2,$$

where N_t is an indicator function for $-\epsilon_t$ (i.e., $N_t = 1$ if $\epsilon_t < 0$ and 0 otherwise) and α_i, γ_i and β_j are nonnegative real parameters satisfying conditions similar to those of the GARCH models. This model uses zero as its threshold to separate the impacts of past shocks as can be seen in the definition of the indicator function N_t. However, in general, other thresholds can also be used.

The conditional heteroscedastic ARMA (CHARMA) model uses second order interaction terms of past shocks to describe the conditional volatility as

$$\epsilon_t = \boldsymbol{\epsilon}_{tm} \boldsymbol{\delta_t} + \eta_t,$$

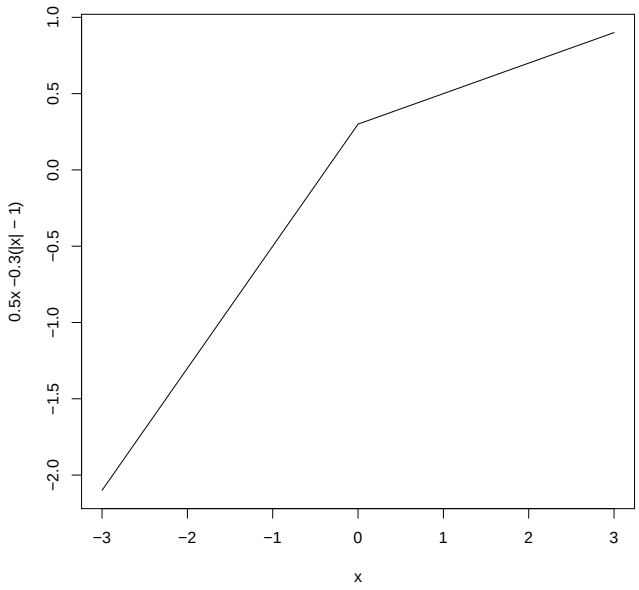

Figure 5.1: An example of the weighted innovation function of the EGARCH model.

where $\epsilon_{tm} = (\epsilon_{t-1}, \ldots, \epsilon_{t-m})$ is a vector of lagged values of ϵ_t available at time $t-1$, $\{\eta_t\}$ is a white noise process with variance σ_η^2 and $\{\boldsymbol{\delta_t}\} = \{(\delta_{t1}, \ldots, \delta_{tm})^T\}$ is a vector white noise process, independent of $\{\eta_t\}$, with covariance matrix $\boldsymbol{\Omega}$. The conditional variance can then be derived as

$$\sigma_t^2 = \sigma_\eta^2 + \boldsymbol{\epsilon}_{mt}\boldsymbol{\Omega}\boldsymbol{\epsilon}_{tm}^T.$$

The random coefficient (RCA) model, as the name suggests, considers time-varying parameters in the conditional mean equation. However, this property also implies conditional heteroscedasticity. Hence, this is a conditional heteroscedastic model also. An RCA (p) model is given by

$$x_t = \phi_0 + \sum_{i=1}^{p} (\phi_i + \delta_{ti})x_{t-i} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process with variance σ_ϵ^2 and $\{\boldsymbol{\delta_t}\} = \{(\delta_{t1}, \ldots, \delta_{tp})^T\}$ is a vector white noise process, independent of $\{\epsilon_t\}$, with covariance matrix $\boldsymbol{\Omega_\delta}$. While the conditional mean of this process is the same as that of an AR(p) process, the conditional variance is given by

$$\sigma_t^2 = \sigma_\epsilon^2 + (x_{t-1}, \ldots, x_{t-p})\boldsymbol{\Omega_\delta}(x_{t-1}, \ldots, x_{t-p})^T,$$

which is the same form as in the CHARMA model with the subtle difference that the past shocks are replaced by the past returns now.

All the models mentioned so far are deterministic in that given the past information, the conditional volatility of the present returns is perfectly determined. The stochastic volatility (SV) model, as the name suggests, is stochastic in nature in that it incorporates randomness in the conditional volatility equation. An SV(p) model is defined as

$$\epsilon_t = \sigma_t\delta_t, \quad (1 - \sum_{i=1}^{p} \alpha_i B^i) ln(\sigma_t^2) = \alpha_0 + \nu_t,$$

where $\{\delta_t\}$ is a white noise process with variance σ_δ^2, $\{\nu_t\}$ is a white noise process, independent of $\{\delta_t\}$, with variance σ_ν^2, B is the back-shift operator, α_0 is a real constant and all zeroes of the associated AR polynomial are greater than 1 in modulus.

In the beginning of this chapter, we saw how joint estimation of the conditional mean and variance equations' parameters is a part of the model building process

for a time series. It was mentioned then that such a joint estimation method is not straightforward. This is now clear since, as is evident from the ARCH and GARCH type models seen just now, the conditional variance equations involve past shocks and variances and hence it is not possible to arrive at an explicit form for the parameter estimates. A natural way to alleviate this problem is involving past returns in the conditional variance equation rather than past shocks and volatilities. The main idea of conditioning the present volatility on past information continues to remain intact. The parameters can also be interpreted more concretely as contributions from specific lagged returns. In the following section, we will look at such a model and analyze its properties.

5.4 A conditional heteroscedastic autoregressive (CHAR) model

In this section, we will introduce a special autoregressive model that incorporates conditional heteroscedasticity as well. We will call this the conditional heteroscedastic autoregressive (CHAR) model. A simple CHAR model of order (1, 1) is defined as follows.

$$x_t = (\alpha + \beta x_{t-1}) + (1 + \gamma x_{t-1})\epsilon_t,$$

where $\{\epsilon_t\}$ is a white noise process with variance σ_ϵ^2. This model can be regarded as an AR(1) model with time-varying conditional variance which is again an AR(1) process with unit mean. We consider the unit mean since any other constant can be absorbed in the variance, σ_ϵ^2, of the associated white noise process $\{\epsilon_t\}$. However, this is different from the GARCH process combined with an AR conditional mean equation. The difference lies in the conditional variance equation. While the GARCH conditional variance equation links the present conditional variance to the past conditional variances and shocks, the CHAR model links the present conditional variance to the past returns directly. This difference has consequences in the model properties and estimation that we will see later. This model is also a conditional heteroscedastic (1, 0, 1, 1) bilinear model.

Another aspect of this model is that positive and negative returns have different impacts on the conditional variance of the future returns. This, as described in the E-GARCH model, is a desired property of volatility models since it is observed in empirical financial data. The difference in impact, on the conditional variance, of the past positive and negative returns can be seen in the CHAR(1, 1) model when one

considers a negative γ. A negative x_{t-1} means the conditional variance of r_t given x_{t-1} is proportional to $(1+\gamma x_{t-1})^2$ which is greater than one. A positive x_{t-1} by the same token, lowers this proportion. By a similar argument, a positive γ reverses this effect. That is, when $\gamma > 0$, positive returns are followed by returns with higher conditional variance.

5.4.1 Analysis

Under stationarity, taking expectations, we have

$$\mu_x = E[x_t] = \alpha + \beta\mu_x \quad \rightarrow \quad \mu_x = \alpha/(1-\beta).$$

Thus, for stationarity, we require $\beta \neq 1$. For the conditional mean, we have

$$\mu_{t|t-1} = E[x_t|x_{t-1}, x_{t-2}, ...] = \alpha + \beta x_{t-1}.$$

Moving towards volatility, the stationary variance of this process can be calculated easily and is given by

$$\sigma_x^2 = \sigma_\epsilon^2 \frac{1+2\gamma\mu_x}{1-(\beta^2+\gamma^2\sigma_\epsilon^2)}.$$

Hence, for this process to be weakly stationary, we need that $\beta^2 + \gamma^2\sigma_\epsilon^2 < 1$. The conditional variance is given by

$$\sigma_{t|t-1}^2 = (1+\gamma x_{t-1})^2\sigma_\epsilon^2.$$

Under non-stationarity, the variance is given by

$$\sigma_t^2 = \sigma_{t-1}^2(\beta^2+\gamma^2\sigma_\epsilon^2) + (1+2\gamma\mu_x)\sigma_\epsilon^2.$$

We can see from this equation that in order for it to have a stationary solution, we must have that $\beta^2 + \gamma^2\sigma_\epsilon^2 < 1$. The above equation can be solved for explicitly by recursively substituting for previous terms. If we start from time index $t_0 = 0$, then the solution has an exponential form and is given by

$$\sigma_t^2 = \theta(\beta^2+\gamma^2\sigma_\epsilon^2)^t + \sigma_\epsilon^2 \frac{1+2\gamma\mu_x}{1-(\beta^2+\gamma^2\sigma_\epsilon^2)}. \tag{5.3}$$

where θ is some arbitrary constant. We can see that this solution either explodes or converges to the stationary solution given before depending on whether $\beta^2+\gamma^2\sigma_\epsilon^2 < 1$ or $\beta^2 + \gamma^2\sigma_\epsilon^2 > 1$ respectively. For $\beta^2 + \gamma^2\sigma_\epsilon^2 = 1$, this has no solution.

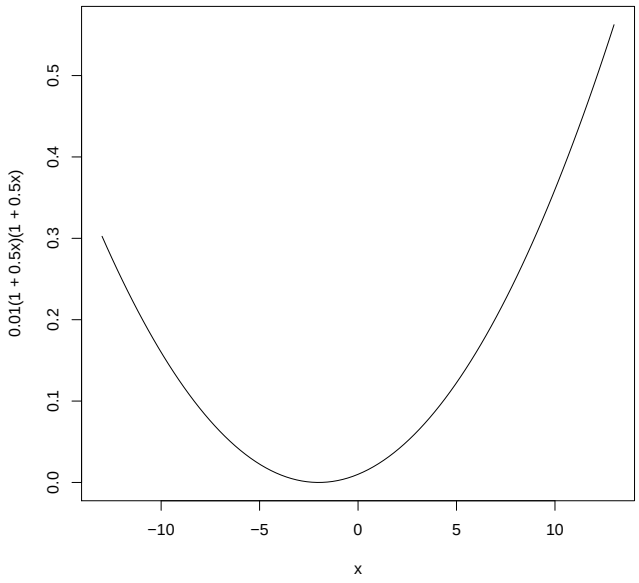

Figure 5.2: An example of the conditional variance of the CHAR(1, 1) model.

As seen, for the process to have zero mean, $\alpha = 0$. The parameter α thus plays a role only in the mean equation. Also, even under non-stationarity, the process converges to stationarity under the appropriate condition : $\beta^2 + \gamma^2\sigma_\epsilon^2 < 1$. Given that one is interested in stationary processes (given its tractable properties) and that this particular bilinear model converges to a stationary process under appropriate conditions, we will limit ourselves to the stationary case. Note, however, that even under stationarity, the conditional variance is not constant.

We now examine the autocovariance of this process. The lag-i autocovariance can be calculated easily as

$$\gamma_i = \beta\gamma_{i-1} \quad \text{with} \quad \gamma_0 = \sigma_x^2$$

This is the same as in an AR(1) process. Note that the stationarity condition implies $|\beta| < 1$ which ensures that the autocorrelation function diminishes as the lag increases. Figures 5.3, 5.4 and 5.5 show simulated stationary zero mean CHAR (1, 1) models with same parameters but of different sample sizes. One can see the conditional heteroscedasticity in the first figure while the later two figures show the stationarity of the mean and variance of this process.

Marginal Distribution

While the conditional distribution of r_t in a CHAR model is the same, F_ϵ, as that of the associated white noise process $\{\epsilon_t\}_{t\in T}$, its marginal distribution, like in the bilinear model, is not easy to compute. This can be demonstrated by looking at the zero mean CHAR(1, 1) model.

$$x_t = ax_{t-1} + (1 + bx_{t-1})\epsilon_t = (a + b\epsilon_t)x_{t-1} + \epsilon_t$$

Expanding x_{t-1} recursively, we get

$$x_t = \sum_{i=0}^{t-1} \epsilon_{t-i}\Pi_{j=0}^{i-1}(a + b\epsilon_{t-j}).$$

Like in the bilinear model, this equation shows that x_t is a sum of random variables that are product of i.i.d random variables. The products involved in the summation are O(t). Yet, if $|b| \ll |a|$, then we can ignore O($b^{(2+i)}$) terms ($i \geq 0$) in comparison to O(ab) terms and see that

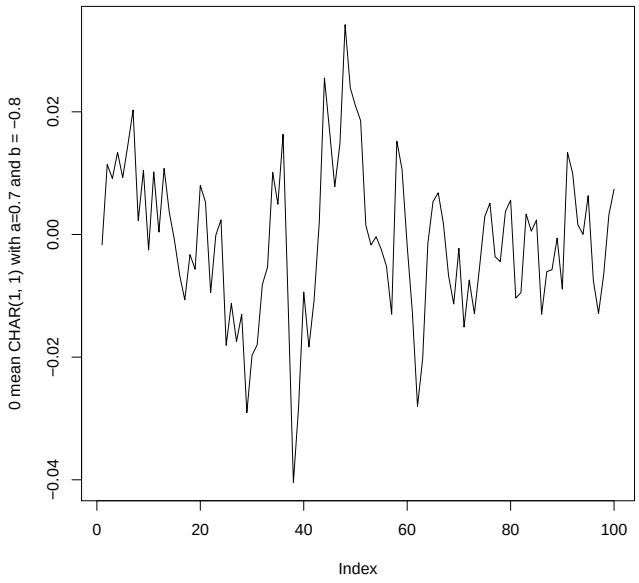

Figure 5.3: An example of a stationary zero mean CHAR(1, 1) data of size 100.

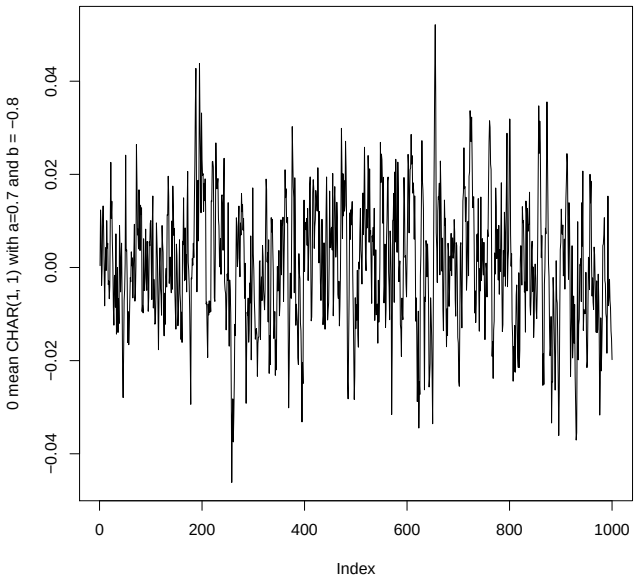

Figure 5.4: An example of a stationary zero mean CHAR(1, 1) data of size 1000.

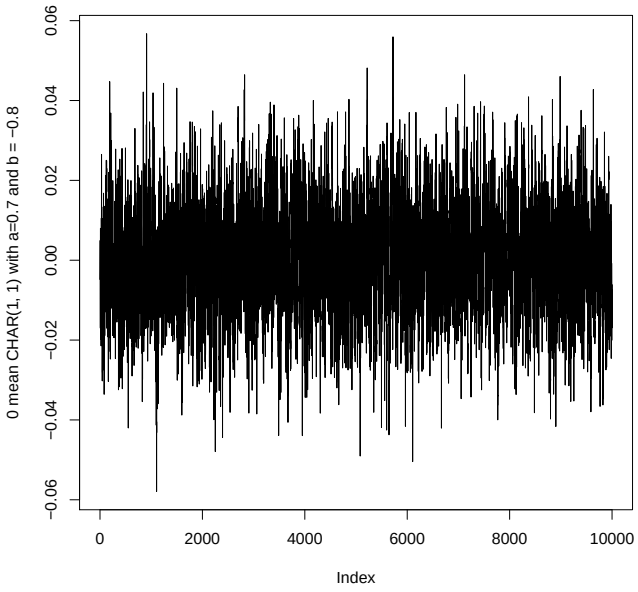

Figure 5.5: An example of a stationary zero mean CHAR(1, 1) data of size 10000.

$$x_t \sim \sum_{i=0}^{t-1} \epsilon_{t-i}(a^i + ab\sum_{j=0}^{i-1} \epsilon_{t-j}).$$

This form of x_t now involves only second order product terms and when $F_\epsilon(.) = \Phi(.)$, the Gaussian distribution, the distribution function for the product is given by

$$F_{\epsilon^2}(x) = \frac{K_0(\frac{|x|}{\sigma_\epsilon^2})}{\pi\sigma_\epsilon^2},$$

where $K_0(.)$ is a modified Bessel function of the second kind. From this, we can see that the marginal distribution function of the zero mean CHAR(1, 1) process is that of a sum of Gaussian and second order products of Gaussian.

However, from the expanded form of x_t we can compute the first two moments of the marginal distribution directly as as given below.

$$E[x_t] = \mu_t = 0$$

and

$$\text{Var}(x_t) = \sigma_t^2 = \sigma_\epsilon^2 \frac{1 - (a^2 + b^2\sigma_\epsilon^2)^t}{1 - (a^2 + b^2\sigma_\epsilon^2)}.$$

As seen above, under the stationarity condition, namely $a^2 + b^2\sigma_\epsilon^2 < 1$, we see that σ_t^2 converges to the stationary variance of

$$\sigma_x^2 = \frac{\sigma_\epsilon^2}{1 - (a^2 + b^2\sigma_\epsilon^2)}.$$

We also see that putting $\theta = -\frac{\sigma_\epsilon^2}{1-(a^2+b^2\sigma_\epsilon^2)}$ and noting that $\mu_x = 0$ in Equation (5.3) gives us the expression we just obtained for the marginal variance.

5.4.2 The general CHAR model

We now move to the general CHAR(p, q) model and analyze its properties. A CHAR(p, q) model is defined as

$$x_t = (\alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i}) + (1 + \sum_{j=1}^{q} \beta_j x_{t-j})\epsilon_t, \tag{5.4}$$

where $\{\epsilon_t\}$ is a white noise process with variance σ_ϵ^2. We can generalize the stationarity condition here as

$$\sum_{i=1}^{p} \alpha_i^2 + \sigma_\epsilon^2 \sum_{j=1}^{q} \beta_j^2 < 1.$$

The exact form of the stationary variance and autocovariances can be calculated using the Yule-Walker equations.

The stationary mean of this model is

$$\mu_x = \alpha_0/(1 - \sum_{i=1}^{p} \alpha_i).$$

Thus, for stationarity, we need $\sum_{i=1}^{p} \alpha_i \neq 1$. The conditional mean is given by

$$\mu_{t|t-1} = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i},$$

which is the same as in an AR(p) process.

An important aspect of this model is to estimate its order. We can use individual AR and ARCH model characterizations to get an idea of p and q. Since this is a nonlinear model in the parameters, an iteratively re-weighted least squares technique can be used to calculate the least squares estimator. Alternately, the Newton Raphson method could also be employed to compute the estimator. Using similar approaches, one can also robustly estimate the parameters. In a latter section, we will look at the least squares estimator and the robust S-estimator for a CHAR model.

5.5 The multivariate CHAR model

We saw in earlier chapters the need for multivariate models in time series analysis due to the presence strong of cross-correlations in multiple time series data. Hence, we can extend the CHAR model to the multivariate case also. The multivariate d-dimensional CHAR(p, q) model is defined as

$$\mathbf{x_t} = (\boldsymbol{\alpha_0} + \sum_{i=1}^{p} \boldsymbol{\alpha_i} \mathbf{x_{t-i}}) + \boldsymbol{\epsilon_t^*}((1, \ldots, 1)^T + \sum_{j=1}^{q} \boldsymbol{\beta_j} \mathbf{x_{t-j}}),$$

where $(\boldsymbol{\alpha_i} : i > 0)$ and $(\boldsymbol{\beta_j} : j > 0)$ are now matrix parameters, $\boldsymbol{\alpha_0}$ is a vector constant and

$$\boldsymbol{\epsilon_t^*} = \begin{pmatrix} \epsilon_{t1} & 0 & \dots & 0 \\ 0 & \epsilon_{t2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \epsilon_{td} \end{pmatrix}$$

where

$$\boldsymbol{\epsilon_t} = \begin{pmatrix} \epsilon_{t1} \\ \vdots \\ \epsilon_{td} \end{pmatrix}$$

is a vector white noise process with covariance matrix $\boldsymbol{\Sigma_\epsilon}$. Note that there are no cross-product terms involving the white noise components in this equation. This is because the concurrent and cross dependencies are taken into consideration by involving cross product terms between returns. Since the vector white noise has its own dependence through $\boldsymbol{\Sigma_\epsilon}$, we will not introduce further dependencies by involving cross product terms between the components of the white noise.

Like in the univariate case, we will start the analysis with the simple multivariate CHAR(1, 1) model defined below.

$$\begin{aligned} \begin{pmatrix} x_{t1} \\ x_{t2} \end{pmatrix} = \Big[\begin{pmatrix} \alpha_{01} \\ \alpha_{02} \end{pmatrix} + \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \end{pmatrix} \Big] + \\ \begin{pmatrix} \epsilon_{t1} & 0 \\ 0 & \epsilon_{t2} \end{pmatrix} \Big[\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \end{pmatrix} \Big]. \end{aligned} \tag{5.5}$$

The stationary mean of this model is given by

$$\boldsymbol{\mu_x} = \begin{pmatrix} \mu_{r1} \\ \mu_{r2} \end{pmatrix} = \begin{pmatrix} 1 - \alpha_{11} & -\alpha_{12} \\ -\alpha_{21} & 1 - \alpha_{22} \end{pmatrix}^{-1} \begin{pmatrix} \alpha_{01} \\ \alpha_{02} \end{pmatrix}.$$

Hence, for this process to have a stationary mean, we must have $(1-\alpha_{11})(1-\alpha_{22}) \neq \alpha_{12}\alpha_{21}$.

The stationarity condition for this process is highly restrictive as will be shown now. For simplicity, we will take a zero mean model; i.e., we will take $\boldsymbol{\alpha_0} = \mathbf{0}$. For this model, the stationary covariance matrix is given by

$$\boldsymbol{\Sigma_X} = \begin{pmatrix} \sigma_{r1}^2 & \rho_r \sigma_{r1} \sigma_{r2} \\ \rho_r \sigma_{r1} \sigma_{r2} & \sigma_{r2}^2 \end{pmatrix},$$

where

$$\begin{pmatrix} \sigma_{r1}^2 \\ \sigma_{r2}^2 \\ \rho_r \sigma_{r1} \sigma_{r2} \end{pmatrix} = [\mathbf{I} - (\boldsymbol{\alpha}^* + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^* \boldsymbol{\beta}_{\mathbf{1}}^*)]^{-1} \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \rho \sigma_1 \sigma_2 \end{pmatrix}$$

where

$$\boldsymbol{\alpha}^* = \begin{pmatrix} \alpha_{11}^2 & \alpha_{12}^2 & 2\alpha_{11}\alpha_{12} \\ \alpha_{21}^2 & \alpha_{22}^2 & 2\alpha_{21}\alpha_{22} \\ \alpha_{11}\alpha_{21} & \alpha_{12}\alpha_{22} & \alpha_{12}\alpha_{21} + \alpha_{11}\alpha_{22} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^* = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \rho \sigma_1 \sigma_2 \end{pmatrix}$$

and

$$\boldsymbol{\beta}_{\mathbf{1}}^* = \begin{pmatrix} \beta_{11}^2 & \beta_{12}^2 & 2\beta_{11}\beta_{12} \\ \beta_{21}^2 & \beta_{22}^2 & 2\beta_{21}\beta_{22} \\ \beta_{11}\beta_{21} & \beta_{12}\beta_{22} & \beta_{12}\beta_{21} + \alpha_{11}\alpha_{22} \end{pmatrix}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}.$$

As seen, for stationarity, we must have $\text{Det}[\mathbf{I} - (\boldsymbol{\alpha}^* + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^* \boldsymbol{\beta}_{\mathbf{1}}^*)] > 0$. This is a very restrictive condition on the parameters, that too, for the most simple 2-dimensional zero mean CHAR(1, 1) model. Hence, although the model is able to capture some interesting properties like conditional heteroscedasticity in a multivariate time series, it imposes very strict conditions on the parameters for stationarity which makes it difficult to deal with. Thus, we will proceed further with the univariate case.

5.6 A zero mean CHAR(1, 1) model identification

In this section we will estimate the parameters of a zero mean CHAR(1,1) model in the standard as well as robust fashions. We will use the Fast-S method to compute the S-estimator. In addition, we will compute the nonlinear least squares estimate and compare the two estimates in the sense of mean squared errors.

Given a time series $\{x_t : t = 1, .., T\}$, a zero mean CHAR(1, 1) model is given by

$$x_t = ax_{t-1} + (1 + bx_{t-1})\epsilon_t \tag{5.6}$$

where $\{\epsilon_t : t = 2, .., T\}$ is the associated white noise process with variance σ_ϵ^2 and kurtosis τ_ϵ^4.

5.6.1 The least squares estimator

For this model, the least squares estimator, $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b})$, of $\boldsymbol{\theta} = (a, b)$ minimizes the sum of squares of residuals given by

$$\sum_{t=2}^{T} r_t^2(\boldsymbol{\theta}), \tag{5.7}$$

where $r_t(\boldsymbol{\theta}) = \frac{x_t - ax_{t-1}}{1+bx_{t-1}}$ are the residuals. Differentiating w.r.t. $\boldsymbol{\theta}$, the least squares estimating equation then is

$$\sum_{t=2}^{T} r_t(\hat{\boldsymbol{\theta}})\mathbf{r}_\mathbf{t}'(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{5.8}$$

Like in the bilinear model nonlinear least squares estimation, the weighted version of the nonlinear least squares estimation for the CHAR(1, 1) model is given by

$$\sum_{t=2}^{T} r_t(\hat{\boldsymbol{\theta}})\mathbf{r}_\mathbf{t}'(\hat{\boldsymbol{\theta}})w_t = \mathbf{0}. \tag{5.9}$$

where $\{w_t : t = 2, \ldots, T\}$ are the weights.

Asymptotic properties

We will now derive the asymptotic covariance of the least squares estimator in the following theorem. However, like in the bilinear model, we will assume consistency of the estimator as supported by our simulations.

Theorem 5. *Consider a stationary zero mean CHAR(1, 1) process defined as in Equation* (5.6)*. Denote the associated white noise process by* $\{\epsilon_t\}$*. Denote the true parameter by*

$$\boldsymbol{\theta_0} = (a_0, b_0).$$

and the least squares estimator by

$$\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b}).$$

Define

$$\alpha_t(a, b) = \frac{x_t}{1 + bx_t}.$$

If

1. $E[\alpha_t(a_0, b_0)] = 0$
2. *The least squares estimator is consistent*
3. σ_ϵ^2 *being the variance and* τ_ϵ^4 *being the kurtosis of the associated white noise process* $\{\epsilon_t : t = 2, .., T\}$ *are well defined and finite,*
4. $\beta = E[\alpha_t^2(a_0, b_0)]$ *is well defined and finite and*
5. $E[\epsilon_t^3] = 0$*,*

then the least squares estimator, as given by Equation (5.8)*, is asymptotically normal with mean* $\boldsymbol{\theta_0}$ *and covariance*

$$\frac{1}{\beta}\begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \frac{\tau_\epsilon^4}{9\sigma_\epsilon^4} \end{pmatrix}$$

Proof. Define $r_{t0} = r_t(a_0, b_0)$ and $\alpha_{t0} = \alpha_t(a_0, b_0)$. The proof follows from the standard Taylor series expansion of the optimization equation given in (5.8). Like before, the assumed consistency of the estimator allows us to do this. From the definition of the residuals

$$r_t = r_t(a, b) = \frac{x_t - ax_{t-1}}{1 + bx_{t-1}},$$

and the fact that $r_{t0} = \epsilon_t$ it can immediately be seen that

$$\frac{\partial r_t}{\partial a}\Big|_{(a_0,b_0)} = -\alpha_{t-1,0}$$

and

$$\frac{\partial r_t}{\partial b}\Big|_{(a_0,b_0)} = -\alpha_{t-1,0}\epsilon_t$$

Let $\mathbf{f} = \sum_{t=2}^{T} r_{t0}\mathbf{r}'_{\mathbf{t}}(a_0, b_0)$. Then, by the above definitions of the partial derivatives, we can write

$$\mathbf{f} = \begin{pmatrix} \sum_{t=2}^{T} \epsilon_t \alpha_{t-1,0} \\ \sum_{t=2}^{T} \epsilon_t^2 \alpha_{t-1,0} \end{pmatrix}.$$

From the definition of $\alpha_t(a, b)$

$$\frac{\partial \alpha_t(a,b)}{\partial a} = 0$$

and

$$\frac{\partial \alpha_t(a,b)}{\partial b} = -\alpha_t^2(a,b).$$

Noting that $r_{t0} = \epsilon_t$ and using the above partial derivatives, the jacobian, $\mathbf{r}''_{\mathbf{t}}(a_0, b_0) = \mathbf{f}'' = \mathbf{J}$ is then given by

$$-\begin{pmatrix} \sum_{t=2}^{T} \alpha_{t-1,0}^2 & 2\sum_{t=2}^{T} \epsilon_t \alpha_{t-1,0}^2 \\ 2\sum_{t=2}^{T} \epsilon_t \alpha_{t-1,0}^2 & 3\sum_{t=2}^{T} \epsilon_t^2 \alpha_{t-1,0}^2 \end{pmatrix}.$$

It is clear from the definition of $\alpha_{t-1,0}$ that it is independent of ϵ_t. By the Taylor expansion of the optimization equation defining the least squares estimator and ignoring the second order terms due to consistency, we have

$$(\hat{a}, \hat{b})^T - (a_0, b_0)^T = \mathbf{J}^{-1}\mathbf{f}.$$

If $E[\alpha_{t0}] = 0$,

$$E[\mathbf{f}] = E\Big[\begin{pmatrix} \sum_{t=2}^{T} \epsilon_t \alpha_{t-1,0} \\ \sum_{t=2}^{T} \epsilon_t^2 \alpha_{t-1,0} \end{pmatrix}\Big] = \mathbf{0}.$$

Now coming to the asymptotic variance, noting that $E[\epsilon_t^3] = 0$, we have

$$\mathbf{\Sigma}_f = E[\mathbf{f}\mathbf{f}^T/(T-1)] = E[\alpha_{t0}^2]\begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \tau_\epsilon^4 \end{pmatrix} = \beta\begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \tau_\epsilon^4 \end{pmatrix}.$$

Hence, by the central limit theorem, we have

$$\mathbf{f}/(T-1) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma_f}).$$

On the other hand, from the definition of the jacobian,

$$\lim_{T\to\infty} \mathbf{J}/(T-1) = \mathbf{\Sigma_J} = E[\mathbf{J}/(T-1)] = E[\alpha_{t0}^2]\begin{pmatrix} 1 & 0 \\ 0 & 3\sigma_\epsilon^2 \end{pmatrix} = \beta\begin{pmatrix} 1 & 0 \\ 0 & 3\sigma_\epsilon^2 \end{pmatrix}. \quad (5.10)$$

Combining the limiting distribution of $\mathbf{f}/(T-1)$ and the limiting value of the jacobian $\mathbf{J}/(T-1)$ and noting that $(\mathbf{\Sigma_J^{-1}})^T = \mathbf{\Sigma_J^{-1}}$, we get that

$$\sqrt{T}[(\hat{a}, \hat{b})^T - (a_0, b_0)^T] \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma_J^{-1}\Sigma_f\Sigma_J^{-1}}).$$

From the definitions of $\mathbf{\Sigma_J}$ and $\mathbf{\Sigma_f}$, we arrive at the desired result (which completes the proof) which is

$$\sqrt{T}[(\hat{a}, \hat{b})^T - (a_0, b_0)^T] \xrightarrow{d} N\left(\mathbf{0}, \frac{1}{\beta}\begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \frac{\tau_\epsilon^4}{9\sigma_\epsilon^4} \end{pmatrix}\right)$$

□

Note that the condition $E[\alpha_t(a_0, b_0)] = 0$ may not be satisfied in many cases. However, when this condition is not met, a bias is introduced only in the estimate of b. In addition, the asymptotic covariance of $\hat{a}$ remains unchanged. The other conditions regarding the second and fourth moments of the white noise being finite and $\beta < \infty$ are also generally met by stationary time series. Thus, the nonlinear least squares estimator of a, the forecast parameter, is asymptotically Normal even if $E[\alpha_t(a_0, b_0)] \neq 0$.

5.6.2 The S-estimator

The S-estimator was defined in Chapter 3. Recalling this definition, we have that the S-estimator, $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b})$, minimizes a robust scale estimate s, which is defined as the solution to the following optimization problem :

Min s^2 subject to

$$1/(T-1)\sum_{t=2}^{T}\rho(r_t(\boldsymbol{\theta})/s) = \xi \tag{5.11}$$

where ξ is a constant, equal to $\mathrm{E}[\rho(X)]$ where X is a random variable with the same distribution as the white noise but with unit variance and $r_t(.)$ are residuals for a given parameter estimate.

Using Lagrange multipliers, we see that the S-estimator satisfies the same necessary condition as those that are satisfied by the classical robust M-estimator, namely,

$$1/(T-1)\sum_{t=2}^{T}\boldsymbol{\rho'}(r_t(\hat{\boldsymbol{\theta}})/s) = \mathbf{0} \tag{5.12}$$

We will use the Fast-S method to compute the S-estimate. The steps of this algorithm pertaining to the zero mean CHAR(1, 1) model follow. Given a starting parameter estimate $\boldsymbol{\theta_n}$ at step n, the improvement step (I-step, which is the core of the algorithm) is as follows.

1. Compute the residuals $\hat{r}(\boldsymbol{\theta_n}) = (\hat{r}_2(\boldsymbol{\theta_n}), \ldots, \hat{r}_T(\boldsymbol{\theta_n}))$.

2. Compute an approximate scale $\hat{s}$ of $\hat{r}(\boldsymbol{\theta_n})$ by applying to Equation (5.11), one step of any iterative algorithm starting from the MAD (median absolute deviation). Here, we use the Newton-Raphson method.

Call the median of the computed residuals $\hat{s}_n$. Then, the one step Newton-Raphson improvement is given by

$$\hat{s}_{n+1} = \hat{s}_n + \frac{\sum_{t=1}^{T}\rho(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_n) - \xi(T-1)}{\sum_{t=1}^{T}\psi(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_n)(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_n^2)}$$

3. Compute the weights

$$w_t = \frac{\psi(\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_{n+1})}{\hat{r}_t(\boldsymbol{\theta_n})/\hat{s}_{n+1}} \tag{5.13}$$

where $\psi = \rho'$.

4. The improved candidate $\boldsymbol{\theta_{n+1}}$ is obtained by a weighted nonlinear least squares with weights defined by (5.13).

As mentioned in the last chapter, in the case of nonlinear models also, the S-estimator in most cases retains a 50% breakdown point.

Asymptotic distribution

Like in the case of the least squares estimator, we establish the asymptotic normality of the S-estimator for the zero mean CHAR(1, 1) model under some regulatory conditions which includes the important assumption of consistency which is supported by simulations.

Theorem 6. *Consider a stationary zero mean CHAR(1, 1) process defined as in Equation* (5.6)*. Suppose the associated white noise process is given by* $\{\epsilon_t\}$ *with variance* σ_ϵ^2*. Denote the parameter as* $\theta = (a, b)$*. Denote the true parameter by*

$$\boldsymbol{\theta_0} = (a_0, b_0).$$

and the S-estimator by

$$\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b}).$$

Define

$$\alpha_t(a, b) = \frac{x_t}{1 + bx_t}.$$

Denote $\rho'(x) = \psi(x)$*. Suppose the following conditions hold :*

1. $\psi(x) = x\gamma(x)$ *where* $1/x$ *is not a factor of* $\gamma(x)$
2. $\gamma'(x) = x\delta(x)$ *where* $1/x$ *is not a factor of* $\delta(x)$
3. $E[\alpha_t(a_0, b_0)] = 0$
4. $E[\epsilon_t^{(2k-1)}\gamma^l(r(a_0, b_0))] = 0 \quad \forall \quad k, l \in \mathbb{N}$
5. $E[\epsilon_t^{(2k-1)}\delta^l(r(a_0, b_0))] = 0 \quad \forall \quad k, l \in \mathbb{N}$
6. $\psi(.)$ *is redescending and bounded.*
7. *The S-estimator is consistent.*
8. $E[\epsilon_t^3] = 0.$
9. $E[\epsilon_t^4] < \infty.$

10. $\beta = E[\alpha_t^2(a_0, b_0)] < \infty$

Define

1. $\psi_t(a, b) = \psi(r_t(a, b))$
2. $\gamma_t(a, b) = \gamma(r_t(a, b))$
3. $\delta_t(a, b) = \delta(r_t(a, b))$
4. $\alpha = E[\gamma_t(a_0, b_0)]$
5. $\sigma^2_{\epsilon\delta} = E[\epsilon_t^2 \delta_t(a_0, b_0)]$
6. $\sigma^2_{\epsilon\gamma} = E[\epsilon_t^2 \gamma_t(a_0, b_0)]$
7. $\sigma^2_{\epsilon\gamma\gamma} = E[\epsilon_t^2 \gamma_t^2(a_0, b_0)]$
8. $\tau^4_{\epsilon\delta} = E[\epsilon_t^4 \delta_t(a_0, b_0)]$
9. $\tau^4_{\epsilon\gamma} = E[\epsilon_t^4 \gamma_t^2(a_0, b_0)]$

These are well defined and finite due to the the variance and kurtosis of the white noise being bounded and the redescending nature of $\psi(.)$.

Then the S-estimator of this model, as given by Equation (5.11)*, is asymptotically normal with mean* $\boldsymbol{\theta_0}$ *and covariance*

$$\frac{1}{\beta} \begin{pmatrix} \frac{\sigma^2_{\epsilon\gamma\gamma}}{(\sigma^2_{\epsilon\delta}+\alpha)^2} & 0 \\ 0 & \frac{\tau^4_{\epsilon\gamma}}{(3\sigma^2_{\epsilon\gamma}+\tau^4_{\epsilon\delta})^2} \end{pmatrix}$$

Proof. Define

1. $r_{t0} = r_t(a_0, b_0)$
2. $\alpha_{t0} = \alpha_t(a_0, b_0)$
3. $\gamma_{t0} = \gamma_t(a_0, b_0)$
4. $\delta_{t0} = \delta_t(a_0, b_0)$.

The proof then follows from the standard Taylor series expansion of the optimization equation given in (5.12). We can do this because we have assumed consistency of the estimator. From the definition of the residuals

$$r_t = r_t(a, b) = \frac{x_t - ax_{t-1}}{1 + bx_{t-1}},$$

and the fact that $r_{t0} = \epsilon_t$ it can immediately be seen that

$$\frac{\partial r_t}{\partial a}\Big|_{(a_0,b_0)} = -\alpha_{t0}$$

and

$$\frac{\partial r_t}{\partial b}\Big|_{(a_0,b_0)} = -\alpha_t\epsilon_t$$

From the definition of $\alpha_t(a, b)$

$$\frac{\partial \alpha_t(a, b)}{\partial a} = 0$$

and

$$\frac{\partial \alpha_t(a, b)}{\partial b} = -\alpha_t^2(a, b).$$

From the definition of $\gamma_t(a, b)$, the conditions in the theorem and noting that $r_{t0} = \epsilon_t$

$$\frac{\partial \gamma_t(a, b)}{\partial a}\Big|_{(a_0,b_0)} = -\delta_{t0}\epsilon_t\alpha_{t-1,0}$$

and

$$\frac{\partial \gamma_t(a, b)}{\partial b}\Big|_{(a_0,b_0)} = -\delta_{t0}\epsilon_t^2\alpha_{t-1,0}.$$

Let $\mathbf{f} = \sum_{t=2}^{T} r_{t0}\mathbf{r}'_\mathbf{t}(a_0, b_0)$. Then, by the above definitions of the partial definitions, we can write

$$\mathbf{f} = \begin{pmatrix} \sum_{t=2}^{T} \epsilon_t\alpha_{t-1}\delta_{t0} \\ \sum_{t=2}^{T} \epsilon_t^2\alpha_{t-1}\delta_{t0} \end{pmatrix}.$$

The jacobian, $\mathbf{r}''_\mathbf{t}(a_0, b_0) = \mathbf{f}'' = \mathbf{J}$ is then given by

$$-\begin{pmatrix} \sum_{t=2}^{T} \alpha_{t-1,0}^2(\gamma_{t0}+\epsilon_t^2\delta_{t0}) & 2\sum_{t=2}^{T}[\epsilon_t\gamma_{t0}\alpha_{t-1,0}^2+\epsilon_t^3\delta_{t0}\alpha_{t-1,0}^2] \\ 2\sum_{t=2}^{T}[\epsilon_t\gamma_{t0}\alpha_{t-1,0}^2+\epsilon_t^3\delta_{t0}\alpha_{t-1,0}^2] & 3\sum_{t=2}^{T}\epsilon_t^2\alpha_{t-1,0}^2+\epsilon_t^4\alpha_{t-1,0}^2\delta_{t0} \end{pmatrix}.$$

It is clear from the definition of $\alpha_{t-1,0}$ that it is independent of ϵ_t. By the Taylor expansion of the optimization equation defining the S-estimator and ignoring the second order terms due to consistency, we have

$$(\hat{a},\hat{b})^T-(a_0,b_0)^T=\mathbf{J}^{-1}\mathbf{f}.$$

Under the conditions mentioned in the theorem,

$$E[\mathbf{f}]=E\Big[\begin{pmatrix} \sum_{t=2}^{T}\epsilon_t\alpha_{t-1,0}\gamma_{t0} \\ \sum_{t=2}^{T}\epsilon_t^2\alpha_{t-1,0}\gamma_{t0} \end{pmatrix}\Big]=\mathbf{0}.$$

Now coming to the asymptotic variance, noting the conditions of the theorem and the initial definitions given in this proof, we have

$$\mathbf{\Sigma}_f=E[\mathbf{f}\mathbf{f}^T/(T-1)]=\beta\begin{pmatrix} \sigma_{\epsilon\gamma\gamma}^2 & 0 \\ 0 & \tau_{\epsilon\gamma}^4 \end{pmatrix}.$$

Hence, by the central limit theorem, we have

$$\mathbf{f}/(T-1)\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma_f}).$$

On the other hand, from the definition of the jacobian,

$$\lim_{T\to\infty}\mathbf{J}/(T-1)=\boldsymbol{\Sigma_J}=E[\mathbf{J}/(T-1)]=\beta\begin{pmatrix} \sigma_{\epsilon\delta}^2+\alpha & 0 \\ 0 & 3\sigma_{\epsilon\gamma}^2+\tau_{\epsilon\delta}^4 \end{pmatrix}. \tag{5.14}$$

Combining the limiting distribution of $\mathbf{f}/(T-1)$ and the limiting value of the jacobian $\mathbf{J}/(T-1)$ and noting that $(\boldsymbol{\Sigma_J^T})^{-1}=\boldsymbol{\Sigma_J^{-1}}$, we get that

$$\sqrt{T}[(\hat{a},\hat{b})^T-(a_0,b_0)^T]\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma_J^{-1}\Sigma_f\Sigma_J^{-1}}).$$

From the definitions of $\boldsymbol{\Sigma_J}$ and $\boldsymbol{\Sigma_f}$, we arrive at the desired result (which completes the proof) which is

$$\sqrt{T}[(\hat{a},\hat{b})^T-(a_0,b_0)^T]\xrightarrow{d}N\left(\mathbf{0},\frac{1}{\beta}\begin{pmatrix} \frac{\sigma_{\epsilon\gamma\gamma}^2}{(\sigma_{\epsilon\delta}^2+\alpha)^2} & 0 \\ 0 & \frac{\tau_{\epsilon\gamma}^4}{(3\sigma_{\epsilon\gamma}^2+\tau_{\epsilon\delta}^4)^2} \end{pmatrix}\right)$$

□

As in the least squares estimator case, the condition $E[\alpha_t(a_0, b_0)] = 0$ may not be satisfied in many cases. However, when this condition is not met, a bias is introduced only in the estimate of b. In addition, the asymptotic covariance of $\hat{a}$ remains unchanged.

Further, the other conditions of the theorem concerning the form of the ρ and ψ functions are met by most commonly used functions such as the Tukey's biweight function and the Welsch function. Finally, the two conditions concerning the zero expectations of the products are also expected to be met by the aforementioned functions since these are even functions in the argument and hence, when the white noise has a symmetric distribution, so will these functions. Thus, products of odd powers of the white noise variable and even powers of the functions $\gamma(.)$ and $\delta(.)$ will be expected to have zero expectation.

The conditions regarding the variables introduced being well defined and finite are expected to be met by most white noise distributions and time series in practical applications.

5.6.3 Examples

Example 3. *In this example also, we will consider the NASDAQ log returns data as it shows both, a strong autocorrelation at lag 1, as well as clear conditional heteroscedasticity. For simplicity, we take a sub-sample of size 1428 starting from the first time index. From this, we take a sample of the first 1000 data points to fit a CHAR(1, 1) model and use the sample of the remainder 428 data points to gauge the quality of the model and fit by computing the median absolute forecast error (MAFE). Like before, we then artificially contaminate the data with an additive outlier by replacing the value at index 500 by the value 10 first, and 100 finally, in two separate contamination scenarios. We refit this data and recompute the MAFE. In all the cases, we compute the conventional least squares (LS) estimate using the multivariate Newton-Raphson method and the robust S-estimate using the Fast-S method. The results, consisting of the estimates and the MAFE, are given in the following table. The sensitivity of the LS method is seen as the contaminant value increases. Also note the highly negative value of $\hat{b}$ which shows that a negative return is followed by a larger conditional variance compared to the conditional variance that follows a positive return.*

Outlier Value	*LS* $(\hat{a}, \hat{b}$ *(standard error), MAFE)*	*S-estimator* $(\hat{a}, \hat{b}$ *(standard error), MAFE)*
N/A	*0.35, -18.10,* 3.30×10^{-6} *(0.0003), (0.009)*	*0.36, -18.47,* 3.29×10^{-6} *(0.0003), (0.011)*
10	*0.44, 108.08,* 3.68×10^{-6} *(0.0031), (0.12)*	*0.33, -26.36,* 3.37×10^{-6} *(0.0003), (0.009)*
100	*-770749, -414, 1287370 (0.0029), (0.29)*	*0.33, -26.31,* 3.37×10^{-6} *(0.0003), (0.098)*

5.6.4 Simulations

Scenario 1

In this section, we will present some performance metrics after running some simulations. For the simulations, for simplicity, we generated 1000 samples of size 100 each of zero mean CHAR(1, 1) data with normal white noise with variance 0.0001, and parameters $a = 0.7, b = -0.8$, that was then contaminated with additive outliers to varying degrees and the Fast-S method was applied to obtain the S-estimator of the parameter a. This parameter was also estimated using the least squares method which is equivalent to the conditional maximum likelihood estimator. Finally, bias and root mean squared errors (RMSE) were calculated for the two estimators (S-estimator and the Least Squares). The following table gives details of the results. The contamination was done by replacing a single data point in the series by 1, 10 and 100 respectively, giving three different contamination scenarios.

Cont. %	Outlier Value	LS of a : Bias and RMSE	S-estimator of a : Bias and RMSE
0	N/A	-0.01, 0.06	-0.01, 0.08
1	1	-0.68, 0.71	-0.02, 0.10
1	10	-0.69, 0.74	-0.02, 0.09
1	100	-0.70, 0.70	-0.01, 0.08

As seen in the table, the least squares estimate shows a marked increase in bias and variance in the case of a single contaminant while the S-estimator remains resistant.

Scenario 2

In this next simulation exercise, we generated 1000 samples of size 1000 each of zero mean CHAR(1, 1) data with normal white noise with variance 0.0001, and parameters $a = 0.7, b = -0.8$, that was then contaminated with additive outliers. Bias

and root mean squared errors (RMSE) were calculated for the S-estimator and the least squares estimator of the forecast parameter a. The contamination was done by replacing a single data point in the series by 100. The following table gives details of the results.

Cont. %	Outlier Value	LS of a : Bias and RMSE	S-estimator of a : Bias and RMSE
0.5	100	-0.69, 0.70	-9.42×10^{-4}, 0.02

5.7 Summary

In this chapter, we saw the concept of conditional heteroscedasticity and its role in time series modeling. Heteroscedasticity is important in time series analysis because many financial and econometric time series data exhibit this characteristic. We took a brief look at the various heteroscedastic models proposed in literature such as the ARCH, GARCH and its variations.

We then defined a special conditional heteroscedastic model that incorporated an autoregressive conditional mean equation. This model , the conditional heteroscedastic autoregressive (CHAR) model, is different from ARCH and GARCH type models in that it associates the conditional standard deviation with past returns and not past shocks and volatilities. This model could be thought of as a conditional heteroscedastic bilinear model with no other second order terms. We saw the stationarity conditions associated with this model and looked at other properties like the autocorrelation functions.

We saw the difficulty in extending this model to the multivariate setup. The primary obstacle was the rather strict conditions on parameters for the model to be stationary. We demonstrated this with the simple multivariate CHAR(1, 1) model.

We then applied the Newton Raphson method to compute the least squares estimate of the parameters of the zero mean CHAR(1, 1) process and could immediately see its lack of resistance to outliers. As a result, we computed the S-estimator, using the Fast-S method, which is a robust estimator, and compared it with the two least squares estimators. The simulation results show the advantage of using robust estimators under even slight contaminations.

CHAPTER 6

Summary

The idea of robust estimation of time series models is central to the aim of this thesis. While robustness and time series modeling have been vastly researched individually in the past, application of robust methods to estimate time series models is still quite open. In addition, with opening up of markets and economies all over the world, the global economy is all the more highly interconnected. In time series analysis, this necessitates building multivariate models.

The first aim of this thesis was to study some prominent linear and nonlinear models in the time series literature. The second aim was to study the multivariate vector autoregressive (VAR) model to understand cross and concurrent correlations. The third aim was to study some simple bilinear models in detail. The fourth aim was to analyze conditional heteroscedasticity in time series models since it is an important aspect in time series modeling. After examining the state of the art in this area, a special bilinear model was studied that incorporated a linear conditional mean equation. We noticed here that even simple conditional heteroscedastic bilinear models can have very strict conditions on the parameters for stationarity. The multivariate representation of this model was also analyzed briefly.

Robustness is essential in modeling financial and econometric data yet underrated. Finally, aspects concerning the robustness of the above mentioned models was studied. In particular, outlier propagation was analyzed and a robust method, the S-

estimator, was used to estimate the parameters of the models and compared to the estimates from the very popular and still widely used, least squares method. The simulation study showed that even under small levels of contamination, the least squares method breaks down easily whereas the S-estimator remains largely unaffected.

Given that the area of robust methods in time series modeling is still in its nascent stages, many further interesting applications remain to be seen in this context. For example, application of the MM-estimators and the more recent multivariate generalized S-estimators by Roelant et al. (2009) [55] are interesting prospects in time series analysis that are yet to be explored in detail. In addition, we saw briefly, the propagation of outliers in time series modeling and how it can break down even robust estimators. Here again, the recent work by Muler et al. (2009) on the bounded innovations propagation (BIP) based models gives us a rich platform from where one can start tackling the problem of outlier propagation. Further, given the slow tilt of balance in favor of nonlinear models as opposed to linear ones, study of extensions of the CHAR type models seems an interesting prospect from the point of view of explaining the ever growing complex behavior of econometric and financial time series, thanks to the increasing globalization of the economy.

Bibliography

[1] S.I. Akamanam, M.B. Rao, and K. Subramanyam. On the ergodicity of some bilinear time series models. *Journal of Time Series Analysis*, 7(5):157–163, 1986.

[2] F. Alqallaf, V.J. Yohai, S. Van Aelst, and R.H. Zamar. Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331, 2009.

[3] T. Amemiya. *Advanced econometrics*. Basil Blackwell, 1986.

[4] C.M. Bach and G. MacKinnon. A maximum likelihood procedure for regression with autocorrelated errors. *Econometrica*, 46(1):51–58, 1978.

[5] M.G. Ben, E.J. Martinez, and V.J. Yohai. Robust estimation in vector autoregressive moving-average models. *Journal of Time Series Analysis*, 20(4):381–399, 1998.

[6] M.G. Ben, A.J. Villar, and V.J. Yohai. Robust estimation in vector autoregressive models based on a robust scale. *Estadística*, 53(160,161):397–434, 2001.

[7] T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[8] A.D. Brunner and G.D. Hess. Potential problems in estimating bilinear time-series models. *Journal of Economic Dynamics and Control 19*, pages 663–681, 1995.

[9] O.H. Bustos and V.J. Yohai. Robust estimates for arma models. *Journal of the American Statsitical Association*, 81(393):155–168, 1986.

[10] B.P. Carlin, N.G. Polson, and D.S. Stoffer. A monte carlo approach to non-normal and nonlinear state space modeling. *Journal of the American Statistical Association*, 87:493–500, 1992.

[11] K.S. Chan and H. Tong. On estimating threshold in autoregressive models. *Journal of Time Series Analysis*, 7:179–190, 1986.

[12] R. Chen and R.S. Tsay. On the ergodicity of tar(1) processes. *Annals of Applied Probability*, 1(4):613–634, 1991.

[13] R. Chen and R.S. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.

[14] R. Chen and R.S. Tsay. Nonlinear additive arx models. *Journal of the American Statistical Association*, 88:955–967, 1993.

[15] R. Christensen. *Linear models for multivariate time series and spatial data*. Springer-Verlag, 1990.

[16] D. Cochrane and G.H. Orcutt. Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44(245):32–61, 1949.

[17] P.J. Cooper. Asymptotic covariance matrix of procedures for linear regression in the presence of first-order autoregressive disturbances. *Econometrica*, 40(2):305–310, 1972.

[18] C. Croux and K. Joossens. Robust estimation of the vector autoregressive model by a least trimmed squares procedure. *Compstat2008*, 2008.

[19] L. Davies. The asymptotics of s-estimators in the linear regression model. *The Annals of Statistics*, 18(4):1651–1675, 1990.

[20] R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[21] M.M. Gabr. Robust estimation of bilinear time series models. *Journal of Communication in Statistics - Theory and Methods*, 27(1):41–53, 1998.

[22] F. Giordano. The variance of cls estimators for a simple bilinear model. *Quaderni di Statistica*, 2, 2000.

[23] U. Granader. On the estimation of regression coefficients in the case of an autocorrelated disturbance. *The Annals of Mathematical Statistics*, 25(2):252–272, 1954.

[24] C.W.J. Granger and A. Andersen. *An introduction to bilinear time series models.* Vandenhoeck and Ruprecht, Gttingen, 1978.

[25] J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[26] J.D. Hamilton. Analysis of time series subject to regime changes. *Journal of Econometrics*, 45:39–70, 1990.

[27] J.D. Hamilton. *Time series analysis.* Princeton University Press, 1994.

[28] D. Kazakos and K. Makki. Robust estimation of multivariate time series. 12^{th} *biennial IEEE conference on electromagnetic field computation*, pages 464–464, 2006.

[29] S. Kim, M.D. McKenzie, and R.W. Faff. Macroeconomic news announcements and the role of expectations : evidence for us bond, stock and foriegn exchange markets. *Journal of Multinational Financial Management*, 14:217–232, 2004.

[30] W.K. Kim, L. Billard, and I.V. Basawa. Estimation of the first order bilinear time series model. *Journal of Time Series Analysis*, 11:215–227, 1990.

[31] J. Kmenta and R.F. Gilbert. Estimation of seemingly unrelated regression with autoregressive disturbances. *Journal of the American Statistical Association*, 65(329):186–197, 1970.

[32] P.A.W. Lewis and J.G. Stevens. Nonlinear modeling of time series using multivariate adaptive regression splines (mars). *Journal of the American Statistical Association*, 86(416):864–877, 1991.

[33] W.K. Li and Y.V. Hui. Robust multiple time series modeling. *Biometrika*, 76(2):309–315, 1989.

[34] J. Liu. Estimation for some bilinear time series. *Journal of Time Series Analysis. Stochastic Models*, 6(4):649–665, 1990.

[35] J. Liu. On stationarity and asymptotic inference of bilinear time series models. *Statistica Sinica 2*, pages 479–494, 1992.

[36] J. Liu and P.J. Brockwell. On the general bilinear time series model. *Journal of Applied Probabbility*, 25(3):553–564, 1988.

[37] H.P. Lopuhaä. On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, 17(4):1662–1683, 1989.

[38] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer, 2007.

[39] R.A. Maronna and V.J. Yohai. Asymptotic behavior of general m-estimates for regression and scale with random carriers. *Journal Probability Theory and Related Fields*, 58(1):7–20, 1981.

[40] C.J. Masreliez. Approximate non-gaussian filtering with linear state and observation relations. *IEEE Trans. Automat. Control*, 20:107–110, 1975.

[41] R.E. McCulloch and R.S. Tsay. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, 88(423):968–978, 1993.

[42] R.E. McCulloch and R.S. Tsay. Statistical analysis of economic time series via markov switching models. *Journal of Time Series Analysis*, 15:523–539, 1994.

[43] R.R. Mohler and Z. Tang. On estimation with bilinear time series. *IEEE ICASSP 86 Tokyo*, 1986.

[44] F.S. Møller, J. Von Frese, and R. Bro. Robust methods for multivariate data analysis. *Journal of Chemometrics*, 19:549–563, 2005.

[45] N. Muler, D. Peña, and V.J. Yohai. Robust estimation for arma models. *The Annals of Statistics*, 37(2):816–840, 2009.

[46] D.B. Nelson. Conditional heteroscedasticity in asset returns : A new approach. *Econometrica*, 59:347–370, 1991.

[47] J.D. Petruccelli and S.W. Woolford. A threshold ar(1) model. *Journal of Applied Probability*, 21:270–286, 1991.

[48] T.D. Pham and L.T. Tran. On the first-order bilinear time series model. *Journal of Applied Probability*, 18(3):617–627, 1981.

[49] D.A. Pierce. Least squares estimation in the regression model with atoregressive-moving average errors. *Biometrika*, 58(2):299–312, 1971.

[50] M.B. Priestley. State-dependent models : A general approach to non-linear time series analysis. *Journal of Time Series Analysis*, 1(1):41–71, 1980.

[51] B. Rao and T.S. Rao. On the existence of some bilinear time series models. *Journal of Time Series Analysis*, 11:95–110, 1990.

[52] T.S. Rao. On the theory of bilinear time series models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(2):244–255, 1981.

[53] T.S. Rao and M.M. Gabr. *Lecture notes in statistics : An introduction to bispectral analysis and bilinear time series models*. Springer-Verlag, 1980.

[54] B.D. Ripley. Statistical aspects of neural networks. *Networks and Chaos Statistical and Probabilistic Aspects*, pages 40–123, 1993.

[55] E. Roelant, S. Van Alest, and C. Croux. Mulivariate generalized s-estimators. *Journal of Multivariate Analysis 100*, 100(5):876–887, 2009.

[56] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[57] P.J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, B:283–297, 1985.

[58] P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[59] P.J. Rousseeuw, K. Van Driessen, S. Van Aelst, and J. Agulló. Robust multivariate regression. *Technometrics*, 46(3), 2004.

[60] P.J. Rousseeuw and V.J. Yohai. Robust regression by means of s-estimators. *Lecture Notes in Statistics, No. 26, Springer-Verlag, Berlin,*, pages 256–272, 1984.

[61] D. Ruppert. Computing s-estimators for regression and multivariate location/dispersion. *Journal of the Computational and Graphical Statistics*, 1(3):253–270, 1992.

[62] D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, 1994.

[63] S. Sakata and H. White. S-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics 103*, pages 5–72, 2001.

[64] A.J. Stromberg and D. Ruppert. Breakdown in nonlinear regression. *Journal of the American Statistical Association*, 87(420):991–997, 1992.

[65] R.D. Synder. Robust time series analysis. *European Journal of Operations Research 9*, pages 168–172, 1982.

[66] T. Teräsvirta. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(208-218), 1994.

[67] H. Tong. On a threshold model. *Pattern recognition and signal processing. NATO ASI Series E: Applied Sc. (29). Sijthoff & Noordhoff, Netherlands*, pages 575–586, 1978.

[68] H. Tong. *Threshold models in non-linear time series analysis.* New York : Springer-Verlag, 1983.

[69] R.S Tsay. *Analysis of financial time series.* Wiley-Interscience, 2005.

[70] A.E. Usoro and C.O. Omekara. Bilinear autoregressive vector models and their application to estimation of revenue series. *Asian Journal of Mathematics and Statistics*, 1(1):50–56, 2008.

[71] S. Van Aelst, J. Agulló, and C. Croux. The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99(3):311–338, 2008.

[72] S. Van Aelst and G. Willems. Multivariate regression s-estimators for robust estimation and inference. *Statistica Sinica 15*, pages 981–1001, 2005.

[73] T.M. Welbourne and A.O. Andrews. Predicting the performance of initial public offerings : Should human resouce management be in the equation ? *The Academy of Management Journal*, 39(4):891–919, 1996.

[74] G.T. Wilson. The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(1):76–85, 1973.

[75] V.J. Yohai and P. Rousseeuw. *Robust regression by means of S-estimators, Lecture Notes in Statistics No. 26.* Springer-Verlag (Berlin/New York), 1984.

[76] V.J. Yohai and M. Salibian-Barrera. A fast algorithm for s-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2):414–427, 2006.

Ravi Ramakrishnan

Avenue d'Epenex 17
1024 Ecublens, Switzerland
Email : ravi.ramakrishnan@mail.com

Summary
Quantitative investment manager with expertise in forecasting and risk modeling specializing in robust multivariate data analysis and non-linear modeling; Over four years of professional experience in software development and management

Professional Experience

2010 - Present	**Banque Cantonale Vaudoise, Lausanne, Switzerland** *Quantitative Investment Manager*
2005 – 2010	**Swiss Federal Institute of Technology, Lausanne, Switzerland** *Research & Teaching Assistant (Dept. Of Applied Statistics – STAP/IMA/SB)*
2007 – 2008	**RouteRank SA, Lausanne, Switzerland** *Technical Consultant*
2004 - 2005	**StorePerform Technologies, Bangalore, India** *Senior Software Engineer*
2002 - 2004	**Tavant Technologies, Bangalore, India** *Software Engineer*
2001	**Swiss Federal Institute of Technology, Lausanne, Switzerland** *Intern* - Worked on ARIADNE, an on-going EU project
1999	**Tata Consultancy Services, Chennai, India** *Intern* – Worked on database porting from UNIX/ORACLE to OS390/DB2

Education

2005 – 2010	**Swiss Federal Institute of Technology, Lausanne, Switzerland** *PhD – Statistics*
2001 – 2002	**Swiss Federal Institute of Technology, Lausanne, Switzerland** *Pre-Doctoral School in Computer Science*
1996 – 2001	**Indian Institute of Technology, Kanpur, India** *Master of Science in Mathematics and Scientific Computing*

Selected Publications

2009	**S-Estimation of a VAR Model** *Conference talk, ICORS 2009, Parma, Italy*
2009	**An efficient heuristic algorithm for the Bottleneck Traveling Salesman Problem** *OPSearch, Vol. 46, No. 3 (2009), pp. 275-288, India*

Technical Skills

Programming Languages	Java 1.6, SQL, Matlab, Python
Operating Systems	Linux, Windows
Databases	PostgreSQL 8.2
Architectures	OOAD, J2EE, Web 2.0
Server Platforms	Apache Tomcat

Personal Information

Date of Birth	13th March 1978
Language Skills	English (Fluent), French (Basic)