Multivariate Time Series Modelling

Suhayl Muhammed Vayej

December, 2012

University of KwaZulu-Natal

Multivariate Time Series Modelling

by

Suhayl Muhammed Vayej
Student number : 206502665

Supervisor: Dr W.H. Moolman
BCom(Stellenbosch) BCom(Hons)(Stellenbosch)
MCom(Natal) DCom(UDW) Diploma in Datametrics (Unisa)

Thesis submitted to the University of KwaZulu-Natal in fulfillment of the requirements for the Masters degree in Statistics in the School of Mathematics, Statistics and Computer Science University of KwaZulu-Natal, 2012.

Declaration

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal under the supervision of Dr W.H. Moolman.

I, Suhayl Muhammed Vayej, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

December 2012

_____ **_____**

Mr Suhayl Muhammed Vayej **Date**

Acknowledgements

I would like to express my thanks and gratitude to the following people for their help towards the completion of this thesis:

- My supervisor, Dr W.H. Moolman for his guidance, advice and for going out of his way to assist me during his retirement.
- Professor D. North, head of the Department of Statistics, for her excellent leadership and scholarship.
- My parents, Muhammed and Shamima for inspiring me, motivating me and always believing in me.
- My sister, Nabila for all her support and encouragement.
- My Dadi, Hawa Bibi Vayej and Nani, Ruby Saloojee for their kindness, blessings and support throughout this project.

Abstract

This research is based on a detailed description of model building for multivariate time series models. Under the assumption of stationarity, identification, estimation of the parameters and diagnostic checking for the Vector Autoregressive (p) (VAR(p)), Vector Moving Average (q) (VMA(q)) and Vector Autoregressive Moving Average (VARMA(p,q)) models are described in detail. With reference to the non-stationary case, the concept of cointegration is explained. Procedures for testing for cointegration, determining the cointegrating rank and estimation of the cointegrated model in the VAR(p) and VARMA(p,q) cases are discussed.

The utility of multivariate time series models in the field of economics is discussed and its use is demonstrated by analysing quarterly South African inflation and wage data from April 1996 to December 2008. A review of the literature shows that multivariate time series analysis allows the researcher to: (i) understand phenomenon which occur regularly over a period of time (ii) determine interdependencies between series (iii) establish causal relationships between series and (iv) forecast future variables in a time series based on current and past values of that variable. South African wage and inflation data was analysed using SAS version 9.2. Stationary VAR and VARMA models were run. The model with the best fit was the VAR model as the forecasts were reliable, and the small values of the Portmanteau statistic indicated that the model had a good fit. The VARMA models by contrast, had large values of the Portmanteau statistic as well as unreliable forecasts and thus were found not to fit the data well. There is therefore good evidence to suggest that wage increases occur independently of inflation, and while inflation can be predicted from its past values, it is dependent on wages.

Summary

This thesis will focus on modelling, inferences and the practical applications of multivariate time series. The aim of this study, is to describe, compare and discuss the practical applications of the following three multivariate time series models in the stationary and non stationary case: (i) The Vector Autoregressive model of order p (VAR(p)) (ii) The Vector Moving Average model of order q (VMA(q)) and (iii) The Vector Autoregressive Moving Average model of order p, q (VARMA(p, q)).

Under the assumption of stationarity, the first four chapters will explore multivariate time series models. In Chapter 1 the basic principles underpinning multivariate time series analysis are described. Chapter 2 addresses the simplest model, the Vector Autoregressive model of order p (VAR(p)) in which an observation at time t is regressed on lagged values of itself. The autocovariance properties as well as forecasting are discussed, followed by the model building stage, which involves identifying the lag order p and estimating the parameters through the use of two methods viz, the method of least squares estimation and the method of maximum likelihood estimation. The concept of diagnostic checking once the model has been estimated is also explained. This is done in order to determine whether or not the model is adequate or not.

The Vector Moving Average model of order q (VMA(q)) model is introduced in Chapter 3. In this model vectors of observations at time t are regressed with lagged values of their error terms. As in the case of the Vector Autoregressive model, the autocovariance and forecasting properties are discussed in addition to the identification of the lag order q. Estimation of the parameters and diagnostic checks are once explained for this model.

In the third model, the Vector Autoregressive Moving Average model of order p, q (VARMA(p, q)) a vector of observations at time t is regressed on both lagged values of themselves as well as lagged values of their error terms. The properties of the model as well as the model building stage are discussed in Chapter 4.

Non-stationarity of multivariate time series is discussed in Chapter 5. The concept of cointegration is defined and procedures for testing for cointegration are explained. Determining the cointegrating rank and the estimation of the cointegrated VAR and VARMA models

Chapter 6 deals with an important technique used for inference in multivariate time series models, Granger-causality tests. The interpretation of these tests when the series are non-stationary is also discussed.

In Chapter 7, the practical applications of multivariate time series are described. In order to illustrate the practical applications of the technique and to emphasise the utility of the different models, a brief review of multivariate time series analysis in the literature was undertaken. Applications in the fields of finance, economics, physical and environmental sciences, social sciences, engineering and medicine are highlighted.

This is followed in Chapter 8 by the author's empirical application. In order to illustrate model selection and goodness of fit, the author analysed quarterly South African wage and inflation data from April 1996 to December 2008 using the program SAS version 9.2. The adequacy of the VAR(p) and VARMA (p, q) models was compared and their forecasting performance was evaluated.

In this application, three models were run, a VAR model and two VARMA models. The VAR model was estimated using the least squares method. Both VARMA models were estimated using maximum likelihood methods and solved using two optimisation techniques. The first model employed the Quasi-Newton method and the second, the Newton-Raphson procedure. The model with the best fit was the VAR model as the forecasts were reliable (there was not much difference between the observed and the predicted values), while the small values of the Portmanteau statistic indicated that there was little serial correlation in the residuals. The VARMA models in contrast, had large values of the Portmanteau statistic as well as unreliable forecasts and were thus were found not to fit the data well.

In conclusion, multivariate time series analysis is a dynamic statistical procedure, which is used extensively to analyse the interrelationships between variables over a period of time and is supported by a large body of literature demonstrating its utility globally in the fields of economics, natural and health sciences.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

A time series is an ordered sequence of values observed through time. In other words a time series refers to the repeated measurements of data items over a period of time. These data items must be well defined and must be measurable at equally spaced time intervals. Time dependent financial and economic data, can be easily constituted into a time series. There is a large body of literature dating from the seminal paper by Sims (1980) firmly establishing the utility of time series analysis to analyse macroeconomic data such as exchange rates, interest rates, growth, inflation etc. Time series analysis has also been used in the disciplines of criminology, meteorology, chemistry, ecology, geology and medicine.

A single time series is referred to as a univariate series. Analysis of such a series provides useful information about the systems which generated the data over a specified period. This statistical method enables the practitioner to understand the underlying structure of the time series by breaking it down into its components. For example, trends in the data may be uncovered using this technique. A very popular application of the univariate method is for macroeconomic forecasting. Forecasting is the use of past values of a variable to predict future values of that variable. It provides a likely or expected future value for the outcome under investigation. It is of tremendous value because it reduces uncertainty and risk associated with the future. This type of information is particularly useful for investors and financial institutions. Scientific forecasting enhances knowledge of the future and the foresight gained allows for improved planning.

However, economic and financial markets globally are dynamic and integrated systems. Financial indicators are commonly dependent on each other. Movements in one domain, for example, income can spread quickly and easily to other domains such as inflation. Furthermore, these relationships are not unidirectional and may be reversed depending on the context within which they occur. Univariate time series analysis does not capture the interactions and co-movements between variables as it is confined to the analysis of a single variable.

Consequently, another method, multivariate time series analysis was developed to analyse two or more time series that are observed simultaneously.

Multiple time series analysis allows for

(i) An understanding of the relationship the variables share with each other.
(ii) The establishment of causal relationships between series.
(iii) The determination of the interdependencies between series.

Econometric models and methods used to investigate the relationship between economic variables such as inflation and income, forward and spot exchange rates, prices and interest rates (Chen & Lee, 1990), monetary and fiscal policy (Ansari, 1996), sales and stock prices (Chien, Lee & Tsai, 2006), etc. belong to vector or multivariate time series analysis in the statistical literature. The availability of computerised software packages and the frequent use of the time series analysis in the published literature have led to a rapid expansion in the modelling of multivariate time series (Athanasopoulos & Vahid, 2008).

In this study we have chosen to describe the following multivariate time series models: Vector Autoregressive model of order p (VAR(p)), the Vector Moving Average model of order q (VMA(q)) and the Vector Autoregressive Moving Average model of order p, q (VARMA(p, q)) for the stationary and nonstationary case.

A key aspect of multivariate time series analysis is the choice of model used to represent the series. Differences in model specifications and parameter estimates can result in very different findings. It is important to select the appropriate model to avoid obtaining spurious results. (Fackler & Krieger, 1986).

The VAR(p)) model, because it is simple, flexible and easy to use is especially popular for forecasting economic data (Escanciano, Lobtato & Zhu, 2010; Kascha, 2010). On the other hand, the VARMA(p, q) model despite its well described theoretical advantages, has rarely been considered as an alternative to the VAR (Athanasopoulos & Vahid, 2008; Dias & Kapetanios, 2011). This is due to the difficulty related to its implementation. Researchers are still plagued by the challenges of identifying and estimating unique VARMA models more than four decades after they were introduced (Lütkepohl & Poskitt, 1996; Raghavan, Athanasopoulos & Silvapulle, 2009; Poskitt, 2011).

In view of the paucity of literature related to the use the VARMA model as compared to the VAR model, this thesis describes the different methodologies proposed in the literature to deal with the complexities in establishing uniquely identified VARMA models. There is also a gap in the literature comparing the forecasting performance of these two models for a specific data set.

The relationship between two economic variables, inflation and wages is currently the subject of much debate locally. South Africa has a highly unionised work force and there have been increasing calls by the Congress of South African Trade Unions (COSATU) for wage increases (Fin24, 2010) to counter the effects of rising inflation. This demand for higher wages has been

resisted by government leading to growing employee dissatisfaction, widespread destruction of property, violence and even death (mail and guardian).

The post apartheid African National Congress led government has endeavoured through its "new growth path " (NGP) to keep inflation at low levels in an effort to contain poverty by 2020 (Nattrass, 2011). There are two main types of inflation; (i) Demand pull inflation where prices are pulled upwards by the demand for goods and (ii) Cost push inflation where the cost of producing goods or services pushes up prices.

The relationship between wages and inflation is well anchored in the econometric literature (Hess & Schweitzer, 2000). There is a perception based on Keynesian economics that higher wages lead to an increase in prices which in turn leads to increasingly higher wages (Ghali, 1999; Todani, 2006). This is known as the wage-price spiral. On the other hand authors such as Mehra (1993) and Jonsson and Palmqvist (2004) report that it is inflation which is actually responsible for wage increases. I have thus used a multivariate time series approach to investigate whether there is indeed a relationship between wages and inflation in South Africa from 1996-2008.

Against this background, the purpose of this study is to explain model selection for multivariate time series models emphasising methods to simplify model building procedures for the VARMA model and to apply the technique to a wage-inflation data set.

1.2 Objectives and Significance of the Study

The main objectives of this study are to:

1. Explain model selection for multivariate time series models by describing and comparing model building procedures for (i) The Vector Autoregressive model of order p (VAR(p)) (ii) The Vector Moving Average model of order q (VMA(q)) and (iii) The Vector Autoregressive Moving Average model of order p, q (VARMA(p, q)) in the stationary and non-stationary case.

2. Summarise recent methodological advances for simplifying the identification and estimation procedures for the VARMA model in the literature.

3. Illustrate the use of multivariate modelling techniques by applying it to South African wage and inflation data.

4. Compare the forecasting performance of the VAR and VARMA models for the above data set.

The significance of this study is that it contributes to the body of evidence for the use of the VARMA model and its comparison with the VAR for econometric forecasting in the international literature. The results of this study will also assist in the local demand for economic data to inform the difficult and tough fiscal and monetary policy decisions which lie ahead for this country.

1.3 Organisation of the Chapters

Chapter one lists the objectives and sets the background against which this thesis is based. Chapters two addresses the properties and model building process of the simplest multivariate time series model, the stationary Vector Autoregressive model of order p (VAR(p)) in which a variable is regressed with past values of itself and other variables. Chapter three discusses the properties and model building process of the stationary Vector Moving Average model of order q (VMA(q)) in which a vector of observations is regressed with past values of their error terms. Chapter four discusses the properties and model building stages of the stationary Vector Autoregressive Moving Average model of order p, q (VARMA(p, q)) in which a vector of observations are regressed with lagged values of themselves and their error terms. Chapter five discusses the properties of non-stationary models and introduces the Vector Error Correction Model (VECM). Chapter six addresses the topic of Granger-causality. Following on the model building process, chapter seven reviews the literature on model selection and utilisation of multivariate time series models. Chapter eight demonstrates the use of the technique by analysing simultaneous inflation and wage series using the VARMAX procedure in SAS (version 9.2) and includes the results and discussion. Chapter nine serves to conclude.

CHAPTER 2

The Vector Autoregressive(p) (VAR(p)) Model

2.1 Introduction

This chapter focuses on the simplest multivariate time series model, the finite vector autoregressive model of order p (VAR(p)) in which an observation at time t is regressed on lagged values of itself and all other variables in the system at p time periods where two variables share a common lag (Fackler & Krieger, 1986). This model is a generalised version of the univariate autoregressive model of order p and is commonly used to determine the interrelationships amongst the different variables in a system. This model is simplistic and flexible and is especially popular in the fields of finance and economics where it is used to determine the relationships between various economic factors and for the forecasting of economic data. It has gained particular attention over the past 30 years (Escanciano, Lobato & Zhu, 2010).

In this section, some basic principles which are relevant to multivariate time series analysis will be defined before the VAR model is described in detail.

A K dimensional multivariate time series $\boldsymbol{y_t}$, is denoted as $\boldsymbol{y_t} = (y_{1,t}, \dots, y_{K,t})$ $t = \dots, -2, -1, 0, 1, 2, \dots$ where each individual component, $y_{i,t}\ i = 1, \dots, K$ is a univariate time series.

Suppose $\boldsymbol{y}_t$ has a constant mean. The autocovariance between $\boldsymbol{y_t}$ and $\boldsymbol{y_{t+h}}$ is defined as

$$\boldsymbol{\Gamma}(h) = \text{Cov}\,(\boldsymbol{y}_t, \boldsymbol{y}_{t+h}) = E(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t+h} - \boldsymbol{\mu})'$$

$$= E\begin{pmatrix} y_{1,t} - \mu_1 \\ y_{2,t} - \mu_2 \\ \vdots \\ y_{K,t} - \mu_K \end{pmatrix} (y_{1,t+h} - \mu_1, y_{2,t+h} - \mu_2\ , \dots, y_{K,t+h} - \mu_K\ )$$

$$= \begin{pmatrix} \Gamma_{11}(h) & \Gamma_{12}(h) & \cdots & \Gamma_{1K}(h) \\ \Gamma_{21}(h) & \Gamma_{22}(h) & & \Gamma_{2K}(h) \\ \vdots & & \ddots & \vdots \\ \Gamma_{K1}(h) & \Gamma_{K2}(h) & \cdots & \Gamma_{KK}(h) \end{pmatrix}.$$

If for the process $\boldsymbol{y}_t$, $E(\boldsymbol{y}_t) = \boldsymbol{\mu}$ for all t and Cov $(\boldsymbol{y}_t, \boldsymbol{y}_{t+h})$ = Cov $(\boldsymbol{y}_{t-h}, \boldsymbol{y}_t)$, then $\boldsymbol{y}_t$ will be known as a stationary process. The covariance of a stationary process is not dependent on the time t but is instead dependent on the time interval h.

The process $\boldsymbol{y}_t$ is strongly stationary if the probability distributions of the vectors $\boldsymbol{y}_t = (y_{1,t1}, \dots, y_{K,tn})$ and $\boldsymbol{y}_t = (y_{1,t1+h}, \dots, y_{K,tn+h})$ are identical for times $t1, \dots, tn$ and for all time lags h (Box, Jenkins & Reinsel, 2008).

A more generalised definition of stationarity known as covariance stationarity/ weak stationarity occurs when a process $\boldsymbol{y}_t$ has finite first and second moments which satisfy the conditions that the mean, $E(\boldsymbol{y}_t) = \boldsymbol{\mu}$, does not depend on time t and if the covariance $E(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t+h} - \boldsymbol{\mu})'$ depends only on the lag h (Tsay, 2005).

A vector white noise process $\boldsymbol{u}_t = (u_{1,t}, \dots, u_{K,t})$ is a sequence of independent random vectors such that the mean $E(\boldsymbol{u}_t)$ is 0 and autocovariance matrix $E(\boldsymbol{u}_t \boldsymbol{u}_{t+h}{}')$ is $\boldsymbol{\Sigma}_u$ for $h = 0$ and 0 for $h \neq 0$.

The covariance between two individual univariate time series is known as the cross covariance and measures how strong the linear dependence is between them. This is expressed in mathematical terms as

$$\gamma_{ij}(h) = E\big(y_{i,t} - \mu_i\big)\big(y_{j,t+h} - \mu_j\big)' \\ = E(y_{i,t-h} - \mu_i)(y_{j,t} - \mu_j)' \,.$$

$h = 0, \pm 1, \pm 2, \dots \qquad i = 1, \dots, K \qquad j = 1, \dots, K$

The $K \times K$ cross correlation matrix for a vector process is defined as
$\boldsymbol{\rho}(h) = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{\Gamma}(h) \boldsymbol{D}^{-\frac{1}{2}}$.

where $\boldsymbol{D} = \text{diag}\,[\gamma_{11}(0), \gamma_{22}(0), \dots, \gamma_{KK}(0)]$ is a diagonal matrix whose ith diagonal element is the variance of the ith process (Wei, 2006) .

i.e. $\boldsymbol{D} = \begin{pmatrix} \gamma_{11}(0) & 0 & \cdots & 0 \\ 0 & \gamma_{22}(0) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{KK}(0) \end{pmatrix}$,

The cross correlation coefficient between two individual univariate time series $y_{i,t}$ and $y_{j,t-h}$ is $r_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}$ which measures the linear dependence of $y_{i,t}$ on $y_{j,t-h}$ $(h > 0)$. If $i = j$, then $r_{ij}(h)$ will be reduced to the autocorrelation function of $y_{i,t}$ (Brockwell & Davis, 1996).

The following are additional properties of cross correlation functions, see e.g. Tsay (2005),

- The element $r_{ij}(0)$ measures the coexisting relationship between $y_{i,t}$ and $y_{j,t}$. If $r_{ij}(0) \neq 0$, then $y_{i,t}$ and $y_{j,t}$ are said to be 'concurrently' correlated.
- $r_{ij}(h)$, $h > 0$ measures the linear dependence on $y_{i,t}$ on the past value $y_{j,t-h}$. If $r_{ij}(h) = r_{ji}(h) = 0$ for all values of $h > 0$, then $y_{i,t}$ and $y_{j,t}$ do not have a linear relationship between each other.
- If for all $h > 0$, but for some $l > 0$, $r_{ji}(l) \neq 0$, then there is an unidirectional relationship between $y_{i,t}$ and $y_{j,t}$ and $y_{i,t}$ does not depend on any of the past values of $y_{j,t}$. $y_{j,t}$ however is dependent on some of the past values of $y_{i,t}$.
- If $r_{ij}(h) \neq 0$ for some $h > 0$ and $r_{ij}(l) \neq 0$ for some $l > 0$, then there is a feedback relationship between $y_{i,t}$ and $y_{j,t}$.

2.2 Model Dynamics

The general univariate autoregressive model of order p is of the form

$$y_{1,t} = c + \varphi_1 y_{1,t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{1,t-p} + u_{1,t} \,. \tag{2.1}$$

The term c is a constant, $y_{1,t-1}, \dots, y_{1,t-p}$ are random variables and $u_{1,t}$ is an error term. The $\varphi_1, \dots, \varphi_p$ coefficients of the univariate AR(p) model measure how dependent an observation $y_{1,t}$ is on its past p values. The error term is assumed to be uncorrelated at different time periods.

The vector autoregressive model (VAR) of order p is a multivariate extension of the model (2.1) in which each individual variable $y_{i,t}$ $i = 1, \dots, K$ is regressed on a constant and p of its own lags as well as p lags of each of the other variables in the system. This can be expressed in mathematical form as

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t \,. \tag{2.2}$$

The random vector $\boldsymbol{y}_t = [y_{1,t}, \dots, y_{K,t}]$ is of dimension $(K \times 1)$, $\boldsymbol{\Phi}_i = \begin{bmatrix} \varphi_{11,i} & \cdots & \varphi_{1K,i} \\ \vdots & \ddots & \vdots \\ \varphi_{K1,i} & \cdots & \varphi_{KK,i} \end{bmatrix}$ $i = 1, \dots, p$ is the ith order parameter matrix of dimension $(K \times K)$, $\boldsymbol{c}_t = (c_1, \dots, c_K)$ is a $(K \times 1)$ intercept vector which denotes the constant terms. $\boldsymbol{u}_t = [u_{1t}, \dots, u_{Kt}]$ is a $(K \times 1)$ vector of residual error terms which in general is assumed to follow a multivariate normal process.

The error vector $\boldsymbol{u}_t$ is also assumed to be white noise, i.e. the components $u_{1t}, \dots, u_{Kt}$ are random vectors such that the mean $E(\boldsymbol{u}_t) = 0$ and the covariance matrix $E(\boldsymbol{u}_t \boldsymbol{u}_t')$ is

$$E(\boldsymbol{u}_t\boldsymbol{u}_t{}') = \boldsymbol{\Sigma}_u = \begin{pmatrix} E({u_{1t}}^2) & E(u_{1t}u_{2t}) & \cdots & E(u_{1t}u_{Kt}) \\ E(u_{1t}u_{2t}) & E({u_{2t}}^2) & & E(u_{2t}u_{Kt}) \\ \vdots & & \ddots & \vdots \\ E(u_{1t}u_{Kt}) & E(u_{2t}u_{Kt}) & \cdots & E({u_{Kt}}^2) \end{pmatrix}.$$

$E(\boldsymbol{u}_t\boldsymbol{u}_t{}')$ is of dimension $(K \times K)$ and is assumed to be a positive definite matrix. In addition $E(\boldsymbol{u}_t\boldsymbol{u}_{t+h}{}') = 0$ for $h \neq 0$ (Box et al., 2008). The terms $u_{1t}, \dots, u_{Kt}$ are often allowed to be correlated contemporaneously, i.e. u_{1t} could be correlated with $u_{2t}, \dots, u_{Kt}$ but not with the past values of either $u_{1t}, \dots, u_{Kt}$ (Chatfield, 2004). This means that if a single effect of $\boldsymbol{u}_t$ is examined while the other components are kept constant, it could lead to results that are contrary to the historical information summarized in $\boldsymbol{\Sigma}_u$ (Agénor & Hoffmaister, 1997).

If the mean $\boldsymbol{\mu}$ is known, then the VAR(p) model in (2.2) can be written in the deviations of mean representation as

$$\boldsymbol{y}_t - \boldsymbol{\mu} = \boldsymbol{\Phi}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\Phi}_2(\boldsymbol{y}_{t-2} - \boldsymbol{\mu}) + \cdots + \boldsymbol{\Phi}_p(\boldsymbol{y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{u}_t\,. \tag{2.3}$$

Before discussing any underlying theory with regards to the VAR(p) model, the VAR(p) model in its simplest form will be considered, i.e. the zero mean VAR(1) model,

$$\boldsymbol{y}_t - \boldsymbol{\Phi}_1 y_{t-1} = \boldsymbol{u}_t \text{ or } (\boldsymbol{I} - \boldsymbol{\Phi}_1 L)\,\boldsymbol{y}_t = \boldsymbol{u}_t\,. \tag{2.4}$$

L is known as the lag/backshift operator which shifts back the vector $\boldsymbol{y}_t$ by one period, i.e. $L\boldsymbol{y}_t = \boldsymbol{y}_{t-1}$. In general, $L^i\boldsymbol{y}_t = \boldsymbol{y}_{t-i}$ if $\boldsymbol{y}_t$ is shifted back i periods.

Consider the zero intercept, bivariate VAR(1) model below,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} - \begin{bmatrix} \varphi_{11,1} & \varphi_{12,1} \\ \varphi_{21,1} & \varphi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} = \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}. \tag{2.5}$$

An equivalent representation of (2.5) which results from the multiplication of the matrices $\begin{bmatrix} \varphi_{11,1} & \varphi_{12,1} \\ \varphi_{21,1} & \varphi_{22,1} \end{bmatrix}$ and $\begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}$ is

$$\begin{aligned} y_{1,t} &= \varphi_{11,1}\, y_{1,t-1} + \varphi_{12,1}\, y_{2,t-1} + u_{1,t} \\ y_{2,t} &= \varphi_{21,1}\, y_{1,t-1} + \varphi_{22,1}\, y_{2,t-1} + u_{2,t}\ . \end{aligned} \tag{2.6}$$

From the use of above equations (2.5) and (2.6), it can easily be seen that each variable $y_{i,t}$ $(i = 1,2)$ does not only involve lagged values of itself but also lagged values of the other variables. The coefficient $\varphi_{12,1}$ shows the linear dependence of $y_{1,t}$ on $y_{2,t-1}$ in the presence of y_{t-1}. If $\varphi_{12,1} = 0$ but $\varphi_{21,1} \neq 0$, it is said that there exists a unidirectional relationship between

$y_{1,t}$ and $y_{2,t}$ (and vice-versa) and if both the coefficients $\varphi_{12,1}$ and $\varphi_{21,1}$ are not equal to 0, then there is a feedback relationship between the series (Tsay, 2005).

The VAR(1) process is always invertible. For the process to be classified as stationary, the zero's of the determinantal polynomial $|\boldsymbol{I} - \boldsymbol{\Phi}_1|$ must lie outside the unit circle. The stationary, no intercept VAR(1) model can be written in the following representation which is known as the infinite moving average form as

$$\boldsymbol{y}_t = (\boldsymbol{I} - \boldsymbol{\Phi}_1 L)^{-1}\boldsymbol{u}_t = \sum_{i=0}^{\infty} \boldsymbol{\psi}_i L^i = \boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \cdots \tag{2.7}$$

The values of $\boldsymbol{\psi}_i$ are obtained from the equation

$$\boldsymbol{\psi}_i = \frac{1}{\boldsymbol{I} - \boldsymbol{\Phi}_1 L}$$

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 L)(\boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \ldots) = \boldsymbol{I} \tag{2.8}$$

Now by comparing each of the lag coefficients of (2.8),

$$\begin{aligned} &\boldsymbol{\psi}_0 = \boldsymbol{I} \\ L &: \boldsymbol{\psi}_1 = \boldsymbol{\Phi}_1 \\ L^2 &: \boldsymbol{\psi}_2 - \boldsymbol{\Phi}_1\boldsymbol{\psi}_1 = 0 \qquad \boldsymbol{\psi}_2 = \boldsymbol{\Phi}_1\boldsymbol{\psi}_1 \\ L^3 &: \boldsymbol{\psi}_3 - \boldsymbol{\Phi}_1\boldsymbol{\psi}_2 = 0 \qquad \boldsymbol{\psi}_3 = \boldsymbol{\Phi}_1\boldsymbol{\psi}_2 \end{aligned}$$

In general for L^i: $\boldsymbol{\psi}_i = \boldsymbol{\Phi}_1\boldsymbol{\psi}_{i-1}$.

If there is an intercept vector $\boldsymbol{c}_t$ in the model, then the VAR(1) model can be written in moving average form by the use of successive recursions as

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{c}_t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \boldsymbol{u}_t \\ &= \boldsymbol{c}_t + \boldsymbol{\Phi}_1(\boldsymbol{c}_t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-2} + \boldsymbol{u}_{t-1}) + \boldsymbol{u}_t \\ &= (\boldsymbol{I} + \boldsymbol{\Phi}_1)\boldsymbol{c}_t + \boldsymbol{u}_t + \boldsymbol{\Phi}_1\boldsymbol{u}_{t-1} + \boldsymbol{\Phi}_1{}^2\boldsymbol{y}_{t-2} \\ &= (\boldsymbol{I} + \boldsymbol{\Phi}_1)\boldsymbol{c}_t + \boldsymbol{\Phi}_1{}^2(\boldsymbol{c}_t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-3} + \boldsymbol{u}_{t-2}) + \boldsymbol{u}_t + \boldsymbol{\Phi}_1\boldsymbol{u}_{t-1} \\ &= \boldsymbol{c}_t\left(\boldsymbol{I} + \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_1{}^2\right) + \boldsymbol{u}_t + \boldsymbol{\Phi}_1\boldsymbol{u}_{t-1} + \boldsymbol{\Phi}_1{}^2\boldsymbol{u}_{t-2} + \boldsymbol{\Phi}_1{}^3\boldsymbol{y}_{t-3}\,. \end{aligned} \tag{2.9}$$

Continuing this procedure up to h,

$$\boldsymbol{y}_t = \sum_{i=0}^{h} \boldsymbol{\Phi}_1{}^i\boldsymbol{u}_{t-i} + \boldsymbol{\Phi}_1{}^{h+1}\boldsymbol{y}_{t-h-1} + \boldsymbol{c}_t(\boldsymbol{I} + \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_1{}^2 + \cdots + \boldsymbol{\Phi}_1{}^h)\,.$$

Reinsel (1997) noted that if the absolute value of the eigenvalues of $\boldsymbol{\Phi}(L)$ is less than one, then the value of $\boldsymbol{\Phi}_1{}^{t+h}$ will converge to 0 as $h \to \infty$. Thus

$$\boldsymbol{y}_t = (\boldsymbol{I} + \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_1{}^2 + \cdots + \boldsymbol{\Phi}_1{}^h)\boldsymbol{c}_t + \boldsymbol{u}_t + \boldsymbol{\Phi}_1\boldsymbol{u}_{t-1} + \boldsymbol{\Phi}_1{}^2\boldsymbol{u}_{t-2} + \ldots \tag{2.10}$$

Lütkepohl (2005) noted that the term $(\boldsymbol{I} + \boldsymbol{\Phi}_1 + {\boldsymbol{\Phi}_1}^2 + \cdots + {\boldsymbol{\Phi}_1}^h)\boldsymbol{c}_t \rightarrow (\boldsymbol{I} - \boldsymbol{\Phi}_1)^{-1}\boldsymbol{c}_t$ as $h \rightarrow \infty$

Thus (2.10) can be written as $\boldsymbol{y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} {\boldsymbol{\Phi}_1}^i \boldsymbol{u}_{t-i}$ where $\boldsymbol{\mu} = (I - \boldsymbol{\Phi}_1)^{-1}\boldsymbol{c}_t$ is the mean of the process. This equation shows that $\text{Cov}(\boldsymbol{u}_t, \boldsymbol{y}_{t-h}) = 0$ for all $h > 0$ since $\boldsymbol{u}_t$ is serially uncorrelated. $\boldsymbol{u}_t$ is thus often referred to as the shock or innovation vector at time t (Tsay, 2005)

It is important to note that the moving average representation does not necessarily have to be in infinite order. The VAR(1) process, (2.4) can also be written in the determinant/adjoint form as

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 L)^{-1} = \frac{(adjoint(\boldsymbol{I} - \boldsymbol{\Phi}_1 L))}{|\boldsymbol{I} - \boldsymbol{\Phi}_1 L|}.$$

Wei (2006) noted that if the matrix $\boldsymbol{\Phi}_1$ is nilpotent (i.e. if there exist integers j, k with $j > k$ such that ${\boldsymbol{\Phi}_1}^k \neq 0$ and ${\boldsymbol{\Phi}_1}^j = 0$) or if the determinantal polynomial $|\boldsymbol{I} - \boldsymbol{\Phi}_1 L|$ is independent of L, then the K dimensional VAR(1) model can be expressed as a finite K dimensional Vector Moving Average model of order q, (VMA (q) model) with $q \leq K - 1$. This is because the determinant of $\boldsymbol{I} - \boldsymbol{\Phi}_1 L$ is a constant and the elements of the adjoint matrix are polynomials in L of order less than or equal to one.

The methodology that has been used for explaining the VAR(1) model can be similarly extended to the VAR(p) model

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t$$
$$\text{or } \boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{u}_t . \qquad (2.11)$$

The VAR(p) process is always invertible and is stationary if the roots of the determinant of $\boldsymbol{I}_K - \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} - \cdots - \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p}$ lie outside the unit circle i.e. if they are greater than one. The stationary VAR(p) process can be written in moving average form by defining the operator

$\boldsymbol{\psi}(L) = \sum_{j=0}^{\infty} \boldsymbol{\psi}_j L^j$ such that $\boldsymbol{\psi}(L)\boldsymbol{\Phi}(L) = \boldsymbol{I}$. The operator $\boldsymbol{\psi}(L)$ is the inverse of $\boldsymbol{\Phi}(L)$.

If (2.11) is pre-multiplied by $\boldsymbol{\psi}(L)$, then

$$\boldsymbol{y}_t = \boldsymbol{\psi}(L)\boldsymbol{c}_t + \boldsymbol{\psi}(L)\boldsymbol{u}_t$$
$$= (\sum_{i=0}^{\infty} \boldsymbol{\psi}_i)\boldsymbol{c}_t + \sum_{i=0}^{\infty} \boldsymbol{\psi}_i \boldsymbol{u}_{t-i}$$
$$= \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\psi}_i \boldsymbol{u}_{t-i} .$$

The mean of $\boldsymbol{y}_t$, $\boldsymbol{\mu}$ is obtained from

$$\boldsymbol{\mu} = \boldsymbol{\Phi}(1)^{-1}\boldsymbol{c}_t = (\boldsymbol{I} - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2 - \cdots - \boldsymbol{\Phi}_p)^{-1}\boldsymbol{c}_t \, .$$

The $\boldsymbol{\psi}_i$ coefficients are obtained from the relation

$$\begin{aligned} &[\boldsymbol{\Phi}(L)]^{-1} = \textstyle\sum_{i=0}^{\infty} \boldsymbol{\psi}_i L^i \\ &(\boldsymbol{I} - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2 - \cdots - \boldsymbol{\Phi}_p L^p)(\,\boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \cdots) = \boldsymbol{I} \\ &\boldsymbol{\psi}_0 + (\boldsymbol{\psi}_1 - \boldsymbol{\psi}_0\boldsymbol{\Phi}_1)L + (\,\boldsymbol{\psi}_2 - \boldsymbol{\psi}_0\boldsymbol{\Phi}_2 - \boldsymbol{\Phi}_1\boldsymbol{\psi}_1)\,L^2 + \cdots = \boldsymbol{I}. \end{aligned} \tag{2.12}$$

Equating each of the lag coefficients

$$\begin{aligned} &\boldsymbol{\psi}_0 = \boldsymbol{I} \\ L\text{:}\quad &\boldsymbol{\psi}_1 = \boldsymbol{\Phi}_1 \\ L^2\text{:}\quad &\boldsymbol{\psi}_2 - \boldsymbol{\Phi}_2 - \boldsymbol{\Phi}_1\boldsymbol{\psi}_1 = 0 \qquad\qquad \boldsymbol{\psi}_2 = \boldsymbol{\Phi}_2 + \boldsymbol{\Phi}_1\boldsymbol{\psi}_1 \\ L^3\text{:}\quad &\boldsymbol{\psi}_3 - \boldsymbol{\Phi}_3 - \boldsymbol{\Phi}_1\boldsymbol{\psi}_2 - \boldsymbol{\Phi}_2\boldsymbol{\psi}_1 = 0 \qquad\qquad \boldsymbol{\psi}_3 = \boldsymbol{\Phi}_3 + \boldsymbol{\Phi}_1\boldsymbol{\psi}_2 + \boldsymbol{\Phi}_2\boldsymbol{\psi}_1 \end{aligned}$$

In general L^i: $\boldsymbol{\psi}_i = \sum_{j=1}^{i} \boldsymbol{\psi}_{i-j}\ \boldsymbol{\Phi}_j \quad i = 1,2, \ldots$

The VAR(p) process can also be expressed in the determinant/adjoint form as

$$|\boldsymbol{\Phi}(L)|\,\boldsymbol{y}_t = \boldsymbol{\Phi}^{+}(L)\,\boldsymbol{u}_t \, . \tag{2.13}$$

$|\boldsymbol{\Phi}(L)|$ is the determinant of $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Phi}^{+}(L)$ is the adjoint matrix of $\boldsymbol{\Phi}(L)$. Reinsel (1997) noted that $\boldsymbol{\Phi}^{+}(L)$ is a $K \times K$ matrix where the elements are polynomials with a maximum degree of $K - 1$ and is obtained from the relation

$$\boldsymbol{\Phi}^{+}(L) = |\boldsymbol{\Phi}(L)|[\boldsymbol{\Phi}(L)]^{-1}. \tag{2.14}$$

The determinantal polynomial,$|\boldsymbol{\Phi}(L)|$ is of maximum order Kp (Wei, 2006). The right hand side of equation (2.13), $\boldsymbol{\Phi}^{+}(L)\,\boldsymbol{u}_t$ is of the form of a MA$(K-1)$ process (Reinsel, 1997). This means that each of the individual components of $\boldsymbol{y}_t$ follow a univariate ARMA process with a maximum order of $Kp, (K-1)p$ .This order is less if there are common factors present between the autoregressive and moving average polynomials (Wei, 2006).

2.3 Autocovariance and Autocorrelation for a VAR(p) Process

An understanding of the autocovariances of a VAR(p) process can be gained by considering the simple example of the stationary VAR(1) process where the mean is known

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \boldsymbol{u}_t \, . \tag{2.15}$$

The process (2.15) can be written in mean adjusted form with $E(\boldsymbol{y}_t) = \boldsymbol{\mu}$ as

$$\boldsymbol{y}_t - \boldsymbol{\mu} = \boldsymbol{\Phi}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{u}_t . \quad (2.16)$$

Tsay (2005) noted that by using the same recursive techniques as demonstrated for the VAR(1) models, it follows that Cov$(\boldsymbol{y}_t, \boldsymbol{u}_t) = \boldsymbol{\Sigma}_u$ and Cov$(\boldsymbol{y}_{t-h}, \boldsymbol{u}_t) = \boldsymbol{\Sigma}_u$ for $h > 0$.

Now, by multiplying both sides of (2.16) by $(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'$ and taking expectations,

$$E[(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] = \boldsymbol{\Phi}_1' E[(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] + E[\boldsymbol{u}_t(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] .$$

If $h = 0$,

$$\begin{aligned} \boldsymbol{\Gamma}(0) &= \boldsymbol{\Phi}_1' E[(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})'] + E[\boldsymbol{u}_t(\boldsymbol{y}_t - \boldsymbol{\mu})'] \\ &= \boldsymbol{\Phi}_1' \boldsymbol{\Gamma}(-1) + \boldsymbol{\Sigma}_u \\ &= \boldsymbol{\Gamma}(1)' \boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}_u . \end{aligned} \quad (2.17)$$

If $h \geq 1$, $E[\boldsymbol{u}_t(\boldsymbol{y}_t - \boldsymbol{\mu})'] = 0$.

Thus for $h \geq 1$,

$$\begin{aligned} \boldsymbol{\Gamma}(h) &= \boldsymbol{\Phi}_1' E[(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] + E[\boldsymbol{u}_t(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] \\ &= \boldsymbol{\Phi}_1' \boldsymbol{\Gamma}(h-1) \\ &= \boldsymbol{\Gamma}(0)(\boldsymbol{\Phi}_1')^h . \end{aligned} \quad (2.18)$$

Thus it follows that $\boldsymbol{\Gamma}(0) = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(1) \boldsymbol{\Phi}_1' + \boldsymbol{\Sigma}_u$ and $\boldsymbol{\Gamma}(1) = \boldsymbol{\Gamma}(0) \boldsymbol{\Phi}_1'$ (Reinsel, 1997).

If the covariance matrix $\boldsymbol{\Gamma}(h)$ is given, then the values of $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Sigma}_u$ can be calculated recursively from

$$\boldsymbol{\Phi}_1' = \boldsymbol{\Gamma}(1)\, \boldsymbol{\Gamma}(0)^{-1}$$
$$\boldsymbol{\Sigma}_u = \boldsymbol{\Gamma}(0) - \boldsymbol{\Gamma}(-1)\boldsymbol{\Gamma}(0)^{-1}\boldsymbol{\Gamma}(-1) = \boldsymbol{\Gamma}(0) - \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}(0) \boldsymbol{\Phi}_1' .$$

The covariance for the VAR (p) process is found in a similar manner. If equation (2.4) is multiplied by $(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'$ on both sides and if expectations are taken then,

$$\begin{aligned} E[(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] &= E(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})' [(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})\boldsymbol{\Phi}_1' + (\boldsymbol{y}_{t-2} - \boldsymbol{\mu})\boldsymbol{\Phi}_2' + \cdots \\ &+ (\boldsymbol{y}_{t-p} - \boldsymbol{\mu})\boldsymbol{\Phi}_p'] + E[(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'\boldsymbol{u}_t] \end{aligned} \quad (2.19)$$

$$\begin{aligned} E[(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] &= \boldsymbol{\Phi}_1' E[(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] \\ &+ \boldsymbol{\Phi}_2' E[(\boldsymbol{y}_{t-2} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] + \cdots + \boldsymbol{\Phi}_p' E[(\boldsymbol{y}_{t-p} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})'] + \boldsymbol{\Sigma}_u . \end{aligned} \quad (2.20)$$

If $h = 0$, (2.20) is simplified to

$$\begin{aligned} E[(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})'] &= \boldsymbol{\Phi}_1' E[(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})'] + \boldsymbol{\Phi}_2' E[(\boldsymbol{y}_{t-2} - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})'] \\ &+ \ldots + \boldsymbol{\Phi}_p' E[(\boldsymbol{y}_{t-p} - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})'] + \boldsymbol{\Sigma}_u \end{aligned}$$

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Phi}_1{}' \boldsymbol{\Gamma}(-1) + \cdots + \boldsymbol{\Phi}_p{}' \boldsymbol{\Gamma}(-p) + \boldsymbol{\Sigma}_u \,. \tag{2.21}$$

Since the process is stationary, $\boldsymbol{\Gamma}(-i) = \boldsymbol{\Gamma}(i)'$.

Thus $\boldsymbol{\Gamma}(0) = \boldsymbol{\Phi}_1{}' \boldsymbol{\Gamma}(1)' + \cdots + \boldsymbol{\Phi}_p{}' \boldsymbol{\Gamma}(p)' + \boldsymbol{\Sigma}_u$.

Similarly for $h = 1, .., p$, (2.20) is simplified to

$$h = 1\text{: } \boldsymbol{\Gamma}(1) - \boldsymbol{\Gamma}(0)\, \boldsymbol{\Phi}_1{}' - \boldsymbol{\Gamma}(1)'\, \boldsymbol{\Phi}_2{}' - \boldsymbol{\Gamma}(2)'\, \boldsymbol{\Phi}_3{}' - \cdots - \boldsymbol{\Gamma}(p-1)'\, \boldsymbol{\Phi}_p{}' = 0 \tag{2.22}$$

$$h = 2\text{: } \boldsymbol{\Gamma}(2) - \boldsymbol{\Gamma}(1)\, \boldsymbol{\Phi}_1{}' - \boldsymbol{\Gamma}(0)\, \boldsymbol{\Phi}_2{}' - \boldsymbol{\Gamma}(1)'\, \boldsymbol{\Phi}_3{}' - \cdots - \boldsymbol{\Gamma}(p-2)' \boldsymbol{\Phi}_p{}' = 0 \tag{2.23}$$

.
.
.

$$h = p\text{: } \boldsymbol{\Gamma}(p) - \boldsymbol{\Gamma}(p-1)\, \boldsymbol{\Phi}_1{}' - \boldsymbol{\Gamma}(p-2)\, \boldsymbol{\Phi}_2{}' - \boldsymbol{\Gamma}(p-3)\, \boldsymbol{\Phi}_3{}' - \cdots - \boldsymbol{\Gamma}(0)\, \boldsymbol{\Phi}_p{}' = 0. \tag{2.24}$$

In the case where $h \geq p$, $\boldsymbol{\Gamma}(h) = \boldsymbol{\Phi}_1{}' \boldsymbol{\Gamma}(h-1) + \cdots + \boldsymbol{\Phi}_p{}' \boldsymbol{\Gamma}(h-p) = 0.$

The equations (2.22) – (2.24) are known as the Yule-Walker equations and can be used to solve for the parameter matrices $\boldsymbol{\Phi}_1$,…, $\boldsymbol{\Phi}_p$ in terms of $\boldsymbol{\Gamma}(0), \ldots, \boldsymbol{\Gamma}(p)$ (Box et al., 2008).

The equations (2.22) – (2.24) can also be expressed in matrix form as

$$\begin{bmatrix} \boldsymbol{\Gamma}(1) \\ \boldsymbol{\Gamma}(2) \\ \vdots \\ \boldsymbol{\Gamma}(p) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}(0) & \boldsymbol{\Gamma}(1)' & \boldsymbol{\Gamma}(2)' & \cdots & \boldsymbol{\Gamma}(p-1)' \\ \boldsymbol{\Gamma}(1) & \boldsymbol{\Gamma}(0) & \boldsymbol{\Gamma}(1)' & \vdots & \boldsymbol{\Gamma}(p-2)' \\ \vdots & & \vdots & \ddots & \vdots \\ & \vdots & & & \\ \boldsymbol{\Gamma}(p-1) & \boldsymbol{\Gamma}(p-2) & \boldsymbol{\Gamma}(p-3) & \cdots & \boldsymbol{\Gamma}(0) \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_1{}' \\ \boldsymbol{\Phi}_2{}' \\ \vdots \\ \boldsymbol{\Phi}_p{}' \end{bmatrix}.$$

Once the values of $\boldsymbol{\Phi}_1$,…, $\boldsymbol{\Phi}_p$ are determined, the value of $\boldsymbol{\Sigma}_u$ is obtained from

$\boldsymbol{\Gamma}(0) - \sum_{i=1}^{p} \boldsymbol{\Gamma}(i)'\boldsymbol{\Phi}_i{}'$.

The autocorrelations for values of $h = 0,1, \ldots, p$ are determined by the relation

$\boldsymbol{\rho}(h) = \boldsymbol{D}^{\frac{-1}{2}}\, \boldsymbol{\Gamma}(h)\, \boldsymbol{D}^{\frac{-1}{2}}$,

where $\boldsymbol{D} = \begin{pmatrix} \gamma_{11}(0) & 0 & \cdots & 0 \\ 0 & \gamma_{22}(0) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{KK}(0) \end{pmatrix}$

is a diagonal matrix with the autocovariances of the VAR(p) process on the diagonal. The autocorrelations are generally easier to work with as compared to the autocovariances as they do not depend on the unit of measurement (Lütkepohl, 2005). The cross correlation function between two individual time series is

$$r_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}} \; i, j = 1, \dots, K.$$

2.4 VAR(p) Order Selection/Identification

The VAR(p) model (2.2) is very extensive and may have a large number of parameters present. Given a VAR(p) model of a finite length $t = 1, \dots, T$, the objective is to try and find a model that has as few parameters that are required to be estimated while at the same time one that suitably represents the dynamic relationships present in the model (Tiao & Tsay, 1983). In this section methods to obtain the correct lag order p are discussed.

The correct order of p is of prime importance as the inclusion of each additional lag reduces the degrees of freedom by the square of the total number of variables present in the system (Fackler & Krieger, 1986). If the model order selected is larger than that of the optimal model order, then the model is said to have been over fitted which can often result in inefficient parameter estimates (de Waele & Broersen, 2003). Over fitting the model is a problem with multivariate models because the number of parameters that are required to be estimated increases at a significantly quicker pace as the order of the model increases i.e. the degrees of freedom are wasted (Enders, 2004). In addition, if the order selected is too small (i.e. the model is under fitted), then it will lead to the dynamics and the effects of variables being ignored which could result in a reduction in the forecasting accuracy of the model (Escanciano et al., 2010).

The lag length selected should be the most parsimonious while at the same time should account for the dynamics of the model. Brandt and Williams (2007) propose the following rules of thumb for selecting lag length that are applicable to seasonal data (monthly or quarterly) only.

(i) The VAR models should generally have enough lags to encompass the full cycle length of the data. Thus, in the case of monthly data the minimum number of lags required in the model should be at least 12, while for quarterly data there should be at least four lags present.

(ii) The lag length should not be more than a quarter of the degrees of freedom for an equation i.e. if r is the number of endogenous variables, p is the lag length and T is the number of observations, then the value of $rp + 1$ should be less than T.

The Yule-Walker equations can also be used to find the order of p. This is done by finding the partial autoregression matrix of order g which is determined from the matrix coefficients $\boldsymbol{\Phi}_{1g}, \boldsymbol{\Phi}_{2g}, \ldots, \boldsymbol{\Phi}_{gg}$ in the equations

$$\boldsymbol{\Gamma}(h) = \sum_{i=1}^{g} \boldsymbol{\Phi}_{ig}\, \boldsymbol{\Gamma}(h-i) \qquad h = 1, \ldots, g. \tag{2.25}$$

These equations (2.25) arise when a VAR model of order g has been fitted to $\boldsymbol{y}_t$. $\boldsymbol{y}_t$ is a VAR(p) process if the partial autoregressive matrices $\boldsymbol{\Phi}_{gg} \quad g = 1,2 \ldots$ are equal to 0 for $g > p$, i.e. they cut off after lag p (Reinsel, 1997).

The above tests however, depend very much on the analyst's discretion. This is because model selection for multivariate time series is significantly more complex than for univariate series as the patterns need to be detected through the matrices of autocovariances (Granger & Newbold, 1986). More formal tests have therefore been developed. The first test involves a sequential procedure used to find the order of p. If g is an upper bound then define the null and alternate hypothesis as

H_0: $\boldsymbol{\Phi}_g = 0$
H_1: $\boldsymbol{\Phi}_g \neq 0$

There have been various methods which have been derived in order to determine a test statistic. Tsay (2005) derived a statistic which was based on a method used by Tiao and Box (1981) and Reinsel (1997). Consider the following VAR(g) model

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_g \boldsymbol{y}_{t-g} + \boldsymbol{u}_t \tag{2.26}$$

where the parameters $\boldsymbol{c}_t, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_g$ have been estimated from the use of either ordinary least squares or maximum likelihood estimation.

If the estimate of $\boldsymbol{\Phi}_g$ is denoted by $\widehat{\boldsymbol{\Phi}}_g$, then the residuals u_t are recursively estimated from

$$\widehat{\boldsymbol{u}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{c}}_t - \widehat{\boldsymbol{\Phi}}_1 \boldsymbol{y}_{t-1} - \cdots - \widehat{\boldsymbol{\Phi}}_g \boldsymbol{y}_{t-g}\ . \tag{2.27}$$

The residual sum of squares and cross products under the null hypothesis H_0: $\boldsymbol{\Phi}_g = 0$ is

$$\text{SS}(g) = \sum_{t=g+1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{c}}_t - \widehat{\boldsymbol{\Phi}}_1 \boldsymbol{y}_{t-1} - \cdots - \widehat{\boldsymbol{\Phi}}_g \boldsymbol{y}_{t-g})(\boldsymbol{y}_t - \hat{\boldsymbol{c}}_t - \widehat{\boldsymbol{\Phi}}_1 \boldsymbol{y}_{t-1} - \cdots - \widehat{\boldsymbol{\Phi}}_g \boldsymbol{y}_{t-g}). \tag{2.28}$$

The test statistic $N(g)$, which makes use of (2.27) and (2.28) is

$$N(g) = -\left(T - g\ - K - \tfrac{3}{2}\right) \ln(|\text{SS}(g)|/|\text{SS}(g-1)\ |).$$

$N(g)$ follows an approximately χ^2 distribution with K^2 degrees of freedom. The null hypothesis (H_0: $\boldsymbol{\Phi}_g = 0$) is rejected for particularly large values of $N(g)$ (Box, Jenkins & Reinsel, 2008). If the null hypothesis is rejected, then the value of g is set as the value of p. If the null hypothesis is not rejected, the hypothesis H_0: $\boldsymbol{\Phi}_{g-1} = 0$ is tested against an alternative of H_1: $\boldsymbol{\Phi}_{g-1} \neq 0 | \boldsymbol{\Phi}_g = 0$. This procedure carries on until there is rejection of the null hypothesis. In general the null hypothesis H_0: $\boldsymbol{\Phi}_{g-i+1} = 0$ is tested against an alternative of H_1: $\boldsymbol{\Phi}_{g-i+1} \neq 0 | \boldsymbol{\Phi}_g = \cdots = \boldsymbol{\Phi}_{g-i+2} = 0$ (Lütkepohl, 2005).

In addition, an alternate likelihood ratio test statistic was derived by Hamilton (1994) by defining the hypotheses

H_0: The model is of order $p = g$
H_1: The model is of order $p = g_1 > g$

The test statistic λ_{LR} makes use of the maximum likelihood function

$$\text{LR} = \frac{-KT}{2}\ln(2\pi) - \frac{T}{2}\ln|\boldsymbol{\Sigma}_u{}^{-1}| - \frac{TK}{2}$$

and is defined as

$$\lambda_{LR} = 2[\ln(\hat{\kappa}) - \ln(\hat{\kappa}_r)]. \qquad (2.29)$$

$\hat{\kappa}$ is the unconstrained maximum likelihood estimator (over the full parameter space) and $\hat{\kappa}_r$ is the constrained or restricted maximum likelihood estimator subject to the restrictions that are stated in the null hypothesis. The statistic λ_{LR} follows a χ^2 distribution with the amount of degrees of freedom being the same as the number of different linear restrictions, $K^2(g_1 - g)$ (Hamilton, 1994). The χ^2 statistic is only accurate asymptotically i.e. for large values where $T \to \infty$ (Lütkepohl, 2005).

Reinsel (1997) noted that the likelihood ratio statistic is not suitable for complex situations, for example, when a low order VAR model is not an adequate representation for the data. It also tends to be spuriously higher in cases where there are a large number of parameters present because the test will then tend to choose incorrect lag lengths resulting in bias in the model (Brandt & Williams, 2007). The statistic is also only applicable if one of the models is a restricted version of the other (Enders, 2004). Finally, the likelihood ratio test statistic is not always satisfactory if the model has been constructed for the specific purposes of forecasting. There have, however been a number of other criteria that have specifically been developed for this particular purpose. These criteria have been developed for univariate models but are easily extended for multivariate models.

The minimum forecasting mean square error is generally used to choose the model order if forecasting is the objective. The one step ahead MSE is calculated from

$\widehat{\boldsymbol{\Sigma}_y}(1) = \frac{T+Kg+1}{T}\boldsymbol{\Sigma}_u$,

where g is the order of the VAR process fitted, T is the sample size and K is the dimension of the time series.

This statistic $\widehat{\boldsymbol{\Sigma}_y}(1) = \frac{T+Kg+1}{T}\boldsymbol{\Sigma}_u$ can be adjusted by the degrees of freedom $\frac{T}{T-Kg-1}$. Fackler and Krieger (1986) noted a statistic which balances the fit of the model with the degrees of freedom used is known as the Final Prediction Error statistic (FPE) and is defined as

$$\text{FPE}(g) = \det\left[\frac{T+Kg+1}{T}\frac{T}{T-Kg-1}\widehat{\boldsymbol{\Sigma}_u}(g)\right]. \tag{2.30}$$

The term $\widehat{\boldsymbol{\Sigma}_u}(g)$ is the maximum likelihood estimate of $\boldsymbol{\Sigma}_u$ when a VAR(g) model is fitted. Equation (2.30) can be simplified to

$$\left[\frac{T+Kg+1}{T-Kg-1}\right]^K \left|\widehat{\boldsymbol{\Sigma}_u}(g)\right| .$$

In order for a suitable value for p to be chosen as the overall model order, various models of orders $g = 1, \ldots, G$ are estimated and recorded with their corresponding FPE values. The order which corresponds to the minimum value of FPE is chosen as the estimate for p.

The Akaike Information Criterion (AIC) is a similar criterion that is based on the FPE and is defined as

$$\begin{aligned}\text{AIC}(g) &= \ln\left|\widehat{\boldsymbol{\Sigma}_u}(g)\right| + \frac{2}{T}(\text{number of free parameters estimated by maximum likelihood}) \\ &= \ln\left|\widehat{\boldsymbol{\Sigma}_u}(g)\right| + \frac{2gK^2}{T} \\ &= T\ln\left|\widehat{\boldsymbol{\Sigma}_u}(g)\right| + 2gK^2. \end{aligned} \tag{2.31}$$

Grubb (1992) noted that the AIC can alternatively also be expressed as

$$\text{AIC}(g) = -2(\text{maximum likelihood}) + 2gK^2 .$$

This AIC criterion (2.31) asymptotically minimises the mean square error for the estimation of the parameters. As in the case of the FPE, the order of p chosen corresponds to the minimum value of the AIC.

An important concept when selecting a model criterion is that of consistency. Lütkepohl (2005) noted that a criterion is consistent if it selects the correct order of the VAR with absolute certainty (probability of one) asymptotically i.e. $lim_{T\to\infty} P(\{\hat{p} = p\}) = 1$. In addition a criterion is strongly consistent if $P[lim_{T\to\infty}\hat{p} = p] = 1$.

This concept can be explained by supposing that there is a VAR criterion say,

$$c_r(g) = \ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + \frac{gc_T}{T}$$

where c_T is the number of freely varying parameters and $\widehat{\boldsymbol{\Sigma}_u}(g)$ is the error covariance matrix. Suppose various models of orders $g = 1, \dots, G$ are fitted and $\hat{p}$ is the order chosen which has minimised $c_r(g)$. $\hat{p}$ will be a consistent estimator if $c_T \to \infty$ and $\frac{c_T}{T} \to 0$ (i.e. converges towards 0) as the value of $T \to \infty$. The estimator $\hat{p}$ will be strongly consistent if all of the above conditions hold and in addition $\frac{c_T}{2\ln\ln T} > 1$ as $T \to \infty$ (Lütkepohl, 2005).

Therefore from the above definitions, the AIC is not consistent because the value of c_T is $2K^2$ and thus as $T \to \infty$, the value of $\frac{c_T}{T}$ does not converge towards 0. It has been reported in the literature that the FPE and the AIC are asymptotically equivalent criteria and since the AIC criterion is not consistent, it follows that the FPE criterion is not consistent either (Lütkepohl, 2005).

The AIC is also minimised at the highest possible order which results in it being biased for finite sample cases. The AIC however, can be modified in order for its efficiency to be improved. The most widely used modification is the corrected AIC or AICC discussed by Hurvich and Tsai (1989). This criterion incorporates the addition of a penalty term $\frac{2(K+1)(K+2)}{T-K-2}$.

$$\text{Thus AICC} = \text{AIC} + \frac{2(K+1)(K+2)}{T-K-2}. \qquad (2.32)$$

Karimi (2011) noted that the AICC criterion (2.32) is extremely efficient compared to the AIC criterion however it is not consistent and suffers from bias in finite samples.

Karimi (2011) defined a more recent modification to the AIC as

$$\text{AIC(modified)} = T\ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + \frac{2gK^2T}{T-(K+1)g}. \qquad (2.33)$$

Simulation studies showed that while this criterion (2.33) led to improved efficiency and had less bias, there was still not enough evidence to suggest that it was consistent. A consistent version of the AIC that can be used for multivariate VAR(p) models with certainty has yet to be determined.

The Kullback information criterion is another criterion which is known to have less bias in finite samples although it is not consistent. It is defined as

$$\text{KIC}(g) = T\ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + 3K^2g.$$

There are however other criteria which have been developed with the primary consideration being given to consistency. The Hannan-Quinn criterion is defined as

$$\begin{aligned}\text{HQ}\ (g) &= \ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + \frac{2\ln\ln T}{T}(\text{number of freely estimated parameters}) \\ &= \ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + \frac{2\ln\ln T}{T}\ (gK^2)\ . \end{aligned} \tag{2.34}$$

The term gK^2 refers to the number of parameters estimated by the log likelihood. The estimator $\hat{p}$ is the order for which the value of HQ (g) at different orders $g = 1, \dots, G$ is at a minimum. The value of c_T is $2K^2\ln(\ln T)$ which means that the HQ (g) criterion is consistent because $\frac{c_T}{T} \to 0$ as $T \to \infty$. In addition, the HQ (g) criterion, (2.34) is strongly consistent for $K > 1$ (Lütkepohl, 2005).

The Bayesian Information Criterion (BIC)/Schwartz Bayesian Criterion (SBC) has also been developed for the specific purpose of forecasting. In this criterion, the penalty term (2 in the AIC) is increased to T and is defined as

$$\begin{aligned}\text{BIC}\ (g) &= \ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + \frac{\ln T}{T}(\text{number of freely estimated parameters}) \\ &= \ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + \frac{\ln T}{T}\ (gK^2)\ . \end{aligned} \tag{2.35}$$

The order which corresponds to the minimum value of BIC (g) is chosen as the estimate for $\hat{p}$. Since $c_T = K^2 \ln T$, the estimator is consistent and the BIC (g) criterion is strongly consistent for any value of K, as $\frac{K^2\ln T}{2\ln\ln T}$ approaches infinity when $T \to \infty$ (Lütkepohl, 2005).

A statistic known as the combined information criterion was defined by de Waele and Broersen (2003) as

$$\text{CIC}\ (g) = T\ln|\widehat{\boldsymbol{\Sigma}_u}(g)| + TK\left(\prod_{i=1}^{g}\frac{1+Kv_i}{1-Kv_i}\right) + g \tag{2.36}$$

where v_i are known as sample variance coefficients that are determined by the use of simulation techniques. These coefficients contain the finite sample behaviour of the estimator. Using simulations, de Waele and Broersen (2003) reported that this technique was accurate in determining the order correctly and does not tend to overfit the model.

The HQ and BIC are generally the preferred criteria for choosing the order in large samples. It should be noted however, that just because HQ and BIC are more consistent, it does not necessarily mean that they are superior overall (Lütkepohl, 2005). Karimi (2011) noted that in many cases efficiency is usually preferred to consistency. Since the AIC and FPE criteria are designed with the primary objective of minimising the forecast error variance, they may not always predict the order of $\hat{p}$ correctly in large samples but they do produce more accurate forecasts (Lütkepohl, 2005).

Studies comparing the performance of criteria such as the AIC, BIC, HQ and the likelihood ratio statistic using simulation techniques have been published.

Lütkepohl (2005) studied 1000 simulations using mean square error forecasts and reported that for small samples $(T = 30)$, the likelihood ratio statistic performed the worst while the BIC criterion performed the best. In larger samples $(T = 100)$ he found that there was no conclusive evidence to suggest that any one of the criteria outperformed any other. He recommended that multiple criteria should be used to in order to determine the order of a VAR(p) model especially if the primary objective is forecasting.

The forecasting performance of unrestricted and restricted VAR models (where the insignificant parameters are set to 0) in which the lag orders were chosen by either the AIC or the BIC was compared by Athanasopoulos and Vahid (2008). For the unrestricted models, the VAR models that were chosen by the BIC criterion performed better than those selected by the AIC, while for longer time horizons $(h > 2)$, the AIC had a better performance. Amongst the restricted VAR models, the AIC outperformed the BIC at all forecast horizons.

A more innovative procedure used by Hatemi and Hacker (2009) showed that the likelihood ratio test combined with the BIC and HQ criteria results in the correct lag order being chosen more frequently.

2.5 Estimation of the VAR $(\boldsymbol{p})$ Model

The K dimensional VAR(p) model $\boldsymbol{y}_1, \dots, \boldsymbol{y}_T$ where $\boldsymbol{y}_t = (y_{1t}, \dots, y_{Kt})$ has the following representation

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t \,.$$

In general the parameters, $\boldsymbol{c}_t$, $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p$ and $\boldsymbol{\Sigma}_u$ are unknown and are required to be estimated. There are various methods of estimation that can be used, the most well known being least squares estimation, Yule - Walker estimation and maximum likelihood estimation.

2.5.1 Least Squares Estimation

Least squares estimation is easy to implement and is quick as no iterations are generally required for the estimates to be obtained. The ordinary least squares estimates are consistent and asymptotically efficient (Enders, 2004). The estimates are also unbiased since all of the regressors are predetermined and the error term is white noise (Ewing, Kruse, Shroeder & Smith, 2007). The least squares estimation procedure was derived by Lütkepohl (2005) as follows

The VAR(p) model for $t = 1, \dots, T$ can be written in the general from of the multivariate linear model

$\mathbf{Y} = \boldsymbol{\varrho}\boldsymbol{X} + \mathbf{U} = \boldsymbol{c}_t + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Y}_{t-i}$ (i.e. the general form of the multivariate linear model). (2.37)

$\mathbf{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_T) = \begin{pmatrix} y_{11} & \cdots & y_{1T} \\ \vdots & \ddots & \vdots \\ y_{K1} & \cdots & y_{KT} \end{pmatrix}$ is a $(K \times T)$ vector.

$\boldsymbol{\varrho} = (\boldsymbol{c}_t\ , \boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p)$ is the $(K \times (Kp + 1))$ vector of parameters.

$\boldsymbol{X}_t = \begin{bmatrix} 1 \\ \boldsymbol{y}_t \\ \vdots \\ \boldsymbol{y}_{t-p-1} \end{bmatrix}$ is a $((Kp + 1) \times 1)$ vector.

$\boldsymbol{X} = (\boldsymbol{X}_0\ , \dots, \boldsymbol{X}_{p-1}\ )$ is a $((Kp + 1) \times T)$ vector.

It thus follows that $\boldsymbol{\varrho}\boldsymbol{X}$ is of dimension $(K \times (Kp + 1), (Kp + 1) \times T)$.

$\mathbf{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_T] = \begin{pmatrix} u_{11} & \cdots & u_{1T} \\ \vdots & \ddots & \vdots \\ u_{K1} & \cdots & u_{KT} \end{pmatrix}$ is a $(K \times T)$ vector of error terms.

The 'vec' operator is one which can transform a matrix by stacking the column vectors below each other. Defining the following for which the 'vec' operator is applied,

$\boldsymbol{y}\ = \text{vec}(\mathbf{Y})$ which is of dimension $(KT \times 1)$.

$\boldsymbol{\varrho}^* = \text{vec}(\boldsymbol{\varrho})$ which is of dimension $((K^2p + K) \times 1)$.

$\mathbf{u} = \text{vec}\ (\mathbf{U})$ which is of dimension $(KT \times 1)$.

The first procedure is to obtain the an estimate for the vector of parameters $\widehat{\boldsymbol{\varrho}}$. This is done by applying the 'vec' operator to the equation (2.37),

$$\begin{aligned} \text{vec}(\mathbf{Y}) &= \text{vec}(\boldsymbol{\varrho}\boldsymbol{X}) + \text{vec}(\mathbf{U}) \\ &= (\boldsymbol{X}\,' \otimes\ \boldsymbol{I}_{\boldsymbol{K}})\text{vec}(\boldsymbol{\varrho}) + \text{vec}(\mathbf{U}) \end{aligned} \qquad (2.38)$$

$$\boldsymbol{y}\ = (\boldsymbol{X}' \otimes\ \boldsymbol{I}_{\boldsymbol{K}})\ \boldsymbol{\varrho}^* + \mathbf{u} \qquad (2.39)$$

$\otimes$ is known as the Kronecker product and is defined in the Appendix.

The covariance matrix of $\mathbf{u}$ is

$$E[\text{vec}(\mathbf{U})\ \text{vec}(\mathbf{U})'] = E\begin{pmatrix}\boldsymbol{u}_1\\ \vdots \\ \boldsymbol{u}_T\end{pmatrix}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)$$

$$= \begin{bmatrix} E(\boldsymbol{u}_1\boldsymbol{u}_1') & \cdots & E(\boldsymbol{u}_1\boldsymbol{u}_T') \\ \vdots & \ddots & \vdots \\ E(\boldsymbol{u}_T\boldsymbol{u}_1') & \cdots & E(\boldsymbol{u}_T\boldsymbol{u}_T') \end{bmatrix}.$$

In order for least square estimates of $\boldsymbol{\varrho}$ to be obtained, the residual sum of squares $\text{SS}(\boldsymbol{\varrho}^*) = \mathbf{u}'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)^{-1}\,\mathbf{u}$ needs to be minimised.

$$\begin{aligned}\text{SS}(\boldsymbol{\varrho}^*) &= \mathbf{u}'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)^{-1}\,\mathbf{u} \\ &= \text{vec}(\mathbf{U})'\boldsymbol{\Sigma}_u{}^{-1}\,\text{vec}(\mathbf{U})' \\ &= \text{vec}(\mathbf{Y} - \boldsymbol{\varrho}\boldsymbol{X})'\boldsymbol{\Sigma}_u{}^{-1}\,\text{vec}(\mathbf{Y} - \boldsymbol{\varrho}\boldsymbol{X}) \\ &= \text{tr}[(\mathbf{Y} - \boldsymbol{\varrho}\boldsymbol{X})'\boldsymbol{\Sigma}_u{}^{-1}(\boldsymbol{Y} - \boldsymbol{\varrho}\boldsymbol{X})'] \\ &= [\boldsymbol{y} - (\boldsymbol{X}'\otimes \boldsymbol{I}_K)\boldsymbol{\varrho}^*]'(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})[\,\boldsymbol{y} - (\boldsymbol{X}'\otimes \boldsymbol{I}_K)\boldsymbol{\varrho}^*] \end{aligned} \quad (2.40)$$

tr() is the trace function of a matrix such that tr[A] refers to the sum of all the diagonal elements for a given matrix A (Tsay, 2005). Equation (2.40) is simplified to

$$\begin{aligned}\text{SS}(\boldsymbol{\varrho}^*) = \boldsymbol{y}'(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y} + \boldsymbol{\varrho}^{*\prime}(\boldsymbol{X}\otimes \boldsymbol{I}_K)(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})(\boldsymbol{X}'\otimes \boldsymbol{I}_K)\boldsymbol{\varrho}^* \\ -2\,\boldsymbol{\varrho}^{*\prime}(\boldsymbol{X}'\otimes \boldsymbol{I}_K)'(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y}.\end{aligned} \quad (2.41)$$

Note : $(\boldsymbol{X}\otimes \boldsymbol{I}_K)(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})(\boldsymbol{X}'\otimes \boldsymbol{I}_K)$
$= (\boldsymbol{X} \otimes \boldsymbol{\Sigma}_u{}^{-1})\,(\boldsymbol{X}'\otimes \boldsymbol{I}_K)$
$= (\boldsymbol{X}\boldsymbol{X}' \otimes \boldsymbol{\Sigma}_u{}^{-1})$

and $(\boldsymbol{X}'\otimes \boldsymbol{I}_K)'(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})$
$= (\boldsymbol{X} \otimes \boldsymbol{\Sigma}_u{}^{-1})$

Thus $\text{SS}(\boldsymbol{\varrho}^*) = \boldsymbol{y}'(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y} + \boldsymbol{\varrho}^{*\prime}(\boldsymbol{X}\boldsymbol{X}'\otimes \boldsymbol{\Sigma}_u{}^{-1})\boldsymbol{\varrho}^* - 2\boldsymbol{\varrho}^{*\prime}(\boldsymbol{X}\otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y}.$ (2.42)

Partially differentiating equation (2.42) with respect to $\boldsymbol{\varrho}^*$ and setting it equal to 0

$$\frac{d\text{SS}(\boldsymbol{\varrho}^*)}{d\boldsymbol{\varrho}^*} = 2\big(\boldsymbol{X}\boldsymbol{X}'\otimes \boldsymbol{\Sigma}_u{}^{-1}\big)\boldsymbol{\varrho}^* - 2(\boldsymbol{X}\otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y} = 0$$

$$\big(\boldsymbol{X}\boldsymbol{X}'\otimes \boldsymbol{\Sigma}_u{}^{-1}\big)\,\boldsymbol{\varrho}^* = (\boldsymbol{X}\otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y}\,. \quad (2.43)$$

If the terms in (2.43) are rearranged, the least squares estimator of $\boldsymbol{\varrho}$, $\widehat{\boldsymbol{\varrho}^*}$ will be

$$\begin{aligned}\widehat{\boldsymbol{\varrho}^*} &= (\boldsymbol{X}\boldsymbol{X}'\otimes \boldsymbol{\Sigma}_u{}^{-1})^{-1}(\boldsymbol{X}\otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\otimes \boldsymbol{\Sigma}_u)\,(\boldsymbol{X}\otimes \boldsymbol{\Sigma}_u{}^{-1})\,\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\otimes \boldsymbol{I}_K)\,\boldsymbol{y}\,. \end{aligned} \quad (2.44)$$

Wei (2006) noted that $\widehat{\boldsymbol{\varrho}^*}$ is distributed joint multivariate normal with mean $\boldsymbol{\varrho}^*$ and covariance matrix $\boldsymbol{\Sigma}_u \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}$.

If the mean is known, then the VAR(p) model can be written in mean adjusted form as

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\Phi}_2\,(\boldsymbol{y}_{t-2} - \boldsymbol{\mu}) + \ldots\ldots + \boldsymbol{\Phi}_p(\boldsymbol{y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{u}_t\,. \tag{2.45}$$

This estimation method is similar to the previous case although there are a few differences. These are;

$\mathbf{Y}$ now comprises of the following terms $\mathbf{Y} = (\boldsymbol{y}_1 - \boldsymbol{\mu}, \ldots\,, \boldsymbol{y}_T - \boldsymbol{\mu})$.

The vector $\boldsymbol{X}_t$, is denoted as $\boldsymbol{Z}_t = \begin{bmatrix} \boldsymbol{y}_t - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{t-p-1} - \boldsymbol{\mu} \end{bmatrix}$ and is of dimension $(Kp \times 1)$.

$\boldsymbol{Z} = (\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_{T-1})$ is of dimension $(Kp \times T)$.

The vector of parameters is defined as $\boldsymbol{\varrho}_0 = (\,\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p)$ and is of dimension $(K \times Kp)$.

vec$(\boldsymbol{\varrho}_0) = \boldsymbol{\varrho}_0{}^*$ is of dimension $(K^2p \times 1)$.

The least squares estimator of (2.45) is

$$\widehat{\boldsymbol{\varrho}_0}^* = ((\boldsymbol{Z}'\boldsymbol{Z})^{-1} \otimes \boldsymbol{I}_K)\mathrm{vec}(\mathbf{Y}).$$

In most cases however, the true mean $\boldsymbol{\mu}$ is unknown in advance. If this is the case, then the vector of sample means $\overline{\boldsymbol{y}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t$ can be used as an estimate for $\boldsymbol{\mu}$ (Lütkepohl, 2005). $\boldsymbol{\mu}$ can also be estimated from $\widehat{\boldsymbol{\mu}} = (\boldsymbol{I} - \widehat{\boldsymbol{\Phi}}_1 - \cdots - \widehat{\boldsymbol{\Phi}}_p)\hat{\boldsymbol{c}}_t\,.$

Karimi (2011) noted that the residual covariance matrix of the estimated VAR(p) model with no intercept is $\widehat{\boldsymbol{y}}_t = \widehat{\boldsymbol{\Phi}}_1\widehat{\boldsymbol{y}}_{t-1} + \ldots + \widehat{\boldsymbol{\Phi}}_p\widehat{\boldsymbol{y}}_{t-p}$ and is of the form $\frac{1}{T-p}(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t)(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t)'$.

2.5.2 Yule – Walker Estimation

A quicker method of obtaining the estimates is through the use of Yule-Walker equations which were discussed earlier. This method is approximately equivalent to that of least squares estimation (Reinsel, 1997).

The Yule-Walker equations imply that the covariance matrix $\boldsymbol{\Gamma}(h)$ is

$$\boldsymbol{\Gamma}(h) = \left[\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p\right]' \begin{pmatrix} \boldsymbol{\Gamma}(h-1)' \\ \vdots \\ \boldsymbol{\Gamma}(h-p)' \end{pmatrix} \; h > 0 \tag{2.46}$$

Or $[\boldsymbol{\Gamma}(1), \dots, \boldsymbol{\Gamma}(p)] = \left[\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p\right]' \begin{bmatrix} \boldsymbol{\Gamma}(0) & \cdots & \boldsymbol{\Gamma}(p-1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}(-p+1) & \cdots & \boldsymbol{\Gamma}(0) \end{bmatrix}$. (2.47)

If the vector of parameters $[\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p]$ is expressed as the main subject of the equation (2.47), then,

$$[\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p] = [\boldsymbol{\Gamma}(1), \dots, \boldsymbol{\Gamma}(p)] \begin{bmatrix} \boldsymbol{\Gamma}(0) & \cdots & \boldsymbol{\Gamma}(p-1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}(-p+1) & \cdots & \boldsymbol{\Gamma}(0) \end{bmatrix}^{-1}.$$

The inverse matrix $\begin{bmatrix} \boldsymbol{\Gamma}(0) & \cdots & \boldsymbol{\Gamma}(p-1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}(-p+1) & \cdots & \boldsymbol{\Gamma}(0) \end{bmatrix}^{-1}$ is estimated by $\frac{\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{Z}}'}{T}$ and $[\boldsymbol{\Gamma}(1), \dots, \boldsymbol{\Gamma}(p)]$ is estimated by $\frac{\widehat{\boldsymbol{y}}_t \widehat{\boldsymbol{Z}}'}{T}$ (Lütkepohl, 2005).

From these estimates the vector of parameters $\left[\widehat{\boldsymbol{\Phi}}_1, \dots, \widehat{\boldsymbol{\Phi}}_p\right]$ is estimated by $\widehat{\boldsymbol{y}}_t \boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}$. Box et al. (2008) noted that the error covariance matrix, $\boldsymbol{\Sigma}_u$ is obtained from the relation

$$\boldsymbol{\Sigma}_u = \boldsymbol{\Gamma}(0) - \sum_{j=1}^{p} \boldsymbol{\Gamma}(-j)\, \widehat{\boldsymbol{\Phi}}_j .$$

The advantage of the Yule - Walker estimation is that it is quick and simple to use. Reinsel (1997) however, noted that estimates obtained from using the Yule - Walker equations suffer from a greater bias than those obtained from using least squares estimation. The least squares estimates also have a better sampling behaviour when the VAR(p) process is near non-stationary.

2.5.3 Maximum Likelihood Estimation

If the VAR(p) process has a known distribution and $\boldsymbol{y}_t$ is assumed to be normally distributed, then maximum likelihood estimation can be used as an alternative procedure to least squares estimation. The main advantage of maximum likelihood estimation is that it is efficient asymptotically. The maximum likelihood estimation procedure below was discussed in detail by Lütkepohl (2005).

Since $\boldsymbol{u}_t \sim N(0, \boldsymbol{\Sigma}_u)$,

It follows that $\mathbf{u} = \text{vec}\,(\mathbf{U}) = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix} \sim N(0, \boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)$.

From the properties of a normal distribution, the probability density of $\mathbf{u}$ is

$$f_u(\mathbf{u}) = \frac{1}{(2\pi)^{\frac{KT}{2}}} |\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u|^{\frac{-1}{2}} \exp\left[\frac{-1}{2}\mathbf{u}'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})\,\mathbf{u}\right]. \tag{2.48}$$

Consider the VAR(p) model written in deviations of the mean form

$$\boldsymbol{y}_t - \boldsymbol{\mu} = \boldsymbol{\Phi}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\Phi}_2(\boldsymbol{y}_{t-2} - \boldsymbol{\mu}) + \cdots + \boldsymbol{\Phi}_p(\boldsymbol{y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{u}_t\,. \tag{2.49}$$

For observations $[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$, this representation (2.49) is equivalent to

$$\begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_T - \boldsymbol{\mu} \end{bmatrix} = \boldsymbol{\Phi}_1 \begin{bmatrix} \boldsymbol{y}_0 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{T-1} - \boldsymbol{\mu} \end{bmatrix} + \boldsymbol{\Phi}_2 \begin{bmatrix} \boldsymbol{y}_{-1} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{T-2} - \boldsymbol{\mu} \end{bmatrix} + \cdots + \boldsymbol{\Phi}_p \begin{bmatrix} \boldsymbol{y}_{1-p} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{T-p} - \boldsymbol{\mu} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}. \tag{2.50}$$

(2.50) can be rewritten as

$$\begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_T - \boldsymbol{\mu} \end{bmatrix} - \boldsymbol{\Phi}_1 \begin{bmatrix} \boldsymbol{y}_0 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_T - \boldsymbol{\mu} \end{bmatrix} - \boldsymbol{\Phi}_2 \begin{bmatrix} \boldsymbol{y}_{-1} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{T-2} - \boldsymbol{\mu} \end{bmatrix} - \cdots - \boldsymbol{\Phi}_p \begin{bmatrix} \boldsymbol{y}_{1-p} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{T-p} - \boldsymbol{\mu} \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{u}_1 &= (\boldsymbol{y}_1 - \boldsymbol{\mu}) - \boldsymbol{\Phi}_1(\boldsymbol{y}_0 - \boldsymbol{\mu}) - \boldsymbol{\Phi}_2(\boldsymbol{y}_{-1} - \boldsymbol{\mu}) - \cdots - \boldsymbol{\Phi}_p(\boldsymbol{y}_{1-p} - \boldsymbol{\mu}) \\ \boldsymbol{u}_2 &= (\boldsymbol{y}_2 - \boldsymbol{\mu}) - \boldsymbol{\Phi}_1(\boldsymbol{y}_1 - \boldsymbol{\mu}) - \boldsymbol{\Phi}_2(\boldsymbol{y}_0 - \boldsymbol{\mu}) - \cdots - \boldsymbol{\Phi}_p(\boldsymbol{y}_{2-p} - \boldsymbol{\mu}) \\ &\vdots \\ \boldsymbol{u}_T &= (\boldsymbol{y}_T - \boldsymbol{\mu}) - \boldsymbol{\Phi}_1(\boldsymbol{y}_{T-1} - \boldsymbol{\mu}) - \boldsymbol{\Phi}_2(\boldsymbol{y}_{T-2} - \boldsymbol{\mu}) - \cdots - \boldsymbol{\Phi}_p(\boldsymbol{y}_{T-p} - \boldsymbol{\mu}) \end{aligned}$$

$$\mathbf{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & 0 \\ -\boldsymbol{\Phi}_1 & \boldsymbol{I}_K & & 0 \\ \vdots & & \ddots & \vdots \\ -\boldsymbol{\Phi}_p & -\boldsymbol{\Phi}_{p-1} & \boldsymbol{I}_K & \\ 0 & & & \\ 0 & & -\boldsymbol{\Phi}_p \cdots & \boldsymbol{I}_K \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_T - \boldsymbol{\mu} \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Phi}_1 & -\boldsymbol{\Phi}_2 & \cdots & -\boldsymbol{\Phi}_p \\ -\boldsymbol{\Phi}_2 & -\boldsymbol{\Phi}_3 & & 0 \\ \vdots & & \ddots & \vdots \\ -\boldsymbol{\Phi}_p & 0 & & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_0 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{-p+1} - \boldsymbol{\mu} \end{bmatrix}. \tag{2.51}$$

In order to determine the probability distribution of $\boldsymbol{y}$, a transformation is required to determine the value of the Jacobian matrix $\frac{dvec(\boldsymbol{U})}{dvec(\boldsymbol{y})'}$, which is a lower triangular matrix with a unit diagonal and consists of the parameters $[\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p]$.

$$\frac{dvec(\boldsymbol{U})}{dvec(\boldsymbol{y})'} = \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & 0 \\ -\boldsymbol{\Phi}_1 & \boldsymbol{I}_K & & 0 \\ \vdots & & \ddots & \vdots \\ -\boldsymbol{\Phi}_p & -\boldsymbol{\Phi}_{p-1} & \boldsymbol{I}_K & \\ 0 & & & \\ 0 & & -\boldsymbol{\Phi}_p \cdots & \boldsymbol{I}_K \end{bmatrix} \tag{2.52}$$

$\mathbf{u}$ can be written as

$$\mathbf{u} = -\tilde{\boldsymbol{\mu}} - (\boldsymbol{Z}' \otimes \boldsymbol{I}_K)\boldsymbol{\varrho}_0{}^*.$$

$\boldsymbol{\varrho}_0{}^* = \text{vec}\,[\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p]$ is the $(K \times Kp)$ vector of parameters and $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\mu}', \ldots, \boldsymbol{\mu}')'$ is the vector of means which is of dimension $(TK \times 1)$.

Since this matrix (2.52) is lower triangular, the determinant $|\frac{dvec(\boldsymbol{U})}{dvec(\boldsymbol{y})'}|$ is one. Thus the probability density of $\boldsymbol{y}$ is

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = |\frac{dvec(\boldsymbol{U})}{dvec(\boldsymbol{y})'}| f_u(\mathbf{u})$$

$$= (2\pi)^{\frac{-KT}{2}} |\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u|^{\frac{-1}{2}}$$
$$\times \exp\,[\,\frac{-1}{2}\,[\boldsymbol{y} - \tilde{\boldsymbol{\mu}} - (\boldsymbol{Z}' \otimes \boldsymbol{I}_K)\boldsymbol{\varrho}_0{}^*]'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})[\,\boldsymbol{y} - \tilde{\boldsymbol{\mu}} - (\boldsymbol{Z}' \otimes \boldsymbol{I}_K)\boldsymbol{\varrho}_0{}^*]. \tag{2.53}$$

The matrix $\boldsymbol{Z}$ is

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{y}_0 - \boldsymbol{\mu} & \boldsymbol{y}_1 - \boldsymbol{\mu} & \cdots & \boldsymbol{y}_{T-1} - \boldsymbol{\mu} \\ \boldsymbol{y}_{-1} - \boldsymbol{\mu} & \boldsymbol{y}_0 - \boldsymbol{\mu} & & \boldsymbol{y}_{T-2} - \boldsymbol{\mu} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{y}_{1-p} - \boldsymbol{\mu} & \boldsymbol{y}_{2-p} - \boldsymbol{\mu} & \cdots & \boldsymbol{y}_{T-p} - \boldsymbol{\mu} \end{bmatrix}.$$

If natural logs are taken on both sides on (2.53), then the resulting log likelihood function is

$$ln\, L(\boldsymbol{\mu}, \boldsymbol{\varrho}_0{}^*, \boldsymbol{\Sigma}_u)$$

$$= -\frac{KT}{2}\, ln(2\pi) - \frac{T}{2}\, ln|\boldsymbol{\Sigma}_u| - \frac{1}{2}[\,\boldsymbol{y} - \tilde{\boldsymbol{\mu}} - (\boldsymbol{Z}' \otimes \boldsymbol{I}_K)\,\boldsymbol{\varrho}_0{}^*]'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})[\,\boldsymbol{y} - \tilde{\boldsymbol{\mu}} - (\boldsymbol{Z}' \otimes \boldsymbol{I}_K)\,\boldsymbol{\varrho}_0{}^*] \tag{2.54}$$

$$= -\frac{KT}{2}\, ln(2\pi) - \frac{T}{2}\, ln|\boldsymbol{\Sigma}_u| - \frac{1}{2}\sum_{t=1}^{T}\, [(\boldsymbol{y}_t - \boldsymbol{\mu}) - \sum_{i=1}^{p} \boldsymbol{\Phi}_i(\boldsymbol{y}_{t-i} - \boldsymbol{\mu})]'$$
$$\times\ \boldsymbol{\Sigma}_u{}^{-1}[(\boldsymbol{y}_t - \boldsymbol{\mu}) - \sum_{i=1}^{p} \boldsymbol{\Phi}_i(\boldsymbol{y}_{t-i} - \boldsymbol{\mu})]$$

$$= -\frac{KT}{2}\, ln(2\pi) - \frac{T}{2}\, ln|\boldsymbol{\Sigma}_u| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i})' \boldsymbol{\Sigma}_u{}^{-1}(\boldsymbol{y}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i})$$
$$+ \boldsymbol{\mu}'(\boldsymbol{I}_K - \sum_{i=1}^{p} \boldsymbol{\Phi}_i)'\boldsymbol{\Sigma}_u{}^{-1} \sum_{t=1}^{T}(\boldsymbol{y}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i})$$
$$-\frac{T}{2}\boldsymbol{\mu}'(\boldsymbol{I}_K - \sum_{i=1}^{p} \boldsymbol{\Phi}_i)'\, \boldsymbol{\Sigma}_u{}^{-1}(\boldsymbol{I}_K - \sum_{i=1}^{p} \boldsymbol{\Phi}_i)\boldsymbol{\mu}\ . \qquad (2.55)$$

The maximum likelihood estimators of $\boldsymbol{\mu}, \boldsymbol{\varrho}_0{}^*$ and $\boldsymbol{\Sigma}_u$ can now be determined by partially differentiating the log likelihood function, $ln\ L(\boldsymbol{\mu},\ \boldsymbol{\varrho}_0{}^*, \boldsymbol{\Sigma}_u)$ with respect to $\boldsymbol{\mu}, \boldsymbol{\varrho}_0{}^*$ and $\boldsymbol{\Sigma}_u$ (Lütkepohl, 2005).

a) For $\boldsymbol{\mu}$:

If (2.55) is partially differentiated with respect to $\boldsymbol{\mu}$,

$$\frac{\partial lnL}{\partial \boldsymbol{\mu}} = (\boldsymbol{I}_K - \sum_{i=1}^{p} \boldsymbol{\Phi}_i)'\, \boldsymbol{\Sigma}_u{}^{-1} \sum_{t=1}^{T}(\boldsymbol{y}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i})$$
$$- T\ (\boldsymbol{I}_K - \sum_{i=1}^{p} \boldsymbol{\Phi}_i)'\, \boldsymbol{\Sigma}_u{}^{-1}(\boldsymbol{I}_K - \sum_{i=1}^{p} \boldsymbol{\Phi}_i)\boldsymbol{\mu}\ . \qquad (2.56)$$

The likelihood function is maximised when (2.56) is equated to 0. The estimator of μ is

$$(\boldsymbol{I}_K - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)'\boldsymbol{\Sigma}_u{}^{-1} \sum_{t=1}^{T}(\boldsymbol{y}_t - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i \boldsymbol{y}_{t-i}) - T\ (\boldsymbol{I}_K - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)'\boldsymbol{\Sigma}_u{}^{-1}(\boldsymbol{I}_K - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)\widehat{\boldsymbol{\mu}} = 0$$

$$T(\boldsymbol{I}_K - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)'\boldsymbol{\Sigma}_u{}^{-1}(\boldsymbol{I}_K - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)\, \widehat{\boldsymbol{\mu}} = (\boldsymbol{I}_K - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)'\boldsymbol{\Sigma}_u{}^{-1} \sum_{t=1}^{T}(\boldsymbol{y}_t - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i \boldsymbol{y}_{t-i})$$

$$\widehat{\boldsymbol{\mu}} = \frac{1}{T}(\boldsymbol{I}_K\ -\ \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i)^{-1} \sum_{t=1}^{T}(\boldsymbol{y}_t - \sum_{i=1}^{p} \widehat{\boldsymbol{\Phi}}_i \boldsymbol{y}_{t-i})\ . \qquad (2.57)$$

b) For the vector of parameters $\boldsymbol{\varrho}_0{}^*$:

If (2.54) is partially differentiated with respect to $\boldsymbol{\varrho}_0{}^*$,

$$\frac{\partial lnL}{\partial \boldsymbol{\varrho}_0{}^*} = (\boldsymbol{Z} \otimes \boldsymbol{I}_K)\ (\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})\ [\ \boldsymbol{y} - \tilde{\boldsymbol{\mu}} - (\boldsymbol{Z}' \otimes \boldsymbol{I}_K)\boldsymbol{\varrho}_0{}^*]$$
$$= (\boldsymbol{Z} \otimes \boldsymbol{\Sigma}_u{}^{-1})(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) - (\boldsymbol{Z}\boldsymbol{Z}' \otimes\ \boldsymbol{\Sigma}_u{}^{-1})\boldsymbol{\varrho}_0{}^*\ . \qquad (2.58)$$

Equating (2.58) to 0,

$$(\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{Z}}' \otimes\ \boldsymbol{\Sigma}_u{}^{-1})\widehat{\boldsymbol{\varrho}_0{}^*} = (\widehat{\boldsymbol{Z}} \otimes \boldsymbol{\Sigma}_u{}^{-1})(\boldsymbol{y} - \widehat{\tilde{\boldsymbol{\mu}}})$$

$$\widehat{\boldsymbol{\varrho}_0{}^*} = \frac{(\widehat{\boldsymbol{Z}} \otimes\ \boldsymbol{\Sigma}_u{}^{-1})(\ \boldsymbol{y} - \widehat{\tilde{\boldsymbol{\mu}}})}{(\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{Z}}' \otimes\ \boldsymbol{\Sigma}_u{}^{-1})}$$

$$= (\ \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{Z}}'{}^{-1}\widehat{\boldsymbol{Z}} \otimes \boldsymbol{I}_K)\ (\boldsymbol{y} - \widehat{\tilde{\boldsymbol{\mu}}})\ . \qquad (2.59)$$

c) For $\boldsymbol{\Sigma}_u$:

Lütkepohl (2005) noted that equation (2.55) can also be written as

$$ln\ L(\boldsymbol{\mu}, \boldsymbol{\varrho}_0{}^*, \boldsymbol{\Sigma}_u) = \frac{-KT}{2} ln(2\pi) - \frac{T}{2} ln|\boldsymbol{\Sigma}_u| - \frac{1}{2} tr[(\mathbf{Y}^0 - \boldsymbol{\varrho}_0 \boldsymbol{Z})\boldsymbol{\Sigma}_u{}^{-1}(\mathbf{Y}^0 - \boldsymbol{\varrho}_0 \boldsymbol{Z})'] \quad (2.60)$$

where $\boldsymbol{Y}^0 = \begin{pmatrix} \boldsymbol{y}_1 - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_T - \boldsymbol{\mu} \end{pmatrix}$.

It follows that by partially differentiating (2.60) with respect to $\boldsymbol{\Sigma}_u$,

$$\frac{\partial lnL}{\partial \boldsymbol{\Sigma}_u} = -\frac{T}{2}\boldsymbol{\Sigma}_u{}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_u{}^{-1}(\widehat{\mathbf{Y}^0} - \widehat{\boldsymbol{\varrho}_0}\widehat{\boldsymbol{Z}})(\widehat{\mathbf{Y}^0} - \widehat{\boldsymbol{\varrho}_0}\hat{Z})'\,\boldsymbol{\Sigma}_u{}^{-1}$$

$$T = (\widehat{\boldsymbol{Y}^0} - \widehat{\boldsymbol{\varrho}_0}\widehat{\boldsymbol{Z}})(\widehat{\boldsymbol{Y}^0} - \widehat{\boldsymbol{\varrho}_0}\widehat{\boldsymbol{Z}})'\,\widehat{\boldsymbol{\Sigma}}_u{}^{-1}$$

$$\widehat{\boldsymbol{\Sigma}}_u{}^{-1} = \frac{1}{T}(\widehat{\boldsymbol{Y}^0} - \widehat{\boldsymbol{\varrho}_0}\widehat{\boldsymbol{Z}})(\widehat{\boldsymbol{Y}^0} - \widehat{\boldsymbol{\varrho}_0}\widehat{\boldsymbol{Z}})'. \quad (2.61)$$

The estimates $\widehat{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{Y}^0}$ are obtained from $\boldsymbol{Z}$ and $\boldsymbol{Y}^0$ respectively by replacing $\boldsymbol{\mu}$ by $\widehat{\boldsymbol{\mu}}$.
Since these equations are nonlinear in the parameters, iterative procedures (such as those discussed in the appendix) need to be employed in order for them to be evaluated. Ma (1997) noted that this log likelihood is very close to being quadratic with respect to the autoregressive parameters which imply that the Newton-Raphson optimisation method is the preferred option for solving these equations.

A modification of the traditional maximum likelihood estimation procedure was proposed by Roy, Fuller and Zhu (2009) who used a VAR(1) representation and a regression type approach in order to find an approximation for the maximum likelihood estimator. This estimator for stationary processes had a limiting distribution that was the same as the ordinary least squares estimator but was not recommended for the non-stationary processes as it had a different limiting distribution.

2.6 Diagnostic Checking of the VAR$(\boldsymbol{p})$ model

In order to check whether a model has a good fit, it is necessary to observe whether the residuals are uncorrelated with each other over a period of time. If the residuals show signs of correlation with each other over a period of time, it means that there is serial correlation present in the model which means that the model does not have a good fit.

The easiest and least computationally burdensome method is to simply plot the residuals over a period of time. This procedure is not recommended as it is not always obvious to detect a

pattern. Brandt and Williams (2007) noted that in many instances, special expertise is required to detect these patterns.

An alternative method is to plot the sample residual autocorrelation function (which shows the correlation of the residuals of the variable with its own past values) and the sample cross correlation functions (which analyses the correlation of a variable with past values of other variables). A model with a good fit will not show any significant evidence of autocorrelation or cross correlations in the residuals (Brandt & Williams, 2007). As a rule of thumb, there is no serial correlation present if the sample residual cross correlations lie between the two standard error limits $\pm \frac{2}{\sqrt{T}}$ (Granger & Newbold, 1986).

There is also the goodness of fit test that can be used in order to test for serial correlation. In order to proceed with this method, assume that $\boldsymbol{u}_t$ is a K dimensional white noise process with covariance matrix $\boldsymbol{\Sigma}_u$. If $\boldsymbol{u}_t$ denotes the vector of residuals of a VAR(p) process, then the residual autocovariance matrices denoted by $\boldsymbol{\Gamma}_u(i)$, are estimated from

$\hat{\boldsymbol{\Gamma}}_u(i) = \frac{1}{T}\sum_{t=i+1}^{T} \boldsymbol{u}_t \boldsymbol{u}_{t-i}{}' \; i = 0, \dots, h$ where $h < T$.

The residual autocovariance matrices, $\hat{\boldsymbol{\Gamma}}_u(i)$ are used to determine the residual autocorrelation function, $\widehat{\boldsymbol{\rho}_u}(i)$. $\widehat{\boldsymbol{\rho}_u}(i)$ is calculated from the relation $\boldsymbol{D}_u{}^{-1}\hat{\boldsymbol{\Gamma}}_u(i)\boldsymbol{D}_u{}^{-1}$ $i = 0, \dots, h$ where $\boldsymbol{D}_u$ is a $(K \times K)$ diagonal matrix which has the square roots of the diagonal elements of $\boldsymbol{\Gamma}_u(0)$ on its diagonal (Reinsel, 1997).

The test procedure in order to test for the significance of residual autocorrelations up to lag h is conducted by defining the hypotheses

$H_0 : \widehat{\boldsymbol{\rho}_u}(1), \dots, \widehat{\boldsymbol{\rho}_u}(h) = 0$ i.e. the residuals are not autocorrelated up to h lags
$H_1 : \widehat{\boldsymbol{\rho}_u}(1), \dots, \widehat{\boldsymbol{\rho}_u}(h) \neq 0$ i.e. the residuals are autocorrelated up to h lags.

These hypotheses can equivalently be expressed as

$H_0 : E[\boldsymbol{u}_t{}'\boldsymbol{u}_{t-i}] = 0 \quad i = 1, \dots, h > p$
$H_1 : E[\boldsymbol{u}_t{}'\boldsymbol{u}_{t-i}] \neq 0 \quad i = 1, \dots, h > p$

In order to find a test statistic, a statistic known as the Portmanteau statistic (denoted by Q_h) was initially derived for univariate time series models though it is easily extendable for multivariate models. This statistic, expressed in terms of residual correlations is

$$Q_h = \sum_{i=1}^{h} tr(\, \widehat{\boldsymbol{\rho}_u}(i)\, \widehat{\boldsymbol{\rho}_u}(0)^{-1}\, \widehat{\boldsymbol{\rho}_u}(-i)\, \widehat{\boldsymbol{\rho}_u}(0)^{-1}) \,. \qquad (2.62)$$

tr() is the trace function discussed earlier. The Q_h statistic (2.62) can also be expressed in terms of residual covariances as

$$
\begin{aligned}
Q_h &= T \sum_{i=1}^{h} tr(\, \widehat{\boldsymbol{\rho}_u}(i)\, \widehat{\boldsymbol{\rho}_u}(0)^{-1}\, \widehat{\boldsymbol{\rho}_u}(-i)\, \widehat{\boldsymbol{\rho}_u}(0)^{-1}) \\
&= T \sum_{i=1}^{h} tr(\, \widehat{\boldsymbol{\rho}_u}(i)\, \widehat{\boldsymbol{\rho}_u}(0)^{-1}\, \widehat{\boldsymbol{\rho}_u}(-i)\, \widehat{\boldsymbol{\rho}_u}(0)^{-1} \widehat{\boldsymbol{D}_u}^{-1} \widehat{\boldsymbol{D}_u}) \\
&= T \sum_{i=1}^{h} tr(\widehat{\boldsymbol{D}_u}\, \widehat{\boldsymbol{\rho}_u}(i) \widehat{\boldsymbol{D}_u}\, \widehat{\boldsymbol{D}_u}^{-1}\, \widehat{\boldsymbol{\rho}_u}(0)^{-1} \widehat{\boldsymbol{D}_u}^{-1} \widehat{\boldsymbol{D}_u}\, \widehat{\boldsymbol{\rho}_u}(-i) \widehat{\boldsymbol{D}_u} \widehat{\boldsymbol{D}_u}^{-1}\, \widehat{\boldsymbol{\rho}_u}(0)^{-1} \widehat{\boldsymbol{D}_u}^{-1}) \\
&= T \sum_{i=1}^{h} tr(\, \hat{\boldsymbol{\Gamma}}_u(i) \hat{\boldsymbol{\Gamma}}_u(0)^{-1} \hat{\boldsymbol{\Gamma}}_u(-i) \hat{\boldsymbol{\Gamma}}_u(0)^{-1}) \,. \qquad (2.63)
\end{aligned}
$$

Q_h follows a χ^2 distribution with $K^2(h-p)$ degrees of freedom. The degrees of freedom is expressed as the number of autocorrelations multiplied by the number of series considered. This approximate distribution is valid provided that the value of h is chosen large enough so that the weights $\boldsymbol{\psi}_j$ in the infinite moving average representation of the VAR(p) model are small for $j > h$ (Reinsel, 1997). Escanciano et al. (2010) recommend that a small value of h should be used when p is small and a larger value of h for higher values of p. The model specified will thus be inadequate for large values of Q_h.

However, the same authors also noted that there are a few limitations of these critical values. The first is that these critical values are accurate only when the number of autocorrelations taken is sufficiently large. Secondly, this limiting distribution is sensitive to h which means that there is a loss of power in the test when h is too large (Pfaff, 2008). Escanciano et al. (2010) have proposed modified critical values which follow a χ^2 distribution with hK^2 degrees of freedom. These critical values take into account the estimation uncertainty by letting the value of h diverge towards infinity.

The Box-Ljung statistic is a modified version of the Portmanteau statistic specifically for use in smaller sample sizes. It accounts for the estimates of serial correlation from the lags 1 to h (Brandt & Williams, 2007). This statistic is defined as

$$
\widetilde{Q_h} = T^2 \sum_{i=1}^{h} (T-i)^{-1}\, tr\big(\hat{\boldsymbol{\Gamma}}_u(i)\, \widehat{\boldsymbol{\Gamma}_u}(0)^{-1} \hat{\boldsymbol{\Gamma}}_u(-i)\, \widehat{\boldsymbol{\Gamma}_u}(0)^{-1}\big). \qquad (2.64)
$$

$\widetilde{Q_h}$ follows the same set of critical values as the Portmanteau statistic. This statistic was tested by Hosking (1980) who conducted a simulation study to conclude that the Box-Ljung statistic outperformed the original Portmanteau statistic for smaller samples.

Lagrange multiplier tests can also be used to test for residual autocorrelation in a VAR(p) process. If the error vector $\boldsymbol{u}_t$ is assumed to be of the form of a VAR(p) model, i.e. $\boldsymbol{u}_t = \boldsymbol{\Lambda}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Lambda}_h \boldsymbol{u}_{t-h} + \boldsymbol{v}_t$ where $\boldsymbol{v}_t$ is a white noise process (Lütkepohl, 2005).

There is no residual autocorrelation present in the model if $\boldsymbol{u}_t = \boldsymbol{v}_t$. The residual covariances of v_t are calculated from $\widetilde{\boldsymbol{\Sigma}}_v = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_t \boldsymbol{v}_t'$. The hypotheses for testing whether there is serial correlation in the residuals are

$H_0 : \boldsymbol{\Lambda}_1 = \cdots = \boldsymbol{\Lambda}_h = 0$ (There is no serial correlation present in the model)
H_1 : At least one of the $\boldsymbol{\Lambda}_j \;\; j \in [1, \ldots, h] \neq 0$

There are various methods which can be used in order to find a suitable Lagrange Multiplier test statistic and its critical values. Brandt and Williams (2007) proposed a step by step procedure by estimating an unrestricted as well as a restricted VAR(p) model. The unrestricted model is of the form

$$\widehat{\boldsymbol{u}}_t = \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\Lambda}_1 \widehat{\boldsymbol{u}}_{t-1} + \cdots + \boldsymbol{\Lambda}_h \widehat{\boldsymbol{u}}_t + \boldsymbol{v}_t \, . \tag{2.65}$$

The second restricted VAR is fitted under the assumption that the null hypothesis $(H_0 : \boldsymbol{\Lambda}_1 = \cdots = \boldsymbol{\Lambda}_h = 0)$ is true. The resulting model is

$$\widehat{\boldsymbol{u}}_t = \ \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + {\boldsymbol{v}_t}^R \, . \tag{2.66}$$

${\boldsymbol{v}_t}^R$ is a white noise process of the restricted model. The error covariance matrix of (2.66) is calculated from $\widetilde{\boldsymbol{\Sigma}}_R = \frac{1}{T}\sum_{t=1}^{T} {\boldsymbol{v}_t}^R {\boldsymbol{v}_t}^{R\prime}$. $\widetilde{\boldsymbol{\Sigma}}_v$ and $\widetilde{\boldsymbol{\Sigma}}_R$ are included in the LM test statistic which is defined as

$$\text{LM} = T[K - tr(\widetilde{\boldsymbol{\Sigma}}_v {\widetilde{\boldsymbol{\Sigma}}_R}^{-1})]. \tag{2.67}$$

The value of K refers to the number of endogenous variables in the system. This LM statistic follows a χ^2 distribution with the degrees of freedom, where hK^2 is the number of restrictions placed on the parameters of the model (2.66) under the null hypothesis that there is no residual correlation present.

As a final measure the information criteria (AIC, BIC, HQ) can be used for the purposes of diagnostic checking if they have not been previously used for selecting the order of the model.

2.7 Forecasting of the VAR(p) Model

It is very useful to be able to forecast future values of a variable under study based on its current and past values. This is true especially when there is not much knowledge available regarding the data generating process of a variable. Forecasting using time series methods is widely used in the fields of economics, finance, engineering, public health and geography. It is of practical value for both policy makers and scientists as it enables and ensures proper planning for the future.

The predictor of a vector of future values, $\boldsymbol{y}_{t+h}$ $h = 1,2, \ldots$ is based on a realisation $(\boldsymbol{y}_s \; s \leq t)$ yields the minimum mean square error matrix $\widehat{\boldsymbol{y}}_t(h) = E(\boldsymbol{y}_{t+h} | \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots)$(Box et al., 2008).

Tsay (2005) noted that the one step ahead forecast at a time origin t is

$$\widehat{\boldsymbol{y}}_t(1) \ = \boldsymbol{c}_t + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \, \boldsymbol{y}_{t+1-i} \tag{2.68}$$

with forecast error $\boldsymbol{u}_t(1) = \ \boldsymbol{y}_{t+1} - \widehat{\boldsymbol{y}}_t(1) \ = \ \boldsymbol{u}_{t+1}$.

The covariance matrix of the forecast error is $\boldsymbol{\Sigma}_u$.

The 2 step ahead forecast is

$\widehat{\boldsymbol{y}}_t(2) = \boldsymbol{c}_t + \boldsymbol{\Phi}_1\widehat{\boldsymbol{y}}_t(1) + \sum_{i=2}^{p}\boldsymbol{\Phi}_i\,\boldsymbol{y}_{t+2-i}$ (2.69)

with forecast error $\boldsymbol{u}_t(2) = \boldsymbol{y}_{t+2} - \widehat{\boldsymbol{y}}_t(2) = \boldsymbol{u}_{t+2} + \boldsymbol{\Phi}_1\boldsymbol{u}_{t+1}$ where the value for $\boldsymbol{y}_{t+1}$ is substituted by its forecast at that particular period $\widehat{\boldsymbol{y}}_t(1)$.

In general, the forecast for h steps ahead is

$\widehat{\boldsymbol{y}}_t(h) = \boldsymbol{c}_t + \sum_{i=1}^{p}\boldsymbol{\Phi}_i\,\widehat{\boldsymbol{y}}_t(h-i)$.

The VMA representation can also be used to calculate forecasts (Enders, 2004). If this is the case, then the forecast for h steps ahead is

$\widehat{\boldsymbol{y}}_t(h) = \boldsymbol{\mu} + \sum_{i=h}^{\infty}\boldsymbol{\psi}_i\boldsymbol{u}_{t+h-i}$. (2.70)

This has a forecast error, $\boldsymbol{u}_t(h) = \boldsymbol{y}_{t+h} - \widehat{\boldsymbol{y}}_t(h) = \boldsymbol{u}_{t+h} + \boldsymbol{\psi}_1\boldsymbol{u}_{t+h-1} + \cdots + \boldsymbol{\psi}_{h-1}\boldsymbol{u}_{t+1}$. (2.71)

The left hand side of the equation (2.71) shows the difference between the observed values in the vector of dependent variables $\boldsymbol{y}_{t+h}$ from the predicted values of a VAR while the right hand side shows the vector moving average representation of these forecasting errors from the current period h back to the time period 1 (Brandt & Williams, 2007).

2.8 Forecast Error Variance Decomposition

There is a method, derived from the use of forecast errors that can be used to interpret the interrelated changes in a VAR model which is known as the forecast error variance decomposition method or innovation accounting. In this method the amount of variation in each of the other variables in the system which is due to the changes in any one of the variables over a period of time is estimated. This procedure determines what portion of the squared forecast error variance of one variable at h time periods ahead is associated with the surprise movements of each variable in the model (Kulshreshtha & Parikh, 2000). This is because the forecast of a VAR model has two components: (a) the predicted paths of the variables in the model and (b) the unexplained innovations (shocks) (Brandt & Williams, 2007). From these two components (a) and (b), it is possible to establish in each equation how much of the variance in the forecast of $\boldsymbol{y}_t$ is due to the past shocks of itself as and how much is due to past shocks of other variables (Enders, 2004). A group of variables $\boldsymbol{y}_t$ are exogenous if the shocks of all the other variables do not explain the forecast error variance of itself at all forecast horizons. On the other hand, if the shocks of the other variables present explain all of the forecast error variance of $\boldsymbol{y}_t$ then $\boldsymbol{y}_t$ is said to be an entirely endogenous group of variables (Enders, 2004). In

general it is typical for a variable to explain a large proportion of its forecast error variance at shorter time horizons and increasingly smaller proportions at longer time horizons (Brandt & Williams, 2007).

The method of forecast error variation is performed by the simplification of the variance for the forecast errors in a VAR model. Recall that the error in forecasting the variance for a group of variables $\boldsymbol{y}_t$, h periods in the future is

$$\begin{aligned}\boldsymbol{y}_{t+h} - \widehat{\boldsymbol{y}}_{t+h} &= \boldsymbol{y}_{t+h} + \boldsymbol{y}_1\boldsymbol{u}_{t\pm h-1} + \boldsymbol{\psi}_2\boldsymbol{u}_{t\pm h-2} + \cdots + \boldsymbol{\psi}_{h-1}\boldsymbol{u}_{t\pm 1} \\ &= \textstyle\sum_{h=0}^{s-1} \boldsymbol{\psi}_h\boldsymbol{u}_{t+s-h}\ .\end{aligned}$$

The moving average coefficients ($\boldsymbol{\psi}_i$) $i\ =\ 1,\ldots,h$ show how the shocks in the VAR model are functions of their past values in this vector moving average (VMA) representation.

The mean square error of the h period forecast is

$$\begin{aligned}\text{Var}(\boldsymbol{y}_{t+h} - \widehat{\boldsymbol{y}}_{t+h}) &= E[(\boldsymbol{y}_{t+h} - \widehat{\boldsymbol{y}}_{t+h})'\,(\boldsymbol{y}_{t+h} - \widehat{\boldsymbol{y}}_{t+h})] \\ &= \boldsymbol{\Sigma}_u + \boldsymbol{\psi}_1\boldsymbol{\Sigma}_u\boldsymbol{\psi}_1' + \boldsymbol{\psi}_2\Sigma_u\boldsymbol{\psi}_2' + \cdots + \boldsymbol{\psi}_{h-1}\boldsymbol{\Sigma}_u\boldsymbol{\psi}_{h-1}'. \qquad (2.72)\end{aligned}$$

As stated previously, the objective of forecast error variance decomposition is similar to that of an ANOVA as it determines how much the variance in a variable is due to its own shocks and how much of the variance is due to innovations of other variables over a period of time. In order for this to be determined, the equation (2.72) is required to be orthogonalised in order to standardize the variance of the shocks so that the relationships amongst the forecast innovations are observed (Brandt & Williams, 2007). This identifies which linear combinations of the forecast innovations are related to each other.

The orthogonal innovations of the forecast are written in the following form (Hamilton, 1994)

$$\boldsymbol{u}_t = e_{1t}a_1 + e_{2t}a_2 + \cdots + e_{Kt}a_K \qquad (2.73)$$

where $a_i\ \ i = 1,\ldots,K$ is the ith column of the covariance matrix of decomposition of residuals.

Now since the e_{it}'s are not correlated, if (2.73) is multiplied by its transpose and if expectations are taken, $\boldsymbol{\Sigma}_u = E(\ \boldsymbol{u}_t\boldsymbol{u}_t'\ )$ can be written in the following form

$$\boldsymbol{\Sigma}_u = a_1a_1'var(e_{1t}) + a_2a_2'var(e_{2t}) + \cdots + a_Ka_K'var(e_{Kt})\ . \qquad (2.74)$$

From (2.74) the mean square error forecast which is orthogonalised is obtained which is written as the sum of the K terms

$$\text{MSE} = \textstyle\sum_{j=1}^{K} var\big(u_{jt}\big)\,[a_ja_j' + \boldsymbol{\psi}_1a_ja_j'\boldsymbol{\psi}_1' + \boldsymbol{\psi}_2a_ja_j'\boldsymbol{\psi}_2' + \cdots + \boldsymbol{\psi}_{h-1}a_ja_j'\boldsymbol{\psi}_{h-1}'\ ]\ . \qquad (2.75)$$

This equation (2.75) shows how much of the jth orthogonalised innovation of the mean square error contributes is due to the h time period ahead forecasts.

In general, the decomposition of the forecast error variance is usually found in a tabular form that indicates the percentage of a variable's forecast error variance that can be attributed to itself as well as the percentage that is attributed to other variables in the model (Brandt & Williams, 2007). This is a limitation as the percentages mean that the forecast error variance cannot always be measured to the scale of the initial variable.

In conclusion, the forecast error variance decomposition procedure is important because it demonstrates how the changes in the variables impact on each other.

2.9 Impulse Response Analysis

2.9.1 Introduction

A shock to the ith variable affects not only itself but also the other endogenous variables in the system due to the lag structure of the VAR. Impulse responses demonstrate the response of present and future values of each of the variables to a one unit increase in the present value of one of the VAR error terms (Stock & Watson, 2001) i.e. they measure the time profile of the effects of shocks at a given point in time based on the expected future values of variables (Kulshreshtha & Parikh, 2000). In order to measure the effects of a shock where the VAR process is stationary, it is necessary to rewrite the model in an infinite vector moving average form as

$$\boldsymbol{y}_t = \boldsymbol{u}_t + \boldsymbol{\psi}_1 \boldsymbol{u}_{t-1} + \boldsymbol{\psi}_2 \boldsymbol{u}_{t-2} + \cdots . \tag{2.76}$$

The values of $\boldsymbol{\psi}_i$ $i = 1,2, \ldots$ are obtained from equating the coefficients of L^i in the equation

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2 - \cdots - \boldsymbol{\Phi}_p L^p)(\boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 L + \boldsymbol{\psi}_2 L^2 + \cdots) = \boldsymbol{I} .$$

The weight of $\boldsymbol{\psi}_i$ measures the impact of the past shock of $\boldsymbol{u}_{t-i}$ on $\boldsymbol{y}_t$ and is known as the impulse response function of $\boldsymbol{y}_t$. If the innovations $\boldsymbol{u}_t$ are uncorrelated, then the impulse response function is easy to interpret as the ith innovation is simply a shock to the ith endogenous variable.

If the innovations $\boldsymbol{u}_t$ are correlated, then a transformation needs to be made in order for them to be uncorrelated (Brandt & Williams, 2007). Suppose there is a lower triangular matrix $\boldsymbol{H}$ in which the diagonal elements are 0 and $\boldsymbol{\Sigma}_u = \boldsymbol{HGH}'$ where $\boldsymbol{G}$ is a diagonal matrix, then as noted by Tsay (2005),

$$\boldsymbol{y}_t = \boldsymbol{HH}^{-1}\boldsymbol{u}_t + \boldsymbol{\psi}_1 \boldsymbol{HH}^{-1}\boldsymbol{u}_{t-1} + \boldsymbol{\psi}_2 \boldsymbol{HH}^{-1}\boldsymbol{u}_{t-2} + \cdots$$

$$= \boldsymbol{H}\,\boldsymbol{u}_t{}^* + \boldsymbol{\psi}_1\boldsymbol{H}\boldsymbol{u}_{t-1}{}^* + \boldsymbol{\psi}_2\boldsymbol{H}\boldsymbol{u}_{t-2}{}^* + \cdots$$
$$= \boldsymbol{\psi}_0{}^*\boldsymbol{u}_t{}^* + \boldsymbol{\psi}_1{}^*\boldsymbol{u}_{t-1}{}^* + \boldsymbol{\psi}_2{}^*\boldsymbol{u}_{t-2}{}^* + \cdots \quad (2.77)$$

where $\boldsymbol{u}_t{}^* = \boldsymbol{H}^{-1}\boldsymbol{u}_t$ and $\boldsymbol{\psi}_0{}^* = \boldsymbol{H}$ and $\boldsymbol{\psi}_i{}^* = \boldsymbol{\psi}_i\boldsymbol{H}$.

The transformed matrices $\boldsymbol{\psi}_i{}^*$are the impulse response functions of $\boldsymbol{u}_t{}^*$.

2.9.2 Error Bands for Impulse Responses

As stated above, impulse responses play an important role with regards to the impact that shocks have on variables (Pesavento & Rossi, 2006). Since moving averages are generally used to describe shocks for stationary autoregressive processes, it is expected that these shocks die off and return towards 0.

It is thus important to know from the moving average representation of a VAR whether the reaction of equation i to a shock in variable j is significantly distant from 0. If the confidence intervals of the responses are significantly distant from 0, then the shocks are said to be statistically significant which means that they have a large impact. If the confidence intervals for the responses include 0 then the time horizon of responses is not statistically significant from 0 and the shocks do not have a major impact (Brandt & Williams, 2007).

The estimation of confidence intervals requires the calculation of the variances of the impulse responses which can pose a significant challenge. This is because the derivation of these variances may be based on prior assumptions of the data. An additional difficulty relating to the construction of confidence intervals is that impulse responses also have a high dimensional parameter space. This implies that the same impulse response functions can be derived from two completely different $\boldsymbol{\Phi}(L)$ operators. Suppose $c_{ij}(t)$ denotes the response of variable i to a one time shock in variable j. The converge probability of a stochastic (random) set of the $c_{ij}(t)$'s is not dependent on only $c_{ij}(t)$ itself and is different for the various $\boldsymbol{\Phi}(L)$ operators which correspond to $c_{ij}(t)$. Thus it is impossible to obtain an exact confidence set for $c_{ij}(t)$. In practice confidence intervals can be justified by the use of asymptotic theory (Brandt & Williams, 2007).

2.9.3 Methods for the Estimation of the Error Bands

There are several methods which have been proposed in order to derive the error bands/impulse responses using asymptotic theory.

a. The Bootstrap Method

This method involves taking unique random samples from the same data set. The first step is to take an initial consistent estimate say $\hat{E}$ and use Monte Carlo simulation to make random draws from a model $E\boldsymbol{\Phi} = \boldsymbol{\Phi}\hat{E}$ in order to obtain estimates of $\hat{E}$ that are conditional on the estimated

coefficients ($\widehat{\boldsymbol{\Phi}}_1$, …,$\widehat{\boldsymbol{\Phi}}_p$). The $\hat{E}$'s are mapped onto the distribution of the estimated impulse responses $c_{ij}\widehat{(t)}$ in order to obtain the 100(1− α) confidence interval which is of the form [$c_{ij,\alpha}(t)$, $c_{ij,1-\alpha}(t)$]. The lower and upper confidence limits are constructed based on the α and 1− α percentiles of the bootstrap estimate. The advantage of this method is that it accounts for the skewness of the distribution of the impulse responses. The limitation however, is that reading the α and the 1− α endpoints from a bootstrap distribution means that $c_{ij}\widehat{(t)}$ has to be assumed to be unbiased. If there is any bias present in $c_{ij}\widehat{(t)}$, it can be difficult to correct.

b. The Bootstrap Correction Method.

A modification to the original bootstrap method known as the bootstrap correction method was proposed by Kilian (1998). This method involves indirectly removing the bias in the model prior to the bootstrapping of the estimate. The biased coefficient estimates are replaced by corrected estimates before the construction of impulse responses. Kilian (1998) determined from the use of simulation techniques that the bootstrap correction method had a better coverage as compared to the standard bootstrap method. In addition, after the inclusion of a time trend, the coverage performance for the standard bootstrap method deteriorated significantly while the bootstrap correction method remained effective. The disadvantage was demonstrated from the use of an empirical example which showed that the bootstrap correction method also led to conclusions that were substantially different from those obtained from using the standard bootstrap method. The computational cost of the bootstrap correction method was also slightly higher.

c. Sims & Zha Method

A similar method in estimating the error bands was used by Sims and Zha (1999) in which an interval was constructed from

$c_{ij}\widehat{(t)} \pm \delta_{ij}(t)$.

$\delta_{ij}(t)$ are functions which define the upper and lower bands of confidence intervals at a time t in the 100$(1-\alpha)$ region. This can be depicted on a graph by plotting 3 functions , $c_{ij}\widehat{(t)} - \delta_{ij}(t)$, $c_{ij}\widehat{(t)}$ and $c_{ij}\widehat{(t)} + \delta_{ij}(t)$ onto a single set of axes which shows both the point estimate of the response as well as the range of uncertainty of the form of the response. The variance of the impulse responses is assumed to be not correlated over a period of time. This results in the variance of the future responses being independent of the past responses with the exception of the variance of the autoregressive parameters.

d. The Simulation Method

The simulation method used by Enders (2004) is advantageous in that it does not take into account any assumptions regarding the distribution of the autoregressive coefficients. The method proceeds as follows.

Consider the univariate AR (p) process

$y_{1,t} = c_1 + \varphi_1 y_{1,t-1} + \cdots + \varphi_p y_{t-p} + u_{1,t}$.

The coefficients $c_1, \varphi_1, \dots, \varphi_p$ are estimated from the use of ordinary least squares estimation and are denoted by $\hat{c}_1, \hat{\varphi}_1, \dots, \hat{\varphi}_p$ while the estimated residuals are saved and denoted by $\hat{u}_{1,t}$.

In a sample of size T, T different numbers are drawn in order to obtain a sequence say $\{u_t\}$. The result of this is a simulated series of length ${u_t}^s$ which has the same properties as an error process. In a similar way, T different random numbers are used to simulate a y_t sequence of length ${y_t}^s$,

$${y_{1,t}}^s = \hat{c}_1 + \hat{\varphi}_1 {y_{t-1}}^s + \cdots + \hat{\varphi}_p {y_{t-p}}^s + {u_{1,t}}^s \ . \qquad (2.78)$$

The simulated series ${y_{1,t}}^s$ can be treated as an AR(p) process. The process is repeated multiple times until a large number of impulse response functions are obtained. These impulse response functions are used to construct the confidence intervals by ordering them from the smallest to the largest and by using the percentiles as the upper and lower class boundaries.

The above discussion can be extended to the bivariate VAR(1) model below

$y_{1,t} = \varphi_{11,1} y_{1,t-1} + \varphi_{12,1} y_{2,t-1} + u_{1,t}$
$y_{2,t} = \varphi_{21,1} y_{1,t-1} + \varphi_{22,1} y_{2,t-1} + u_{2,t}$.

In this model, there is a possibility that the residuals of the regression $(u_{1,t}$ and $u_{2,t})$ are correlated and as a result, they need to be drawn so that the error structure is maintained. In order for this to be done, the correlation coefficient is required to be taken into account when the random numbers are drawn to construct the ${u_{1t}}^s$ and ${u_{2t}}^s$ sequences.

2.9.4 Limitations of Impulse Responses

The interpretation of moving average representations as well as the error bands for impulse responses should be treated cautiously. This is because the width of the confidence interval has the potential to increase dramatically as there are a finite number of parameters estimated in the VAR representation. This in turn, affects the accuracy and precision of the impulse responses (Brandt & Williams, 2007). In addition, the impulse responses may show signs of model instability in small samples which can result in extremely large error bands.

2.10 Conclusion

The VAR(p) model is the most widely used model for the modelling of multivariate time series. This is largely due to the fact that it is easy to specify because only one lag order needs to be chosen. The estimation of the parameters is also simpler as the method of least squares estimation can be used in addition to that of maximum likelihood estimation. The VAR(p) model also produces reliable forecasts.

CHAPTER 3

The Vector Moving Average(q) (VMA(q)) Model

3.1 Introduction

The VAR process discussed earlier is useful in understanding the associations between different variables. However, it cannot be used to establish the relationship between a group of variables $\boldsymbol{y}_t$ and their shocks/innovations at different time periods. A finite multivariate time series model which only takes into account the relationship between $\boldsymbol{y}_t$ and its various shocks at q time lags is known as the vector moving average model of order q.

3.2 Model Dynamics

The finite VMA(q) model has the representation,
$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{u}_t + \boldsymbol{\Theta}_1\boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t-q} = \boldsymbol{c}_t + \boldsymbol{\Theta}(L)\boldsymbol{u}_t \quad t = 1, \ldots, q$ where
$\boldsymbol{\Theta}(L) = \boldsymbol{I} + \boldsymbol{\Theta}_1 L + \boldsymbol{\Theta}_2 L^2 + \cdots + \boldsymbol{\Theta}_q L^q$. $\boldsymbol{u}_t$ is a the error vector with zero mean and covariance matrix $E(\boldsymbol{u}_t \boldsymbol{u}_t{}') = \boldsymbol{\Sigma}_u$.

The VMA(q) process is always stationary as the autocorrelations and the mean vector are independent of time. Thus, the constant vector $\boldsymbol{c}_t$ generally refers to the mean vector of $\boldsymbol{y}_t$ for a VMA(q) model (Tsay, 2005). The VMA(q) process is invertible if the roots of the determinant of the polynomial $\boldsymbol{\Theta}(L)$ lie outside the unit circle. An invertible VMA process can be written in the VAR representation as

$\boldsymbol{\Pi}(L)\boldsymbol{y}_t = \boldsymbol{u}_t$.

The operator $\boldsymbol{\Pi}(L)$ is defined as

$$\boldsymbol{\Pi}(L) = [\boldsymbol{\Theta}(L)]^{-1} = \frac{1}{|\boldsymbol{\Theta}(L)|}\boldsymbol{\Theta}^{+}(L) = \sum_{i=0}^{\infty}\boldsymbol{\Pi}_i L^i \ . \tag{3.1}$$

The elements of the adjoint matrix $\boldsymbol{\Theta}^{+}(L)$ are polynomials in L and have a maximum order of $(K - 1)q$ (Wei, 2006). In order to obtain the coefficients $\boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2, \ldots$ note that

$$\boldsymbol{\Pi}(L) = \frac{1}{\boldsymbol{\Theta}(L)} \tag{3.2}$$

Multiplying (3.2) throughout by $\boldsymbol{\Theta}(L)$,

$$\begin{aligned} &\boldsymbol{\Pi}(L)\,\boldsymbol{\Theta}(L) = \boldsymbol{I} \\ &(\boldsymbol{\Pi}_0 - \boldsymbol{\Pi}_1 L - \boldsymbol{\Pi}_2 L^2 - \cdots)(\boldsymbol{I} + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q) = \boldsymbol{I} \\ &\boldsymbol{\Pi}_0 + (\boldsymbol{\Theta}_1 - \boldsymbol{\Pi}_1)L + (\boldsymbol{\Theta}_2 - \boldsymbol{\Pi}_2 - \boldsymbol{\Theta}_1\boldsymbol{\Pi}_1)L^2 + \cdots = \boldsymbol{I}\,. \end{aligned} \tag{3.3}$$

The $\boldsymbol{\Pi}_i$ weights are calculated by comparing the lag coefficients in (3.3)

$$\begin{aligned} &L: && \boldsymbol{\Theta}_1 - \boldsymbol{\Pi}_1 = 0 && \boldsymbol{\Pi}_1 = \boldsymbol{\Theta}_1 \\ &L^2: && \boldsymbol{\Theta}_2 - \boldsymbol{\Pi}_2 - \boldsymbol{\Theta}_1\boldsymbol{\Pi}_1 = 0 && \boldsymbol{\Pi}_2 = \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1\boldsymbol{\Pi}_1 \\ &L^3: && \boldsymbol{\Theta}_3 - \boldsymbol{\Pi}_3 - \boldsymbol{\Theta}_1\boldsymbol{\Pi}_2 - \boldsymbol{\Theta}_2\boldsymbol{\Pi}_1 = 0 && \boldsymbol{\Pi}_3 = \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1\boldsymbol{\Pi}_2 - \boldsymbol{\Theta}_2\boldsymbol{\Pi}_1 \end{aligned} \tag{3.4}$$

In general for L^j: $\boldsymbol{\Pi}_j = \boldsymbol{\Theta}_j - \boldsymbol{\Theta}_1\boldsymbol{\Pi}_{j-1} - \cdots - \boldsymbol{\Theta}_{j-1}\boldsymbol{\Pi}_1$ with $\boldsymbol{\Pi}_j = 0$ for $j > q$.

Du Frutos and Serrano (1997) noted that if the VMA(q) model is invertible, then $\boldsymbol{\Pi}_j$ will tend towards 0 as $j \to \infty$ and thus a long but finite VAR(h_T) model is a good approximation for the model (3.1). The choice of h_T is dependent on the data and is usually set between log T and $\sqrt{T}$.

The moving average coefficients can be explained by considering the following VAR(1) process without an intercept term,

$$\boldsymbol{y}_t = (\boldsymbol{I} + \boldsymbol{\Theta}_1 L)\,\boldsymbol{u}_t\,.$$

Suppose $\boldsymbol{y}_t$ represents a bivariate series, $K = 2$,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + \begin{bmatrix} \theta_{11,1} & \theta_{12,1} \\ \theta_{21,1} & \theta_{22,1} \end{bmatrix} \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix}. \tag{3.5}$$

This model (3.5) is known as the finite memory model and depends only on its current and past shocks. The $\theta_{12,1}$ coefficient shows the linear dependence of $y_{1,t}$ on $u_{2,t-1}$ in the presence of $u_{1,t-1}$. If $\theta_{12,1} = 0$, then $y_{1,t}$ will be independent of the lagged values of $u_{2,t}$. On the contrary if $\theta_{21,1} = 0$, then $y_{2,t}$ will be independent of the lagged values of $u_{1,t}$. A unidirectional relationship occurs if either $\theta_{12,1} = 0$ and $\theta_{21,1} \neq 0$ or if $\theta_{21,1} = 0$ and $\theta_{12,1} \neq 0$ and a feedback relationship occurs if both values of $\theta_{12,1}$ and $\theta_{21,1}$ are not equal to zero (Tsay, 2005).

3.3 Autocovariances and Autocorrelations

Since $\boldsymbol{u}_t$ has no serial correlations, it follows that $\text{Cov}(\boldsymbol{y}_t, \boldsymbol{u}_t) = \boldsymbol{\Sigma}_u$ (Tsay, 2005). This can be used in the explanation of the autocovariances for the VMA(1) model,

$\boldsymbol{y}_t = \boldsymbol{\mu} + (\boldsymbol{I} + \boldsymbol{\Theta}_1 L)\, \boldsymbol{u}_t$.

$$\begin{aligned}\boldsymbol{\Gamma}(0) &= E[(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})'] \\ &= E[(\boldsymbol{I} + \boldsymbol{\Theta}_1 L)\, \boldsymbol{u}_t]\,[(\boldsymbol{I} + \boldsymbol{\Theta}_1 L)\, \boldsymbol{u}_t]' \\ &= E(\boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1})(\boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1})' \\ &= E(\boldsymbol{u}_t \boldsymbol{u}_t') + \boldsymbol{\Theta}_1' E(\boldsymbol{u}_t \boldsymbol{u}_{t-1}') + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_t' \boldsymbol{u}_{t-1}) + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_{t-1} \boldsymbol{u}_{t-1}')\boldsymbol{\Theta}_1' \end{aligned} \quad (3.6)$$

Since $\boldsymbol{u}_t$ is white noise, $E(\boldsymbol{u}_t \boldsymbol{u}_{t-1}')$ and $E(\boldsymbol{u}_t' \boldsymbol{u}_{t-1}) = 0$.
Equation (3.6) is thus simplified to $\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma}_u + \boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1'$.

In general the covariance for time interval h is

$$\begin{aligned}\boldsymbol{\Gamma}(h) &= E(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})' \\ &= E(\boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1})(\boldsymbol{u}_{t-h}' + \boldsymbol{\Theta}_1' \boldsymbol{u}_{t-h-1}') \\ &= E(\boldsymbol{u}_t \boldsymbol{u}_{t-h}') + \boldsymbol{\Theta}_1' E(\boldsymbol{u}_t \boldsymbol{u}_{t-h-1}') + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_{t-1} \boldsymbol{u}_{t-h}') + \boldsymbol{\Theta}_1' E(\boldsymbol{u}_{t-1} \boldsymbol{u}_{t-h-1}')\boldsymbol{\Theta}_1 .\end{aligned}$$

Since $E(\boldsymbol{u}_t \boldsymbol{u}_t') = \boldsymbol{\Sigma}_u$ and $E(\boldsymbol{u}_t \boldsymbol{u}_{t-h}') = 0$ for $h > 0$

$$\boldsymbol{\Gamma}(h) = \begin{cases} \boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_u & h = 1 \\ 0 & h > 1 \end{cases} \quad (3.7)$$

The method of obtaining the covariance for the VMA(1) process can be extended to that of the VMA(q) process. The autocovariance for a VMA(q) process is

$$\begin{aligned}\boldsymbol{\Gamma}(0) &= E(\boldsymbol{u}_t - \boldsymbol{\mu})(\boldsymbol{y}_t - \boldsymbol{\mu})' \\ &= E(\boldsymbol{I} + \boldsymbol{\Theta}(L)\, \boldsymbol{u}_t)\,(\boldsymbol{I} + \boldsymbol{\Theta}(L)\, \boldsymbol{u}_t)' \\ &= E(\boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q})\,(\boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q})' \\ &= E(\boldsymbol{u}_t \boldsymbol{u}_t') + \boldsymbol{\Theta}_1\, E(\boldsymbol{u}_{t-1} \boldsymbol{u}_{t-1}')\boldsymbol{\Theta}_1' + \cdots + \boldsymbol{\Theta}_q E(\boldsymbol{u}_{t-q} \boldsymbol{u}_{t-q}')\boldsymbol{\Theta}_q' .\end{aligned}$$

Since $\boldsymbol{u}_t$ is white noise, $E(\boldsymbol{u}_t' \boldsymbol{u}_{t-h}) = 0$ for $h > 0$.

Thus $\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma}_u + \boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1' + \cdots + \boldsymbol{\Theta}_q \boldsymbol{\Sigma}_u \boldsymbol{\Theta}_q'$.

The covariance time interval h is

$$\begin{aligned}\boldsymbol{\Gamma}(h) &= E(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})' \\ &= E(\boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q})(\boldsymbol{u}_{t-h} + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-h-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-h-q})' \\ &= \boldsymbol{\Theta}_h E(\boldsymbol{u}_{t-h} \boldsymbol{u}_{t-h}') + \boldsymbol{\Theta}_{h+1} E(\boldsymbol{u}_{t-h-1} \boldsymbol{u}_{t-h-1}')\boldsymbol{\Theta}_1' + \cdots + \boldsymbol{\Theta}_q E\big(\boldsymbol{u}_{t-q} \boldsymbol{u}_{t-q}'\big)\boldsymbol{\Theta}_{q-1}'.\end{aligned} \quad (3.8)$$

Since $E(\boldsymbol{u}_t\boldsymbol{u}_t{}')=\boldsymbol{\Sigma}_u$ and $E(\boldsymbol{u}_t\boldsymbol{u}_{t-h}{}')=0$ for $h>0$, $\boldsymbol{\Gamma}(h)$ is simplified to

$$\boldsymbol{\Gamma}(h)=\begin{cases}\sum_{i=0}^{q-h}\boldsymbol{\Theta}_{i+h}{}'\boldsymbol{\Sigma}_u\boldsymbol{\Theta}_i & h=0,1,\dots q\\ 0 & h>q\end{cases}. \tag{3.9}$$

The autocorrelation function for the VMA(q) process is similar to that of a VAR(p) with the exception of the covariance matrix obtained in (3.9).

Thus, as for the VAR(p) model, the autocorrelations, $\boldsymbol{\rho}(h)\ \ h=0,1,\dots,q$ are obtained from the relation

$$\boldsymbol{\rho}(h)=\boldsymbol{D}^{-1}\,\boldsymbol{\Gamma}(h)\,\boldsymbol{D}^{-1}.$$

$\boldsymbol{D}=\begin{pmatrix}\gamma_{11}(0) & 0 & \cdots & 0\\ 0 & \gamma_{22}(0) & & 0\\ \vdots & & \ddots & \vdots\\ 0 & 0 & \cdots & \gamma_{KK}(0)\end{pmatrix}$ is a diagonal matrix with covariances on the diagonal.

The cross correlation between two different series is $y_{i,t}$ and $y_{j,t-h}$ is

$$r_{ij}(h)=\frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}$$

3.4 Identification of a VMA(q) Model

The VMA model has not been studied as extensively as that of the VAR or VARMA models and orders of q are usually derived as a special case of a VARMA$(0,q)$ model. It should be emphasised that for $h>q$, $\boldsymbol{\Gamma}(h)=0$, and as a result the cross correlation matrices $r_{ij}(h)$ are 0 for $h>q$, i.e. they cut off after lag q. Wei (1990) noted that the pattern of these cross correlation matrices for higher dimensional processes can be difficult to detect since the matrices are if this is the case.

3.5 Estimation of a VMA(q) Model

Maximum likelihood estimation is the method that is used most often to estimate the parameters of the VMA model. This is because of its asymptotic efficiency in a correctly specified model (Galbraith, Ullah & Zinde-Walsh, 2002). Since the process is assumed to be stationary, the problem of an intercept term can be solved by subtracting the sample mean prior to estimation (Lütkepohl & Claessen, 1997).

The maximum likelihood function for the simple zero mean and invertible VMA(1) process $\boldsymbol{y}_t = \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1}$ where $\boldsymbol{\Theta}_0 = \boldsymbol{I}$ is derived first based on the method used by Lütkepohl (2005).

Defining $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} = \begin{pmatrix} \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & 0 & \cdots & & 0 \\ 0 & \boldsymbol{\Theta}_1 & & & & \\ \vdots & & \ddots & & & \vdots \\ 0 & & \cdots & & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{pmatrix} \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$

$$= \widetilde{\boldsymbol{\Theta}_1}\, \mathbf{u}\,.$$

$\widetilde{\boldsymbol{\Theta}_1} = \begin{bmatrix} \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & 0 & \cdots & & 0 \\ 0 & \boldsymbol{\Theta}_1 & & & & \\ \vdots & & \ddots & & & \vdots \\ 0 & & \cdots & & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{bmatrix}$ is a $(KT \times K(T+1))$ matrix.

$\mathbf{u} = \text{vec}(\mathbf{U}) = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$.

$\boldsymbol{u}_t$ is white noise and follows a normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_u$

i.e. $\begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix} \sim N(0, \boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)$.

Since $\boldsymbol{y} = \widetilde{\boldsymbol{\Theta}_1}\, \mathbf{u}$,

$E(\boldsymbol{y}) = \widetilde{\boldsymbol{\Theta}_1} E(\mathbf{u}) = \widetilde{\boldsymbol{\Theta}_1} \times 0 = 0$

and var $(\boldsymbol{y}) =$ var $(\widetilde{\boldsymbol{\Theta}_1}\, \mathbf{u}) = \widetilde{\boldsymbol{\Theta}_1}\, (\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u) \widetilde{\boldsymbol{\Theta}_1}'$.

Thus $\boldsymbol{y} \sim N\,(0, \widetilde{\boldsymbol{\Theta}_1}\, (\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u) \widetilde{\boldsymbol{\Theta}_1}')$. (3.10)

From the use of the property that $\boldsymbol{y}$ is normally distributed in (3.10), the likelihood function $L(\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_u \,|\, \boldsymbol{y})$ is proportional to

$$|\widetilde{\boldsymbol{\Theta}_1}\, (\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u) \widetilde{\boldsymbol{\Theta}_1}'|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}\, \boldsymbol{y}' [\widetilde{\boldsymbol{\Theta}_1}\, (\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u) \widetilde{\boldsymbol{\Theta}_1}']^{-1} \boldsymbol{y}\right\}. \quad (3.11)$$

The maximum likelihood function is 'exact ' if none of the terms are set towards 0 and is conditional if the insignificant terms are set to zero. The conditional likelihood function is useful because it results in a substantial reduction in the number of moving average parameters which are needed to be estimated, especially for stationary processes. The conditional methods often however, lead to the model producing unreliable and inefficient parameter estimates especially

for smaller samples and for the seasonal case (Hillmer & Tiao, 1979). The exact maximum likelihood estimation procedure, despite being more computationally burdensome reduces the bias in the model significantly by providing more accurate parameter estimates particularly when some of the zero's of the determinantal polynomial $|\boldsymbol{\Theta}(L)|$ are close to the unit circle or if the eigenvalues of $\boldsymbol{\Theta}_1$ are close to one (Tiao & Tsay, 1983). Reinsel (1997) recommended that the conditional maximum likelihood procedure should be used for the estimation in the initial stages of model building and the exact maximum likelihood procedure should be used for the estimation of the latter stages. This is because if the MA(1) process is invertible, then as t tends towards infinity, the value of $\boldsymbol{\Theta}_1$ will converge towards 0.

In the case for the conditional likelihood procedure where $\boldsymbol{u}_0 = 0$, $\boldsymbol{y}$ can be partitioned as $\boldsymbol{y} = \overline{\overline{\boldsymbol{\Theta}_1}}\,\mathbf{u}$ where

$\overline{\overline{\boldsymbol{\Theta}_1}} = \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & & 0 \\ \boldsymbol{\Theta}_1 & & & & \\ & \ddots & & & \vdots \\ & \cdots & & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{bmatrix}$ is a $(KT \times KT)$ matrix where the column $\begin{bmatrix} \boldsymbol{\Theta}_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ has been

removed from $\widetilde{\boldsymbol{\Theta}_1}$.

$\mathbf{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix} \sim N(0, \boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)$ is of dimension $(KT \times 1)$.

Recall $\boldsymbol{y} = \overline{\overline{\boldsymbol{\Theta}_1}}\,\mathbf{u}$.

Thus $E(\boldsymbol{y}) = \overline{\overline{\boldsymbol{\Theta}_1}}\,E(\mathbf{u}) = \overline{\overline{\boldsymbol{\Theta}_1}} \times 0 = 0$

The variance is var$(\boldsymbol{y})$ = var$(\overline{\overline{\boldsymbol{\Theta}_1}}\,\mathbf{u}) = \overline{\overline{\boldsymbol{\Theta}_1}}\,(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)\overline{\overline{\boldsymbol{\Theta}_1}}'$.

Thus $\boldsymbol{y} \sim N\,(0, \widetilde{\boldsymbol{\Theta}_1}\,(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}')$ (3.12)

The likelihood function is calculated from (3.12) as

$|\,\overline{\overline{\boldsymbol{\Theta}_1}}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)\overline{\overline{\boldsymbol{\Theta}_1}}'|^{\frac{-1}{2}} \exp\,\{\frac{-1}{2}\,\boldsymbol{y}'\,[\,\overline{\overline{\boldsymbol{\Theta}_1}}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)\,\overline{\overline{\boldsymbol{\Theta}_1}}']^{-1}\,\boldsymbol{y}\}$

$= |\overline{\overline{\boldsymbol{\Theta}_1}}|^{\frac{-1}{2}}|\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u|^{\frac{-1}{2}}|\overline{\overline{\boldsymbol{\Theta}_1}}'|^{\frac{-1}{2}} \exp\,\{\frac{-1}{2}\,\boldsymbol{y}'\,\overline{\overline{\boldsymbol{\Theta}_1}}^{-1\prime}\big(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1}\big)\overline{\overline{\boldsymbol{\Theta}_1}}^{-1}\,\boldsymbol{y}\}$. (3.13)

The determinant of $\overline{\overline{\boldsymbol{\Theta}_1}}$ is equal to 1 as the matrix $\overline{\overline{\boldsymbol{\Theta}_1}}$ is lower triangular with identity matrices in the main diagonal. Thus (3.13) is simplified to

$|\boldsymbol{\Sigma}_u|^{\frac{-T}{2}} \exp\{\frac{-1}{2}\boldsymbol{y}'\,\overline{\overline{\boldsymbol{\Theta}_1}}^{-1\prime}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})\overline{\overline{\boldsymbol{\Theta}_1}}^{-1}\boldsymbol{y}\}$. (3.14)

The inverse $\overline{\overline{\boldsymbol{\Theta}_1}}^{-1}$ is

$$\overline{\overline{\boldsymbol{\Theta}_1}}^{-1} = \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & 0 & 0 \\ \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & 0 & 0 \\ -\boldsymbol{\Theta}_1 & \boldsymbol{I}_K & \ldots & 0 & 0 \\ -\boldsymbol{\Theta}_1{}^2 & -\boldsymbol{\Theta}_1 & & \vdots & \vdots \\ \vdots & & \ddots & & \vdots \\ -(\boldsymbol{\Theta}_1)^{T-1} & -(\boldsymbol{\Theta}_1)^{T-2} & \cdots & -\boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{bmatrix}. \qquad (3.15)$$

The zero mean, invertible VMA(1) process $\boldsymbol{y}_t = \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1}$ $t = \pm 0, \pm 1, \pm 2, \ldots$ can be written in infinite vector autoregression form by noting that $\boldsymbol{u}_t = \boldsymbol{y}_t - \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1}$. Now from the use of successive recursions

$$\begin{aligned} \boldsymbol{u}_t &= \boldsymbol{y}_t - \boldsymbol{\Theta}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-2}) \\ &= \boldsymbol{y}_t - \boldsymbol{\Theta}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Theta}_1{}^2 \boldsymbol{u}_{t-2} \\ &\;\;\vdots \\ &= \boldsymbol{y}_t - \boldsymbol{\Theta}_1 \boldsymbol{y}_{t-1} + \cdots + (-\boldsymbol{\Theta}_1)^h \boldsymbol{y}_{t-h} + (-\boldsymbol{\Theta}_1)^{h+1} \boldsymbol{u}_{t-h-1}\,. \end{aligned}$$

Lütkepohl (2005) noted that if $i \to \infty$, then the value of $(-\boldsymbol{\Theta}_1)^i$ will converge towards 0.
Thus $\boldsymbol{u}_t = \boldsymbol{y}_t + \sum_{i=1}^{\infty} (-\boldsymbol{\Theta}_1)^i \boldsymbol{y}_{t-i}$
or $\boldsymbol{y}_t = \boldsymbol{u}_t - \sum_{i=1}^{\infty} (-\boldsymbol{\Theta}_1)^i \boldsymbol{y}_{t-i}$. (3.16)

Equation (3.16) is of the form of an infinite VAR representation with $\boldsymbol{\Pi}_i = -(-\boldsymbol{\Theta}_1)^i$. Thus it follows that $\overline{\overline{\boldsymbol{\Theta}_1}}^{-1}$ can be expressed as

$$\overline{\overline{\boldsymbol{\Theta}_1}}^{-1} = \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & 0 \\ -\boldsymbol{\Pi}_1 & \boldsymbol{I}_K & & 0 \\ \vdots & & \ddots & \vdots \\ -\boldsymbol{\Pi}_{T-1} & -\boldsymbol{\Pi}_{T-2} & \cdots & \boldsymbol{I}_K \end{bmatrix}.$$

From the use of substitutions the VMA(1) process can be rewritten as

$$\boldsymbol{y}_t + \sum_{i=1}^{t-1} (-\boldsymbol{\Theta}_1)^i \boldsymbol{y}_{t-i} + (-\boldsymbol{\Theta}_1)^t \boldsymbol{u}_0 = \boldsymbol{u}_t\,. \qquad (3.17)$$

Thus following the result in (3.15) the conditional likelihood function with $\boldsymbol{u}_0 = 0$ is

$$L(\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_u|\boldsymbol{y}) = |\boldsymbol{\Sigma}_u|^{\frac{-T}{2}} \exp\{\tfrac{-1}{2}\sum_{t=1}^{T} \boldsymbol{u}_t\,'\boldsymbol{\Sigma}_u\boldsymbol{u}_t\}\,. \tag{3.18}$$

The theory that has been used for estimating the VMA(1) process can be extended to that of the VMA(q). As discussed earlier, the VMA(q) process with a zero mean is

$$\boldsymbol{y}_t = \boldsymbol{u}_t + \boldsymbol{\Theta}_1\boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t-q}$$

An initial derivation of the likelihood function was proposed by Osborn (1977) and Hillmer and Tiao (1979) in which it was noted that $\boldsymbol{u}_t$ can be expressed as a function of the starting residuals $\boldsymbol{u}_{1-q}, \ldots, \boldsymbol{u}_0$ and observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$.

By defining $\ddot{\boldsymbol{u}}_t = (\boldsymbol{u}_{1-q}\,', \ldots, \boldsymbol{u}_0\,')$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)'$, $\boldsymbol{u}_t$ can be expressed as a linear combination of $\ddot{\boldsymbol{u}}_t$ and $\boldsymbol{y}$.

$$\boldsymbol{u}_t = \mathbf{R}\,\boldsymbol{y} + \boldsymbol{S}\,\ddot{\boldsymbol{u}}_t \tag{3.19}$$

where $\mathbf{R}$ is of dimension $K(T+q) \times KT$ and $\boldsymbol{S}$ is of dimension $K(T+q) \times Kq$.

The values of $\ddot{\boldsymbol{u}}_t$ are estimated from

$$\widehat{\ddot{\boldsymbol{u}}_t} = -(\boldsymbol{S}'\boldsymbol{I}_{T+q} \otimes {\boldsymbol{\Sigma}_u}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)^{-1}\mathbf{R}\,\boldsymbol{y}\,. \tag{3.20}$$

From the use of (3.19) and (3.20), the exact likelihood function for the VMA(q) process is

$$L(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}_u|\boldsymbol{y}) = (2\pi)^{-KT/2}\left|\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u\right|^{\frac{-1}{2}} \left|\boldsymbol{S}'(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)^{-1}\boldsymbol{S}\right|^{\frac{-1}{2}}$$
$$\times \exp\{\tfrac{-1}{2}(\mathbf{R}\,\boldsymbol{y}_t + \boldsymbol{S}\,\widehat{\ddot{\boldsymbol{u}}_t})\,(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)^{-1}\,(\mathbf{R}\,\boldsymbol{y}_t + \boldsymbol{S}\,\widehat{\ddot{\boldsymbol{u}}_t})'\}. \tag{3.21}$$

Hillmer and Tiao (1979) noted that the exact likelihood function is extremely difficult to compute when the order of q increases significantly.

A similar procedure was adopted by Lütkepohl (2005). He used a partition similar to that used for the estimation of the VAR(1) model

Partitioning $\boldsymbol{y}$ as, $\boldsymbol{y} = \widetilde{\boldsymbol{\Theta}_q}\begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$

where $\widetilde{\boldsymbol{\Theta}_q} = \begin{bmatrix} \boldsymbol{\Theta}_q & \boldsymbol{\Theta}_{q-1} & \cdots & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & 0 & & 0 \\ 0 & \boldsymbol{\Theta}_q & & \boldsymbol{\Theta}_2 & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & \cdots & 0 \\ \vdots & & \ddots & \vdots & & & & \vdots \\ & & \cdots & & & & & \\ 0 & 0 & & \boldsymbol{\Theta}_q & & \ldots & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{bmatrix}$ is a $(KT \times K(T+q))$ matrix.

The error vector $\begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix} \sim N(0, \boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)$.

$$\boldsymbol{y} = \widetilde{\boldsymbol{\Theta}_q} \begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$$

$$E(\boldsymbol{y}) = E(\widetilde{\boldsymbol{\Theta}_q} \begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix})$$

$$= \widetilde{\boldsymbol{\Theta}_q} E \begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$$

$$= 0.$$

$$\text{Var}(\boldsymbol{y}) = \widetilde{\boldsymbol{\Theta}_q} E(\begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{1-q} \\ \vdots \\ \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}') \widetilde{\boldsymbol{\Theta}_q}'$$

$$= \widetilde{\boldsymbol{\Theta}_q} \left(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u\right) \widetilde{\boldsymbol{\Theta}_q}'.$$

Thus $\boldsymbol{y} \sim N(0, \widetilde{\boldsymbol{\Theta}_q} \left(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u\right) \widetilde{\boldsymbol{\Theta}_q}')$. (3.22)

Thus it follows from (3.22) that the exact likelihood function is

$$L(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}_u|\boldsymbol{y}) \propto |\widetilde{\boldsymbol{\Theta}_q}\,(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_q}'|^{\frac{-1}{2}} \exp\,\{\frac{-1}{2}\boldsymbol{y}'[\widetilde{\boldsymbol{\Theta}_q}\,(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_q}']^{-1}\boldsymbol{y}\}.$$

The conditional maximum likelihood function is obtained by regarding the values of $\boldsymbol{u}_{-q+1} = \cdots = \boldsymbol{u}_0$ as fixed numbers. Thus the conditional likelihood function is

$$L(\boldsymbol{\Theta}_1,\dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}_u|\boldsymbol{y}) \propto |\,\overline{\overline{\boldsymbol{\Theta}_q}}(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)\overline{\overline{\boldsymbol{\Theta}_q}}'|^{\frac{-1}{2}} \exp\,\{\frac{-1}{2}\boldsymbol{y}'[\,\overline{\overline{\boldsymbol{\Theta}_q}}(\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u)\overline{\overline{\boldsymbol{\Theta}_q}}']^{-1}\boldsymbol{y}\}$$

$$\overline{\overline{\boldsymbol{\Theta}_q}} = \begin{bmatrix} \boldsymbol{I}_K & 0 & \dots & \dots & \dots & 0 & 0 \\ \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & & & & 0 & 0 \\ \boldsymbol{\Theta}_2 & \boldsymbol{\Theta}_1 & & & & 0 & 0 \\ \vdots & & \ddots & \vdots & & & \\ \boldsymbol{\Theta}_q & \boldsymbol{\Theta}_{q-1} & & & & \vdots & \vdots \\ & \boldsymbol{\Theta}_q & & & & & \\ & & & & & & \\ 0 & 0 & \cdots & \boldsymbol{\Theta}_q & \dots & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{bmatrix} \text{ is a } (KT \times K(T+q)) \text{ matrix}$$

$$L(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q, \Sigma_u|\boldsymbol{y}) \propto = |\overline{\overline{\boldsymbol{\Theta}_q}}|^{\frac{-1}{2}}|\,\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u|^{\frac{-1}{2}}|\overline{\overline{\boldsymbol{\Theta}_q}}'|^{\frac{-1}{2}} \exp\,\{\frac{-1}{2}\boldsymbol{y}'\,\overline{\overline{\boldsymbol{\Theta}_q}}'^{-1}\big(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1}\big)\overline{\overline{\boldsymbol{\Theta}_q}}^{-1}\,\boldsymbol{y}\}. \tag{3.23}$$

Since $\overline{\overline{\boldsymbol{\Theta}_q}}$ is lower triangular with identity matrices on its diagonal, its determinant will be one. (3.23) can now be simplified to

$$\begin{aligned} L(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}_u|\boldsymbol{y}) &= |\boldsymbol{\Sigma}_u|^{\frac{-T}{2}} \exp\,\{\frac{-1}{2}\boldsymbol{y}'[\,\overline{\overline{\boldsymbol{\Theta}_q}}'^{-1}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})\overline{\overline{\boldsymbol{\Theta}_q}}^{-1}]\boldsymbol{y}\} \\ &= |\boldsymbol{\Sigma}_u|^{\frac{-T}{2}} \exp\left\{\frac{-1}{2}\sum_{t=1}^{T}\boldsymbol{u}_t{}'\,\boldsymbol{\Sigma}_u{}^{-1}\boldsymbol{u}_t\right\}. \end{aligned} \tag{3.24}$$

The inverse of $\overline{\overline{\boldsymbol{\Theta}_q}}$ is

$$\overline{\overline{\boldsymbol{\Theta}_q}}^{-1} = \begin{bmatrix} \boldsymbol{I}_K & 0 & \cdots & 0 \\ -\boldsymbol{\Pi}_1 & \boldsymbol{I}_K & & 0 \\ \vdots & & \ddots & \vdots \\ -\boldsymbol{\Pi}_{T-1} & -\boldsymbol{\Pi}_{T-2} & \cdots & \boldsymbol{I}_K \end{bmatrix}.$$

The $\boldsymbol{\Pi}_i$ coefficients are obtained from the VAR representation as has been explained for the VMA (1) process. Similarly, the conditional maximum likelihood approximation for VMA (q) processes is precise for large sample sizes.

There have been more recent developments with regards to estimation of the VMA model by methods other than that of maximum likelihood estimation. In particular, an estimator developed by Galbraith et al. (2002) used a VAR approximation in order to estimate the VMA

coefficient matrices. This method is more robust for the detection of misspecification than maximum likelihood estimation but it is biased for finite samples.

3.6 Diagnostic Checking of a VMA(q) Model

Testing for the adequacy of the VMA(q) model is similar to the testing for adequacy of a VAR(p) model. The residuals can be checked to see if they constitute white noise. The Portmanteau statistic $Q_h = T\sum_{i=1}^{h} tr(\hat{\boldsymbol{\Gamma}}_u(i)\,\widehat{\boldsymbol{\Gamma}_u}(0)^{-1}\hat{\boldsymbol{\Gamma}}_u(-i)\,\widehat{\boldsymbol{\Gamma}_u}(0)^{-1})$ is used in a manner similar to that of the VAR(p) model with the notable exception being that the degrees of freedom used for Q_h is now $K^2(h-q)$. As with the VAR(p) process, the Box-Ljung statistic can also be used to test the adequacy of the VMA(q) model.

3.7 Forecasting VMA(q) Processes

The forecasts for a VMA(q) model are similar to those performed for a univariate MA(q) process. It is important to note however that $E(\boldsymbol{u}_{t+j}) = \boldsymbol{u}_{t+j}$ for $j \leq 0$ and $E(\boldsymbol{u}_{t+j}) = 0$ for $j > 0$

Consider the VMA(q) process,

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1\boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t-q}\ .$$

The one step ahead forecast is

$$\begin{aligned}\hat{\boldsymbol{y}}_t(1) &= \boldsymbol{\mu} + E(\boldsymbol{u}_{t+1}) + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_t) + \boldsymbol{\Theta}_2 E(\boldsymbol{u}_{t-1}) + \cdots + \boldsymbol{\Theta}_q E(\boldsymbol{u}_{t+1-q}) \\ &= \boldsymbol{\mu} + \boldsymbol{\Theta}_1\boldsymbol{u}_t + \boldsymbol{\Theta}_2\boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t+1-q}\ .\end{aligned}$$

The two step ahead forecast is

$$\begin{aligned}\hat{\boldsymbol{y}}_t(2) &= \boldsymbol{\mu} + E(\boldsymbol{u}_{t+2}) + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_{t+1}) + \boldsymbol{\Theta}_2 E(\boldsymbol{u}_t) + \cdots + \boldsymbol{\Theta}_q E(\boldsymbol{u}_{t+2-q}) \\ &= \boldsymbol{\mu} + \boldsymbol{\Theta}_2\boldsymbol{u}_t + \boldsymbol{\Theta}_3\boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t+2-q}\ .\end{aligned}$$

In general for $h \leq q$

$$\begin{aligned}\hat{\boldsymbol{y}}_t(h) &= \boldsymbol{\mu} + E(\boldsymbol{u}_{t+h}) + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_{t+h-1}) + \boldsymbol{\Theta}_2 E(\boldsymbol{u}_{t+h-2}) + \cdots + \boldsymbol{\Theta}_q E(\boldsymbol{u}_{t+h-q}) \\ &= \boldsymbol{\mu} + \boldsymbol{\Theta}_1\boldsymbol{u}_{t+h-1} + \boldsymbol{\Theta}_2\boldsymbol{u}_{t+h-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{u}_{t+h-q}\ .\end{aligned}$$

$\hat{\boldsymbol{y}}_t(h) = 0$ for $h > q$.

3.8 Conclusion

The finite VMA(q) model is not as widely used as the VAR(p) model because it only includes lagged shocks of the variables and it is thus more difficult to determine the interrelationships amongst the variables unless the model is invertible. The estimation procedure is also more complicated than that of the VAR(p) model.

CHAPTER 4

The Vector Autoregressive Moving Average (p, q) (VARMA (p, q)) Model

4.1 Introduction

The VAR(p) model is used to determine the interdependence among two or more series however it does not take into account the effect of innovations or shocks at different time lags and neither is it very parsimonious. The VARMA(p, q) model which relates a group of variables $\boldsymbol{y}_t$ to past values of itself and that of other variables as well as past values of shocks of itself and other variables is discussed in the following section.

4.2 Model Dynamics

The VARMA model of autoregressive order p and moving average order q is represented as

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q} \,. \tag{4.1}$$

The model (4.1) applies for all t greater than an initial time origin. $\boldsymbol{y}_t$ is a $(K \times 1)$ vector, $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ are $(K \times K)$ autoregressive coefficient matrices, $\boldsymbol{c}_t$ is a $(K \times 1)$ vector of constants and $\boldsymbol{u}_t$ is a vector white noise process with $E(\boldsymbol{u}_t) = 0$ and covariance matrix $E(\boldsymbol{u}_t \boldsymbol{u}_t{}') = \boldsymbol{\Sigma}_u$.

The model (4.1) can alternatively be expressed in backshift/lag order representation as

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p)\, \boldsymbol{y}_t = \boldsymbol{c}_t + (\boldsymbol{I} + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q)\, \boldsymbol{u}_t$$
$$\text{or } \boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{\Theta}(L)\boldsymbol{u}_t \,.$$

The diagonal elements of $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ are known as the autoregressive and moving average structures in each series respectively while the off diagonal elements of these matrices refer to the causal effects between the different pairs of series and shocks.

Each element of $\boldsymbol{y}_t$ shows how a current value of a particular series is related to its own past values as well as the past values of the other series. If all the series are unrelated to each other then $\boldsymbol{\Phi}(L), \boldsymbol{\Theta}(L)$ and $\boldsymbol{\Sigma}_u$ will all be diagonal matrices and as a result, each individual series can be represented by an independent ARMA model (Montgomery & Moe, 2002).

Suppose there is bivariate series, $K = 2$. If both of the matrices $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ are upper triangular, then it is said that the previous values of the second series (y_2) have an effect on the first series (y_1), while there is no effect of the previous values of the y_1 on y_2 at the same time. If both the matrices $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ are instead lower triangular, then there is a unidirectional relationship among the series. Finally, there is a dynamic relationship among the series if the matrices $\boldsymbol{\Theta}(L)$, $\boldsymbol{\Theta}(L)$ and $\boldsymbol{\Sigma}_u$ are all populated.

The VARMA (p, q) process can be both stationary and invertible. It is stationary if all the roots of the determinantal polynomial $|\boldsymbol{\Phi}(L)|$ lie outside the unit circle and is invertible if all the roots of the determinantal polynomial $|\boldsymbol{\Theta}(L)|$ lie outside the unit circle.

If the model (4.1) is invertible, then the VARMA(p, q) process can be written as

$$\boldsymbol{\Pi}(L)\boldsymbol{y}_t = \left(\boldsymbol{I} - \boldsymbol{\Theta}_1 - \cdots - \boldsymbol{\Theta}_q\right)^{-1}\boldsymbol{c}_t + \boldsymbol{u}_t,$$

where the coefficients of $\boldsymbol{\Pi}_i$ are $(K \times K)$ matrices that are obtained from the autoregressive relation

$$\boldsymbol{I} - \sum_{i=1}^{\infty} \boldsymbol{\Pi}_i L^i = [\boldsymbol{\Theta}(L)]^{-1}\,\boldsymbol{\Phi}(L)\,. \qquad (4.2)$$

If equation (4.2) is multiplied from the left by $\boldsymbol{\Theta}(L)$, then

$$\left(\boldsymbol{I} + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q\right)\left(\boldsymbol{I} - \sum_{i=1}^{\infty} \boldsymbol{\Pi}_i L^i\right) = \boldsymbol{\Phi}(L) = \boldsymbol{I} - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p$$

with $\boldsymbol{\Theta}_i = 0$ for $i > q$.

Thus values of $\boldsymbol{\Pi}_i$ are obtained by comparing lag coefficients. In general the value of $\boldsymbol{\Pi}_i$ is calculated from

$$\boldsymbol{\Pi}_i = \boldsymbol{\Phi}_i + \boldsymbol{\Theta}_i - \sum_{j=1}^{i} \boldsymbol{\Theta}_{i-j}\,\boldsymbol{\Pi}_j \quad i = 1,2,\ldots \quad .$$

Wei (2006) noted that since $\frac{1}{|\boldsymbol{\Theta}(L)|}\boldsymbol{\Theta}^{+}(L)\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{u}_t$ ($\boldsymbol{\Theta}^{+}(L)$ is the adjoint matrix), then if the determinantal polynomial $|\boldsymbol{\Theta}(L)|$ is independent of L, the process can be represented as a finite VAR$(\ddot{p})$ model with $\ddot{p} \leq (K - 1)q$.

If the VARMA(p, q) process is stationary, it can be written as

$$\begin{aligned}\boldsymbol{y}_t &= \boldsymbol{\Phi}(1)^{-1}\boldsymbol{c}_t + \boldsymbol{\psi}(L)\,\boldsymbol{u}_t \\ &= \left(\boldsymbol{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p\right)^{-1}\boldsymbol{c}_t + \boldsymbol{\psi}(L)\,\boldsymbol{u}_t \\ &= \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\psi}_i \boldsymbol{u}_{t-i}\,.\end{aligned}$$

The $\boldsymbol{\psi}_i$ $i = 1,2,\ldots$ matrices are obtained from the moving average relation

$$[\boldsymbol{\Phi}(L)]^{-1}\,\boldsymbol{\Theta}(L) = \sum_{i=0}^{\infty} \boldsymbol{\psi}_i L^i \qquad (4.3)$$

If (4.3) is multiplied from the left by $\boldsymbol{\Phi}(L)$,

$\boldsymbol{\Theta}(L) = \boldsymbol{\Phi}(L)(\sum_{i=0}^{\infty} \boldsymbol{\psi}_i L^i)$.

The values of $\boldsymbol{\psi}_i$ are obtained by comparing lag coefficients. In general the value of $\boldsymbol{\psi}_i$ is calculated from

$\boldsymbol{\psi}_i = \boldsymbol{\Theta}_i + \sum_{j=1}^{i} \boldsymbol{\Phi}_j \boldsymbol{\psi}_{i-j}$ $i = 1,2, \ldots$ with $\boldsymbol{\Phi}_j = 0$ for $j > p$ and $\boldsymbol{\Theta}_i = 0$ for $i > q$.

The process is a VMA($\ddot{q}$) model of order at most $\ddot{q} \leq (K-1)p$ if the determinantal polynomial $|\boldsymbol{\Phi}(L)|$ is independent of L (Wei, 2006).

In a univariate time series model, if the autoregressive and moving average polynomials are non-degenerate (i.e. if $\varphi(L) \neq 1$ and $\theta(L) \neq 1$), then the orders of $[\varphi(L)]^{-1}$ and $[\theta(L)]^{-1}$ will be infinite and a finite order AR(p) process will correspond to an infinite order moving average process while a finite MA(q) process will correspond to an infinite autoregressive process (Wei, 2006). In multivariate models this is not necessarily true since the inverse $[\boldsymbol{\Phi}(L)]^{-1}$ can be written in the form of a determinant and an adjoint matrix as

$[\boldsymbol{\Phi}(L)]^{-1} = \frac{1}{|\boldsymbol{\Phi}(L)|} \boldsymbol{\Phi}^{+}(L)$.

Since the order of $\boldsymbol{\Phi}^{+}(L)$ is a finite order autoregressive matrix polynomial, then conditional on the value of $|\boldsymbol{\Phi}(L)|$ being constant, the inverse matrix $[\boldsymbol{\Phi}(L)]^{-1}$ will also have a finite order.

4.3 Autocovariance and Autocorrelations

An understanding of the autocovariances for the general VARMA (p, q) model can be gained by considering the simpler zero mean VARMA $(1,1)$ model

$$\boldsymbol{y}_t - \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} = \boldsymbol{u}_t + \boldsymbol{\Theta}_1 u_{t-1} \quad . \tag{4.4}$$

Multiplying (4.4) by $\boldsymbol{y}_{t-h}$ and taking expectations gives

$E[\boldsymbol{y}_{t-h}(\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\boldsymbol{\Phi}_1)'] = E[\boldsymbol{y}_{t-h}(\boldsymbol{u}_t + \boldsymbol{u}_{t-1}\boldsymbol{\Theta}_1)']$.

Note that $E[\boldsymbol{y}_t(\boldsymbol{u}_{t-1}\boldsymbol{\Theta}_1)']$
$= E[(\boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1})(\boldsymbol{u}_{t-1}\boldsymbol{\Theta}_1)']$
$= \boldsymbol{\Phi}_1 E(\boldsymbol{y}_{t-1}\boldsymbol{u}_{t-1}')\boldsymbol{\Theta}_1' + \boldsymbol{\Theta}_1' E(\boldsymbol{u}_t \boldsymbol{u}_{t-1}') + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_{t-1}\boldsymbol{u}_{t-1}')\boldsymbol{\Theta}_1'$. (4.5)

Since $E(\boldsymbol{u}_t \boldsymbol{u}_t') = \boldsymbol{\Sigma}_u, E(\boldsymbol{y}_t \boldsymbol{u}_t') = \boldsymbol{\Sigma}_u$ and $E(\boldsymbol{u}_t \boldsymbol{u}_{t-1}') = 0$,

$E[\boldsymbol{y}_t(\boldsymbol{u}_{t-1}\boldsymbol{\Theta}_1)'] = \boldsymbol{\Phi}_1 \boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1' + \boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1'$

$$= (\boldsymbol{\Phi}_1 - \boldsymbol{\Theta}_1)\,\boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1{}' \qquad (4.6)$$

$h = 0:\ E(\boldsymbol{y}_t \boldsymbol{y}_t{}') - E(\boldsymbol{y}_t \boldsymbol{y}_{t-1}{}') - \boldsymbol{\Phi}_1{}' = E(\boldsymbol{u}_t \boldsymbol{y}_t{}') + E(\boldsymbol{y}_t(\boldsymbol{u}_{t-1}\boldsymbol{\Theta}_1)')$

$\boldsymbol{\Gamma}(0) - \boldsymbol{\Gamma}'(1)\boldsymbol{\Theta}_1{}' = \boldsymbol{\Sigma}_u + (\boldsymbol{\Phi}_1 - \boldsymbol{\Theta}_1)\,\boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1{}'$

$h = 1:\ E(\boldsymbol{y}_t \boldsymbol{y}_{t-1}{}') - E(\boldsymbol{y}_{t-1} \boldsymbol{y}_{t-1}{}') - \boldsymbol{\Phi}_1{}' = E(\boldsymbol{y}_{t-1} \boldsymbol{u}_t{}') + \boldsymbol{\Theta}_1{}' E(\boldsymbol{y}_{t-1} \boldsymbol{u}_{t-1}{}')$

$\boldsymbol{\Gamma}(1) = \boldsymbol{\Gamma}(0)\boldsymbol{\Phi}_1{}' + \boldsymbol{\Sigma}_u \boldsymbol{\Theta}_1{}'$

$h \geq 2\ :\ \boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}(h-1)\,\boldsymbol{\Phi}_1{}'$

The autocovariances of a stationary VARMA (p, q) process are found in a similar manner. Recall that this model is of the form

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q}\ . \qquad (4.7)$$

If (4.7) is multiplied by $\boldsymbol{y}_{t-h}$ and if expectations are taken,

$E(\boldsymbol{y}_t - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})' = \boldsymbol{\Phi}_1 E(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})' + \cdots + \boldsymbol{\Phi}_p E(\boldsymbol{y}_{t-p} - \boldsymbol{\mu})(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})' + E(\boldsymbol{u}_t(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})') + \cdots + \boldsymbol{\Theta}_q E(\boldsymbol{u}_{t-q}(\boldsymbol{y}_{t-h} - \boldsymbol{\mu})')\ .$

Reinsel (1997) noted that this is equivalent to

$\boldsymbol{\Gamma}(h) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(h-j)\,\boldsymbol{\Phi}_j{}' + \mathrm{Cov}(\boldsymbol{y}_{t-h}, \boldsymbol{u}_t) + \sum_{j=1}^{q} \mathrm{Cov}\big(\boldsymbol{y}_{t-h}, \boldsymbol{u}_{t-j}\big)\,\boldsymbol{\Theta}_j{}'\ .$

From the use of the infinite moving average representation $\boldsymbol{y}_{t-h} = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\psi}_i\,\boldsymbol{u}_{t-h-i}$, it then follows that

$E((\boldsymbol{y}_{t-h} - \boldsymbol{\mu})\boldsymbol{u}_{t-i}{}') = \boldsymbol{\psi}_{j-h}\boldsymbol{\Sigma}_u\ .$

Thus the covariance for a VARMA(p, q) process is

$$\boldsymbol{\Gamma}(h) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(h-j)\boldsymbol{\Phi}_j{}' + \sum_{j=h}^{q} \boldsymbol{\psi}_{j-h}\boldsymbol{\Sigma}_u \boldsymbol{\Theta}_j{}' \qquad h = 0,1,\ldots,q\ . \qquad (4.8)$$

Since $E(\boldsymbol{u}_{t-q}\boldsymbol{y}_{t-h}{}') = 0$ for $h > q$, it follows that

$$\boldsymbol{\Gamma}(h) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(h-j)\,\boldsymbol{\Phi}_j{}' \qquad h > q\ . \qquad (4.9)$$

Using the relation that $\boldsymbol{\Gamma}(-h) = \boldsymbol{\Gamma}(h)'$, the first $p+1$ equations are used to solve for the components $\boldsymbol{\Gamma}(0), \ldots, \boldsymbol{\Gamma}(p)$. This is followed by the covariances $\boldsymbol{\Gamma}(p+1)$, $\boldsymbol{\Gamma}(p+2)$ being solved recursively (Brockwell & Davis, 1996).

The autocorrelation matrices for the VARMA(p, q) process are obtained in a similar manner as that of the VAR and VMA processes, i.e. from the representation $\boldsymbol{\rho}(h) = \boldsymbol{D}^{-1}\,\boldsymbol{\Gamma}(h)\,\boldsymbol{D}^{-1}$ where $\boldsymbol{D}$ is a diagonal matrix with square roots of the diagonal elements of $\boldsymbol{\Gamma}(0)$ on the diagonal.

4.4 Unique Representations of VARMA $(\boldsymbol{p}, \boldsymbol{q})$ Models

4.4.1 Uniqueness

The K dimensional zero mean stationary and invertible process VARMA (p, q) process is represented as

$\boldsymbol{y}_t = \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q}$ or
$\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{u}_t$. (4.10)

This standard model (4.10) is not always easy to identify because it is not unique. Box et al. (2008) noted that it is possible to have two VARMA(p, q) representations with different autoregressive and moving average operators which result in the same infinite moving average representation of $\boldsymbol{y}_t$, eg, the operator $\boldsymbol{\psi}(L) = \sum_{i=0}^{\infty} \boldsymbol{\psi}_i L^i$ can be expressed as $\boldsymbol{\psi}(L) = \boldsymbol{\Phi}(L)^{-1}\boldsymbol{\Theta}(L) = \boldsymbol{\Phi}^*(L)^{-1}\,\boldsymbol{\Theta}^*(L)$ where $\boldsymbol{\Phi}^*(L)$ is an autoregressive operator and $\boldsymbol{\Theta}^*(L)$ is a moving average operator.

In order to further illustrate this property consider the example of the VMA(1) process given by Granger and Newbold (1986) where

$y_{1,t} = u_{1,t} + \theta_1 u_{2,t-1}$ (4.11)
$y_{2,t} = u_{2,t}$.

This process (4.11) can be written in an equivalent VAR(1) form as

$y_{1,t} - \theta_1 y_{2,t-1} = u_{1,t}$
$y_{2,t} = u_{2,t}$.

Thus, in order to ensure that the VARMA (p, q) representation is unique, there needs to be restrictions placed on the AR and MA operators. Dufour and Pelletier (2008) noted that for a given $\boldsymbol{\psi}(L)$, there should be only one set of operators $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ which generates the same moving average representation.

A more generalised model can be considered in practice, with the attachment of the non-identity coefficient matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Theta}_0$ to $\boldsymbol{y}_t$ and $\boldsymbol{u}_t$. Lütkepohl (2005) noted that this generalised VARMA (p, q) model with the matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Theta}_0$ attached has the following representation.

$\boldsymbol{\Phi}_0 \boldsymbol{y}_t = \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\Theta}_0 \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q}$. (4.12)

This representation measures the instantaneous effects of some of the variables and assists in the development of unique structures for VARMA (p, q) models. (4.10) can be rewritten in the standard VARMA (p, q) representation as

$$\left(\boldsymbol{\Phi}_0 - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p\right)\boldsymbol{y}_t = (\boldsymbol{\Theta}_0 + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q)\, \boldsymbol{u}_t$$
$$\boldsymbol{y}_t = (\boldsymbol{\Phi}_0 - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p)^{-1} (\boldsymbol{\Theta}_0 + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q)\, \boldsymbol{u}_t \,. \quad (4.13)$$

4.4.2 Unimodular and Left Coprime

It is difficult to obtain a uniquely parameterised representation of a VARMA (p, q) if the operator has a finite order inverse because the multiplication of one operator may cancel out the effects of another operator even if the finite order of the process is maintained. Suppose $\boldsymbol{M}(L)$ is an operator which is a common factor of both $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ but which does not change the structure of the process if cancelled out. If the operator $\boldsymbol{M}(L)$ has a finite order inverse, then its determinant will be a nonzero constant and $\boldsymbol{M}(L)$ will then be known as a unimodular operator (Lütkepohl, 2005).

The operators $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ are defined as left coprime if they have a representation which has no common factors in the autoregressive and moving average parts except for the unimodular operators. This property results in no further cancellation. This concept can be further explained that if the matrices $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ are left coprime, then once the value of $\boldsymbol{\Phi}(L)^{-1}\boldsymbol{\Theta}(L)$ is computed, the elements of $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ do cancel each other out. The representation $\boldsymbol{\Phi}(L)^{-1}\boldsymbol{\Theta}(L)$ will then be said to be irreducible (Reinsel, 1997). If the parameters $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Theta}_i$ are uniquely determined from the matrices $\boldsymbol{\psi}_i$ of the operator $\boldsymbol{\psi}(L)$ in the infinite moving average representation, then the model is said to be identifiable (Reinsel, 1997).

Two representations of a process $\boldsymbol{y}_t$ are known as equivalent if they give rise to the same covariance matrix structure of a process (Reinsel, 1997). There are two forms of VARMA modelling which give rise to equivalent representations. These forms impose restrictions which are needed for uniqueness and are known as the final equations form and the echelon form.

4.4.3 Final Equations Form Representation

Under the assumption of a stationary, invertible VARMA(p, q) process $\boldsymbol{\Phi}(L)\, \boldsymbol{y}_t = \boldsymbol{\Theta}(L)\, u_t$ with a zero mean, operators $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ that are left coprime and $\boldsymbol{\Sigma}_u$ which is nonsingular ,the VARMA representation is in final autoregressive equations form if $\boldsymbol{\Theta}_0 = \boldsymbol{I}_K$ and $\varphi(L)\, \boldsymbol{I}_K = \boldsymbol{\Phi}(L)$ where $\varphi(L) = 1 - \varphi_1 L - \cdots - \varphi_p L^p$ is a one dimensional scalar operator (Lütkepohl, 2005). This model is identifiable if $\varphi_p \neq 0$.

Dufour and Pelletier (2008) have argued that this representation is not advisable as the operator $\varphi(L)\boldsymbol{y}_t$ contains lagged values of y_{1t} but not lagged values of $y_{2t}, \ldots, y_{Kt}$. The interaction between different variables is modelled through the moving average part of the equation and often results in complications. The final equations representation also generally requires more parameters than the other representations in order for the same stochastic

process to be obtained which results in the final equations representation not being very efficient (Kascha, 2010).

The VARMA(p, q) process has a final moving average representation if $\boldsymbol{\Theta}(L) = \theta(L)\boldsymbol{I}_K$ where $\theta(L) = I + \theta_1 L + \cdots + \theta_q L^q$ is a scalar polynomial. This representation has a closer resemblance to a finite order VAR than the final autoregressive equations form. The limitation of this representation is that the moving average operator is the same across all the equations which implies that the it has the potential to lead to a high moving average order which reduces the parsimony of the model (Dufour & Pelletier, 2008).

A further representation for the VARMA(p, q) model is the diagonal moving average representation where the operator $\boldsymbol{\Theta}(L)$ is

$$\boldsymbol{\Theta}(L) = \text{diag}\,[\theta_{ii}(L)] = \boldsymbol{I}_K + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q$$
$$\theta_{ii}(L) = 1 + \theta_{ii,1} L + \cdots + \theta_{ii,q_i} L^{q_i} \neq 0 \text{ and } q = max(q_i) \quad i = 1, \ldots, K. \tag{4.14}$$

This representation is easier to specify in that it is not necessary to change the off diagonal elements in the autoregressive and moving average operators. It is a natural extension of a VAR model and results in a more parsimonious representation (Dufour & Pelletier, 2008). For a stationary VARMA process of diagonal moving average representation, the polynomial operators $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ are uniquely identified if the matrices $\boldsymbol{\Phi}(z)$ and $\boldsymbol{\Theta}(z)$ are of the form

$$\boldsymbol{\Phi}(z) = \boldsymbol{I} - \boldsymbol{\Phi}_1 z - \cdots - \boldsymbol{\Phi}_p z^p \text{ and } \boldsymbol{\Theta}(z) = \boldsymbol{I} + \boldsymbol{\Theta}_1 z + \cdots + \boldsymbol{\Theta}_q z^q$$
where $\Theta(z) = \text{diag}\,[\theta_{11}(z), \ldots, \theta_{kk}(z)]$ and $\theta_{ii}(z) = 1 + \theta_{ii,1} z + \cdots + \theta_{ii,q_i} z^{q_i} \neq 0$.

In a similar manner, Dufour and Pelletier (2008) noted that the VARMA(p, q) model has a diagonal VAR representation if $\boldsymbol{\Phi}(L) = \text{diag}[\varphi_{ii}(L)] = \boldsymbol{I}_K - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p$ where
$\varphi_{ii}(L) = 1 - \varphi_{ii,1} z - \cdots - \varphi_{ii,p_i} L^{p_i} \quad p = max(p_i) \qquad i = 1, \ldots, K.$

4.4.4 Echelon Form Representation

The VARMA(p, q) representation (4.12) is in echelon form if the lag operators $\boldsymbol{\Phi}(L) = [\varphi_{ki}(L)]$ $k, i = 1, \ldots, K$ and $\boldsymbol{\Theta}(L) = |\theta_{ki}(L)|$ $k, i = 1, \ldots, K$ are left coprime and if the operators in the kth row of $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ ie $\varphi_{ki}(L)$ i $= 1, \ldots, K$ and $\theta_{kj}(L)$ $j = 1, \ldots, K$ are of degree p_k (Lütkepohl, 2005).

This form imposes the following restrictions

$$\varphi_{kk}(L) = 1 - \sum_{j=1}^{p_k} \varphi_{kk,j} L^j \qquad k = 1, \ldots, K$$

$$\varphi_{ki}(L) = -\sum_{j=p_k - p_{ki}-1}^{p_k} \varphi_{ki,j} L^j \qquad k, i = 1, \ldots, K \ (k \neq i)$$

$$\theta_{ki}(L) = \sum_{j=0}^{p_k} \theta_{ki,j} L^j \qquad k, i = 1, \ldots, K. \tag{4.15}$$

The row degrees ($p_1, \dots, p_K$) are known as the Kronecker indices and are unique for a given VARMA(p, q) process (Box et al., 2008). These Kronecker indices are the maximum row degrees of each equation in a VARMA model and are the maximum degree of both the polynomials $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ (Lütkepohl & Poskitt, 1996). The number of Kronecker indices present in the model is the same as the dimension of the process. The sum of these Kronecker indices $\sum_{k=1}^{K} p_k$ is known as the McMillan degree and is interpreted as the number of independent linear combinations of the present and past vectors $\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots$ required to optimally predict all of the future values within the ARMA structure (Box et al., 2008). The McMillan degree is also a measure of the overall complexity of the VARMA(p, q) model (Kascha, 2010).

The McMillan degree can also be computed from the matrix $\boldsymbol{H}_K$ below

$$\boldsymbol{H}_K = \begin{bmatrix} \boldsymbol{\Gamma}(1) & \boldsymbol{\Gamma}(2) & \cdots & \boldsymbol{\Gamma}(h) \\ \boldsymbol{\Gamma}(2) & \boldsymbol{\Gamma}(3) & & \boldsymbol{\Gamma}(h+1) \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Gamma}(h) & \boldsymbol{\Gamma}(h+1) & \cdots & \boldsymbol{\Gamma}(2h-1) \end{bmatrix}$$

$\boldsymbol{H}_K$ is known as the Hankel matrix and is of dimension $Kh \times Kh$. The rank of $\boldsymbol{H}_K$ is the McMillan degree (Tsay, 1989). $\boldsymbol{y}_t$ will have a finite VARMA(p, q) representation if and only if rank$(\boldsymbol{H}_K)$ is finite (Reinsel, 1997).

The values of p_{ki} in the operator $\varphi_{ki}(L)$ are calculated from the following relation

$$p_{ki} = \begin{cases} min(p_k + 1, p_i) & k \geq i \\ min(p_k, p_i) & k < i \end{cases}.$$

The term p_{ki} represents the number of free coefficients in $\varphi_{ki}(L)$ i.e. the number of coefficients still required to be estimated in each operator $\varphi_{ki}(L)$ for $k \neq i$ of $\boldsymbol{\Phi}(L)$ (Dias & Kapetanios, 2011) while p_{kk} is the number of free coefficients on the kth diagonal of $\boldsymbol{\Phi}(L)$ as well as the order of the polynomials on the corresponding row of $\boldsymbol{\Theta}(L)$ (Dufour & Jouini, 2008). The matrix that is formed by the Kronecker indices $[p_{ki}]$ $k = 1, \dots, K$ implies that there are $\sum_{k=1}^{K} \sum_{i=1}^{K} p_{ki}$ free autoregressive coefficients and $K \sum_{k=1}^{K} p_k$ free moving average coefficients while the maximum number of freely varying parameters in the model is $2K \sum_{k=1}^{K} p_k$ (Lütkepohl & Poskitt, 1996).

An echelon by definition is the certain positioning of an array in the form of steps and is the positioning of the nonzero parameters in this context. The position of the freely varying parameters is dependent when $p_i \leq p_k$. The Kronecker indices impose a number of zero restrictions on the coefficient matrices which is sufficient for the echelon form model to be unique (Lütkepohl & Claessen, 1997).

The echelon form model is complicated because the choice of p_{ki} depends on the diagonal $i = j$. The difficulties are also compounded due to the summation subscript of $j = p_k - p_{ki} - 1$ in the operator $\varphi_{ki}(L)$ (Dufour & Pelletier, 2008).

Despite the echelon form representation being more complicated, it is usually preferred to the final equations form as it involves relatively fewer free parameters and is therefore more parsimonious. The computational difficulties which occur when maximum likelihood estimation is used are greatly reduced due to the parsimony in the model (Lütkepohl, 2005).

4.5 Specification of VARMA(p, q) Models

As with VAR(p) and VMA(q) models it is of prime importance to correctly specify the values of p and q as the number of parameters increase dramatically as p and q get larger which in turn makes statistical analysis difficult. Unlike for univariate processes, there is no specific method to specify vector ARMA(p, q) models.

It is difficult to read or to detect the order of the VARMA(p, q) model from the autocorrelation function, the partial autocorrelation function and the cross correlation function especially if there are more than two time series used because there will be a large number of parameters involved (Lütkepohl & Poskitt, 1996). Grubb (1992) noted that the identification of the cut offs of these functions can be subjective and several different models might have to be fitted until a suitable model is found. It is also important to note that when trying to determine whether a VARMA(p, q) model is parsimonious, it is not always sufficient that only the minimal orders are chosen for the autoregressive and moving average parameters (Lütkepohl & Poskitt, 1996). There are various methods for specifying VARMA(p, q) models, each with its own set of advantages and disadvantages. In the following section, methods of specifying the two main forms of VARMA modelling i.e. the final equations form and the echelon form as well as an alternate method of specifying the VARMA(p, q) model known as the Scalar Component Method will be discussed.

4.5.1 Specification of the Final Equations Form

In the specification of the final autoregressive equations form, the main aim is to find the orders of p and q in the equation

$$\varphi(L)\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{u}_t \,. \tag{4.16}$$

The specification procedure described below was discussed in detail by Lütkepohl (2005). The mean μ is assumed to have been removed prior to the specification of this stage.
$\boldsymbol{y}_t = (y_{1t}, \dots, y_{Kt})$ is a K dimensional system, $\varphi(L) = I - \varphi_1 L - \cdots - \varphi_p L^p$ is a one dimensional scalar operator and $\boldsymbol{\Theta}(L) = \boldsymbol{I} + \boldsymbol{\Theta}_1 L + \cdots + \boldsymbol{\Theta}_q L^q$.

This representation (4.16), i.e. $\varphi(L)\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{u}_t$ implies that each individual component has the univariate ARIMA representation $\varphi(L)y_{kt} = \overline{\theta_k(L)}v_{kt}\ \ k = 1, \ldots, K$ where v_{kt} is univariate white noise and $\overline{\theta_k(L)}$ is an operator which has a maximum degree of q. This means that each individual component series $y_{it}, i = 1, \ldots, K$ will have the same autoregressive operator while the degree of the moving average operator will be at most q. Thus, in order for the final equations representation to be specified, it is important that each of the univariate models $\varphi_k(L)y_{kt} = \theta_k(L)v_{kt}$ are specified first. The operators $\varphi_k(L)$ and $\theta_k(L)$ are defined as

$\varphi_k(L) = I\ - \varphi_{k1}L - \cdots - \varphi_{kp_k}L^{p_k}$
$\theta_k(L) = I + \ \theta_{k1}L + \cdots + \theta_{kq_k}L^{q_k}$.

Following this, a common $\varphi(L)$ operator needs to be determined. This is performed by taking the product of all the individual AR polynomials $\varphi(L) = \varphi_1(L), \ldots, \varphi_K(L)$. This operator $\varphi(L)$ has a degree of $p = \sum_{i=1}^{K} p_i$.

The moving average operator $\overline{\theta_k(L)} = \theta_k(L)\prod_{i=1}^{K}\varphi_i(L)\ i \neq k$ which is of degree $q_k + \sum_{i=1}^{K} p_i$ is specified next. The maximum degree of all the individual operators is chosen as q which is the degree of the joint moving average operator. If there are common factors present in $\varphi_K(L)$ then the degree of $\varphi(L)$ will be less than $\sum_{i=1}^{K} p_i$ and the degree of the joint moving average operator will be less than the maximum of $(q_k + \sum_{i=1}^{K} p_i)$.

The final autoregressive equations form leads to VARMA models which contain a large number of parameters. This can result in imprecise parameter estimates (Dufour & Pelletier, 2002). Restrictions can be imposed on the autoregressive and moving average operators ($\varphi(L)$ and $\boldsymbol{\Theta}(L)$) but this will make the modelling procedure very tedious.

4.5.2 Specification of Echelon Forms

The main purpose of specifying a K dimensional echelon form representation of dimension K is to determine the K Kronecker indices and to impose further restrictions on the parameters (Lütkepohl, 2005). Reinsel (1997) noted that it is the knowledge of the Kronecker indices that is used to locate a special structure from amongst the autoregressive and moving average parameter matrices which in turn leads to specification of echelon form VARMA models. Dufour and Pelletier (2002) noted that the choice of lag orders of p and q in echelon form is significantly more complicated in the VARMA case than that of the univariate ARMA case. This is because the number of parameters is higher and the diagonal elements need to be considered when p_{ki} is chosen. The summation subscript in the operator $\varphi_{ki}(L)$ is different across the rows and columns which further complicates the representation (Dufour & Pelletier, 2008).

If the data generating process of a K dimensional multivariate time series has an echelon VARMA(p, q) representation with $p_k \leq p_{max}\ k = 1, \ldots, K$ (where p_{max} is a number chosen prior to specification), then it is possible to calculate the maximum likelihood estimates for all

of the Kronecker indices. The estimates $(\hat{p}_1, \ldots, \hat{p}_K)$ which optimise the maximum log likelihood function are chosen. This procedure requires many computations and is generally not used. However, the underlying theory is often used as the basis for other methods.

Suppose the maximum likelihood estimator of the white noise residual covariance matrix is denoted as $\widetilde{\boldsymbol{\Sigma}_u}(\hat{p}_1, \ldots, \hat{p}_K)$. A possible criterion Cr $(\hat{p}_1, \ldots, \hat{p}_K)$ which can be minimised over all sets of Kronecker indices is

$$\text{Cr}\ (\hat{p}_1, \ldots, \hat{p}_K) = \ln|\ \widetilde{\boldsymbol{\Sigma}_u}(\hat{p}_1, \ldots, \hat{p}_K)\ | + \frac{c_T v(\hat{p}_1, \ldots, \hat{p}_K)}{T}. \tag{4.17}$$

c_T is a function of size T and $v(\hat{p}_1, \ldots, \hat{p}_K)$ refers to the number of coefficient parameters which are implied by p in the echelon form (Dufour & Jouini, 2008). Since only the maximum of the log likelihood (or $\ln|\ \widetilde{\boldsymbol{\Sigma}_u}(\hat{p}_1, \ldots, \hat{p}_K)\ |$) is required, the problem of overspecification is greatly reduced. The criterion Cr $(\hat{p}_1, \ldots, \hat{p}_K)$ is a consistent estimator for a set of Kronecker indices $(\hat{p}_1, \ldots, \hat{p}_K)$ as $c_T \to 0$ and $\frac{c_T}{T} \longrightarrow 0$ when $T \to \infty$. As discussed previously for VAR processes, the BIC is a strongly consistent criterion, the HQ criterion is consistent and the AIC criterion is not consistent. A modification to the AIC criterion for echelon form VARMA models has recently been developed by Boubacar Mainassara (2010). This criterion led to an improvement in the consistency of the AIC particularly in the 'weak case' (where u_t is not assumed to be not identically and independently distributed).

The procedure using information criteria has a significant drawback in that the log likelihood estimator $\ln|\ \widetilde{\boldsymbol{\Sigma}_u}(\hat{p}_1, \ldots, \hat{p}_K)|$ needs to be maximised a fair number of times since it is nonlinear in the parameters. This means that iterative optimisation procedures which are time consuming need to be employed. Taking this into account, a 5 step procedure using linear least squares estimation was proposed for univariate models by Hannan and Rissanen (1982) and was extended to the vector case by Lütkepohl and Poskitt (1996). The procedure consists of the following steps

1. Under the assumption that the VARMA (p, q) model is stationary and invertible, the first step is to fit a VAR process of order h_T to the data and obtain the estimated residual vectors $\widehat{\boldsymbol{u}_t}(h_T)\ h_T = 1, \ldots, T$ from the relation $\widehat{\boldsymbol{u}_t}(h_T) = \boldsymbol{y}_t - \sum_{i=1}^{h_T} \boldsymbol{\pi}_i\, \boldsymbol{y}_{t-i}{}^{h_T}$. The choice of h_T can be derived from using a suitable criterion such as the AIC and should ideally be higher than the largest Kronecker index but at the same time not be exceptionally large (Hannan & Rissanen, 1982). Lütkepohl and Poskitt (1996) suggested that the value of h_T should be between $\ln(T)$ and $\sqrt{T}$. There are also criteria which have been developed for this specific purpose such as that which has recently been developed by Kascha and Trenkler (2011). In this method the value of h_T is chosen from $h_T = \max(\max\ (p_T, q_T) + 1, (\ln T)^{1.25})$ where p_T and q_T are the upper bounds for p and q respectively. The BIC criterion should be treated with caution for this stage as it is known to select a very small lag order for finite samples. These estimated residuals $\widehat{\boldsymbol{u}_t}(h_T)$ will be good estimates of the true residuals if h_T approaches infinity as $T \to \infty$ (Bartel & Lütkepohl, 1998).

2. In this stage, linear least squares estimation is used to fit echelon VARMA models under different sets of Kronecker indices $(p_1, \dots, p_K)$ where $\boldsymbol{y}_{t-i}$ and $\widehat{\boldsymbol{u}_t}(h_T)$ are used as the regressor variables (Box et al., 2008). The optimal model is selected on the basis of model criteria such as the AIC, BIC or HQ. Athanasopoulos, Poskitt and Vahid (2012) recommend from the use of Monte Carlo simulation experiments that the BIC is known to outperform the AIC for larger samples and HQ while for smaller samples, the HQ criterion generally outperforms the AIC and the BIC.

   The minimum information criterion (MINIC) is a criterion similar to the BIC for VARMA model and is defined as

   $$\text{MINIC} = \ln |\widetilde{\boldsymbol{\Sigma}_u}(p_T, q_T)| + (p_T + q_T)K^2 \frac{\ln T}{T}.$$

   $\widetilde{\boldsymbol{\Sigma}_u}(p_T, q_T)$ denotes the estimated error covariance matrix. Granger and Newbold (1986) suggested that various combinations of (p_T, q_T) should be tried and the order (p_T, q_T) which corresponds to the minimum value of the MINIC criterion is chosen as the correct order for (p, q).

   There has also been a recent method developed by Dufour and Pelletier (2008) which involves the assumption of a diagonal moving average equations form for $\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{u}_t$. This is done by regressing

   $$y_{kt} = \sum_{i=1}^{p_T} \sum_{j=1}^{K} \varphi_{kj,i}\, y_{j,t-i} + u_{kt} + \sum_{j=1}^{q_T} \theta_{kk,j}\, \hat{u}_{k,t-j}\ .$$

   The criterion used is

   $$\text{Duf-Pe} = \ln \mid \widetilde{\boldsymbol{\Sigma}_u}(p_T, q_T) \mid + \frac{K \ln(T)^{1+\nu_K}}{T}, \tag{4.18}$$

   where ν_K is an integer which is generally set to 0.2. Kascha and Trenkler (2011) used a similar criterion defined as

   $$\text{KRA-TRE} = \ln |\widetilde{\boldsymbol{\Sigma}_u}(p_T, q_T)| + \frac{K \ln(T - \max(p_T, q_T)^{1+\nu_K}}{T - \max(p_T, q_T)}\ . \tag{4.19}$$

   As with the MINIC criterion, the order (p, q) which is selected is that which corresponds to the minimum value of the criteria (4.18) and (4.19).

3. The optimum echelon form VARMA model selected in stage 2 can be re-estimated using maximum likelihood estimation in order to obtain efficient estimates of the parameters.

4. The parameters must now be tested for significance by using t ratios or by performing χ^2 tests. Zero restrictions should be placed on the coefficients if necessary in order to obtain a more parsimonious model.

5. The model should now be tested for adequacy by performing a residual analysis and checking for serial correlation.

The above procedure is very complicated and time consuming. The second step in particular is known to be very tedious especially if the dimension K of the system or the maximum Kronecker index p_K is very large. There have been various methods proposed in order to simplify this second stage. Hannan and Kavaliers (1984) proposed a method which carries the initial assumption that all of the Kronecker indices are identical. The last Kronecker index p_K is varied while the rest of the indices are fixed. The same procedure is repeated with p_{K-1} and so forth.

Dufour and Jouini (2008) suggested a modification to stage 2 which involves the formulation of a restriction matrix for all possible sets of Kronecker indices in any dimension of the VARMA process. This procedure improves accuracy while reduces over parameterisation at the same time.

4.5.3 Identification using Scalar Components

There is another method of identifying VARMA(p,q) models which is known as the scalar component method (SCM). This method was first proposed by Tiao and Tsay (1989) and was further developed by Athanasopoulos and Vahid (2006). This involves checking to observe that if there are any linear combinations of the variables which depend on fewer than p autoregressive lags and q moving average lags. These linear combinations should have consistent but not necessarily efficient estimates. The remaining parameters of the structure are then estimated conditional on the estimates of the linear combinations. The choice of p and q (the overall tentative order) is based on the zero sample canonical correlations between $(\boldsymbol{y}_t', \dots, \boldsymbol{y}_{t-g}')$ and $(\boldsymbol{y}_{t-1-j}', \dots, \boldsymbol{y}_{t-1-j-h}')$ for combinations of g and j (Athanasopoulos & Vahid, 2008). The smallest integer values for g and j, for which there are K zero canonical correlations are the ones that identify the orders of p and q (Reinsel, 1997).

For the K dimensional VARMA model described in (4.1), the linear combination $\varpi'\boldsymbol{y}_t$ is said to follow a SCM (p_1, q_1) process if

$$\begin{aligned}
&\varpi'\boldsymbol{\Phi}_{p_1} \neq 0 \text{ for } 0 \leq p_1 \leq p \\
&\varpi'\boldsymbol{\Theta}_{q_1} \neq 0 \text{ for } 0 \leq q_1 \leq q \\
&\varpi'\boldsymbol{\Phi}_{l} = 0 \text{ for } p_1 + 1 \leq l \leq \dots p \\
&\varpi'\boldsymbol{\Theta}_{l} = 0 \text{ for } q_1 + 1 \leq l \leq \dots q
\end{aligned} \tag{4.20}$$

(Tiao & Tsay, 1989) .

The random variable $\varpi' \boldsymbol{y}_t$ is dependent only on the lags 1 to p_1 of all the variables and the lags 1 to q_1 of the innovations of the system (Athanasopoulos & Vahid, 2006). The process starts with the SCM$(0,0)$ model ($p_1 = 0, q_1 = 0$) and K linear combinations of orders (p_i, q_i) $i = 1, \dots, K$ are sought in such a way that the orders $p_i + q_i$ are at a minimum (Reinsel 1997). Suppose that $\check{\mathbf{N}} = (\varpi_1, \dots, \varpi_K)'$ where $\varpi_1, \dots, \varpi_K$ are the coefficients of the various linear combinations. A consistent estimator for $\check{\mathbf{N}}$ is constructed by using the estimated canonical covariates which correspond to the canonical correlations that are not significant. The resulting model if (4.1) is rotated by $\check{\mathbf{N}}$ is

$$\begin{aligned}\check{\mathbf{N}}\boldsymbol{y}_t &= \check{\mathbf{N}}\boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \cdots + \check{\mathbf{N}}\boldsymbol{\Phi}_p\boldsymbol{y}_{t-p} + \check{\mathbf{N}}\boldsymbol{u}_t + \check{\mathbf{N}}\boldsymbol{\Theta}_1\boldsymbol{u}_{t-1} + \cdots + \check{\mathbf{N}}\boldsymbol{\Theta}_q\boldsymbol{u}_{t-q} \\ &= \breve{\boldsymbol{\Phi}}_1\boldsymbol{y}_{t-1} + \cdots + \breve{\boldsymbol{\Phi}}_p\boldsymbol{y}_{t-p} + \breve{\boldsymbol{u}}_t + \breve{\boldsymbol{\Theta}}_1\breve{\boldsymbol{u}}_{t-1} + \cdots + \breve{\boldsymbol{\Theta}}_q\breve{\boldsymbol{u}}_{t-q}, \end{aligned} \quad (4.21)$$

where $\breve{\boldsymbol{\Phi}}_i = \check{\mathbf{N}}\boldsymbol{\Phi}_i\breve{\boldsymbol{u}}_t = \check{\mathbf{N}}\boldsymbol{u}_t$ and $\breve{\boldsymbol{\Theta}}_i = \check{\mathbf{N}}\boldsymbol{\Theta}_i\check{\mathbf{N}}^{-1}$.

The transformation (4.21) implies that there are fewer estimated parameters which results in a more parsimonious model (Grubb, 1992). Tsay (1989) noted that there are still redundant parameters present which occur if the model of a SCM is embedded in another SCM. Tsay (1989) thus proposed a 'rule of elimination' where the number of redundant parameters $\mathcal{M}_i$ is calculated from

$$\mathcal{M}_i = \sum_{v=1}^{K} \max[0, (p_i - p_{v,} q_i - q_v)] \,. \quad (4.22)$$

The value of $\mathcal{M}_i$ can be used to calculate the total number of parameters needed to be estimated in the model. This is defined as

$$K\sum_{i=1}^{K}(p_i + q_i) - \sum_{i=1}^{K}(\mathcal{M}_i) + \frac{K(K+1)}{2} \,. \quad (4.23)$$

This procedure was once again extended by Athanasopoulos and Vahid (2006) who suggested that further restrictions should be placed on $\check{\mathbf{N}}$ in order for the number of free parameters to be determined. In this instance, the restrictions which account for the redundant parameters are set on the moving average rather than the autoregressive coefficients. This procedure leads to a uniquely identified VARMA representation which can be estimated by maximum likelihood estimation.

The scalar component method generally requires more computations than other methods of specification which may lead to difficulties due to the evaluation of a large number of eigenvalues (Dufour & Jouini, 2008). Lütkepohl and Poskitt (1996) noted that unlike the echelon form representation, the scalar component method can be problematic with regards to the asymptotic inference if the transformation $\check{\mathbf{N}}$ is data dependent. Athanasopoulos et al. (2012) found that scalar component models have a better forecasting performance than echelon form models. They are also more flexible because the maximum autoregressive order does not have to be the same as that of the moving average component. The echelon form also demonstrated signs of being over parameterised when compared to scalar component models. The echelon

form models are however, more practical for application purposes and as a result, further research needs to be conducted in order for more refined echelon form models to be obtained (Athanasopoulos et al., 2012).

4.6 Estimation of the VARMA$(\boldsymbol{p}, \boldsymbol{q})$ Model

The methodology of the maximum likelihood estimation procedure which was discussed for the VMA(q) model can be extended to the VARMA(p, q) model.

4.6.1 Maximum Likelihood Estimation of the VARMA$(\boldsymbol{p}, \boldsymbol{q})$ Model

The easiest case of the VARMA (p, q) representation is the stationary and invertible VARMA $(1,1)$ representation with a zero mean that is of the following form and will be considered first. The estimation procedure is based on the method used by Lütkepohl (2005).

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_1 \, \boldsymbol{y}_{t-1} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_t$$

Under the assumption of a sample of size T, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ can be represented as

$$\boldsymbol{y}_1 = \boldsymbol{\Phi}_1 \boldsymbol{y}_0 \; + \boldsymbol{u}_1 + \boldsymbol{\Theta}_1 \boldsymbol{u}_0$$
$$\boldsymbol{y}_2 = \boldsymbol{\Phi}_1 y_1 \; + \boldsymbol{u}_2 + \boldsymbol{\Theta}_1 \boldsymbol{u}_1$$

.
.
.
.

$$\boldsymbol{y}_T = \boldsymbol{\Phi}_1 \boldsymbol{y}_{T-1} \; + \boldsymbol{u}_{T-1} + \boldsymbol{\Theta}_1 \boldsymbol{u}_{T-1} \, . \qquad (4.24)$$

(4.24) can be expressed in matrix form as

$$\boldsymbol{\eta}_1 \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \widetilde{\boldsymbol{\Theta}_1} \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix},$$

where $\boldsymbol{\eta}_1 = \begin{bmatrix} \boldsymbol{I}_K & & \cdots & \cdots & 0 & 0 \\ -\boldsymbol{\Phi}_1 & \boldsymbol{I}_K & \cdots & & 0 & 0 \\ 0 & -\boldsymbol{\Phi}_1 & & & 0 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & & & 0 & 0 \\ \vdots & 0 & \ddots & \vdots & 0 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & & & \boldsymbol{I}_K & 0 \\ 0 & 0 & \cdots & 0 & -\boldsymbol{\Phi}_1 & \boldsymbol{I}_K \end{bmatrix}$ and

$\widetilde{\boldsymbol{\Theta}_1} = \begin{pmatrix} \boldsymbol{\Theta}_1 & \boldsymbol{I}_K & 0 & \cdots & & 0 \\ 0 & \boldsymbol{\Theta}_1 & & & & \\ \vdots & & \ddots & & & \vdots \\ 0 & & \cdots & & \boldsymbol{\Theta}_1 & \boldsymbol{I}_K \end{pmatrix}$ as was previously discussed for the VMA(q) process.

Now since $\begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$ is a white noise process with a zero mean and a variance $(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)$,

$$E(\boldsymbol{\eta}_1 \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}) = E(\widetilde{\boldsymbol{\Theta}_1} \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix})$$

$$\boldsymbol{\eta}_1 E \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \widetilde{\boldsymbol{\Theta}_1} \, E \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_T \end{bmatrix}$$

$$\boldsymbol{\eta}_1 E \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$E \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} = \boldsymbol{\eta}_1^{-1} \begin{bmatrix} \boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus $\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} \sim N(\boldsymbol{\eta}_1^{-1} \begin{bmatrix} \boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \boldsymbol{\eta}_1^{-1} \widetilde{\boldsymbol{\Theta}_1} \, (\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u) \widetilde{\boldsymbol{\Theta}_1}' \boldsymbol{\eta}_1'^{-1})$. (4.25)

From the use of (4.25), the exact likelihood function is of the form

$$L(\boldsymbol{\Phi}_1, \boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_u|\boldsymbol{y}, \boldsymbol{y}_0) \propto |\boldsymbol{\eta}_1{}^{-1}\widetilde{\boldsymbol{\Theta}_1}(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}'\boldsymbol{\eta}_1{}^{-1\prime}|^{\frac{-1}{2}}$$

$$\times \exp\{\frac{-1}{2}(\boldsymbol{y} - \boldsymbol{\eta}_1{}^{-1}\begin{bmatrix}-\boldsymbol{\Phi}_1\boldsymbol{y}_0\\0\\\vdots\\0\end{bmatrix})'\boldsymbol{\eta}_1'[\widetilde{\boldsymbol{\Theta}_1}(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}']^{-1}\boldsymbol{\eta}_1(\boldsymbol{y} - \boldsymbol{\eta}_1{}^{-1}\begin{bmatrix}-\boldsymbol{\Phi}_1\boldsymbol{y}_0\\0\\\vdots\\0\end{bmatrix})\}.$$

$$= |\boldsymbol{\eta}_1{}^{-1}\widetilde{\boldsymbol{\Theta}_1}(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}'\boldsymbol{\eta}_1{}^{-1\prime}|^{\frac{-1}{2}}$$

$$\times \exp\{\frac{-1}{2}(\boldsymbol{\eta}_1\boldsymbol{y} - \begin{bmatrix}-\boldsymbol{\Phi}_1\boldsymbol{y}_0\\0\\\vdots\\0\end{bmatrix})'[\widetilde{\boldsymbol{\Theta}_1}(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}']^{-1}(\boldsymbol{\eta}_1\boldsymbol{y} - \begin{bmatrix}-\boldsymbol{\Phi}_1\boldsymbol{y}_0\\0\\\vdots\\0\end{bmatrix})\} \qquad (4.26)$$

Since $|\boldsymbol{\eta}_1| = 1$,

$$L(\boldsymbol{\Phi}_1, \boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_u|\boldsymbol{y}, \boldsymbol{y}_0) \propto |\widetilde{\boldsymbol{\Theta}_1}(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}'|^{\frac{-1}{2}}$$

$$\times \exp\{\frac{-1}{2}(\boldsymbol{\eta}_1\boldsymbol{y} - \begin{bmatrix}-\boldsymbol{\Phi}_1\boldsymbol{y}_0\\0\\\vdots\\0\end{bmatrix})'[\widetilde{\boldsymbol{\Theta}_1}(\boldsymbol{I}_{T+1} \otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_1}']^{-1}(\boldsymbol{\eta}_1\boldsymbol{y} - \begin{bmatrix}-\boldsymbol{\Phi}_1\boldsymbol{y}_0\\0\\\vdots\\0\end{bmatrix})\}$$

The conditional likelihood can be found by setting $\boldsymbol{u}_0 = \boldsymbol{y}_0 = 0$. Lütkepohl (2005) noted that the likelihood function is

$$L(\boldsymbol{\Phi}_1, \boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_u) \propto |\boldsymbol{\Sigma}_u|^{\frac{-T}{2}} \exp\{\frac{-1}{2}(\boldsymbol{\eta}_1\overline{\overline{\boldsymbol{\Theta}_1}}^{-1}\boldsymbol{y})'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u{}^{-1})\overline{\overline{\boldsymbol{\Theta}_1}}^{-1}\boldsymbol{\eta}_1\boldsymbol{y}\}. \qquad (4.27)$$

Recall that a zero mean VARMA process can be represented in VAR form as

$\boldsymbol{y}_t = -\sum_{i=1}^{\infty}\boldsymbol{\Pi}_i\boldsymbol{y}_{t-i}$ where $\boldsymbol{\Pi}_i = \boldsymbol{\Phi}_i + \boldsymbol{\Theta}_i - \sum_{j=1}^{i}\boldsymbol{\Theta}_{i-j}\boldsymbol{\Pi}_j$. $\qquad i = 1,2, \ldots$

In a stationary and invertible zero mean VARMA $(1,1)$ process, the summation operator is 0, as the lower bound in this operator exceeds the upper bound

Thus $\boldsymbol{\Pi}_1 = \boldsymbol{\Phi}_1 + \boldsymbol{\Theta}_1$

$\boldsymbol{\Pi}_2 = \boldsymbol{\Phi}_2 + \boldsymbol{\Theta}_2 - \boldsymbol{\Pi}_1\boldsymbol{\Theta}_1 = -\boldsymbol{\Phi}_1\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_1{}^2$

.

.

.

$\boldsymbol{\Pi}_i = (-1)^{i-1}(\boldsymbol{\Theta}_1{}^i + \boldsymbol{\Theta}_1{}^{i-1}\boldsymbol{\Phi}_1)$.

The $\boldsymbol{\Pi}_i$ coefficients can be used to obtain $\boldsymbol{u}_t = \boldsymbol{y}_t - \sum_{i=1}^{t-1} \boldsymbol{\Pi}_i \boldsymbol{y}_{t-i}$ and thus (4.27) can alternatively be represented as

$$L(\boldsymbol{\Phi}_1, \boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_u | \boldsymbol{y}) = |\boldsymbol{\Sigma}_u|^{-T/2} \exp\{ \tfrac{-1}{2} \textstyle\sum_{t=1}^{T} \boldsymbol{u}_t{}' \boldsymbol{\Sigma}_u{}^{-1} \boldsymbol{u}_t \} .$$

The above discussion can be generalised for the zero mean stationary and invertible VARMA (p, q) case i.e.

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t + \boldsymbol{\Theta}_1 \boldsymbol{u}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{u}_{t-q} . \tag{4.28}$$

The process (4.28) can be rearranged as follows,

$$\boldsymbol{\eta}_p \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Phi}_p \boldsymbol{y}_{-p+1} \\ \vdots \\ \boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ \vdots \\ 0 \end{bmatrix} = \widetilde{\boldsymbol{\Theta}_q} \begin{pmatrix} \boldsymbol{u}_{-q+1} \\ \vdots \\ \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{pmatrix},$$

where $\boldsymbol{\eta}_p = \begin{bmatrix} \boldsymbol{I}_K & & \cdots & & & 0 & 0 \\ -\boldsymbol{\Phi}_1 & \boldsymbol{I}_K & & & & 0 & 0 \\ -\boldsymbol{\Phi}_2 & -\boldsymbol{\Phi}_1 & \ddots & & & 0 & 0 \\ \vdots & \vdots & & & & \vdots & \vdots \\ -\boldsymbol{\Phi}_p & -\boldsymbol{\Phi}_{p-1} & & \ddots & & 0 & 0 \\ 0 & -\boldsymbol{\Phi}_p & \ddots & \vdots & \ddots & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots & \vdots \\ 0 & 0 & & & & \boldsymbol{I}_K & 0 \\ 0 & 0 & \cdots & -\boldsymbol{\Phi}_p & \cdots & -\boldsymbol{\Phi}_1 & \boldsymbol{I}_K \end{bmatrix}$.

The matrix $\widetilde{\boldsymbol{\Theta}_q}$ is defined in the same way as the VMA (q) process.

Let $\boldsymbol{\Upsilon}^0 = \begin{bmatrix} \boldsymbol{\Phi}_p \boldsymbol{y}_{-p+1} \\ \vdots \\ \boldsymbol{\Phi}_1 \boldsymbol{y}_0 \\ \vdots \\ 0 \end{bmatrix}'$, it follows that

$$\boldsymbol{\eta}_p \boldsymbol{y} - \boldsymbol{\Upsilon}^0 = \widetilde{\boldsymbol{\Theta}_q} \begin{pmatrix} \boldsymbol{u}_{-q+1} \\ \vdots \\ \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_T \end{pmatrix}.$$

Thus $\boldsymbol{y} \sim N(\boldsymbol{\eta_p}^{-1}\boldsymbol{\Upsilon}^0, \boldsymbol{\eta_p}^{-1}\widetilde{\boldsymbol{\Theta}_q}\,(\,\boldsymbol{I}_{T+q} \otimes \boldsymbol{\Sigma}_u\,)\widetilde{\boldsymbol{\Theta}_q}'\boldsymbol{\eta_p}'^{-1})$. (4.29)

The exact likelihood function is

$L(\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p\,, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q\,, \boldsymbol{\Sigma}_u|\boldsymbol{y})\;\alpha = |\boldsymbol{\eta_p}^{-1}\widetilde{\boldsymbol{\Theta}_q}(\boldsymbol{I}_{T+q}\;\otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_q}'\boldsymbol{\eta_p}'^{-1}|^{\frac{-1}{2}}$

$\times \exp\{\frac{-1}{2}\,(\boldsymbol{y}' - \;\boldsymbol{\eta_p}^{-1}\boldsymbol{\Upsilon}^0\,)\,\boldsymbol{\eta_p}'[\widetilde{\boldsymbol{\Theta}_q}\,(\boldsymbol{I}_{T+q}\;\otimes \boldsymbol{\Sigma}_u)\widetilde{\boldsymbol{\Theta}_q}'^{-1}](\boldsymbol{y} - \boldsymbol{\eta_p}[\boldsymbol{\Upsilon}^0])\}$. (4.30)

These equations can be solved from the use of iterative methods discussed in Appendix A.

The conditional likelihood function is obtained by assuming $\boldsymbol{y}_{-p+1} = \cdots \boldsymbol{y}_0 = \boldsymbol{u}_{-q+1} = \cdots = \boldsymbol{u}_0 = 0$.

Thus

$L(\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p\,, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q\,, \boldsymbol{\Sigma}_u|\boldsymbol{y}) = |\boldsymbol{\Sigma}_u|^{-T/2} - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{u}_t{\Sigma_u}^{-1}\boldsymbol{u}_t'$.

4.6.2 Least Squares Estimation for the VARMA(p, q) Model

The method of maximum likelihood estimation is very effective when used for small sample sizes. However, for larger samples its effectiveness is very much diminished. There has unfortunately not been much research published for the estimation procedures regarding larger sample sizes. In addition, Kascha (2010) noted that maximum likelihood estimation can have various numerical problems. There has recently been an iterative least squares estimation procedure proposed by Dias and Kapetanios (2011) where the matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Theta}_0$ are not identity matrices. This model is of the form

$$\boldsymbol{\Phi}_0\boldsymbol{y}_t = \boldsymbol{\Phi}_1\boldsymbol{y}_{t-1} + \cdots + \;\boldsymbol{\Phi}_p\boldsymbol{y}_{t-p} + \boldsymbol{\Theta}_0\boldsymbol{u}_t \;+\; \boldsymbol{\Theta}_1\boldsymbol{u}_{t-1} + \cdots + \;\boldsymbol{\Theta}_q\boldsymbol{u}_{t-q}\;. \tag{4.31}$$

The VARMA(p, q) model (4.31) in this instance is assumed to be stationary, invertible and uniquely defined as in the echelon form

(4.31) can be rearranged as

$$\boldsymbol{Y} = \boldsymbol{\varrho}\;\boldsymbol{X} + \boldsymbol{U}\;, \tag{4.32}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1\;, \ldots, \boldsymbol{y}_T)$ is of dimension $(K \times T)$.

$\boldsymbol{U} = (\boldsymbol{u}_1\;, \ldots, \boldsymbol{u}_T)$ is of dimension $(K \times T)$.

$\boldsymbol{\varrho} = [(\boldsymbol{I}_K - \boldsymbol{\Phi}_0), \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p, (\boldsymbol{\Theta}_0 - \boldsymbol{I}_K),\; \boldsymbol{\Theta}_1, \ldots,\; \boldsymbol{\Theta}_q]$ is the $(K \times K(p + q + 2))$ vector of parameters.

$\boldsymbol{X}_t = [\boldsymbol{Y}_t, \boldsymbol{Y}_{t-1}, \ldots, \boldsymbol{U}_t, \boldsymbol{U}_{t-1}, \ldots, \boldsymbol{U}_{t-p}]$.

$\boldsymbol{X} = [\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T]$ is the $(K(p+q+2) \times T)$ matrix of regressors.

If the vec operator is applied to (4.35), then

$$\text{vec}\,(\mathbf{Y}) = (\boldsymbol{X}' \otimes \boldsymbol{I}_K)\text{vec}(\boldsymbol{\varrho}') + \text{vec}(\boldsymbol{U}) \tag{4.33}$$

Suppose there is now an echelon form transformation, i.e. a transformation in which some of the elements in the vector of parameters, $\boldsymbol{\varrho}$ are equated towards 0. This implies that $\boldsymbol{\Phi}_0 = \boldsymbol{\Theta}_0$ and $\boldsymbol{\varrho}$ will now be of dimension $(K \times K(p+q+1))$. In this case, $\boldsymbol{\varrho}$ can be allowed to be written as a linear combination of a matrix and a vector $\mathbf{R}\boldsymbol{\lambda}$ where **R** is a matrix consisting of 0's and ones implied by this transformation and is of dimension $(K^2(p+q+1) \times K^2(p+q+1))$. $\boldsymbol{\varrho}^* = \text{vec}(\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q)$ is a $(K^2(p+q+1) \times 1)$ matrix which contains all of the $K^2(p+q+1)$ parameters that are required to be estimated. Following this, equation (4.33) can be rewritten as

$$\text{vec}\,(\mathbf{Y}) = (\boldsymbol{X}' \otimes \boldsymbol{I}_K)\mathbf{R}\,\boldsymbol{\varrho}^* + \text{vec}(\boldsymbol{U}). \tag{4.34}$$

There is still a limitation as $\boldsymbol{X}$ contains values of the error terms which are not fully observed. A consistent estimator $\widehat{\boldsymbol{U}}_0$ thus needs to be computed. This is performed by expressing the VARMA(p,q) model in a VAR representation as the model is assumed to be invertible.

$$\widehat{\boldsymbol{U}}_{0,t} = \boldsymbol{y}_t - \textstyle\sum_{i=1}^{p} \boldsymbol{\Pi}_i \boldsymbol{y}_{t-i} \tag{4.35}$$

From this, the new matrix of regressors is obtained by plugging $\widehat{\boldsymbol{U}}_{0,t}$ into the original matrix $\boldsymbol{X}_t$. The resulting matrix of regressors is $\boldsymbol{X}_0$.

The initial estimator for $\boldsymbol{\varrho}^*$, $\widehat{\boldsymbol{\varrho}}_1{}^*$ is

$$\widehat{\boldsymbol{\varrho}}_1{}^* = [\,\boldsymbol{R}'\,(\boldsymbol{X}_0\,\boldsymbol{X}_0{}')\otimes \boldsymbol{I}_K)\boldsymbol{R}]^{-1}\boldsymbol{R}(\boldsymbol{X}_0{}'\otimes\ \boldsymbol{I}_K)\,\text{vec}\,(\mathbf{Y})\,. \tag{4.36}$$

The parameter matrices, $\widehat{\boldsymbol{\Phi}}_{1,0},\ldots,\widehat{\boldsymbol{\Phi}}_{1,p}$, $\widehat{\boldsymbol{\Theta}}_{1,0}, \ldots, \widehat{\boldsymbol{\Theta}}_{1,q}$ can now be computed recursively from a new set of residuals $\widehat{\boldsymbol{U}}_{1,t}$ where the first subscript represents the iteration number and the second subscript denotes the lag order.

$$\begin{aligned}\widehat{\boldsymbol{U}}_{1,t} &= \boldsymbol{y}_t - \widehat{\boldsymbol{\Phi}}_{1,0}{}^{-1}\widehat{\boldsymbol{\Phi}}_{1,1}\boldsymbol{y}_{t-1} - \cdots - \widehat{\boldsymbol{\Phi}}_{1,0}{}^{-1}\widehat{\boldsymbol{\Phi}}_{1,p}\boldsymbol{y}_{t-p} - \widehat{\boldsymbol{\Phi}}_{1,0}{}^{-1}\widehat{\boldsymbol{\Theta}}_{1,1}\widehat{\boldsymbol{U}}_{1,t-1} - \cdots \\ &\quad - \widehat{\boldsymbol{\Phi}}_{1,0}{}^{-1}\widehat{\boldsymbol{\Theta}}_{1,q}\widehat{\boldsymbol{U}}_{1,t-q}\,.\end{aligned} \tag{4.37}$$

In general, the estimator at the jth iteration, $\widehat{\boldsymbol{\varrho}}_j{}^*$ is

$$\widehat{\boldsymbol{\varrho}}_j{}^* = [\,\boldsymbol{R}'\,(\boldsymbol{X}_{j-1}\,\boldsymbol{X}_{j-1}{}')\ \otimes\ \boldsymbol{I}_K)\boldsymbol{R}]^{-1}\boldsymbol{R}(\boldsymbol{X}_{j-1}{}'\otimes\ \boldsymbol{I}_K)\,\text{vec}\,(\mathbf{Y})$$

$$\begin{aligned}\widehat{\boldsymbol{U}}_{j,t} &= \boldsymbol{y}_t - \widehat{\boldsymbol{\Phi}}_{j,0}{}^{-1}\widehat{\boldsymbol{\Phi}}_{j,1}\boldsymbol{y}_{t-1} - \cdots - \widehat{\boldsymbol{\Phi}}_{j,0}{}^{-1}\widehat{\boldsymbol{\Phi}}_{j,p}\boldsymbol{y}_{t-p} - \widehat{\boldsymbol{\Phi}}_{j,0}{}^{-1}\widehat{\boldsymbol{\Theta}}_{j,1}\widehat{\boldsymbol{U}}_{j,t-1} - \cdots \\ &\quad -\widehat{\boldsymbol{\Phi}}_{j,0}{}^{-1}\widehat{\boldsymbol{\Theta}}_{j,q}\widehat{\boldsymbol{U}}_{j,t-q}\,.\end{aligned} \tag{4.38}$$

The jth iteration in which the vector of parameter estimated converges, i.e. $||\widehat{\boldsymbol{U}}_{t,j+1} - \widehat{\boldsymbol{U}}_{t,j}|| <$ (a pre-specified constant) terminates the procedure. The least squares estimates at that stage are taken as $\widehat{\boldsymbol{\varrho}}_j{}^*$. Dias and Kapetanios (2011) noted that this estimator is feasible for higher dimensional models in contrast to that of maximum likelihood estimation.

4.7 Diagnostic Checking of the VARMA(p, q) Model

Diagnostic checking for the VARMA (p, q) model is very similar to that of the VAR (p) and VMA (q) models in that the model is suitable if the residuals are not serially correlated. Tests such as the Portmanteau test are applicable for the VARMA model as well.

4.7.1 Portmanteau Test

Lütkepohl (2005) noted that the Portmanteau test for the VARMA (p, q) model is similar to the VAR (p) case with the exception being that the standard errors are approximated by $\frac{1}{\sqrt{T}}$.

The Box-Ljung statistic, $\widetilde{Q_h} = T^2 \sum_{i=1}^{h} (T-i)^{-1}\, tr(\hat{\boldsymbol{\Gamma}}_u(i)\, \widehat{\boldsymbol{\Gamma}_u}(0)^{-1} \hat{\boldsymbol{\Gamma}}_u(-i)\, \widehat{\boldsymbol{\Gamma}_u}(0)^{-1})$ is also applicable where $\boldsymbol{\Gamma}_u(i)$ is derived from the estimated residuals of the VARMA (p, q) model. The statistic $\widetilde{Q_h}$ follows a χ^2 distribution with $K^2(h-p-q)$ degrees of freedom. The fitted model is inadequate for particularly large values of $\widetilde{Q_h}$. If there are constraints imposed on the parameter coefficients, then $\widetilde{Q_h}$ will follow a χ^2 distribution with $K^2 h - \ b$ degrees of freedom and where b is the number of unrestricted parameters which are estimated in the model (Box et al, 2008).

4.7.2 Other Methods of Diagnostic Checking

A final diagnostic check can be performed by over fitting the model on purpose. A VARMA $(p+s, q)$ or VARMA$(p, q+s)$ model is fitted and successively reduced by restricting all of the parameters which are not significant to 0. The alternate hypothesis when testing for a VARMA $(p+s, q)$ model is $\boldsymbol{\Phi}_{p+1} = \cdots = \ \boldsymbol{\Phi}_{p+s} = 0$ while the alternative hypothesis when testing for a VARMA$(p, q+s)$ model is $\boldsymbol{\Theta}_{q+1} = \cdots = \boldsymbol{\Theta}_{q+s} = 0$. This method makes use of the "score" vector as the test statistic, i.e. the vector of first order partial derivatives with respect to the model parameters of the alternative model which is evaluated at the maximum likelihood estimates of the original model. This score statistic has a χ^2 distribution with sK^2degrees of freedom. The null hypothesis is rejected for particularly large values of this statistic. It is important to remember for this case that a rejection of the null hypothesis does not mean that the alternative hypothesis is accepted; rather it means that the model fitted is inadequate (Reinsel, 1997).

4.8 Forecasting the VARMA(p, q) Model

The concept of forecasting for the VARMA (p, q) model is similar to that performed for the VAR (p) model. Consider the vector ARMA (p, q) model,

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{u}_t + \sum_{i=1}^{q} \boldsymbol{\Theta}_i \boldsymbol{u}_{t-i} , \tag{4.39}$$

where the intercept $\boldsymbol{c}_t = \left(\boldsymbol{I} - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2 - \cdots - \boldsymbol{\Phi}_p\right)\boldsymbol{\mu}$.

It is necessary to assume that the components of the vector $\boldsymbol{u}_t$ are mutually independent of each other (Box et al., 2008). If conditional expectations are applied to both sides of the relation (4.39), then by making use of the property that since the future white noise process $\boldsymbol{u}_{t+h}$ $h > 0$ is independent of the past and present values $\boldsymbol{y}_t$, $\boldsymbol{y}_{t-1}$, … , the value of $E(\boldsymbol{u}_{t+h}|$ $\boldsymbol{y}_t$,$\boldsymbol{y}_{t-1}$, …) is 0.

The minimum mean square error predictor of the h step ahead forecast $\boldsymbol{y}_{t+h}$ is

$$\begin{aligned}\widehat{\boldsymbol{y}}_t(h) &= E\left(\boldsymbol{y}_{t+h}|\boldsymbol{y}_t, \boldsymbol{y}_{t-1,}\ldots\right) \\ &= \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \widehat{\boldsymbol{y}}_t(h-1) + \cdots + \boldsymbol{\Phi}_p \widehat{\boldsymbol{y}}_t(h-p) + \boldsymbol{\Theta}_1 E(\boldsymbol{u}_{t+h-1}) + \cdots + \boldsymbol{\Theta}_q E\left(\boldsymbol{u}_{t+h-q}\right).\end{aligned} \tag{4.40}$$

$h = 1, \ldots q$, $\widehat{\boldsymbol{y}}_t(h) = \boldsymbol{y}_{t+h}$ for $h \leq 0$ and $E(\boldsymbol{u}_{t+j}) = \boldsymbol{u}_{t+j}$ for $h \leq 0$ (Wei, 2006).

If $h > q$, then (4.40) becomes

$$\widehat{\boldsymbol{y}}_t(h) = \boldsymbol{c}_t + \boldsymbol{\Phi}_1 \widehat{\boldsymbol{y}}_t(h-1) + \cdots + \boldsymbol{\Phi}_p \widehat{\boldsymbol{y}}_t(h-p). \tag{4.41}$$

Lütkepohl (2005) noted that the forecast at period h, $\widehat{\boldsymbol{y}}_t(h)$ can also be calculated by using the infinite VAR representation as $\widehat{\boldsymbol{y}}_t(h) = \sum_{i=1}^{\infty} \boldsymbol{\Pi}_i \widehat{\boldsymbol{y}}_t(h-i)$.

$\widehat{\boldsymbol{y}}_t(h)$ can also be obtained from the infinite VMA representation. The future value $\boldsymbol{y}_{t+h}$ can be expressed in this form as $\boldsymbol{y}_{t+h} = \sum_{i=0}^{\infty} \boldsymbol{\psi}_i \boldsymbol{u}_{t+h-i}$. Since $E(\boldsymbol{u}_{t+h}|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots) = 0$ $h > 0$, the minimum mean square error predictor is $\widehat{\boldsymbol{y}}_t(h) = \sum_{i=1}^{\infty} \boldsymbol{\psi}_i \boldsymbol{u}_{t+h-i}$ (Box et al., 2008). The vector of forecasting errors, $\boldsymbol{y}_{t+h} - \widehat{\boldsymbol{y}}_t(h)$ is normally distributed with a zero mean and covariance matrix $\sum_{i=0}^{h-1} \boldsymbol{\psi}_i \boldsymbol{\Sigma}_u \boldsymbol{\psi}_i{}'$ (Tiao & Tsay, 1983).

The white noise sequence $\boldsymbol{u}_{t+h-i}$ $i = 1,2,\ldots,q$ of the model (4.41) needs to be generated recursively by using the past values $\boldsymbol{y}_t$, $\boldsymbol{y}_{t-1}$, … from the equation

$$\boldsymbol{u}_s = \boldsymbol{y}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{y}_{s-i} - \sum_{i=1}^{q} \boldsymbol{\Theta}_i \boldsymbol{u}_{s-i} . \tag{4.42}$$

This sequence can be obtained by using suitable starting values such as $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{1-q}$ and $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{1-p}$ (Box et al., 2008).

4.9 Conclusion

Although the VAR(p) model can be used to determine the interdependence among two or more series, it does not take into account the effect of innovations or shocks at different time lags and neither is it very parsimonious. The VMA(q) model on the other hand does take into account the various shocks and innovations at different time periods but cannot be used to determine the relationships among the various time series at different time periods

The VARMA(p, q) model produces the most precise estimates and the best forecasts of all of the multivariate time series models. It is once again not as widely used as the VAR(p) model, as the estimation procedure is complicated and tedious. The specification procedure can also be problematic because the standard VARMA(p, q) model is not unique. Further research needs to be undertaken in order for the model building procedure to be simplified.

CHAPTER 5

Non-stationary Models

As defined in chapter one, a process is stationary if the covariance does not depend on time t but is instead dependent on the time interval h. In practice however, it is quite common for a time series to have variations and trends in the data. In this chapter the case in which there are unit roots present (i.e. the non-stationary case) will be discussed. The VMA model will not be discussed in this chapter as it is assumed to be always stationary.

5.1 The Integrated Process

A process is known as an integrated process when a unit root has an effect on the autoregressive operator. This occurs when the mean and variance of a process is not stationary. If the variation is very unpredictable, there is said to be a stochastic trend present while if the trends are more predictable, it is said that there is a deterministic term present (Maddala & Kim 1999). As has been stated previously, a VAR (p) or VARMA (p, q) process is stationary if the determinant of ($\boldsymbol{I} - \boldsymbol{\Phi}_1 L - \cdots - \boldsymbol{\Phi}_p L^p$) has no roots in or outside a unit circle. Consider the case of univariate AR(1) model, $y_t = \varphi_1 y_{t-1} + u_t$ which is stationary if

$$1 - \varphi_1 L \neq 0 \text{ for } |L| \leq 1, |\varphi_1| < 1 . \tag{5.1}$$

In the borderline case where $\varphi_1 = 1$, equation (5.1) becomes

$$y_t = y_{t-1} + u_t . \tag{5.2}$$

This model (5.2) is known as the AR(1) model, with a unit root present because the root of the AR(1) equation is equal to one (Maddala & Kim, 1999). It is also more commonly known in the literature as a random walk model. By repeat substitution this model (5.2) can be expressed as the sum of all the disturbances and innovations (Lütkepohl, 2005).

$$\begin{aligned} y_t &= y_{t-1} + u_t \\ &= y_{t-2} + u_{t-1} + u_t \\ &= y_{t-3} + u_{t-2} + u_{t-1} + u_t \\ &\ \ \vdots \\ &= y_0 + u_1 + \cdots + u_t \end{aligned}$$

$$= y_0 + \sum_{i=1}^{t} u_i \ . \tag{5.3}$$

If there is a nonzero constant term in (5.3), then the model is of the form $y_t = v_0 + y_{t-1} + u_t$ and will be known as a random walk with drift. This model now has a deterministic trend in the mean.

Lütkepohl (2005) showed that the mean of the random walk model is $E(y_t) = y_0$ and the variance is var$(y_t) = t{\sigma_u}^2$. In contrast to (5.1) in which the variance converges to a constant, the variance of the random walk model increases as $t \to \infty$ (Maddala & Kim, 1999). The correlation between y_t and y_{t+s} is

$$\text{Corr}(y_t, y_{t+s}) = \frac{E[(\sum_{i=1}^{t} u_i)(\sum_{i=1}^{t+s} u_i)]}{\sqrt{[t{\sigma^2}_u (t+s){\sigma^2}_u]}} \tag{5.4}$$

for any integer $s > 0$. Now the value of (5.4) converges towards one as $t \to \infty$. This demonstrates that the correlation between y_t and y_s is strong even if the time interval s is large.

A similar scenario to that of a random walk can also occur for higher order processes. Consider the univariate AR(p) model,

$$y_t = v_0 + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + u_t \ . \tag{5.5}$$

If there is just one unit root in the process (5.5), then the behaviour will be similar to that of a random walk i.e. the variance increases linearly and the correlation between the variables is strong when they are h time periods apart. If however, one of the roots is deep inside the unit circle (significantly less than one), then the variances will diverge towards infinity at a significantly quicker rate (Lütkepohl, 2005).

A non-stationary time series can be made stationary by either transforming the data (by using a log transform or root transformation) or by differencing the data. Differencing is obtained by subtracting the previous value y_{t-1} from the current value y_t or by multiplying y_t by $(1 - L)$ where L is the lag operator.

If a non-stationary univariate process y_t has its $(d-1)$th difference non-stationary but if its dth difference, $(1 - L)^d y_t$ is stationary, then y_t is said to be integrated of order d or $I(d)$(Wei, 2006). A process which is $I(1)$ can be made stationary by taking the first differences of the original process. The process is integrated of order 0 ($I(0)$) once it has reached stationarity.

The following properties for integrated variables were noted by Engle and Granger (1987)

If y_t is an $I(0)$ variable with a mean of 0, then the variance of y_t will be finite, a shock will only have a temporary effect on the value of y_t and the autocorrelations will decrease quickly in magnitude. The result of this is that the sum of these autocorrelations will always be finite.

On the contrary if $y_t \sim I(1)$ with a zero mean, then the variance of y_t will diverge towards infinity as $t \to \infty$, and an innovation will have a permanent effect on the value of y_t. The autocorrelations for this case converge towards 1 as $t \to \infty$.

5.2 The Integrated Variable – Vector Case

The theory used to describe the integrated variable for univariate models can also be extended to that of multivariate models by considering a K dimensional VAR(p) process with a unit root and without a deterministic term,

$$\boldsymbol{\Phi}(L)\, \boldsymbol{y}_t = \boldsymbol{u}_t\,. \tag{5.6}$$

For the process (5.6) to be stationary, it has to be differenced. The differencing operator

$$\boldsymbol{D}(L) = \begin{bmatrix} (1-L)^{d_1} & & \cdots & \\ \vdots & (1-L)^{d_2} & \ddots & \vdots \\ & & \cdots & (1-L)^{d_m} \end{bmatrix}$$ was used by Wei (1990) in order to

difference the process where $(d_1, \dots, d_m)$ are nonnegative integers.

Lütkepohl (2005) used a different approach by proposing a differencing stage in which each component of y_t is differenced by the same operator. If the left hand side of (5.6) is multiplied by the adjoint of $\boldsymbol{\Phi}(L)$, $\boldsymbol{\Phi}^{+}(L)$

$$|\boldsymbol{\Phi}(L)|\boldsymbol{y}_t = \boldsymbol{\Phi}^{+}(L)\, \boldsymbol{u}_t\,. \tag{5.7}$$

All of the components of the autoregressive process (5.7) have the same operator, i.e. the determinant $|\boldsymbol{\Phi}(L)|$. The process (5.7) can thus be said to have been written in a similar manner as that of a univariate process. Now supposing that $|\boldsymbol{\Phi}(L)|$ has d unit roots, then the autoregressive operator can now be rewritten as

$$|\boldsymbol{\Phi}(L)|\boldsymbol{y}_t = \varphi(L)(1-L)^d \boldsymbol{y}_t = \varphi(L)\Delta^d \boldsymbol{y}_t\,. \tag{5.8}$$

The process $\Delta^d \boldsymbol{y}_t$ is a stationary process. Thus the relation (5.8) shows that each component of (5.7) can be made stationary upon differencing. This implies that a VAR(p) process is non-stationary only because of the presence of unit roots. However, this does not always mean that a process with d unit roots should be differenced d times as there might be common factors present which means that some terms may cancel each other out.

Thus it is not always suitable to fit multivariate VAR models after differencing all of the component series as there is a possibility of over differencing which leads to complications in

model fitting. These complications include model representations which are noninvertible (Box et al., 2008). Chatfield (2004) noted that this is a problem particularly if different degrees of differencing are used for each individual series. Differencing the series can also result in a rank deficiency in the matrix coefficients (Saidi, 2007). This happens particularly in the case of cointegration which will be discussed later in chapter 5.4.

5.3 Testing for Non-stationarity : The Dickey - Fuller Test

In most analyses, it is unknown whether the variables are integrated or stationary. Pre-tests for unit roots are often required in order to determine whether the series are stationary or not (Toda & Yamamoto, 1995). The most widely used and well established test for non-stationarity in a series is the augmented Dickey - Fuller test. It is an extension of the Dickey - Fuller test with the exception that the autocorrelation in a time series is removed prior to the testing for a unit root by the addition of extra lags of the dependent variable.

The model used when testing the null hypothesis of non-stationarity against the alternative that a series is stationary is

$$y_t - y_{t-1} = v_0 + v_1 t + \pi y_{t-1} + \varphi_1{}^* \Delta y_{t-1} + \cdots + \varphi_{p-1}{}^* \Delta y_{t-p+1} + u_t \,. \quad (5.9)$$

y_t in this model represents a univariate time series where $\Delta y_t = y_t - y_{t-1}$ refers to the differenced series. The term v_0 is a constant while v_1 is the coefficient of a time trend. The model (5.9) can be extended to allow for moving average terms in u_t.

There are 3 cases to consider when testing for a unit root:

Testing for a unit root with drift and a deterministic time trend

$$y_t - y_{t-1} = v_0 + v_1 t + \pi y_{t-1} + \varphi_1{}^* \Delta y_{t-1} + \cdots + \varphi_{p-1}{}^* \Delta y_{t-p+1} + u_t \,. \quad (5.10)$$

Testing for a unit root with drift ($v_1 = 0$)

$$y_t - y_{t-1} = v_0 + \pi y_{t-1} + \varphi_1{}^* \Delta y_{t-1} + \cdots + \varphi_{p-1}{}^* \Delta y_{t-p+1} + u_t \,. \quad (5.11)$$

Testing for a unit root with zero mean ($v_0 = v_1 = 0$)

$$y_t - y_{t-1} = \pi y_{t-1} + \varphi_1{}^* \Delta y_{t-1} + \cdots + \varphi_{p-1}{}^* \Delta y_{t-p+1} + u_t \,. \quad (5.12)$$

The tests are carried out by testing the null hypothesis,
$H_0 : \pi = 0$ (There is a unit root present)
$H_1 : \pi < 0$ (There are no unit roots present).

The test statistic is computed as $\frac{\hat{\pi}}{SE(\hat{\pi})}$ and is compared with the critical values of a Dickey - Fuller table. These critical values are functions of Brownian motion and are different for each of the models (5.10), (5.11) and (5.12). If the test statistic is less than the critical value, then the

null hypothesis is rejected and it can thus be concluded that there is no unit root present and hence the series is stationary. Therefore the insignificant parameters of unit root tests will be revealed if lagged values are included in the regression of (5.9). The test statistic used in SAS is known as the τ(tau) statistic.

An additional modification of the tests occurs when the estimator $\hat{\pi}$ is such that $T(\hat{\pi} - 1)$ has a large sample distribution but is not stationary. The test in this instance is known as the normalised bias (studentised) statistic and is known in SAS by the symbol ρ (rho).

5.4 Cointegration and the VECM Model

The concept of cointegration was defined by Engle and Granger (1987) as "If each element of a vector time series $\boldsymbol{y}_t$ achieves stationarity after differencing d times but if a linear combination, say $\boldsymbol{\Omega}'\boldsymbol{y}_t$ of all of the unit root series is stationary (i.e. integrated of order 0), then there is cointegration present in the model. $\boldsymbol{\Omega}$ is known as the cointegrating vector". This is the same as saying that cointegration occurs if there exists a linear combination of various individual non-stationary time series which result in a single stationary time series. The individual components of $\boldsymbol{y}_t$ can be integrated many times however for simplicity it is usually assumed that each series is $I(1)$ (unless otherwise specified by a rank test). The number of unit roots in a cointegrated time series is always less than the dimension of the process. Thus for a K dimensional unit root non-stationary time series there can only be cointegration present if there are less than K unit roots in the system.

The concept of cointegration is concerned with the long term behaviour amongst the components of partially non-stationary time series i.e. if $\boldsymbol{y}_t$ is non-stationary where the determinantal polynomial $|\boldsymbol{\Phi}(L)| = 0$ has $K^* < K$ (where K is the full rank of the cointegrating matrix) unit roots which are equal to one and where all of the other are roots outside the unit circle. This is an indication of a feature known as a common trend (the trend occurs simultaneously for all the series) (Ahn, 1997). Thus cointegration can be seen as the grouping of a few univariate non-stationary time series which are "moving together" (Brockwell & Davis, 1996). This implies that each of the individual components, y_{it} share common non-stationarity parts which result in similar behaviour over a period of time. These common trends can be eliminated if there are linear combinations of all the components of y_{it} (Box et al., 2008).

Cointegration implies, that even if there are many instances which cause permanent changes in each of the individual elements of $\boldsymbol{y}_t$, there is still the possibility of a long run relation existing if all of the elements are put together. For instance if $\boldsymbol{y}_t = (y_{1t}, \dots, y_{Kt})$ represents all of the variables of interest, then the long run equilibrium relation is $\boldsymbol{\Omega}'\boldsymbol{y}_t = \Omega_1 y_{1t} + \cdots + \Omega_K y_{Kt} = 0$. The cointegrating vector for this example is $\boldsymbol{\Omega} = (\Omega_1, \dots, \Omega_K)$. The stationary linear combinations are interpreted as the long run stable equilibrium relations among $\boldsymbol{y}_t$.

The cointegrating vector is not unique as it yields a further cointegrating vector when it is multiplied by a nonzero constant eg, if the cointegrating vector $\boldsymbol{\Omega}$ is multiplied by a nonzero scalar b, then the vector $b\boldsymbol{\Omega}$ will also still be a cointegrating vector. Similarly if $\Omega'\boldsymbol{y}_t$ is stationary, then the linear combination, $b\boldsymbol{\Omega}'\boldsymbol{y}_t$ will also be stationary.

The number of cointegrating factors is defined as the number of different linear combinations which are stationary. The components of a vector $\boldsymbol{y}_t$ are cointegrated of order (d, b) if all the components of $\boldsymbol{y}_t$ are $I(d)$ and if there exists a vector $\boldsymbol{\Omega} \neq 0$ such that $\boldsymbol{z}_t = \boldsymbol{\Omega}'\boldsymbol{y}_t \sim I(d-b)$. In the example given in Dolado, Gonzalo and Marmol (1999), consider two series y_{1t} and y_{2t} which are both $I(d)$. A linear combination of these two variables under normal circumstances is also $I(d)$. However, if there exists a vector say $(1,-\Omega)$ such that a linear combination say $\boldsymbol{z}_t = y_{1t} - \Omega\, y_{2t}$ is $I(d-b)$ where $d, b > 0$, then it is said that y_{1t} and y_{2t} are cointegrated variables of order (d, b) with a cointegrating vector $(1,-\Omega)$.

An example of cointegration in a bivariate series can be explained in the following example used by Hamilton (1994) as well as by Saikonnen and Lütkepohl (1996).

Consider the following bivariate system,

$$y_{1t} = \beth y_{2t} + u_{1t} \quad (5.13)$$
$$y_{2t} = y_{2,t-1} + u_{2t}, \quad (5.14)$$

where $\boldsymbol{u}_t = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$ is strictly stationary with $E(\boldsymbol{u}_t) = 0$ and positive definite covariance matrix $\boldsymbol{\Sigma}_u = E(\boldsymbol{u}_t \boldsymbol{u}_t')$. The error series u_{1t} and u_{2t} are uncorrelated and are white noise processes.

(5.14) can be rewritten as
$(1-L)y_{2t} = u_{2t}$ or
$$\Delta y_{2t} = u_{2t} \quad (5.15)$$

If equation (5.13) is differenced, then from the use of relation (5.15),

$$\Delta y_{1t} = \beth \Delta y_{2t} + \Delta u_{1t} = \beth u_{2t} + u_{1t} - u_{1,t-1} \,. \quad (5.16)$$

y_{2t} consists of $I(1)$ variables which are not cointegrated while y_{1t} and y_{2t} are cointegrated (Saikonnen & Lütkepohl, 1996).

The right side of (5.16) has a univariate MA(1) representation
$\Delta y_{1t} = v_t + \theta_1 v_{t-1}$ where v_t is a white noise process .

Thus although both of the individual series y_{1t} and y_{2t} are not stationary, the linear combination of y_{1t} and y_{2t} , $y_{1t} - \beth y_{2t}$ is stationary.

If a cointegrated VAR or VARMA model is differenced, it may lead to the model not being invertible. This can lead to problems regarding estimation. An alternate representation which overcomes much of the difficulty in the estimation of cointegrated VARMA models is known as the Vector Error Correction Model (VECM). The VECM model separates the cointegration and long run/equilibrium relations from the short term dynamics (Lütkepohl & Claessen, 1997). This model is of the form (under the assumption of reduced rank K*)

$$\Delta \boldsymbol{y}_t = \boldsymbol{\pi} \boldsymbol{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^{*} \Delta \boldsymbol{y}_{t-i} + \boldsymbol{u}_t + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{u}_{t-j} \qquad (5.17)$$

The autoregressive coefficient matrices $\boldsymbol{\Phi}_i^{*}$ are functions of the coefficient matrices $\boldsymbol{\Phi}_i$ where

$$\boldsymbol{\Phi}_i^{*} = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i \quad j = 1, \ldots, p-1$$
$$\boldsymbol{\pi} = \boldsymbol{\Phi}_p + \boldsymbol{\Phi}_{p-1} + \cdots + \boldsymbol{\Phi}_1 - \boldsymbol{I} = -\boldsymbol{\Phi}(1).$$

The VECM model is especially convenient as the number of unit roots in the autoregressive operator $\boldsymbol{\Phi}(L)$ is incorporated in the term $\boldsymbol{\pi} \boldsymbol{y}_{t-1}$ in equation (5.17) and therefore the type of non-stationarity which is observed in this model is dependent on the behaviour of $\boldsymbol{\pi}$ (Box et al., 2008). The term $\boldsymbol{y}_{t-1}$ shows the extent of disequilibrium in the variables of the previous period. Thus the VECM shows that changes in one variable are not only dependent on the changes of other variables and its own past changes but also on the extent of the equilibrium between levels of the variable (Dolado et al., 1999). This property means that the VECM can allow a number of variables to adjust simultaneously at different rates in response to a short run equilibrium (Kulshreshtha & Parikh, 2000). In order to distinguish the VECM from the usual VARMA model, the latter version is sometimes known as the levels version (Lütkepohl, 2004).

The VECM model (5.17) can also be written as

$$\boldsymbol{\Phi}^{*}(L)(1-L)\boldsymbol{y}_t = -\boldsymbol{\Phi}(1)\boldsymbol{y}_{t-1} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{u}_{t-j},$$

where $\boldsymbol{\Phi}^{*}(L) = \boldsymbol{I} - \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^{*} L^i$.

Under the assumptions on $\boldsymbol{\Phi}(L)$, $\boldsymbol{\Phi}^{*}(L)$ is a stationary operator with all of the roots of $|\boldsymbol{\Phi}^{*}(L)|$ outside the unit circle (Reinsel, 1997).

5.5 Cointegrated VAR(p) Models

The concept of cointegration and error correction is easier to understand in the simple VAR case or the VARMA $(p, 0)$ case. Consider the following K dimensional VAR(p) series, $\boldsymbol{y}_t$ with the inclusion of a possible time trend,

$$\boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t. \tag{5.18}$$

$\boldsymbol{v}_t = \boldsymbol{v}_0 + \boldsymbol{v}_1 t$ where $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ are constant vectors of dimension K and the error term $\boldsymbol{u}_t$ is assumed to be Gaussian (Normally) distributed (Tsay, 2005).

The error correction representation for the VAR(p) model is derived as follows (Wei, 2006)

The operator $\boldsymbol{\Phi}(L)$ can be rewritten as

$\boldsymbol{\Phi}(L) = (\boldsymbol{I} - (\boldsymbol{\Phi}_1 + \cdots + \boldsymbol{\Phi}_p)L) - (\boldsymbol{\Phi}_1{}^* L + \cdots + \boldsymbol{\Phi}_{p-1}{}^* L^{p-1})(1 - L)$,
where $\boldsymbol{\Phi}_i{}^* = -(\boldsymbol{\Phi}_{i+1} + \cdots + \boldsymbol{\Phi}_p)$ for $i = 1,2,\ldots,p-1$.

As a result, (5.18) can be rewritten as

$(\boldsymbol{I} - (\boldsymbol{\Phi}_1 + \ldots + \boldsymbol{\Phi}_p)L)\, \boldsymbol{y}_t - (\boldsymbol{\Phi}_1{}^* L + \cdots + \boldsymbol{\Phi}_{p-1}{}^* L^{p-1})\, \Delta \boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{u}_t$

or $\boldsymbol{y}_t = \boldsymbol{v}_t + \left((\boldsymbol{\Phi}_1 + \ldots + \boldsymbol{\Phi}_p)L\right) \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1{}^* \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^* \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t$. (5.19)

If $\boldsymbol{y}_{t-1}$ is subtracted from both sides of (5.19), then

$$\Delta \boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{\pi} y_{t-1} + \boldsymbol{\Phi}_1{}^* \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^* \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t \tag{5.20}$$
$$\boldsymbol{\pi} = \left(\boldsymbol{\Phi}_1 + \cdots + \boldsymbol{\Phi}_p\right) - \boldsymbol{I}\,.$$

If $\boldsymbol{y}_t$ contains unit roots, then $|\boldsymbol{\Phi}(1)| = 0$ and $\boldsymbol{\pi} = -\boldsymbol{\Phi}(1)$ will be nonsingular. There are 3 cases to consider regarding the rank of π.

Case 1: Rank $(\boldsymbol{\pi}) = 0$. This case means that $\boldsymbol{\pi} = 0$ which results in no cointegration being present in the model. In this case a linear combination of the $I(1)$ variables which is stationary does not exist. The VECM is reduced to

$$\Delta \boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{\pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1{}^* \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^* \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t\,. \tag{5.21}$$

The non-stationarity for this model can be removed by taking differences or using a log/root transformation. Thus the appropriate model which should be used is a VAR in first differences which does not involve any long run elements (Harris, 1995).

Case 2: Rank $(\boldsymbol{\pi}) = K$. This case is known as the full rank case with $|\boldsymbol{\Phi}(1)| \neq 0$. The vector process $\boldsymbol{y}_t$ contains no unit roots (is $I(0)$) and is therefore stationary. As a result of this, information cannot be gathered from using the VECM model (Tsay, 2005). $\boldsymbol{y}_t$ should thus be studied directly and modelled in levels, not differences and there is no need for a VECM representation to be used.

Case 3: If $0 <$ Rank $(\boldsymbol{\pi}) = K^* < K$. This is known as the partial non-stationary case where there are K^* distinct linear combinations of $\Delta\boldsymbol{y}_t$ that are stationary (Mauricio, 2006). The $\boldsymbol{\pi}$ matrix is of reduced rank and is required to be reparameterised as $\boldsymbol{\pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(K \times K^*)$ matrices of rank K, i.e. full column rank. The VECM is of the form

$$\Delta\boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1^*\Delta\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}^*\Delta\boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t \,. \tag{5.22}$$

This equation (5.22) implies that there are $K^* < K$ linear combinations of $\boldsymbol{y}_t$ which are stationary (Morin, 2010). The coefficient $\boldsymbol{\alpha}$ is a matrix of adjustment coefficients and measures how quickly $\Delta\boldsymbol{y}_t$ reacts to the equilibrium implied by $\boldsymbol{\beta}'\boldsymbol{y}_t$ (Morin, 2010). The coefficients $\boldsymbol{\Phi}_i^*$ and $\boldsymbol{\alpha}$ are short run stationary parameters because of their association with the stationary processes $\Delta\boldsymbol{y}_{t-i}$ and $\boldsymbol{\beta}'\boldsymbol{y}_{t-1}$. The parameter $\boldsymbol{\beta}$ is non-stationary as it is associated with the process $\boldsymbol{y}_{t-1}$ (Ahn, 1997). $\boldsymbol{\beta}$ also contains the K^* cointegrating vectors and shows the long run relationship between the jointly determined variables. The term $\boldsymbol{\beta}'\boldsymbol{y}_{t-1}$ can be regarded as a compensatory term for the overdifferenced system $\boldsymbol{y}_{t-1}$ as the vector error correction representation shows that $\boldsymbol{y}_{t-1}$ is unit root non-stationary and $\Delta\boldsymbol{y}_t$ is stationary (Tsay, 2005). Thus the only way in which $\Delta\boldsymbol{y}_t$ can be related to $\boldsymbol{y}_{t-1}$ is through the stationary series $\boldsymbol{\beta}'\boldsymbol{y}_{t-1}$.

It is important to realize that this decomposition of $\boldsymbol{\pi}$ as a product of two $(K \times K^*)$ matrices i.e. $\boldsymbol{\pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ is not a unique representation. $\boldsymbol{\pi}$ can also be reparametrized as $\boldsymbol{\pi} = \boldsymbol{\alpha}^*\boldsymbol{\beta}^{*\prime}$ with $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}\boldsymbol{W}'$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{W}^{-1}$ where $\boldsymbol{W}$ is a $(K^* \times K^*)$ matrix (Lütkepohl, 2005). Restrictions can be imposed on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in order for them to have unique cointegrating relations. An implication of the model (5.22) is that while all of the K^* component series have non-stationary behaviour, there are K linear combinations of $\boldsymbol{y}_t$ that are stationary which results in a reduced dimensionality of the non-stationarity K^* terms.

In conclusion, the rank of $\boldsymbol{\pi}$ in the VECM tests the number of cointegrating vectors in the model. This result will be further elaborated on later.

5.6 Specification of the Cointegrated VAR(p) Model

5.6.1 Choosing the Order of p

In practice, the order of p is generally unknown and is needed to be chosen prior to the construction of the tests. If the lag order is not chosen correctly, then the tests can potentially have big size distortions. The sample autocorrelation, partial autocorrelation and cross

correlation functions are of little use when identifying the error covariance structure (Singh, Yadavalli & Peiris, 2002). Lütkepohl and Saikonnen (1999) showed that criteria such as the AIC and BIC can be used to find the order of p as they are asymptotically valid. The BIC criterion in addition is also more consistent than the AIC. The FPE criterion should not be used for cointegrated processes however as the use of the forecast mean square error is difficult to justify in the non-stationary case. Nielsen (2006) showed that the likelihood based tests as well as the information criteria can be used regardless of there being unit roots present and can be used in the presence of deterministic terms.

Qu and Perron (2007) on the contrary, showed that the standard information criteria tests can lead to a model which can distort the size of the model. They provided a modification to the AIC in which the error term $\boldsymbol{u}_t$ was made more sensitive to the lag order. The modified statistic was defined as

$$\text{MAIC} = \ln|\boldsymbol{\Sigma}_u(g)| + \frac{2(r_q) + gK^2}{T-(\max lag\ order)-1} \tag{5.23}$$

The term r_q is the extra term of the likelihood ratio test of K cointegrating vectors against the alternative that there are more than K cointegrating vectors present. The above authors noted that this statistic led to improvements in the size of the cointegration tests.

A similar and recent modification to the AIC as well as the HQ and BIC criteria was proposed by Athanasopoulos, de Carvalho Guillén, Issler and Vahid (2011) who included the full rank of π as a parameter required to be selected.

5.6.2 Specification of the Deterministic Function

The general VECM model is assumed to be of the form $\boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{x}_t$ where $\boldsymbol{x}_t$ is the stochastic part that has a VECM representation without any deterministic terms and $\boldsymbol{v}_t$ is the deterministic term. The inclusion of a deterministic term for a non-stationary process is more problematic than for the stationary models as it changes some of the assumptions and results (Lütkepohl & Claessen, 1997). Thus in the next section I will discuss various methods for the specification of the deterministic term. The 5 different cases mentioned by Harris (1995) and Tsay (2005) are,

Case 1 : In the unlikely case, in which there is no deterministic term present i.e. $\boldsymbol{v}_t = 0$, all the component series of $\boldsymbol{y}_t$ are $I(1)$ without drift and the stationary series $\boldsymbol{\beta}'\boldsymbol{y}_t$ will have a zero mean.

Case 2 : $\boldsymbol{v}_t = \boldsymbol{v}_0 = \boldsymbol{\alpha}\boldsymbol{n}_0$ where $\boldsymbol{n}_0$ is a $(K^* \times 1)$ dimensional vector. This case occurs when there are no linear trends in the levels of the data. The deterministic term can be absorbed into the cointegrating relation. The resulting VECM is

$$\Delta\boldsymbol{y}_t = \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{n}_0) + \boldsymbol{\Phi}_1{}^*\Delta\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^*\Delta\boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t\ . \tag{5.24}$$

The components of $\boldsymbol{y}_t$ are $I(1)$ without drift and $\boldsymbol{\beta}'\boldsymbol{y}_t$ has a nonzero mean, $\boldsymbol{n}_0$.

Case 3 : $\boldsymbol{v}_t = \boldsymbol{v}_0$ where $\boldsymbol{v}_0$ is nonzero. This occurs when there are linear trends in the data. The component series of $\boldsymbol{y}_t$ are $I(1)$ with drift and $\boldsymbol{\beta}'\boldsymbol{y}_t$ might have a nonzero mean.

Case 4 : $\boldsymbol{v}_t = \boldsymbol{v}_0 + \boldsymbol{\alpha}\boldsymbol{n}_1 t$ where $\boldsymbol{n}_1$ is a nonzero vector. This occurs when there is no time trend in the short run model. The resulting VECM is

$$\Delta\boldsymbol{y}_t = \boldsymbol{v}_0 + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{n}_1 t) + \boldsymbol{\Phi}_1{}^*\Delta\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^*\Delta\boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t \,. \tag{5.25}$$

In this case the components of $\boldsymbol{y}_t$ are $I(1)$ with drift $\boldsymbol{v}_0$ and $\boldsymbol{\beta}'\boldsymbol{y}_t$ has a linear time trend which is related to $\boldsymbol{n}_1 t$.

Case 5 : $\boldsymbol{v}_t = \boldsymbol{v}_0 + \boldsymbol{v}_1 t$. There is a time component that is included in the regression for this case because $\boldsymbol{v}_1$ is nonzero. This occurs if the constant and slope of the trend are unrestricted and if there are quadratic trends in $\boldsymbol{y}_t$. The components of $\boldsymbol{y}_t$ are $I(1)$ are determined by the drift and also have a quadratic time trend while the term $\boldsymbol{\beta}'\boldsymbol{y}_t$ has a linear time trend. If $\boldsymbol{\pi}\boldsymbol{v}_1 = 0$, then the deterministic linear trend will be orthogonal to the cointegrating relations while if $\boldsymbol{\pi}\boldsymbol{v}_1 \neq 0$, the linear trend will be a part of the cointegrating relations (Demetrescu, Lütkepohl & Saikonnen, 2009).

5.7 Maximum Likelihood Estimation for the Cointegrated VAR$(\boldsymbol{p})$ Model

Two maximum likelihood estimation procedures for a cointegrated VAR(p) model will be discussed in this section.

The first method is a maximum likelihood estimation procedure which does not take into account deterministic terms. This procedure is carried out under the assumptions of normality, that the long run restriction of rank$(\boldsymbol{\pi})$ is K and the short run restriction is the rank of $\boldsymbol{\Phi}_1{}^*, \ldots, \boldsymbol{\Phi}_p{}^*$ is K^*. If this is the case, then the VECM can be written in the form

$$\Delta\boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{\mathcal{E}}\,[\boldsymbol{F}_1\Delta\boldsymbol{y}_{t-1} + \boldsymbol{F}_2\Delta\boldsymbol{y}_{t-2} + \cdots + \boldsymbol{F}_p\Delta\boldsymbol{y}_{t-p}] + \boldsymbol{u}_t \,. \tag{5.26}$$

$\boldsymbol{\mathcal{E}}$ is a K x K^* matrix of reduced rank K^*. The estimation of (5.26) is obtained by using a partial canonical correlation analysis between $\Delta\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ that is conditional on $\Delta\boldsymbol{y}_{t-1}, \ldots, \Delta\boldsymbol{y}_{t-p+1}$. The following step by step procedure was proposed by Athanasopoulos et al. (2011)

a) If $\boldsymbol{\beta}$ is known, then $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{F}_i\ i = 1, \ldots, p$ can be estimated by using a reduced rank regression of $\Delta\boldsymbol{y}_t$ on its lagged values $\Delta\boldsymbol{y}_{t-1}, \ldots, \Delta\boldsymbol{y}_{t-p}$ while controlling for $\boldsymbol{y}_{t-1}$. The first procedure is to estimate the values of $[\widehat{\boldsymbol{F}_1}, \ldots, \widehat{\boldsymbol{F}_p}]$ from a reduced rank regression of

$\Delta \boldsymbol{y}_t$ on $(\Delta \boldsymbol{y}_{t-1}, \dots, \Delta \boldsymbol{y}_{t-p})$ after controlling for $\boldsymbol{y}_{t-1}$. These estimates $[\widehat{\boldsymbol{F}_1}, \dots, \widehat{\boldsymbol{F}_p}]$ are the coefficients of the canonical variates which correspond to the K largest squared partial canonical correlations between $\Delta \boldsymbol{y}_t$ and $(\Delta \boldsymbol{y}_{t-1}, \dots, \Delta \boldsymbol{y}_{t-p})$.

b) The next procedure is to compute the partial canonical correlations between $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ after controlling for $[\widehat{\boldsymbol{F}_1} \Delta \boldsymbol{y}_{t-1} + \widehat{\boldsymbol{F}_2} \Delta \boldsymbol{y}_{t-2} + \cdots + \widehat{\boldsymbol{F}_p} \Delta \boldsymbol{y}_{t-p}]$. The K^* canonical variates of $\boldsymbol{\beta}' \boldsymbol{y}_{t-1}$ which correspond to the K^* largest squared partial canonical correlations are taken as the estimates for the cointegrating relationships. $\Delta \boldsymbol{y}_t$ is then regressed on $\boldsymbol{\beta}' \boldsymbol{y}_{t-1}$ and $[\widehat{\boldsymbol{F}_1} \Delta \boldsymbol{y}_{t-1} + \widehat{\boldsymbol{F}_2} \Delta \boldsymbol{y}_{t-2} + \cdots + \widehat{\boldsymbol{F}_p} \Delta \boldsymbol{y}_{t-p}]$ and the logarithm of the determinant of the residual variance, $\ln |\boldsymbol{\Sigma}_u|$ is computed from this regression.

c) The partial canonical correlations between $\Delta \boldsymbol{y}_t$ and $(\Delta \boldsymbol{y}_{t-1}, \dots, \Delta \boldsymbol{y}_{t-p})$ are computed conditional on $\boldsymbol{\beta}' \boldsymbol{y}_{t-1}$. $\Delta \boldsymbol{y}_t$ is then regressed on $\boldsymbol{\beta}' \boldsymbol{y}_{t-1}$ and $[\widehat{\boldsymbol{F}_1} \Delta \boldsymbol{y}_{t-1} + \widehat{\boldsymbol{F}_2} \Delta \boldsymbol{y}_{t-2} + \cdots + \widehat{\boldsymbol{F}_p} \Delta \boldsymbol{y}_{t-p}]$. This regression is used to compute the value of $\ln |\boldsymbol{\Sigma}_u|$. If the value of $\ln |\boldsymbol{\Sigma}_u|$ is different from that obtained in step b), then it is necessary to revisit that particular step. If the value of $\ln |\boldsymbol{\Sigma}_u|$ is not different then the process is terminated and the estimates are chosen accordingly. The value of $\ln |\boldsymbol{\Sigma}_u|$ becomes smaller at each step until it reaches a minimum value. The values of $\widehat{\boldsymbol{\mathcal{E}}}$ and $\widehat{\boldsymbol{F}_1}, \dots, \widehat{\boldsymbol{F}_p}$ which correspond to this minimum value are taken as maximum likelihood estimates of $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{F}_i \; i = 1, \dots, p$

It is important to note that if there are additional constraints present, iterative procedures will be needed to be implemented in order for the maximum likelihood estimates to be obtained.

In the second estimation method in which there is a deterministic term present, the deterministic term can be written as $\boldsymbol{v}_t = \boldsymbol{v}\boldsymbol{s}_t$ where $\boldsymbol{s}_t = (1, t)'$ and $\boldsymbol{v}_t$ depends on the specification performed in section 5.6.2. The VECM model (5.22) is simplified to

$\Delta \boldsymbol{y}_t = \boldsymbol{v}\boldsymbol{s}_t + \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1^* \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p^* \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t$ (Tsay, 2005) .

The first step is to estimate two multiple linear regressions in which the terms $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ are regressed from the use of ordinary least squares on the remaining set of regressors, $\Delta \boldsymbol{y}_{t-1}, \dots, \Delta \boldsymbol{y}_{t-p+1}$.

$$\Delta \boldsymbol{y}_t = \boldsymbol{\gamma}_0 \boldsymbol{s}_t + \boldsymbol{\gamma}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\gamma}_{t-p+1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{f}_t \quad (5.27)$$

$$\boldsymbol{y}_{t-1} = \boldsymbol{\Xi}_0 \boldsymbol{y}_t + \boldsymbol{\Xi}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Xi}_{t-p+1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{g}_t \quad (5.28)$$

The residuals from equations (5.27) and (5.28), $\boldsymbol{f}_t$ and $\boldsymbol{g}_t$ are used to calculate the sample covariance matrices $\boldsymbol{S}_{00}$, $\boldsymbol{S}_{01}$, $\boldsymbol{S}_{11}$ as

$$\boldsymbol{S}_{00} = \frac{1}{T-p} \sum_{t=p+1}^{T} \hat{\boldsymbol{f}}_t \hat{\boldsymbol{f}}_t'$$

$$\boldsymbol{S}_{01} = \frac{1}{T-p} \sum_{t=p+1}^{T} \hat{\boldsymbol{f}}_t \hat{\boldsymbol{g}}_t'$$

$\boldsymbol{S}_{11} = \frac{1}{T-p}\sum_{t=p+1}^{T} \widehat{\boldsymbol{g}}_t \widehat{\boldsymbol{g}}_t{}'$.

The parameter estimates are under the restriction that the model is not of full rank with $\boldsymbol{\pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ and rank($\boldsymbol{\pi}$) $= K^*$. The eigenvalue equation $|\lambda \boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}{}^{-1}\boldsymbol{S}_{01}| = 0$ is solved in order to obtain the eigenvalues $1 > \hat{\lambda}_1 > \cdots > \hat{\lambda}_p$ (Lütkepohl 2004). These eigenvalues measure the largest squared canonical correlations between the residuals $\boldsymbol{g}_t$ and $\boldsymbol{f}_t$.

Let the matrix of eigenvectors be $\boldsymbol{e} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K]$ and denote the eigenvalue/eigenvector pairs by ($\hat{\lambda}_i, \boldsymbol{e}_i$). These eigenvectors are normalised such that $\hat{\boldsymbol{e}}'\boldsymbol{S}_{ii}\hat{\boldsymbol{e}} = \boldsymbol{I}$ (Tsay 2005). The size of $\hat{\lambda}_i$ measures how strongly the cointegrating relations $\hat{\boldsymbol{e}}'\boldsymbol{y}_t$ are correlated with the stationary section of the model (Harris, 1995).

The unconstrained maximum likelihood estimator of the cointegrating vector $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{K*}]$ and the estimate of the normalised cointegrating vector is $\boldsymbol{S}_{01}\widehat{\boldsymbol{\beta}}$ (Morin, 2010).

The likelihood function $L_{max}{}^{\frac{-2}{T}}$ is

$$L_{max}{}^{\frac{-2}{T}} \propto |\boldsymbol{S}_{00} - \boldsymbol{S}_{01}\boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{S}_{10}| \tag{5.29}$$

$$= \frac{|\boldsymbol{S}_{00}||\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{S}_{10}\boldsymbol{S}_{00}{}^{-1}\boldsymbol{S}_{01}\,\boldsymbol{\beta}|}{|\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta}|}$$

$$= \frac{|\boldsymbol{S}_{00}||\,\boldsymbol{\beta}'(\boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}{}^{-1}\boldsymbol{S}_{01})\,\boldsymbol{\beta}|}{|\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta}|}\,. \tag{5.30}$$

Johansen and Juselius (1990) expressed the likelihood function (5.30) as a function of the eigenvalues that is based on the K^* cointegrating vectors as

$$L_{max}{}^{\frac{-2}{T}} \propto |\,\boldsymbol{S}_{00}\,| \prod_{i=1}^{K*}(1 - \hat{\lambda}_i)\,. \tag{5.31}$$

The values of $\hat{\lambda}_i$ are the canonical correlations obtained by solving the likelihood equation (5.29). The short run effects $\boldsymbol{\Phi}_i{}^*$ can be estimated once an estimate of $\boldsymbol{\pi}$ has been fixed. Under normal Gaussian innovations, the estimates of $\boldsymbol{\Phi}_i{}^*$ are asymptotically normal and efficient. Lütkepohl (2004) noted that in practice, it is recommended that additional restrictions should be placed on the parameters in order for the dimensions of the parameter space to be reduced.

5.8 Testing the Order of Cointegration

In this section, I will discuss the ways in which to test the order of cointegration. Let the null hypothesis be defined such that the rank of $\boldsymbol{\pi}$ is K^*. Under this assumption, the VECM is

$$\Delta \boldsymbol{y}_t = \boldsymbol{v}\boldsymbol{s}_t + \boldsymbol{\pi}\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1{}^*\Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^*\Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t$$
$$t = p+1, \ldots, T . \qquad (5.32)$$

Defining the null hypothesis

H_0 : Rank($\boldsymbol{\pi}$) $= K^*$/ there are K^* cointegrating vectors present
H_1 : Rank($\boldsymbol{\pi}$) $> K^*$/ there are more than K^* cointegrating vectors present

The first rank K^* in which the null hypothesis is not rejected is chosen as the cointegrating rank.

From the use of the equations (5.27) and (5.28), the term $\boldsymbol{\pi}$ is related to the covariance between $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ after adjusting for the effects of $\boldsymbol{s}_t$ and $\Delta \boldsymbol{y}_{t-i}$ $i = 1, \ldots, p-1$. Using multivariate linear regression, the adjusted series for $\boldsymbol{y}_{t-1}$ and $\Delta \boldsymbol{y}_t$ are $\widehat{\boldsymbol{g}}_t$ and $\widehat{\boldsymbol{f}}_t$. Tsay(2005) noted that the equation of interest used for the cointegration test is

$$\widehat{\boldsymbol{f}}_t = \boldsymbol{\pi}\widehat{\boldsymbol{g}}_t + \boldsymbol{u}_t .$$

A likelihood ratio test statistic $LK_{tr}(K^*)$ which follows from the maximum likelihood estimation procedure was derived by Johansen (1988) and further elaborated by Johansen and Juselius (1990) was defined as

$$LK_{tr}(K^*) = -(T-p)\sum_{i=K^*+1}^{K} \ln(1-\hat{\lambda}_i) \qquad (5.33)$$
$$K^* = 0, 1, 2, \ldots, K-2, K-1 .$$

Reinsel (1997) noted that the likelihood ratio statistic (5.33) can be expressed equivalently as

$$LK_{tr}(K^*) = -(T-p)\sum_{i=K^*+1}^{K} \ln(1-p_i(p)^2) . \qquad (5.34)$$

The term $p_i(p)^2$ is the $(K-K^*)$th smallest sample partial canonical correlation between $\Delta \boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ given $\Delta \boldsymbol{y}_{t-1}, \ldots, \Delta \boldsymbol{y}_{t-p+1}$.

This statistic (5.34) is known as the trace statistic. The further the eigenvalues are from 0, then the more negative $\ln(1-\hat{\lambda}_i)$ will be and hence the larger the statistic (5.34) will be. i.e. as the sample size gets larger, the statistic (5.34) diverges towards infinity. This results in the test being consistent. If the rank of $\boldsymbol{\pi}$ is indeed K^* , then the value of $\hat{\lambda}_i$ should be small for $i > K^*$ which results in a smaller value for $LK_{tr}(K^*)$. Since there are unit roots present in the system, the likelihood function is not ${\chi_i}^2$ distributed but is instead a multivariate Dickey-Fuller distribution or a function of standard Brownian motions as there is an unmodelled trend found

in the residuals (Tsay, 2005). The distribution was derived by Johansen and Juselius (1990) as well as by Reinsel (1997) as

$$\text{tr}\Big(\int_0^1 (\boldsymbol{B}_{K-K^*}(u)(K-K^*)\boldsymbol{B}_{K-K^*}(u))' \times \Big(\int_0^1 \boldsymbol{B}_{K-K^*}(u)\big(K-K^*\big)\boldsymbol{B}_{K-K^*}(u)\Big)^{\prime-1} \Big(\int_0^1 (\boldsymbol{B}_{K-K^*}(u)(K-K^*)\boldsymbol{B}_{K-K^*}(u)\Big) . \quad (5.35)$$

$\boldsymbol{B}_{K-K*}(u)$ is the $K-K*$ dimensional standard Brownian motion process. This asymptotic distribution depends only on $K-K*$ and the order p of the model. In addition these critical values can be modified in the event of no deterministic term or trend present in the model (Lütkepohl, 2005).

An alternative sequential procedure has also been proposed by Johansen (1988). In this procedure, the null and alternative hypotheses are defined as

H_0 : Rank($\boldsymbol{\pi}$) $= K^*$
H_1 : Rank($\boldsymbol{\pi}$) $= K^* + 1$

The first step is to test rank($\boldsymbol{\pi}$) $= 0$, i.e. there is no cointegration present against the alternative rank($\boldsymbol{\pi}$) $= 1$, that there is one cointegrating relation. If the null hypothesis for $K^* = K-1$ cointegrating relationships is rejected, then it is concluded that there are $K^* = K$ cointegrating relationships present.

The likelihood ratio test statistic for this procedure is known as the maximum eigenvalue statistic and is calculated from

$$LK_{tr}(K^*) = -(T-p)\ln(1-\hat{\lambda}_{K^*+1}) . \quad (5.36)$$

The nearer the value of $\hat{\lambda}_{K^*+1}$ is to 0, the smaller the $LK_{tr}(K^*)$ statistic will be. As in the previous likelihood ratio test, these critical values are nonstandard and are needed to be calculated by using Monte Carlo simulation techniques.

An asymptotically equivalent version of this statistic (5.36) was given by Reinsel (1997) as

$$LK_{tr}(K^*) = -T\ln(1-p_{i+1}(p)^2) , \quad (5.37)$$

where $p_i(p)^2$ is the $(K-K*)$th sample partial canonical correlation between $\Delta\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ given $\Delta\boldsymbol{y}_{t-1}, \dots, \Delta\boldsymbol{y}_{t-p+1}$.

It is important to note that the power of the tests is dependent on the deterministic term in the model and an over specified deterministic term can have a large effect on the power of the test. Thus when testing for a specific cointegrating rank it is important to test for the deterministic term first and following this, to test for cointegration by using the deterministic

term selected from the initial test. If the deterministic term is under specified, then it is likely to terminate the sequence very quickly which will result in the rank chosen being too small. In other words the test will be terminated too early if the likelihood ratio test ($LK_{tr}(K^*)$) is performed when there is a trend present. The test statistic will then need to be modified by using a regression which takes into account the presence of a deterministic term. Demetresou et al. (2009) developed a test statistic $\widetilde{LK}_{tr}(K^*)$ which is consistent, rejects all the false null hypotheses and takes into account the presence of a deterministic term. The limitation of this modified statistic is that it has a reduced power in small samples and tends to choose a cointegrating rank that is too small when there is no trend term present.

If there is uncertainty regarding the deterministic term, the cointegrating rank can be chosen by performing tests that are based on different models with different possible deterministic terms. The rank can be chosen by taking into account all the different test results (Demetresou et al., 2009). Thus when the cointegrating rank is needed to be determined, both of the likelihood ratio statistics $LK_{tr}(K^*)$ and $\widetilde{LK}_{tr}(K^*)$, should be used. If none of the above tests rejects the null hypothesis, then the value of the cointegrating rank chosen is K^*. If the null hypothesis is rejected, the ranks $K^* + 1$, are tested until there is an acceptance of the null hypothesis.

Simulation studies were performed by the Demetresou et al. (2009) in order to compare this procedure with a procedure in which the cointegrating rank was chosen based on the pre-test for a deterministic term. They found that the first procedure was more effective in choosing the correct rank for sample sizes between 100 and 250. However, this could not be justified for larger sample sizes.

Information criteria such as the AIC and the BIC may also be used when specifying the rank. These criteria need to be modified in order to reflect the number of unknown parameters with a specified rank K in cointegrated form (Reinsel, 1997). Chao and Phillips (1999) developed a statistic which estimates both the lag order p and the cointegrating rank of the full model K simultaneously. This statistic also takes into account the non-stationarity of the regressors associated with the parameters. The limitation of this statistic is that the series observed did not have a deterministic term.

Athanasopoulos et al. (2011) demonstrated that the standard information criteria can also be modified simultaneously to choose p and K.These modified criteria are

$$\text{AIC}\ (p, K) = T \sum_{i=K-K*+1}^{K} \ln(1 - \hat{\lambda}_i(p)) + 2(K*(K - K*) + K*Kp)$$

$$\text{BIC}\ (p, K) = T \sum_{i=K-K*+1}^{K} \ln(1 - \hat{\lambda}_i(p)) + \ln T(K*(K - K*) + K*Kp)$$

$$\text{HQ}\ (p, K) = T \sum_{i=K-K*+1}^{K} \ln(1 - \hat{\lambda}_i(p)) + 2 \ln(\ln T(K*(K - K*) + K*Kp)) .$$

The value of p is the number of lagged differences in the VECM and $\hat{\lambda}_i$ is the number of squared canonical correlations between $\Delta \boldsymbol{y}_t$ and $[\Delta \boldsymbol{y}_{t-1}, \ldots, \Delta \boldsymbol{y}_{t-p+1}, \ \boldsymbol{\beta y}_{t-1}]$.

Athanasopoulos et al. (2011) also suggested that the linear influence of $\boldsymbol{y}_{t-1}$ from $\Delta\boldsymbol{y}_t$ should be removed prior to the testing of the model, and that the Hannan-Quinn (HQ(p, K)) is the preferred criterion that should be used to determine p as well as the full rank K of the model. This statistic can further be modified to find values of the reduced rank $K*$ conditional on the values p and K which were chosen in the initial step.

5.9 Comparison of the VAR and VECM Models for Cointegration

It is often necessary to determine whether the VAR or the VECM model should be used when there is cointegration present. The general consensus is that the VECM model should be used if the main focus is on cointegrating relationships while the VAR model should be used if the central focus of interest is on determining the causal relationships and short term dynamics of the model (Brandt & Williams, 2007).

In the field of economics, the VECM model is more appropriate when there are multiple variables which have the possibility of related trends. The VECM also generates better forecasts for non-stationary models especially when there is cointegration present in the model (Dolado, Gonzalo & Marmol, 1999). However these models also have their limitations. Variables which appear to have stochastic and deterministic trends in the short run may not have these features in the long run. There is also the shortcoming that the error correction method may suffer from a limited application in certain areas.

5.10 Cointegration in the VARMA$(\boldsymbol{p}, \boldsymbol{q})$ Model

The methodology on cointegration previously discussed for the VAR model can be extended to the VARMA (p, q) model. The advantage of using cointegrated VARMA (p, q) models as compared to cointegrated VAR models is that they are more parsimonious and have an improved forecasting performance (Bartel & Lütkepohl, 1998). Lütkepohl and Claessen (1997) suggested an estimation procedure which inverts the moving average component and uses a finite VAR model approximation.

5.10.1 Estimation of the Cointegrated VARMA $(\boldsymbol{p}, \boldsymbol{q})$ Model

Recall that the zero mean stationary and invertible VARMA (p, q) model is of the form

$$\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{u}_t \, .$$

$\boldsymbol{u}_t$ is assumed to be an independent white noise process with a zero mean and covariance matrix $\boldsymbol{\Sigma}_u$. In order for the model to be identifiable, the model is often of the form used in (4.13) where the vectors $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Theta}_0$ are assumed to be nonsingular (Lütkepohl, 2004).

If $\boldsymbol{y}_t$ is cointegrated with rank $K^* = K - d$, then the standard VARMA (p, q) model can be rewritten in the VECM form as

$$\Delta \boldsymbol{y}_t = \boldsymbol{\pi} \boldsymbol{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^* \, \Delta \boldsymbol{y}_{t-i} + \boldsymbol{u}_t + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{u}_{t-j} \quad . \tag{5.38}$$

$\boldsymbol{y}_t$ is a non-stationary process , $\Delta \boldsymbol{y}_t$ is a stationary process and π has reduced rank K^*.

A full rank estimation procedure using conditional maximum likelihood was first described by Yap and Reinsel (1995). The first step is to define the $K^2(p+q)$ matrix of unknown parameters $\boldsymbol{\varrho}^*$ as

$\boldsymbol{\varrho}^*$ = vec $[\boldsymbol{\pi}, \boldsymbol{\Phi}_1^*, \dots, \boldsymbol{\Phi}_{p-1}^*, \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q]$.

The estimator of $\boldsymbol{\varrho}^*$ is

$\widehat{\boldsymbol{\varrho}}^*$ = vec $[\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\Phi}}_1^*, \dots, \widehat{\boldsymbol{\Phi}}_{p-1}^*, \widehat{\boldsymbol{\Theta}}_1, \dots, \widehat{\boldsymbol{\Theta}}_q]$.

This estimator (5.39) maximises the log likelihood function

$$L(\widehat{\boldsymbol{\varrho}}^*) = \frac{-T}{2} \log|\boldsymbol{\Sigma}_u| - \frac{1}{2} \sum_{t=1}^{T} \boldsymbol{u}_t{}' \boldsymbol{\Sigma}_u{}^{-1} \boldsymbol{u}_t \, . \tag{5.39}$$

The partial derivatives of (5.39) with respect to $\boldsymbol{\varrho}^*$ are

$$\frac{dL(\widehat{\boldsymbol{\varrho}}^*)}{d\boldsymbol{\varrho}^*} = \; -\sum_{t=1}^{T} \frac{\partial \boldsymbol{u}_t{}'}{\partial \varrho^*} \boldsymbol{\Sigma}_u{}^{-1} \boldsymbol{u}_t \quad . \tag{5.40}$$

These normal equations of (5.40) are nonlinear in the parameters and iterative procedures such as the Newton-Raphson procedure discussed in Appendix A are required to be employed. The $(i+1)$th iteration of $\widehat{\boldsymbol{\varrho}}^*$ is

$$\widehat{\boldsymbol{\varrho}}^{*(i+1)} = \widehat{\boldsymbol{\varrho}}^{*(i)} + [\sum_{t=1}^{T} \frac{\partial \boldsymbol{u}_t{}'}{\partial \boldsymbol{\varrho}^*} \boldsymbol{\Sigma}_u{}^{-1} \frac{\partial \boldsymbol{u}_t}{\partial \boldsymbol{\varrho}^*}]^{-1} [\sum_{t=1}^{T} \frac{\partial \boldsymbol{u}_t{}'}{\partial \boldsymbol{\varrho}^*} \boldsymbol{\Sigma}_u{}^{-1} \boldsymbol{u}_t \,] \, . \tag{5.41}$$

$\widehat{\boldsymbol{\varrho}}^{*(i)}$ is the estimate at the ith iteration and $\widehat{\boldsymbol{\Sigma}}_u = \; T^{-1} \sum_{t=1}^{T} \widehat{\boldsymbol{u}}_t \widehat{\boldsymbol{u}}_t{}'$.

$\widehat{\boldsymbol{u}}_t$ is computed recursively from

$$\widehat{\boldsymbol{u}}_t = \Delta \boldsymbol{y}_t - \widehat{\boldsymbol{\pi}} \boldsymbol{y}_{t-1} - \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^* \, \Delta \boldsymbol{y}_{t-i} - \sum_{j=1}^{q} \widehat{\boldsymbol{\Theta}}_j \, \widehat{\boldsymbol{u}}_{t-j} \, .$$

The estimation procedure discussed is not limited to the full rank case and can be extended to the reduced rank case. In the reduced rank model, $\boldsymbol{\pi}$ can be written as $\boldsymbol{\pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(K \times K^*)$ matrices. The model in the VECM representation is

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i{}^{*}\, \Delta \boldsymbol{y}_{t-i} + \boldsymbol{u}_t + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{u}_{t-j} \,. \tag{5.42}$$

The vector $\boldsymbol{\beta}$ can be normalised as $\boldsymbol{\beta} = [\boldsymbol{I}, \boldsymbol{\beta}_0]$ where $\boldsymbol{\beta}_0$ is a matrix of dimension $(K^* \times (K - K^*))$.

Let $\boldsymbol{\tau} = \text{vec}\,[\boldsymbol{\alpha}, \widehat{\boldsymbol{\Phi}}_1{}^{*}, \dots, \widehat{\boldsymbol{\Phi}}_{p-1}{}^{*}, \widehat{\boldsymbol{\Theta}}_1, \dots, \widehat{\boldsymbol{\Theta}}_q]$ and $\boldsymbol{\tau}^* = [\boldsymbol{\beta}_0, \boldsymbol{\tau}\,']'$ where $\boldsymbol{\tau}^*$ is the vector of unknown parameters, then if a similar Newton-Raphson procedure to that performed for the full rank model is used, an estimate for $\boldsymbol{\tau}^*$ is obtained from

$$\hat{\boldsymbol{\tau}}^{*(i+1)} = \hat{\boldsymbol{\tau}}^{*(i)} + \left[\sum_{t=1}^{T} \frac{\partial \boldsymbol{u}_{t}'}{\partial \boldsymbol{\tau}^*} \boldsymbol{\Sigma}_u{}^{-1} \frac{\partial \boldsymbol{u}_t}{\partial \boldsymbol{\tau}^*}\right]^{-1} \left[\sum_{t=1}^{T} \frac{\partial \boldsymbol{u}_{t}'}{\partial \boldsymbol{\tau}^*} \boldsymbol{\Sigma}_u{}^{-1} \boldsymbol{u}_t\right] . \tag{5.43}$$

$\hat{\boldsymbol{\tau}}^{*(i)}$ is the estimate of $\boldsymbol{\tau}^*$ at the ith iteration and $\widehat{\boldsymbol{\Sigma}}_u = T^{-1} \sum_{t=1}^{T} \widehat{\boldsymbol{u}}_t \widehat{\boldsymbol{u}}_t{}'$.

Lütkepohl and Claessen (1997) showed that the stationary echelon form can be extended to that of a cointegrated VARMA (p, q) process as the representation of the process is unique. The main difference as compared to the echelon form for stationary models is that the roots of the $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ operators are reversed. This form is known as the 'reverse echelon' form and occurs if the moving average parameter is set to 0 after a restriction is placed on either the autoregressive or moving average parameters (Athanasopoulos et al., 2012).

The reverse echelon form satisfies the following restrictions

$$\theta_{kk}(L) = 1 + \sum_{i=1}^{p_k} \theta_{kk,j} L^i \qquad k = 1, \dots, K$$

$$\theta_{ki}(L) = -\sum_{i=p_k - p_{ki}-1}^{p_k} \theta_{ki,i} L^i \qquad k, i = 1, \dots, K\ (k \neq i)$$

$$\varphi_{ki}(L) = \sum_{i=0}^{p_k} \varphi_{ki,i} L^i \qquad k, i = 1, \dots, K$$

$$\varphi_{ki,i} = \theta_{ki,i} \text{ for } k, i = 1, \dots, K \,. \tag{5.44}$$

The reverse echelon form implies that the autoregressive operator is unrestricted with the exception of the restrictions imposed by the Kronecker indices and the zero order matrices $\boldsymbol{\Phi}_0 = \boldsymbol{\Theta}_0$. In addition, it also implies that there are additional restrictions placed on the moving average coefficient matrices (Lütkepohl, 2004).

The p_{ki} terms are calculated in the same way as the stationary VARMA (p, q) echelon form as

$$p_{ki} = \begin{cases} min(p_k + 1, p_i) & k \geq i \\ min(p_k, p_i) & k < i \end{cases} . \qquad k, i = 1, \dots, K \tag{5.45}$$

The estimator p is the maximum of all the Kronecker indices $p_1, \dots, p_K$ (Lütkepohl & Claessen, 1997). The maximum number of freely parameters is $2(K \sum_{k=1}^{K} p_k)$ (Athanasopoulos et al., 2012).

The echelon VARMA (p, q) model in vector error correction form is

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^* \Delta \boldsymbol{y}_{t-i} + \boldsymbol{u}_t + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{u}_{t-j} \quad (5.46)$$

$\boldsymbol{\Phi}_j^* = \sum_{i=j+1}^{p} \boldsymbol{\Phi}_i \qquad i = 1, \ldots, p-1$
and $\boldsymbol{\alpha}\boldsymbol{\beta}' = \boldsymbol{\Phi}_p + \boldsymbol{\Phi}_{p-1} + \cdots + \boldsymbol{\Phi}_1 - \boldsymbol{I} = \boldsymbol{\Phi}(1)$.

Unlike Yap and Reinsel (1995), Lütkepohl (2005) noted that the estimation of this process (5.46) is similar to that of maximum likelihood estimation with the exception that the estimated vector of parameters $\widehat{\boldsymbol{\varrho}}^*$ contains the free parameters of the VARMA (p, q) model in the reverse echelon form.

The exact maximum likelihood estimation method is less common for cointegrated VARMA (p, q) models and has only recently been given attention. Mauricio (2006) used a stationary VARMA (p, q) model and obtained estimates of the parameters by manipulating the VECM. The first step in this method is to choose an initial guess for the parameters (through say a conditional likelihood estimation method) for every parameter and represent this in the vector of parameters $\widehat{\boldsymbol{\varrho}}$ which contains all the estimates. The next step is to update the vector $\widehat{\boldsymbol{\varrho}}$ numerically using nonlinear optimisation of the log likelihood function in the stationary VARMA (p, q) model. Mauricio (2006) concluded that the exact maximum likelihood estimation procedure has the ability to reveal features in the model which cannot be found when the conditional maximum likelihood estimation procedure is used especially when the nature of non-stationarity in the data is unclear.

5.10.2 Specification of the Cointegrated VARMA $(\boldsymbol{p}, \boldsymbol{q})$ Model

The specification of the Kronecker indices was first described in Lütkepohl and Claessen (1997) in which multivariate least squares is applied to fit a VAR(h_T) process in order to obtain the residuals $\widehat{\boldsymbol{u}}_t(h_T)$. The value of h_T should ideally be between log T and $\sqrt{T}$, where T refers to the sample size. This is followed by the fitting of reverse echelon VARMA models of the form

$$\boldsymbol{y}_t = \boldsymbol{v}_t + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + (\boldsymbol{\Theta}_0 - \boldsymbol{I}_K)\widehat{\boldsymbol{u}}_t(h_T) + \boldsymbol{\Theta}_1 \widehat{\boldsymbol{u}}_{t-1}(h_T) + \cdots + \boldsymbol{\Theta}_p \widehat{\boldsymbol{u}}_{t-p}(h_T) + \boldsymbol{u}_t. \quad (5.47)$$

Under different sets of Kronecker indices $p_1, \ldots, p_K$, the model which optimises a model selection criterion such as the AIC, BIC or HQ is chosen as the specified model.

5.10.3 Testing for Cointegration in the VARMA $(\boldsymbol{p}, \boldsymbol{q})$ Model

The likelihood test statistic is defined by $\mathrm{L}(R) = \frac{|SS|}{|SS_0|}$ where $\mathrm{SS} = \sum_{t=1}^{T} \widehat{\boldsymbol{u}}_t \widehat{\boldsymbol{u}}_t{}'$ is the residual sum of squares for the unrestricted model and SS_0 is the sum of squares for the restricted

model under the restriction rank($\boldsymbol{\pi}$) $= K*$. The test was developed by Yap and Reinsel (1995) and is based on the null hypothesis rank($\boldsymbol{\pi}$) $= K*$ against the alternative hypothesis that rank ($\boldsymbol{\pi}$) $= K$. The critical values are based on asymptotic distributions that are functions of Brownian motion discussed in the cointegrated VAR case (Saikonnen & Lütkepohl, 1996). The sequential likelihood ratio tests are also permissible where the null hypothesis rank (π) $= K*$ is tested against the ranks $K-1, K-2, \dots, 0$. The smallest rank in which the null hypothesis is not rejected is taken as the rank of K.

5.11 Model Diagnostics

Diagnostic checking for non-stationary multivariate time series models is very similar to that performed for stationary models. Recall the VARMA model in error correction representation without any deterministic terms is

$$\Delta \boldsymbol{y}_t = \boldsymbol{\pi}\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1{}^*\Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^*\Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{u}_{t-j} \ . \tag{5.48}$$

The autocovariance and autocorrelation matrices are calculated in the same way as that for level VARMA models

$$\hat{\boldsymbol{\Gamma}}_u(i) = \frac{1}{T}\sum_{t=i+1}^{T} \boldsymbol{u}_t \, \boldsymbol{u}_{t-i}{}' \qquad i = 0,1, \dots, h$$

The Portmanteau statistic and the LM are also applicable for non-stationary models. The Portmanteau statistic Q_h is the same as that for the stationary VARMA (p, q) model

$$Q_h = T \sum_{i=1}^{h} tr(\hat{\boldsymbol{\Gamma}}_u(i) \, \widehat{\boldsymbol{\Gamma}_u}(0)^{-1} \hat{\boldsymbol{\Gamma}}_u(-i) \, \widehat{\boldsymbol{\Gamma}_u}(0)^{-1}) \ . \tag{5.49}$$

Unlike for stationary models however, the degrees of freedom are different in that Q_h follows approximately a $\chi^2(hK^2 - K^2(p-1) - K^2(q-1) - KK^*)$ distribution. This is because of the number of degrees of freedom are adjusted relative to the stationary VAR case (Lütkepohl, 2005). This is important as the stationary VARMA (p, q) critical values, $(hK^2 - pK^2 - qK^2)$ result in the null hypothesis being rejected too often.

The LM statistic is calculated in a similar way by using the auxilliary regression model

$$\hat{\boldsymbol{u}}_t = \boldsymbol{\pi}\boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_1{}^*\Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}{}^*\Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\Lambda}_1\hat{\boldsymbol{u}}_{t-1} + \dots + \boldsymbol{\Lambda}_h\hat{\boldsymbol{u}}_{t-h} + \boldsymbol{\epsilon}_t$$
$$t = 1, \dots, T \tag{5.50}$$

Defining the following hypotheses

H_0: $\boldsymbol{\Lambda}_1 = \cdots = \boldsymbol{\Lambda}_h = 0$
H_1 : $\boldsymbol{\Lambda}_j \neq 0$ for at least $j \, \epsilon \, \{1, \dots, h\}$

The LM statistic is computed from (5.50) as LM $= T(h - \widetilde{\boldsymbol{\Sigma}}_u \, \widetilde{\boldsymbol{\Sigma}}_R{}^{-1})$
where $\widetilde{\boldsymbol{\Sigma}}_u$ is the error covariance matrix for the unrestricted model (5.50) and $\widetilde{\boldsymbol{\Sigma}}_R$ is the error covariance matrix obtained when the null hypothesis H_0: $\boldsymbol{\Lambda}_1 = \cdots = \boldsymbol{\Lambda}_h = 0$ is true.

The LM statistic follows a χ^2 distribution with hK^2 degrees of freedom. This distribution of the LM statistic is unaffected by the presence of variables that are integrated.

5.12 Forecasting

The forecasting procedure for non-stationary VAR(p) and VARMA (p, q) models is similar to that for stationary models with the exception that the forecast error covariance matrices become increasingly unbounded as the time horizon increases. This results in the forecasts becoming uncertain in the distant future (Lütkepohl, 2004).

Engle and Yoo (1987) noted that a level VAR does not suffer from misspecification like that of a VAR in first differences although the estimation procedures do not fully estimate the parameters which are near the unit circle. They recommend that the model should be in error correction form and that long run constraints should be imposed before forecasts are performed. Reinsel (1997) noted as well that the cointegrated model in error correction form produces more accurate forecasts than models which suffer from 'over differencing'.

The forecasting performance of cointegrated VAR(p) and VARMA(p, q) models can be compared with each other by using mean square errors. This has been done by Lütkepohl and Claessen (1997) who found that a cointegrated VARMA(p, q) model has a significantly better forecasting performance as compared to that of a cointegrated VAR(p) model.

A more recent study by Kascha and Trenkler (2011) compared the forecast properties of cointegrated VAR, VARMA models and a random walk. Using mean square prediction errors, they concluded that the cointegrated VAR and VARMA models were very effective when forecasting at small time horizons but for longer time horizons, it was the random walk that was found to be more effective.

5.13 Impulse Response Analysis for Non-stationary Models

The methods discussed earlier for impulse responses are not always effective in the presence of unit roots/deterministic terms. Gospodinov (2004) proposed a method which takes into account the presence of large roots in the autoregressive dynamics of a process (i.e. roots on or near the unit circle). This involves a likelihood ratio statistic for a sequence of null hypotheses which imposes restrictions on the values of the impulse responses. The limiting distribution of the likelihood ratio statistic for models that were nearly non-stationary was derived and the acceptance region was inverted in order for interval estimates of the statistics of interest to be

obtained. From the use of Monte Carlo simulations this method was compared to the bootstrap method and was found to have a significantly better performance.

Pesavento and Rossi (2006) also proposed a similar method to construct impulse responses when there is a unit root is present. This method was developed for the purpose of longer horizons but can be modified in order to accommodate shorter time horizons. The confidence interval was constructed by inverting the acceptance region of different unit root tests such as the Dickey-Fuller test. This method was compared to a test in which the unit root was removed prior to the construction of the impulse responses.

5.14 Conclusion

The presence of non-stationarity in the model means that the data needs to be transformed or differenced in order for it to be stationary. If there is cointegration present, then the data should not be differenced but instead be remodelled in the VECM representation so that the cointegrating relations can be absorbed into the model. This representation can be used to determine the various interrelationships in the system.

CHAPTER 6

Granger-Causality

6.1 Introduction to Granger-Causality in the VAR Model

An integral aspect of multivariate time series modelling is the determination of causality among the series.

The following questions generally arise with regards to causality (Brandt & Williams, 2007)

a) What value does a variable y_1 have in predicting other variables y_2 in a system of equations?
b) Is the variable y_1 exogenous in a time series model with respect to the other variables?

Granger and Newbold (1986), state that a few assumptions must be made before these questions are answered:

(i) The determination of causality is only possible when the past causes the present or future and hence the future cannot be responsible for causing the past.
(ii) A cause contains unique information which is not available elsewhere.

A statement which takes into account both a models forecasting ability and causality was proposed by Granger and Newbold (1986). They state that if a variable say y_1 improves the forecasting performance of another variable say y_2 then it is said that y_1 has Granger-caused y_2. In other words y_1 has Granger-caused y_2 if the current value of the variable y_2 is predicted more accurately by the past values of both y_1 and y_2 rather than by using the past values of y_2 alone. On the contrary, if y_1 has not been helpful in improving the forecast performance of y_2, then it has been said to have failed to Granger-cause y_2 .

The above statement can be expressed in the form of linear functions and mean square errors by saying that y_1 fails to cause y_2 if

$$MSE\ [E(\ y_{2,t+h}|y_{2,t}\ ,\ y_{2,t-1},\dots)] = MSE\ \left[E\left(y_{2,t+h}\middle|\ y_{2,t},\ y_{2,t-1},\dots,y_{1,t},\ y_{1,t-1},\dots\right)\right]. \qquad (6.1)$$

The relation (6.1) is interpreted as y_1 fails to cause y_2 if the variance of the forecast error of y_2 obtained from using past values of y_1 and itself is the same as the variance of forecast error y_2 obtained by only using past values of itself and not y_1.

It is important to note that if the condition of Granger-causality holds, it does not necessarily mean that one variable causes another variable, rather it means that the forecasting ability of one variable is improved if the other variables are included.

The concept of Granger-causality and forecasting are explained in the example used in Hamilton (1994). Consider the following bivariate model,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \varphi_{11,1} & \varphi_{12,1} \\ \varphi_{21,1} & \varphi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varphi_{11,2} & \varphi_{12,2} \\ \varphi_{21,2} & \varphi_{22,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \cdots +$$

$$\begin{bmatrix} \varphi_{11,p} & \varphi_{12,p} \\ \varphi_{21,p} & \varphi_{22,p} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}. \quad (6.2)$$

The error terms u_{1t} and u_{2t} are independent and identically distributed and are white noise while c_1 and c_2 are constant terms.

In this example, y_2 has not Granger-caused y_1 if the coefficient matrices $\begin{bmatrix} \varphi_{11,j} & \varphi_{12,j} \\ \varphi_{21,j} & \varphi_{22,j} \end{bmatrix}$ are lower triangular for all values of $j > 0$,

i.e. the matrices are of the form $\begin{bmatrix} \varphi_{11,j} & 0 \\ \varphi_{21,j} & \varphi_{22,j} \end{bmatrix}$ for $j > 0$.

Thus if y_2 has not Granger-caused y_1, then the model (6.2) will be of the form

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \varphi_{11,1} & 0 \\ \varphi_{21,1} & \varphi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varphi_{11,2} & 0 \\ \varphi_{21,2} & \varphi_{22,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \cdots +$$

$$\begin{bmatrix} \varphi_{11,p} & 0 \\ \varphi_{21,p} & \varphi_{22,p} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}. \quad (6.3)$$

(6.3) can be expressed in backshift/lag order notation form as

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} \begin{bmatrix} \varphi_{11}(L) & \varphi_{12}(L) \\ \varphi_{21}(L) & \varphi_{22}(L) \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \quad (6.4)$$

$$\varphi_{11}(L) y_{1,t} = \varphi_{11,1} y_{1,t-1} + \cdots + \varphi_{11,p} y_{1,t-p}$$

The one step ahead forecast of $y_{2,t}$ which is conditional on its past and present values is

$$\hat{E}(y_{2,t+1} | y_{2,t}, y_{2,t-1}, \dots, y_{1,t}, y_{1,t-1}, \dots) = c_1 + \varphi_{11,1} y_{2,t} + \varphi_{11,2}\, y_{2,t-1} + \cdots + \varphi_{11,p}\, y_{2,t-p+1} + u_{1,t+1} \quad (6.5)$$

It can easily be seen from (6.5) that the forecast of $y_{2,t}$ is dependent only on its own lagged values and not lagged values of $y_{1,t}$.

Similarly the two step ahead forecast of $y_{2,t}$ is

$$\hat{E}(\, y_{2,t+2}|y_{2,t+1}, y_{2,t}, y_{2,t-1}, \dots, y_{1,t}, y_{1,t-1}, \dots) = c_1 + \varphi_{11,1}\, y_{2,t+1} + \varphi_{11,2}\, y_{2,t} + \cdots + \varphi_{11,p}\, y_{2,t-p+2} + u_{1,t+2} \quad (6.6)$$

As with the one step ahead forecast, the two step ahead forecast is based only on lagged values of itself and not lagged values of $y_{1,t}$. From the principals of mathematical induction, the same is true for a h based ahead forecast and in general y_2 does not Granger-cause y_1 for $\varphi_{11,j} = 0$ $j = 1, \dots, p$.

This methodology can be extended into the multivariate case when there are more than 2 time series. For example, consider the trivariate model below

$$\begin{bmatrix} \varphi_{11}(L) & \varphi_{12}(L) & \varphi_{13}(L) \\ \varphi_{21}(L) & \varphi_{22}(L) & \varphi_{23}(L) \\ \varphi_{31}(L) & \varphi_{32}(L) & \varphi_{33}(L) \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \end{bmatrix}.$$

For this model, y_i does not Granger-cause y_j if and only if $\varphi_{12}(L) = 0$. In general for a model with K variables, y_i does not Granger-cause y_j if and only if $\varphi_{ji}(L) = 0$. Boudjellaba, Dufour and Roy (1992) noted that the results of a causality analysis for bivariate models do not necessarily correspond to those models of a dimension larger than 2.

6.2 Testing for Granger-Causality in Stationary VAR(p) Models

Various methods used for testing Granger-causality have been described in the literature for both stationary and non-stationary VAR(p) models. The methods used for testing for Granger-causality for stationary models will be discussed first.

a. The Granger Regression Method

The most common method used for testing Granger-causality is the Granger regression method.

Consider the bivariate model of lag length p

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \varphi_{11,1} & \varphi_{12,1} \\ \varphi_{21,1} & \varphi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varphi_{11,2} & \varphi_{12,2} \\ \varphi_{21,2} & \varphi_{22,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \cdots +$$

$$\begin{bmatrix} \varphi_{11,p} & \varphi_{12,p} \\ \varphi_{21,p} & \varphi_{22,p} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}.$$

Suppose that one wants to determine whether y_1 Granger-causes y_2. In this case, the equation where y_2 at time t is regressed on past values of itself and y_1 is considered, i.e.

$$y_{2,t} = c_2 + \varphi_{22,1}y_{2,t-1} + \varphi_{22,2}y_{2,t-2} + \cdots + \varphi_{22,p}y_{2,t-p} + \varphi_{21,1}y_{1,t-1} + \varphi_{21,2}y_{1,t-2} + \cdots + \varphi_{21,p}y_{1,t-p} + u_{2t}$$

$$= c_2 + \sum_{i=1}^{p} \varphi_{22,i}y_{2,t-i} + \sum_{i=1}^{p} \varphi_{21,i}y_{1,t-i} + u_{2t} \quad (6.7)$$

The error term u_{2t} is assumed to be white noise. The choice of the order of p is usually obtained from the information criteria using the methods described in section 2.4 and is usually set quite high. Under the assumption that all the variables in a VAR are stationary, a F test or a χ^2 test can be perfomed in order to test the null hypothesis of non-causality, i.e. there are no causal relationships among the variables. y_1 is said to have Granger-caused y_2 if the coefficients of lagged values of itself are zero when y_2 is regressed on the lagged values of itself and y_1. Thus the null and alternate hypotheses are defined as

H_0: y_1 does not Granger-cause y_2 if $\varphi_{21,1} = \varphi_{21,2} = \cdots = \varphi_{21,p} = 0$
H_1: y_1 does Granger-cause y_2 if $\varphi_{21,1} \neq 0$ or $\varphi_{21,2} \neq 0$ or … or $\varphi_{21,p} \neq 0$.

In the F test, two regression models are run, an unrestricted model such as the model used in (6.7) and a restricted model which is subject to the constraints that are stated in the null hypothesis i.e. $\varphi_{21,1} = \varphi_{21,2} = \cdots = \varphi_{21,p} = 0$

This restricted model is of the form

$$y_{2,t} = c_0 + \varphi_{22,1}y_{2,t-1} + \varphi_{22,2}y_{2,t-2} + \cdots + \varphi_{22,p}y_{2,t-p} + v_{2t}\,. \quad (6.8)$$

The first step is to compute the residual sum of squares of the unrestricted model (6.7) as

RSS(unrestricted) $= \sum_{t=1}^{T} \hat{u}_{2t}{}^2$.

The residual sum of squares for the restricted model (6.8) is

RSS(restricted) $= \sum_{t=1}^{T} \hat{v}_{2t}{}^2$.

The residual sum of squares for both the restricted and unrestricted models are used to compute the F_1 statistic as

$$F_1 = \frac{\frac{RSS(restricted) - RSS(unrestricted)}{p}}{\frac{RSS(unrestricted)}{T - 2p - 1}}$$

$$= \frac{\sum_{t=1}^{T} \hat{v}_{2t}{}^2 - \sum_{t=1}^{T} \hat{u}_{2t}{}^2/p}{\sum_{t=1}^{T} \hat{u}_{2t}{}^2/T - 2p - 1}\,. \quad (6.9)$$

This statistic (6.9) follows a F distribution with $p, T - 2p - 1$ degrees of freedom. If the F_1 statistic is greater than $p, T - 2p - 1$ at a 5 percent level of significance, then the null hypothesis that y_1 does not Granger-cause y_2 is rejected and it is concluded that y_1 does in fact Granger-cause y_2. The Granger regression method is advantageous in that it is unaffected by the ordering of the VAR system (Friedman & Shachmurove, 1997).

The test to check if y_2 Granger-causes y_1 is performed in a similar manner by considering the regression of y_1 with lagged values of itself and y_2 under the assumption of an autoregressive process with p lags

i.e. $y_{1,t} = c_1 + \varphi_{11,1} y_{1,t-1} + \cdots + \varphi_{11,p} y_{1,t-p} + \varphi_{12,1} y_{2,t-1} + \cdots + \varphi_{12,p} y_{2,t-p} + u_{1,t}$. (6.10)

The hypotheses are defined as

H_0: y_2 does not Granger-cause y_1 if $\varphi_{12,1} = \varphi_{12,2} = \ldots = \varphi_{12,p} = 0$
H_1: y_2 does Granger-cause y_1 if $\varphi_{12,1} \neq 0$ or $\varphi_{12,2} \neq 0$ or … or $\varphi_{12,p} \neq 0$.

The restricted model that is subject to the constraints stated in the null hypothesis is

$y_{1,t} = c_1 + \varphi_{11,1} y_{1,t-1} + \cdots + \varphi_{11,p} y_{1,t-p} + v_{1,t}$.

This rest of the procedure is performed in a similar manner to that which was conducted for testing whether y_1 Granger-causes y_2 by using the residual sum of squares of the restricted and unrestricted models in order to compute the statistic (6.9) and to compare it to the critical values of a F distribution with $p, T - 2p - 1$ degrees of freedom.

An asymptotically equivalent test which uses the χ^2 distribution as critical values was given by Hamilton (1994) where

$$F_2 = \frac{T[RSS(restricted) - RSS(unrestricted)]}{RSS(unrestricted)} \sim \chi^2(p) \, . \quad (6.11)$$

The null hypothesis of non-causality that y_1 does not Granger-cause y_2 at a 5 percent level of significance is rejected if the value of F_2 is greater than the critical values of a $\chi^2(p)$ distribution.

Although the F tests are simple to use, their power is limited because there are generally a large number of lags in the variables of a VAR model. This occurs especially when the numerator and denominator degrees of freedom approach the same value (Brandt & Williams, 2007).

b. The Sims Method

An alternative test for Granger-causality known as the Sims method is based on the premise that the results from a present test cannot be caused by the future. This test states that if y_1 is projected linearly as an equation of the past, present and future values of y_2, then a joint test for significance on the coefficients of the future values of y_2 can determine whether y_1 Granger-causes y_2 .

This is expressed mathematically by considering the equation

$$y_{2,t} = c_1 + \sum_{j=0}^{\infty} a_j y_{1,t-j} + \sum_{j=1}^{\infty} b_j y_{1,t+j} + \epsilon_{2t} . \quad (6.12)$$

The term ϵ_{2t} refers to the error of $y_{2,t}$. y_1 is said to have not Granger-caused y_2 if the coefficients of all future $y_{1,t}$ terms are zero ($b_j = 0$ for $j = 1,2, \dots$) and if the coefficients of all the past $y_{1,t}$ terms are nonzero.

The limitation of equation (6.12) is that the error term ϵ_{2t} is generally autocorrelated with the dependent variables which results in the hypothesis $b_j = 0$ not being valid. A possible solution that has been proposed by Geweke, Meese and Dent (1983) which makes a slight modification to this equation by the addition of lagged values of y_2 to the model. The resulting equation is

$$y_{2,t} = c_1 + \sum_{j=1}^{\infty} \zeta_j y_{1,t-j} + \sum_{j=0}^{\infty} a_j y_{1,t-j} + \sum_{j=1}^{\infty} b_j y_{1,t+j} + \omega_{2t} . \quad (6.13)$$

In this instance, the error term ω_{2t} is uncorrelated with the dependent variables and is white noise (Granger & Newbold, 1986). The null hypothesis for non-causality is that y_1 does not Granger-cause y_2 if $b_1 = b_2 = \cdots = 0$. The same procedure as that done for the Granger regression method can now be used.

c. The Wald Test

A Wald test was developed by Granger and Newbold (1986) where the statistic GRNW is defined as

$$\text{GRNW} = T\, \boldsymbol{f}(\widehat{\boldsymbol{\varrho}})' \, \boldsymbol{\Sigma}_f{}^{-1} \boldsymbol{f}(\widehat{\boldsymbol{\varrho}}) . \quad (6.14)$$

Suppose $\boldsymbol{\varrho}$ refers to the vector of autoregressive parameters where the unconstrained estimator of $\boldsymbol{\varrho}$ is $\widehat{\boldsymbol{\varrho}}$. The set of $\boldsymbol{\varrho}$ non-causality constraints on the model parameters is denoted by $\boldsymbol{f}(\widehat{\boldsymbol{\varrho}})$ which has an estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\varrho}$. $T^{\frac{-1}{2}}\, \boldsymbol{f}(\widehat{\boldsymbol{\varrho}})$ has a zero mean with an estimated covariance matrix

$$\widehat{\boldsymbol{\Sigma}}_f = [\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\varrho}'}]\, \widehat{\boldsymbol{\Sigma}}_{\varrho}\, [\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\varrho}'}]' . \quad (6.15)$$

$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\varrho}'}$ is a $p \times p$ matrix in which the (i, j)th element is the partial derivative of the ith element of $\boldsymbol{f}(\widehat{\boldsymbol{\varrho}})$ with respect to the jth element of $\widehat{\boldsymbol{\varrho}}$. Under the null hypothesis of non-causality, the statistic follows a χ^2 distribution with p degrees of freedom. It is necessary that $[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\varrho}'}]$ must be of full rank however since this is not always the case, a sequential procedure can be used (Boudjellaba et al., 1992).

6.3 Testing for Granger-Causality in Non-stationary VAR $(\boldsymbol{p})$ Models

Testing for Granger-causality in stationary models only has been discussed thus far. The testing for Granger-causality in the non-stationary case (i.e. the instance in which there are unit roots in the VAR model) is as important. This will be discussed next.

a. The F Test

If one or more of the variables present have a unit root, the test statistics for the model parameters will have a non-standard distribution since they are related to the case when an $I(1)$ variable is regressed on a stationary variable. Enders (2004) noted that the F test used for the stationary case can be used if the non-stationary causal variable can be made to appear in first differences. This follows from the theory based on Sims, Stock and Watson (1990) in which the distribution of the F test depends on the nuisance parameters.

The models that are tested for a bivariate series are,

$$\Delta y_{1,t} = v_0 + \sum_{i=1}^{p} {\varphi_{11,i}}^* \Delta y_{1,t-i} + \sum_{i=1}^{p} {\varphi_{12,i}}^* \Delta y_{2,t-i} + \pi_1 y_{1,t-1} + u_{1,t}$$

$$\Delta y_{2,t} = v_0 + \sum_{i=1}^{p} {\varphi_{22,i}}^* \Delta y_{2,t-i} + \sum_{i=1}^{p} {\varphi_{21,i}}^* \Delta y_{1,t-i} + \pi_2 y_{2,t-1} + u_{2,t}\ .$$

v_0 is a deterministic term, π_1 and π_2 are error correction terms, while $u_{1,t}$ and $u_{2,t}$ are white noise error terms.

Oxley and Greasley (1998) noted that $\Delta y_{2,t}$ Granger-causes $\Delta y_{1,t}$ if H_0: ${\varphi_{12,1}}^* = {\varphi_{12,2}}^* = \cdots = {\varphi_{12,p}}^*$ is rejected against the alternative of H_1: at least one ${\varphi_{12,i}}^* \neq 0$ for $1 = 1, \ldots, p$ or if $\pi_1 \neq 0$.

Similarly $\Delta y_{1,t}$ Granger-causes $\Delta y_{2,t}$ if H_0: ${\varphi_{21,1}}^* = {\varphi_{21,2}}^* = \cdots = {\varphi_{21,p}}^*$ is rejected against the alternative of H_1: at least one ${\varphi_{21,i}}^* \neq 0$ for $1 = 1, \ldots, p$ or if $\pi_2 \neq 0$.

The distribution of this test cannot be tabulated but can be computed numerically as the nuisance parameters are able to be estimated consistently. This can only be possible if there is no cointegration present in the model otherwise there will be a rank deficiency in matrix coefficients due to the non-invertibility of moving average components (Saidi, 2007). It is thus

not advisable to test for Granger-causality in the event that cointegration is present in the model.

b. Wald Test

Wald tests for causality were proposed by Toda and Phillips (1993) in cointegrated systems in which two level VAR models were constructed using $I(1)$ variables as well as a vector error correction model (estimated by using maximum likelihood). For the case of the level VAR models constructed using $I(1)$ variables, in order for a causality test to be valid asymptotically, a condition that requires sufficient cointegration to be present needs to be satisfied. This is with respect to the variables whose causal effects are being tested. This condition involves placing a rank restriction on a sub matrix of the cointegrating matrix $\boldsymbol{\pi}$. This rank condition suffers from simultaneous equation bias and as a result, the Wald test should not be used in level VAR's.

Toda and Phillips (1993) also conducted Wald tests for Vector Error Correction models (estimated using maximum likelihood estimation). In this case there is also a rank condition for sufficiency based on a sub matrix of the cointegrating matrix. The estimates of a Vector Error Correction model are generally easier to construct than the case of level VAR estimation when cointegration is present and as a result, the Wald tests for Granger-causality are asymptotically valid χ^2 tests. Oxley & Greasley (1998) noted that if the rank condition fails, the distribution will then be a mixture of a χ^2 and a nonstandard distribution. Kumar, Webber and Perry (2009) suggested that the Wald tests are useful for determining the short run causal effects but recommend that the long run causal effects should be determined by testing for the significance of the lagged error term $\boldsymbol{\pi}$.

Chigira and Yamamoto (2003) noted the shortcomings of the Toda and Phillips (1993) method of cointegrated level VAR models that were constructed by using $I(1)$ variables and proposed two different approaches in order to test for Granger-causality. As in the study by Toda and Phillips (1993) they found that if the rank condition is not satisfied, the relevant matrix in the Wald statistic will be degenerate (there is an inability to obtain the rank) and this will result in the Wald statistic having an asymptotically non-standard distribution. Two procedures were used in order for the rank of this matrix to be obtained, one which was referred to as the lag augmented VAR approach and the other referred to as the generalised inverse approach. The two methods were compared to each other using the Monte Carlo simulation method and the generalised inverse method was found to have greater power in obtaining the rank of the matrix.

c. The Brandt and Williams Method

The method proposed by Brandt and Williams (2007) for testing for Granger-causality in non-stationary models is based on a similar method initially used by Toda and Yamamoto (1995). According to this method, if there are T variables in the model, then the maximum number of roots that are present (maximum order of integration that might be suspected) in the model is

T_1 which is a value less than or equal to T. In order to test for causality, a model with $p + T_1$ lags is estimated and a Granger-causality test (such as the Hamilton method used in the stationary case) is performed on that particular model. As is the case for the Hamilton method the test statistic follows a χ^2 distribution with p degrees of freedom.

This procedure has a major advantage in that there does not have to be any prior knowledge of the cointegration properties of the system and is applicable even if there is no evidence of cointegration present or if the rank conditions are not specified. The limitation is that there is a loss of power since the VAR's are deliberately over fitted. This limitation is dependent on the size of the model. If there are many variables and the lag length is small, then the addition of just one lag might lead to a large amount of inefficiency in the parameter estimates. If there are a small number of variables and a large lag length, then the cost of adding a few more lags will not affect the model greatly. Brandt and Williams (2007) have therefore suggested that this method should not be used on its own, but should rather be used to compliment other methods.

6.4 Granger-Causality for VARMA(p, q) Models

A VAR(p) model which requires the estimation of a large number of parameters may result in a loss of power. In this instance it is worth testing for Granger-causality for VARMA(p, q) models. Testing for Granger-causality in a VARMA(p, q) model however is usually more complicated than for pure VAR(p) models (Boudjellaba et al., 1992). Consider the example of a bivariate, no intercept VARMA(p, q) model described in Granger and Newbold (1986)

$$\begin{bmatrix} \varphi_{11}(L) & \varphi_{12}(L) \\ \varphi_{21}(L) & \varphi_{22}(L) \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \theta_{11}(L) & \theta_{12}(L) \\ \theta_{21}(L) & \theta_{22}(L) \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}, \tag{6.16}$$

where $\varphi_{ij}(L) = \varphi_{ij,1}L - \cdots - \varphi_{ij,p}L^p$ and $\theta_{ij}(L) = \theta_{ij,1}L - \cdots - \theta_{ij,q}L^q$

Under the assumption that the process is invertible, (6.16) can be rewritten as

$$|\varphi(L)| \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \varphi_{11}(L) & \varphi_{12}(L) \\ \varphi_{21}(L) & \varphi_{22}(L) \end{bmatrix}^{-1} \begin{bmatrix} \theta_{11}(L) & \theta_{12}(L) \\ \theta_{21}(L) & \theta_{22}(L) \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$= \begin{bmatrix} \varphi_{22}(L)\theta_{11}(L) - \varphi_{21}(L)\theta_{21}(L) & -[\varphi_{12}(L)\theta_{22}(L) - \varphi_{22}(L)\theta_{12}(L)] \\ -[\varphi_{21}(L)\theta_{11}(L) - \varphi_{11}(L)\theta_{21}(L)] & \varphi_{11}(L)\theta_{22}(L) - \varphi_{21}(L)\theta_{12}(L) \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{6.17}$$

The direction of causality between y_1 and y_2 is the same as that between $|\varphi(L)|y_{1t}$ and $|\varphi(L)|y_{2t}$. Thus y_2 does not cause y_1 if and only if

$$\varphi_{12}(L)\theta_{22}(L) - \varphi_{22}(L)\theta_{12}(L) = 0. \tag{6.18}$$

If the above condition (6.19) holds but y_1 does not Granger-cause y_2, then the variable y_1 is said to be exogenous (Box et al., 2008).

In general for a stationary, invertible and K dimensional VARMA (p, q) process, $\boldsymbol{\Phi}(L)\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{u}_t$, y_i does not Granger-cause y_j if and only if $\det(\varphi_i(z), \Theta_{(j)}(z)) = 0$ where $\varphi_i(z)$ is the ith column of the matrix $\boldsymbol{\Phi}(z)$ and $\Theta_{(j)}(z)$ is the matrix $\boldsymbol{\Theta}(z)$ without its jth column (Boudjellaba et al., 1992).

The Wald test derived by Granger and Newbold (1986) that was discussed for VAR(p) models can also be used to test for causality in the VARMA(p, q) model with the main difference being that $\boldsymbol{\varrho}$ is now the vector of autoregressive and moving average parameters, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\varrho}'}$ is a $(p+q) \times (p+q)$ matrix instead of a $p \times p$ matrix and the Wald statistic follows a $\chi^2(p+q)$ distribution instead of a $\chi^2(p)$ distribution.

A log likelihood test has also been developed by Boudjellaba et al. (1992). If $L(\widehat{\boldsymbol{\varrho}})$ denotes the log likelihood function of $\boldsymbol{\varrho}$ and if $\widetilde{\boldsymbol{\varrho}}^*$ is the maximum likelihood estimate of $\widehat{\boldsymbol{\varrho}}$ under the restrictions that $f(\widehat{\boldsymbol{\varrho}}) = 0$, then the likelihood ratio test statistic is

$$\text{LR} = 2[\, L(\widehat{\boldsymbol{\varrho}}) - L(\widetilde{\boldsymbol{\varrho}}^*)]. \qquad (6.19)$$

The above authors showed that the likelihood ratio test statistic is asymptotically equivalent to the Wald statistic and follows a χ^2 distribution with $p+q$ degrees of freedom.

6.5 Granger-Causality at Long Forecast Horizons

With reference to longer forecast horizons, it is important to note that if there has not been any causality at horizon 1, it does not necessarily mean that causality has not occurred at horizon 2. This was discussed in detail in the paper by Dufour and Taamouri (2010). They propose that a simulation based technique should be used in order to obtain causality measures for any time horizons that are greater than 1. In this technique, the first step required is to simulate a large sample by using an unconstrained model in which the parameters are already known. The next step is to simulate a sample from a constrained model in which the condition for non-causality has already been imposed. The variance – covariance matrices for these unconstrained and constrained forecasting errors at horizon h are then computed and compared in order to detect Granger-causality.

6.6 Granger-Causality and Confounding Variables

Granger-causality can also be tested in the event of a confounding variable, i.e. when two variables appear to be the cause of each other but it is in fact a third and hidden variable that is

responsible for the causation of both the variables. Asghar (2008) used different Granger-causality tests to test the performance of a confounding variable. In all the simulation experiments it was observed that the association between the two variables was in fact due to the presence of another variable.

6.7 Limitations of Granger-Causality

A finding of Granger-causality in the model needs to be met with some caution because:

a. Granger-causality is only applicable if the coefficients of the lagged variables are nonzero, i.e. it does not explicitly specify the direction of the relationship. It is possible that there could be more than one direction of causality in the model, i.e. y_1 causes y_2 as well as y_2 causes y_1. This is known as a feedback relationship and is not uncommon especially in the financial sector (Brandt & Williams, 2007). If there is no evidence of a causal relationship between y_1 and y_2 as well as y_2 and y_1, then the variables y_1 and y_2 are said to be independent of each other. It is therefore important to understand the theory upon which the finding is based before assuming Granger-causality.

b. Although Granger-non-causality infers that the past values of a series are not predictive of each other, it does not necessarily mean that the different series in a multiple time series are uncorrelated. If there is a correlation between the series, then even if no Granger-causality takes place, the innovations/shocks will still be highly correlated.

c. The problem of misspecification should also be considered. If there are too few lags included in the VAR model, then the VAR estimates will be biased and inefficient. The lags omitted will lead to the residuals being serially correlated. This results in the null hypothesis for Granger-non-causality being rejected more times than it should and Granger-causality may be assumed in the model when it is not actually present (Brandt & Williams, 2007).

d. It is also possible that there have been too many lags included in the VAR model. If this is the case, the resulting estimates will not be efficient, even in the event that they are unbiased. This leads to the possibility of a type 1 error, i.e. failing to reject the null hypothesis when it is in fact true. In contrast to the case where too few lags are fitted however, this result does not affect the testing for non-causality.

e. A causal model is not necessarily the best model when it comes to analysing the overall fit of the model. A study by Lanne and Saikonnen (2009) showed that non-causal models can explain the data better than those models where Granger-causality is present.

6.8 Conclusion

Granger-causality is useful for determining the direction of causality in a multivariate process especially for bivariate series. It is also useful for the prediction of future values of a variable based on its past values.

CHAPTER 7

A Review of the Literature on Model Selection and the Application of Multivariate Time Series Modelling

Data generated from many clinical, political science, geographical, economics and environmental science studies can be constituted into a time series. Multivariate time series analysis is of particular use in the fields of economics because time dependent financial and economic data is frequently available to calculate future risk. In order to illustrate practical applications of the technique and to emphasise the utility of the different models, a brief literature review of studies which employed multivariate time series modelling was undertaken.

7.1 Applications of Multivariate Time Series in the Discipline of Economics

The most widespread application of multivariate time series analysis in economics is in the field of macro economics.

James, Koreisha and Povteh (1985) used a VARMA model to investigate the relationship between stock returns, real output and nominal interest rates. They found a strong association between stock returns, expected real output and the growth rate in the monetary base. They also found that expected changes in the real output activity and money supply growth are important predictors of the changes in the expected values of inflation.

Fackler and Krieger (1986) studied the sensitivity of the forecasting performance of various univariate and multivariate time series models such as the ARIMA, VAR and VARMA models on macroeconomic variables such as money stock, interest rate, GDP deflator and total domestic non financial liabilities. They made use of the forecasting errors for each model and came to the conclusion that there is evidence to suggest that the VARMA technique had the potential to outperform the ARIMA and VAR models in terms of forecasting.

Haden and VanTassel (1988) used a VAR model in order to determine if there were any relationships present between the different variables (price of milk, price of daily ration, number of cows, production per head , price of daily cows) of the US dairy sector at different time lags. The price of milk responded quickly to a shock on itself. However, it took a few months for the number of dairy cows to show a significant response to the milk price.

Chen and Lee (1990) used a VARMA model to investigate the dynamic relationship between prices and interest rates. The authors found some evidence that prices are influenced by interest rates but could not find any evidence that interest rates are influenced by prices.

Grubb (1992) compared various VAR and VARMA models at three different locations by making use of monthly US flour price data from August 1972 to November 1980. His objective was to determine if there was any relationship amongst the various locations. The forecast errors for each model at different time lags showed that the VAR(2) model had the best fit.

Simkins (1995) compared the performance of two different VAR models to forecast macroeconomic time series, one of which was unrestricted and the other subjected to business cycle restrictions. Using the Theli U statistics, he found that the business cycle restrictions only lead to a small improvement in the forecasting accuracy of the model.

Stergiou and Christou (1995) compared the forecasting performance of three models:, the regression model, univariate and multivariate time series, to analyse annual fisheries catches. They found that the regression model fitted the data better than both the time series models, as they had a smaller mean percentage error. However, in terms of actual forecasting performance, there was inconclusive evidence to suggest any model was better.

Ansari (1996) used a vector autoregressive approach in order to compare the effects of monetary policy with that of fiscal policy. He investigated the effects of each policy on the GDP, money supply, government expenditure and national income and by modelling them as VAR equations and using F tests to test for the significance of the lagged polynomials. The fiscal policy was found to yield more significant results and was hence more effective.

Kulshreshtha and Parikh (2000) used a vector autoregressive model in order forecast coal demand in India, as well as to determine whether there was a relationship between the coal price and power, cement and steel. Their findings were that the coal prices were exogenous in all the sectors (does not depend on any variables) except cement.

Veenstra and Haralambides (2001) used a multivariate autoregressive model to forecast sea-borne trade flows in four community markets (The crude oil, iron ore, grain and coal markets) on the major trade routes. They also tested whether a vector autoregressive model was accurate in producing long term seaborne trade flow estimates. They came to the conclusion that the model was suitable because the mean square forecasting errors were small.

Tahir and Ghani (2004) applied a VAR technique on yearly Bahraini data (1971 – 2002) in order to determine the interrelationships between five macroeconomic variables (Money supply, GDP, government expenditures, oil prices and CPI) .They found that all of the macroeconomic variables are linked, and that they all influence each other to a certain extent. They also used impulse responses in order to see whether fiscal or monetary policy is more effective. The

impulse response curves showed evidence that fiscal policy is more effective in the short run while monetary policy is more effective in the long run.

Chien, Lee and Tsai (2006) made use of a VARMA model to determine if there is an association between the monthly values of sales and stock prices of Taiwan. The VARMA model suggested that there was a uni-directional relationship between sales figures and the stock prices but that there was no evidence of the stock prices influencing sales.

Hanson (2006) used a vector autoregressive model to compare different monetary policy regimes in the USA pre 1984 and post 1984 in the USA. He found that most of the changes in volatility were attributed to breaks in the non policy portion of the structural VAR.

Papaikonomou and Pires (2006) used a vector autoregressive model to determine if there was sufficient evidence to suggest that the US output expectations were unbiased. They used impulse responses in order to conclude that the expectations were unbiased in the long run.

Kargbo (2007) compared four different time series models, the VAR, ARIMA, Euler Granger single equation and Vector Error Correction Models to forecast agricultural exports and imports in South Africa. Using the Theli U statistic, he found that the univariate ARIMA and Euler Granger methods outperformed that of the VAR and VECM

Marcucci and Quagliariello (2008) used a vector autoregressive model in order to analyse the extent to which macroeconomic shocks affect the banking sector. They constructed impulse response functions and found that shocks have significant impacts on the banking sector.

Raghavan et al. (2009) compared three different multivariate time series models, the VAR, SVAR (Structured VAR) and VARMA models in order to analyse the Malaysian monetary policy. The variables used were money supply, effective rate, industrial production, consumer price index and overnight interbank rate. The impulse response curves were constructed for each model in order to determine the effect of a shock on each variable. Of all the models in the study, the VARMA model produced the results which were the most consistent with prior theoretical expectations.

Gupta, Jurgilas and Kabundi (2010) used a factor augmented vector autoregressive approach in order to determine the impact of monetary policy on the real house price growth in South Africa from the first quarter of 1980 to the final quarter of 2006. They made use of impulse response functions, and found that in general a monetary policy shock triggers a negative response in house price inflation. However, there were a range of varied responses amongst the various sectors of the housing market. The luxury, large middle and medium middle segments showed a significant response to a shock in the monetary policy while the response of the small middle and affordable housing segments was minimal.

7.2 Applications of Multivariate Time Series in the Discipline of Natural Sciences

Hagnell (1991) compared two separate models, a VAR model and a VARMA model to study the relationships between fertility, mortality for ages 20 - 50 years (which they referred to as adult mortality), nuptiality and real wages in Sweden in the period 1751 -1850. A VARMA (1,1) model was identified by analysing the cross correlation function. In order to reduce the number of parameters in the model, he also fitted a restricted VARMA (1,1) model by setting the insignificant parameters in the original model to zero. The results indicated that fertility is influenced by past real wages and past nuptiality. Adult mortality was found to depend only on real wages while nuptiality was influenced by both lagged values of real wages and lagged values of adult mortality. The real wages were found to be exogenous or independent of the other variables. A VAR(2) model was also identified and as in the previous case of the VARMA$(1,1)$ model, the insignificant parameters were set to zero in order to obtain a more parsimonious model. This model yielded results similar to that of the VARMA (1,1) model with the exception that adult mortality was influenced by past nuptiality. Using mean square errors, the author showed that the VARMA model provides more accurate forecasts than the VAR.

Chin (1995) used a multivariate time series in order to determine the variation in the monthly and annual rainfall in South Florida. For monthly rainfall, it appeared that the deviations were caused by regional scale phenomena that had a temporal structure while the majority of the variance for annual rainfall was associated with regional scale phenomena that were randomly and normally distributed.

Lu (2001) used a vector autoregressive approach in order to describe the dynamics of the US population between 1910 and 1990, and to investigate whether there were any associations between the total population, birth rate, immigration and GDP per capita variables. The results indicated that the population of the USA was dependent on its historical population as well as the past values of birth rate and immigration, but did not show any signs of dependency on GDP per capita.

Grimaldi, Tallerini and Serinaldi (2005) applied the commonly used VAR(1) model as well as the more optimal VAR (p) models and in order to analyse daily rainfall series. The study showed that rainfall series can be simulated by modelling and that it is more favourable to use the general VAR (p) models instead of the basic VAR(1) one.

Gan (2006) used a vector autoregressive modelling approach in order to investigate the causality relationships amongst wildfire, the El Nino Southern Oscillation (ENSO), timber harvest and urban sprawl. The author found that an individual factor may not affect wildfire activity when acting alone but can be significantly influential when coupled with other factors. There was also evidence of a feedback effect of wildfire activity on the other variables. Impulse responses were also constructed and the results revealed that the wildfire activity was more responsive to a shock in the urban population density than that of the other variables.

Ewing, Riggs and Ewing (2007) used a VAR model to investigate the dynamic relationships of a predator prey system as well as to analyse the responses of this system to unexpected shocks. The population densities of predator and prey were studied and the results indicated that there were significant responses of the population density growth of the predator to shocks in the growth rate of the prey (and vice versa). This result is not generally found when normal regression modelling is used.

Ewing et al. (2007) used a VAR model to examine whether there was a relationship between the wind speeds at a particular location at different heights (13, 33, 70 and 160 feet). The wind speeds at each height were regressed on the lagged periods of the other heights. The wind speed at 13 feet was shown to be the most dependent on the current and previous wind speeds at other heights, while the wind speeds at 33 feet and 160 feet were shown to be the most independent of the speeds at other heights. They also constructed impulse response curves to determine the response of the wind speed at each height to a random shock. Impulse responses showed that the wind speeds at 70 feet were the most vulnerable to a shock.

7.3 Applications of Multivariate Time Series in Other Disciplines

This study is used to illustrate the versatility of the method. Enders and Sandler (1993) used a vector autoregression analysis in order to determine if there was any relationship amongst the various modes of attack used by terrorists. They found that some modes of attack are substitutes in that they fulfill a similar purpose while others are complements in that they enhance each other's effectiveness. They also studied the effectiveness of six policies designed to counter terrorism and found that there is evidence to suggest that while some policies have an effect in reducing the number of incidents for one mode of attack, they can actually lead to an increase the number of incidents for another mode of attack. A possible explanation put forward by the authors is that when terrorists find one mode of attack is not successful, they resort to other measures.

7.4 Strengths of the VARMA(p, q) model

The VARMA (p, q) model has several strengths. These are:

a. It is possible that for a data generating process, the order p in a VAR(p) model could be very large which in turn results in a large number of parameters that are required to be estimated. This leads to parameter estimates which are imprecise. The VARMA(p, q) model is able to represent this same data generating process in a more parsimonious manner (Lütkepohl, 2005). This is because VARMA(p, q) models on the other hand have the ability to summarise the high order autoregressive lags into low order lagged shocks.

Athanasopoulos et al. (2012) recommend that in order to avoid misspecification, any modelling of microeconomic time series should include the moving average dynamics of the process even when it is assumed that the components of the time series follow finite VAR's.

b. The performance of the VAR(p) model deteriorates with large sample sizes. This can lead to a large number of parameter estimates which are not significantly different from 0. The VARMA(p, q) models, by virtue of being more parsimonious, are able to improve the efficiency of the estimated parameters while at the same time not taking away the important associations among the variables (Fackler & Krieger, 1986). The model dimension of the VARMA(p, q) model can be reduced by setting some of the parameters to 0 (Dias & Kapetanios, 2011). A simulation study by the above authors showed that for large samples and forecast horizons, the VARMA(p, q) model outperforms the VAR(p).

c. The VARMA(p, q) model is preferred to that of the VAR(p) when analysing financial and economic theory as the VAR(p) model fails to uncover the true impulse responses by generally producing results contrary to the underlying economic theory (Kascha, 2010). Dufour and Pelletier (2004) conducted a study in order to find out whether the VAR(p) model or the VARMA(p, q) model is more accurate in the determination of a shock to output, price level and the federal fund rate. They used linear methods to estimate the VARMA(p, q) model and constructed impulse response curves with a one standard deviation confidence interval band. They noted that while the shapes of the impulse response curves were similar, the width of the confidence bands for the VARMA(p, q) models was much smaller which indicates a greater accuracy for the VARMA(p, q) model.

d. There have also been published studies which compared the forecasting performance of the VAR(p) and VARMA(p, q) models. Hagnell (1991) found that by using mean square errors based on the differences between forecasts and actual observations, the VARMA(p, q) model produced superior forecasts compared to that of the VAR(p) model. A comparison between the forecasting performances of the VARMA(p, q) models with unrestricted and restricted VAR models (The restrictions being that the insignificant parameters were set to 0) with lag orders chosen by the AIC and BIC was conducted by Athanasopoulos and Vahid (2008). These authors used mean square errors for each model and found out that the forecasting performance of the VARMA(p, q) model was significantly better than for all the other models. This is important, because if the parameter estimates of a VAR(p) model are imprecise, they can have a major impact on the forecasting performance of the model (Lütkepohl, 2004).

e. Raghavan et al. (2009) also noted that VARMA(p, q) models generally produce more reliable impulse response functions than that of VAR(p) models.

7.5 Weaknesses of the VARMA(p, q) model

There are some distinct weaknesses in the use of the VARMA(p, q) model which may account for it receiving less attention. These are:

a. Challenges when it comes to identifying the unique VARMA(p, q) representations (Raghavan et al., 2009). This means that, if more than three time series are analysed, it can become difficult to find the orders of the operators due to the large number of autocorrelation, cross-correlation and partial autocorrelation functions present (Lütkepohl & Poskitt, 1996). The VAR (p) model is easier to specify as only one lag order needs to be chosen (Dufour & Pelletier, 2008) and

b. Dufour and Pelletier (2004) noted that VARMA(p, q) models are complicated by estimation difficulties. The standard estimation methods (maximum likelihood estimation and least squares estimation) usually require nonlinear optimisation and are not always feasible because the number of parameters in the model can increase quickly. This was confirmed in the authors own analysis which was described earlier in this thesis. The VARMA (p, q) model required a large number of iterations which the program was unable to run for shorter time periods. In addition, in order to simplify modelling and estimation, researchers tend to approximate a VARMA (p, q) model of order which is much higher than selected by AIC/BIC which can lead to a loss of information and the reliability of the impulse responses.

CHAPTER 8

Application of Multivariate Time Series Analysis

8.1 Application of Multivariate Time Series for the Analysis of South African Wage and Inflation Data

Multivariate time series analysis was applied to data collected by the author in order to illustrate the use of the method. A topical and often controversial issue in the field of economics is the study of the relationship between wages and inflation. There is a perception based on Keynesian economics that higher wages lead to an increase in prices which in turn leads to increasingly higher wages (Todani, 2006). This is known as the wage- price spiral. On the other hand authors such as Jonsson and Palmqvist (2004) have found evidence to suggest that it is inflation which is actually responsible for wage increases. South Africa has a highly unionized work force and there have been increasing calls by the Congress of South African Trade Unions (COSATU) for wage increases (Fin24, 2010). This has been met by resistance from the government who argue that wage increases will impact negatively on the economy by driving inflation upward. This is a highly emotive issue and the debate will benefit from evidence. I have thus used a multivariate time series approach to investigate whether there is indeed a relationship between wages and inflation in South Africa from 1996-2008.

Source of data

In order to proceed with this analysis, the domestic inflation rate and the average change in gross earnings over a fixed time period were required. Quarterly changes in gross earnings were obtained from the Statistics South Africa (Stats SA) electronic data base extending from the second quarter of 1996 to the fourth quarter of 2008. This survey is called the Quarterly Labour force survey (Formally known as the Survey of Employment and Earnings) and is published quarterly. It reflects the percentage change in the gross earnings from the previous quarter as well as the percentage change in earnings of each sector from the previous quarter.

In order to calculate the inflation rate, I needed to find a data set consisting of consumer price index (CPI) values. The CPI is a yardstick of the general prices in the economy. These CPI values are calculated monthly by Stats SA and they represent the cost of a basket of goods and services bought by a typical South African household. The monthly inflation rate is calculated by

comparing a month's CPI with the corresponding month of the previous year. This is done by calculating the difference between the CPI index of any month in the year and the CPI index of the corresponding month in the previous year and then dividing by the CPI index of that particular month. In other words, the inflation rate for January 2011 is

$$\frac{CPI(Jan\ 2011) - CPI(Jan\ 2010)}{CPI(Jan\ 2010)}$$

However, for my analysis I required the quarterly inflation rate because the inflation rate had to be aligned with the percentage changes in gross earnings which were calculated quarterly. The quarterly inflation rate was calculated by taking the difference between the current average CPI index for a quarter and the average CPI index of the corresponding quarter of the previous year and dividing it by the average CPI index of that quarter. For example, the inflation rate for the first quarter of 2011 is calculated as

$$\frac{\left[\frac{CPI(Jan\ 2011) + CPI(Feb\ 2011) + CPI(Mar\ 2011)}{3}\right] - \left[\frac{CPI(Jan\ 2010) + CPI(Feb\ 2010) + CPI(Mar\ 2010)}{3}\right]}{\frac{CPI(Jan\ 2010) + CPI(Feb\ 2010) + CPI(Mar\ 2010)}{3}}$$

The statistical package SAS version 9.2 was used to analyse the data by use of the VARMAX procedure. The graphs were performed by the use of Microsoft Excel 2007.

8.2 Results and discussion

The first procedure is to do an analysis of each variable individually. The first variable that will be considered is that of inflation. Figure 8.1 illustrates the time series plot of inflation over the time period specified.

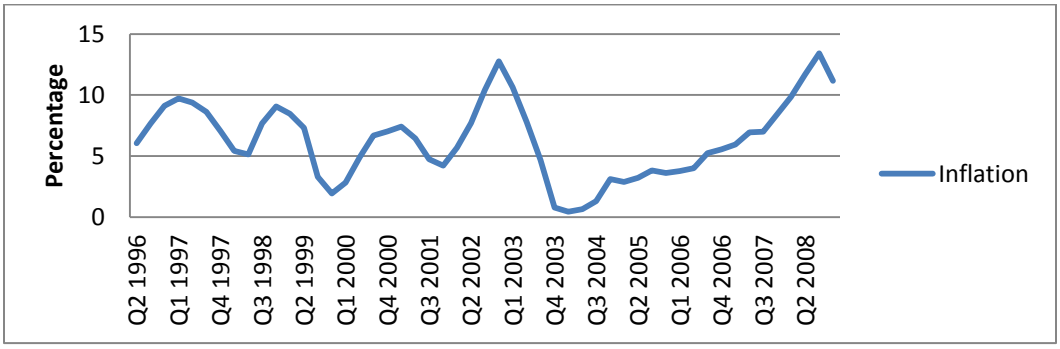

Figure 8.1 Time series plot of inflation

The inflation rate fluctuated between 2 to 10 percent from 1996 to 2001, but for the most part it was a stable at around 6 percent. However from the first quarter of 2001 there was a sharp

increase for the next year when the rate reached a peak of over 12%. The rate then declined sharply until the last quarter of 2003 reaching an all time low of just above 0%. For the years 2003 to 2008 there was a gradual but steady increase reaching a peak of almost 14% in the 2nd quarter of 2008 after which there was evidence of a decline. The summary statistics for inflation are presented in the table below.

N	51
Minimum Value	0.437
Maximum Value	13.404
Mean	6.329
Median	6.441
Variance	9.803
Standard Deviation	3.131
Skewness	0.136
Kurtosis	- 0.432

Table 8.1: Summary statistics for inflation

The inflation rate ranges from a minimum of 0.437% (1st quarter of 2004) to a maximum of 13.404% (3rd quarter of 2008). The average inflation rate over the period concerned was a moderately high figure (mean = 6.329).The amount of standard deviation in the model (3.131%) indicates that there is a fair amount of variability. This is especially prominent from the period of the 3rd quarter of 2001 up until the 1st quarter of 2004. The skewness coefficient (0.136) indicates that the distribution of values is slightly positively skewed and implies that a slight majority of the values lie to the left of the mean. The kurtosis coefficient (-0.432) is negative which indicates that the tails of the distribution of values are thinner than that of a normal distribution and in addition have a lower and wider peak around the mean.

The second variable which has been analysed is that of wage increases. The time series plot of wage increases over the period specified is illustrated in Figure 8.2 below.

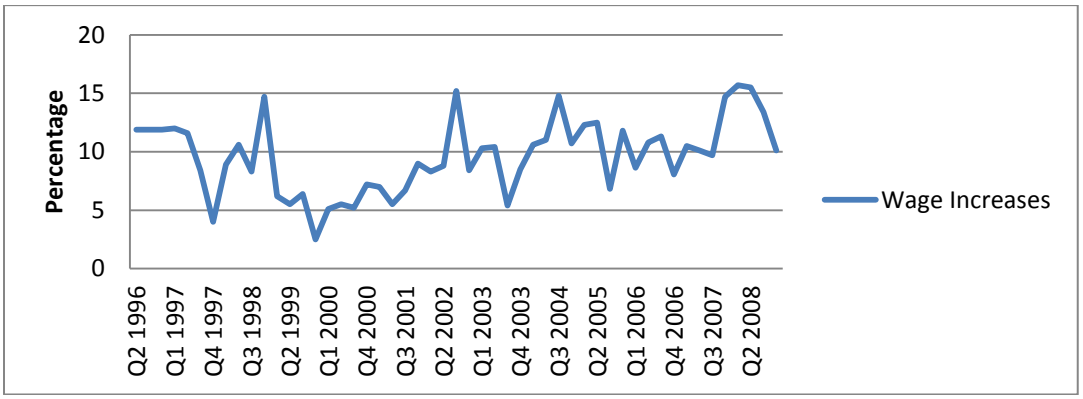

Figure 8.2 Time series plot of wage increases

The pattern for wage increases demonstrates less noticeable trends than that for inflation. There is a significant drop in the rate of wage increases from the end of 1998 until the end of 1999. The summary statistics for wage increases are revealed in Table 8.2 below.

N	51
Minimum Value	2.5
Maximum Value	15.7
Mean	9.614
Median	10.1
Variance	10.143
Standard Deviation	3.185
Skewness	0.013
Kurtosis	- 0.535

Table 8.2 Summary statistics for wage increases

The wage increase percentage ranges from a minimum of 2.5% (4th quarter of 1999) to a maximum of 15.7% (1st quarter of 2008). The mean and median figures (9.614 and 10.1 percent respectively) indicate that on average, wages increase at a significantly quicker rate than inflation. A possible reason for this could have been mounting pressure from organized labour to increase wages at the time. The standard deviation (3.185) indicates that the variability for wage increases is similar to that for inflation. The skewness coefficient (0.013) is very slightly above 0 observations which indicate that the distribution is almost symmetrical. The kurtosis coefficient (— 0.535) demonstrates that the distribution of wage increases has a lower and wider peak as well as thinner tails when compared to inflation.

We then superimposed the values of inflation on the rate of wage increases in order to determine if there was a relationship between the two series. Figure 8.3 shows the time series plots of both the series simultaneously.

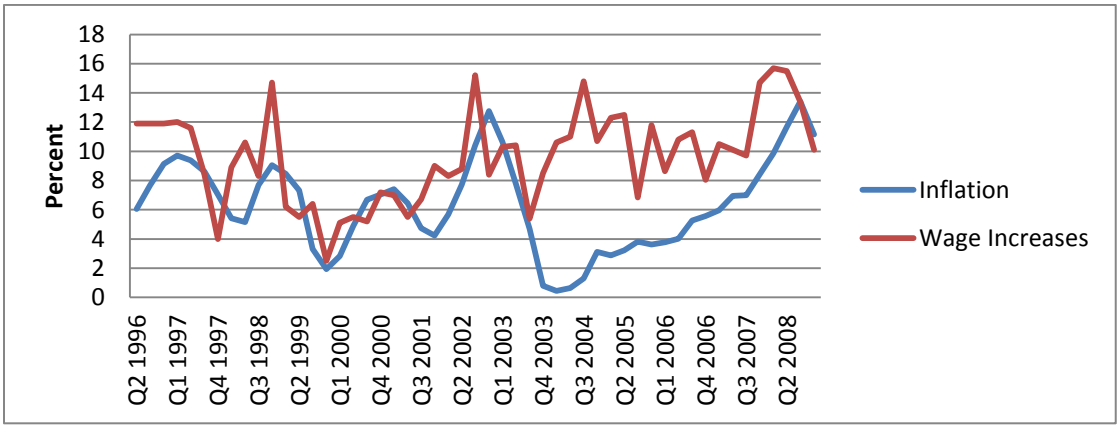

Figure 8.3: Time series plots of inflation and wage increases

There appears to be a clear relationship between inflation and wage increases from the1st quarter of 1997 up to the 3rd quarter of 2003 where both series appear to move in the same direction. The inflation rate decreases at a quicker rate than wage increases around 2003. Wage increases fluctuate around the 10% level from 2004 – 2008 while the rate of inflation shows a gradual increase around the same period.

In order to investigate the relationship between inflation and wage increases further, it was decided to perform a correlations analysis between the two variables at different time lags. y_1 denotes the inflation rate while y_2 refers to the rate of wage increases.

Lag	**Variable**	y_1	y_2
0	y_1	1	0.285
	y_2	0.285	1
1	y_1	0.843	0.087
	y_2	0.448	0.495
2	y_1	0.520	- 0.104
	y_2	0.376	0.408
3	y_1	0.195	- 0.211
	y_2	0.156	0.342
4	y_1	- 0.026	- 0.153
	y_2	- 0.054	- 0.056
5	y_1	- 0.079	- 0.055
	y_2	- 0.195	0.160

Table 8.3: Correlations between inflation and wage increases

The correlation between inflation with itself at lag 1 (r = 0.843) is very strong and positive. It weakens considerably thereafter with each increase in lag but is still positive at lags 2 and 3 (r = 0.520 and 0.195 at lags 2 and 3 respectively). The correlation between inflation and itself is very small and slightly negative at lags 4 and 5 (r = − 0.026 and – 0.079 at lags 4 and 5 respectively).

There are very little signs of any significant cross correlation between inflation and wage increases at all time lags.

The correlation between wage increases and itself is not as strong as that of between inflation and itself at lag 1 (r = 0.495). However there is a positive correlation with itself at all time lags barring lag 4 when it shows signs of a slight negative correlation.

There are signs of a positive correlation between wage increases and inflation at lag 1 (r = 0.448). This is a surprising finding as the correlation between wage increases and inflation at lag 0 is 0.285. This indicates that wage increases are more closely related to the previous

value of inflation rather than the current value and that a change in the inflation rate may prompt a demand in the increase of wages.

The next procedure is to determine if there is a unit root in the data. This is important because if there is one or more unit roots present, it could influence the Granger-causality test results (Brandt & Williams, 2007). The testing for unit roots in the data is performed by using the Dickey-Fuller test under the null hypothesis that there is a unit root present in the data.

Variable	Type	Rho	Pr < Rho	Tau	Pr < Tau
y_1	Zero Mean	- 6.12	0.0829	1.71	0.0827
	Single Mean	- 49.27	0.0004	- 4.42	0.0008
	Trend	- 50.76	< .0001	- 4.36	0.0058
y_2	Zero Mean	- 1.14	0.4462	- 0.77	0.3778
	Single Mean	- 15.74	0.0225	- 2.74	0.0748
	Trend	- 21.28	0.0297	- 3.33	0.0736

Table 8.4: Dickey-Fuller unit root tests

In view of the fact that I am including a deterministic term in the model, the results indicate that for inflation (y_1), the null hypothesis that a unit root is present for the mean and trend terms is rejected at a 5 percent level of significance ($\rho = -50.76$, $p < 0.0001$). For wage increases (y_2), the null hypothesis that a unit root is present is rejected ($\rho = -21.28$, $p < 0.0001$) . It can thus be concluded that there is no unit root present in the data.

The specification of the order of the model is performed by using the Minimum information criterion (discussed in chapter 4.6.2).

Lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6
AR 0	4.449	4.487	3.983	3.143	3.047	2.958	3.180
AR 1	2.866	3.113	2.956	2.426	2.373	2.267	2.572
AR 2	2.208	2.541	2.530	2.061	2.059	2.245	2.602
AR 3	2.109	2.292	2.350	2.229	2.314	2.574	3.007
AR 4	2.052	2.381	2.394	2.387	2.706	3.004	3.446
AR 5	2.204	2.204	2.403	2.665	3.076	3.253	3.821
AR 6	2.266	2.641	2.892	3.217	3.732	4.058	4.639
AR 7	2.318	2.767	3.289	3.831	4.496	5.049	5.883

Table 8.5: Minimum information criterion

The Minimum information criterion table demonstrates that a VAR(4) model has the lowest value (2.052). However, the value for the VARMA $(2,4)$ model is only marginally higher (2.061). We have decided to run both models in order to determine whether similar conclusions can be drawn for both models.

The selection of the VAR(4) model can be confirmed from the methods employed by Tiao & Box (1981) in which the partial autocorrelation functions for both of the variables are determined simultaneously in a VAR model.

```
 Schematic Representation of Partial Autocorrelations
Variable/
Lag          1    2    3    4    5    6    7    8    9   10    11

y1           +    -    .    .    .    .    .    .    .    .    .
y2           +    .    .    -    .    .    .    .    .    .    .

             + is > 2 * std error,  - is < -2 *std error,  . is between
```

The above representation shows that it is clear that the appropriate model is of an order of at most 4. Figures 8.4 and 8.5 illustrate the graphical depictions of the partial autocorrelation functions for each series when they are analysed independently of each other i.e. as two univariate time series.

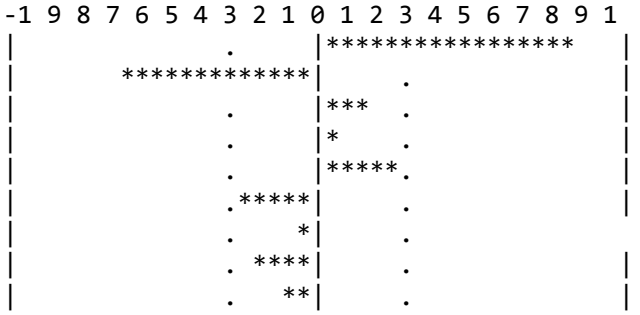

Figure 8.4 Partial autocorrelation function for inflation

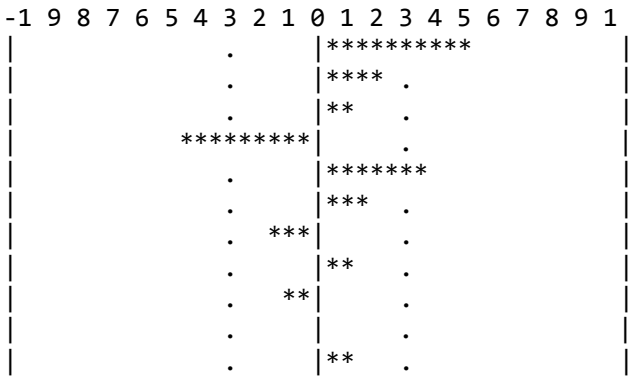

Figure 8.5 Partial autocorrelation function for wage increases

The results show that the inflation series should have a maximum autoregressive order of 2 while the wage increase series should have a maximum order of 5.

The VAR(4) model is the first model which we have estimated and is performed by using the least squares estimation procedure discussed in chapter 2.6.1. This is the default estimation procedure used in SAS for the estimation of VAR models. The parameters and their standard errors are listed in Table 8.6.

Equation	**Parameter**	**Estimate**	**Standard Error**	**t- value**	**Pr > |t|**
y_1	Constant	1.398	0.788	1.77	0.086
	$y_{1,t-1}$	1.398	0.172	8.12	0.000
	$y_{2,t-1}$	0.218	0.066	3.31	0.002
	$y_{1,t-2}$	- 0.586	0.298	-1.97	0.057
	$y_{2,t-2}$	- 0.086	0.068	-1.26	0.216
	$y_{1,t-3}$	- 0.073	0.295	-0.25	0.805
	$y_{2,t-3}$	- 0.129	0.068	-1.92	0.063
	$y_{1,t-4}$	0.104	0.159	0.65	0.517
	$y_{2,t-4}$	- 0.05	0.077	0.65	0.519
y_2	Constant	3.797	1.871	2.03	0.049
	$y_{1,t-1}$	0.114	0.409	0.28	0.782
	$y_{2,t-1}$	0.376	0.156	2.41	0.021
	$y_{1,t-2}$	- 0.291	0.708	- 0.41	0.684
	$y_{2,t-2}$	0.312	0.162	1.92	0.062
	$y_{1,t-3}$	0.413	0.701	- 0.59	0.559
	y_{2t-3}	0.359	0.161	2.23	0.032
	$y_{1,t-4}$	0.506	0.378	1.34	0.188
	$y_{2,t-4}$	- 0.397	0.183	- 2.17	0.037

Table 8.6: Parameter estimates for inflation and wage increases in the VAR(4) model

The above results indicate that the relationship between the current value of y_1 (inflation) and its previous value at lag 1 is statistically significant (t $= 8.12$, $p = 0.0001$). The relationship between inflation and wage increases at lag 1 is also statistically significant (t $= 3.3$, $p = 0.002$). There is no significant evidence to suggest an association between inflation and its past

values as well as past values of wage increases at larger time lags ($p > 0.05$ at lags 2, 3 and 4). This suggests that inflation is significantly influenced from recent events.

The relationship between wage increases (y_2) and its own past values at all lags are statistically significant (t $= 2.41, 1.92, 2.23, 2.17$ and $p = 0.021, 0.062, 0.032, 0.037$ at lags 1, 2, 3, 4 respectively). There is a positive association between wages and its past values at lags 1, 2 and 3 and a significant negative association at lag 4. The relationship between wage increases and past values of inflation at all time horizons is not statistically significant (t $=$ $0.28, -0.41, -0.59$, 1.34 and $p = 0.782, 0.684, 0.559, 0.188$ at lags 1, 2, 3, 4 respectively. This proves that wage increases (y_2) is a more exogenous variable than inflation (y_1).

We have furthermore, conducted a Granger-causality Wald test (discussed in chapter 6.3) under the null hypothesis that inflation and wage increases do not cause each other. Test 1 tests the alternate hypothesis that inflation is caused by wage increases and Test 2 tests the alternate hypothesis that wage increases are caused by inflation.

Test	Degrees of Freedom	Chi-square	P-Value
1	4	14.83	0.0051
2	4	6.55	0.1617

Table 8.7: Granger-causality Wald tests for the VAR(4) model

The null hypothesis of non-causality that inflation is not caused by wage increases is rejected ($\chi^2(4) = 14.83$, $p = 0.0051$). This is consistent with the view expressed in Keynesian economic theory.

Test 2 tests whether wage increases are influenced by inflation. The results for test 2 show that there is an acceptance for the null hypothesis of non-causality ($\chi^2(4) = 6.55$, $p = 0.1617$).This means that there is good evidence to suggest that wage increases are not caused by inflation. This result is again consistent with the estimation procedure. It should be noted however, that Granger-causality tests have been proven to be biased if there is a possibility that the series is non-stationary (Hamilton, 1994).

The next stage in the model building procedure is to determine whether the VAR(4) model is an adequate model. In order to do this, we ran the schematic representation of cross correlation of the residuals as well as the Portmanteau test up until lag 12.

```
              Schematic Representation of Cross Correlations of Residuals
Variable/
Lag           0     1     2     3     4     5     6     7     8     9     10    11    12

y1            ++    ..    ..    ..    -.    ..    ..    ..    ..    ..    ..    ..    ..
y2            ++    ..    ..    ..    ..    ..    ..    ..    ..    ..    ..    ..    ..
```

$+$ is $>$ 2 $*$std error, $-$ is $<$ -2 $*$std error, . is between

This representation demonstrates very little evidence of serial correlation in the residuals which indicates that the model is adequate. Table 8.8 reveals the results of the Portmanteau test.

Up to lag	Degrees of Freedom	Chi-square	p-value
5	4	14.44	0.0060
6	8	15.93	0.0434
7	12	16.11	0.1862
8	16	16.97	0.3875
9	20	17.8	0.6004
10	24	22.32	0.5602
11	28	25.46	0.6025
12	32	28.89	0.6249

Table 8.8: Portmanteau statistics for the VAR(4) model

The Portmanteau test is conducted under the null hypothesis that there is no cross correlation in the residuals. The results show that the model is adequate due to the presence of relatively low Chi-square values and large *p* values with the exception of up to lags 5 and 6.

In view of the fact that one of the main purposes of multivariate time series is forecasting, we have conducted the 3 step ahead forecasts for the 1st, 2nd and 3rd quarters of 2009 (observations 52 – 54) with their 95% confidence intervals. In order to determine how accurate the forecasts for this model are, we ran a test to determine the difference between the actual forecasts for observations 50 and 51 and their predicted values based on the trends of observations 1–49. Tables 8.9 and 8.10 show the 3 step ahead forecasts for inflation and wage increases respectively.

Variable	Obs	Forecast	Standard Error	95% Confidence Limits		Actual	Residual
y_1	50	11.686	1.057	9.615	13.758	13.404	1.718
	51	10.068	2.023	6.104	14.032	11.155	1.087
	52	7.349	2.738	1.981	12.716		
	53	4.622	3.066	- 1.387	10.633		
	54	2.911	3.139	- 3.241	9.063		

Table 8.9: 3 step ahead forecasts for inflation generated from the use of the VAR(4) model

Variable	Obs	Forecast	Standard Error	95% Confidence Limits		Actual	Residual
y_2	50	14.461	2.511	9.540	19.383	13.400	-1.061
	51	11.985	2.698	6.697	17.273	10.100	-1.885
	52	10.046	2.942	4.279	15.812		
	53	9.339	3.308	2.856	15.822		
	54	9.141	3.371	2.533	15.749		

Table 8.10: 3 step ahead forecasts for wage increases generated from the use of the VAR(4) model

The differences between the actual and forecasted values of inflation for observations 50 and 51 are very small (residuals of 1.718 and 1.087) respectively which indicates that the forecasting accuracy of this model is very good. This is confirmed by the relatively small range between the lower and upper values of the 95% confidence limits. The forecasts for observations 52, 53 and 54 indicate that a significant dip in the inflation rate is expected for the following 3 quarters.

The results for wage increases also show that the differences between actual and forecasted values for observations 50 and 51 are small (residuals of -1.061 and -1.885 respectively) which indicates that the forecasts are precise. A dip in the wage increases for observations 52, 53 and 54 is also expected, though this decrease is not expected to be as sharp as the drop in the inflation rate. A possible explanation for this phenomenon is that wage increases are only dependent on their own past values and inflation is dependent on its own past value as well as the past values of wage increases.

In the final procedure for this part of the analysis, a forecast error decomposition analysis was performed for inflation and wage increases respectively. For this we needed to determine how much the values in the model fitted differed from the actual values of the vector of endogenous variables. If the innovations in one variable do not help explain the variation in the other variable, then both variables are said to be exogenous of each other (Brandt & Williams, 2007).

Variable	Lead	y_1	y_2
y_1	1	1	0
	2	0.934	0.066
	3	0.896	0.104
	4	0.884	0.116
	5	0.879	0.121
	6	0.875	0.125
	7	0.843	0.157
	8	0.799	0.201
	9	0.758	0.242
	10	0.742	0.258
	11	0.743	0.257
	12	0.748	0.252

	13	0.750	0.250
	14	0.751	0.249
	15	0.749	0.251

Table 8.11: Forecast error decomposition analysis for inflation

Variable	**Lead**	y_1	y_2
y_2	1	0.094	0.906
	2	0.105	0.895
	3	0.097	0.903
	4	0.081	0.919
	5	0.115	0.885
	6	0.136	0.864
	7	0.149	0.851
	8	0.149	0.851
	9	0.151	0.849
	10	0.158	0.842
	11	0.166	0.834
	12	0.167	0.833
	13	0.166	0.834
	14	0.167	0.833
	15	0.169	0.831

Table 8.12: Forecast error decomposition analysis for wage increases

Since inflation is first in the ordering of the variables, the variance decomposition assumes that the first period contains all of the variance of the forecasts which are attributed to inflation, and none which are attributed to wage increases. For both variables, the forecast errors are explained mainly by themselves and most of the variance in prediction is due to their own shocks. After a period of 15 quarters (3.75 years), approximately 25 % of the forecast error in inflation is attributed to shocks/innovations in wage increases while for the same period, approximately 17 % of the forecast variation in wage increases is attributed to innovations in inflation. It is important to note that the proportion of variance of y_1(inflation) explained by itself has decreased at a quicker rate than y_2. This indicates that unexpected changes in the wage rate have a more profound effect on inflation than inflation has on wage increases. This finding is in contrast to that reported by Agénor and Hoffmaister (1997) who used variance decompositions to demonstrate that wage increases are greatly influenced by its own shocks at short time horizons while inflation is influenced to a greater extent by its own shocks at longer time horizons.

In conclusion, the current value of inflation is dependent on its previous value as well as on the previous value of wage increases. The current value of wage increases is highly dependent on its own past values and there is evidence to suggest that higher inflation rates do not lead to an increase in wages.

A regression of the VARMA(2,4) model was then performed in order to determine whether it would yield results similar to that of the VAR(4) model and if it fits the data better. The default procedure used for estimating the VARMA model in SAS is maximum likelihood estimation. As discussed earlier, the maximum likelihood equations are nonlinear in the parameters and need to be solved using iterative methods. The first method used to solve these equations is the Quasi-Newton method which is where successive gradient matrices are computed in order to obtain the Hessian matrix. Table 8.13 shows the parameter estimates for inflation and wage increases estimated from the use of the Quasi-Newton method

Equation	**Parameter**	**Estimate**	**Standard Error**	**t- value**	**Pr > |t|**
y_1	Constant	3.815	0.000		1
	$y_{1,t-1}$	0.598	0.353	1.69	0.097
	$y_{2,t-1}$	0.199	0.121	1.64	0.108
	$y_{1,t-2}$	- 0.216	0.243	- 0.89	0.379
	$y_{2,t-2}$	0.187	0.070	2.68	0.010
	$u_{1,t-1}$	- 0.863	0.412	- 2.09	0.042
	$u_{2,t-1}$	- 0.154	0.189	- 0.81	0.419
	$u_{1,t-2}$	- 0.740	0.422	- 1.76	0.086
	$u_{2,t-2}$	- 0.096	0.098	- 0.98	0.333
	$u_{1,t-3}$	0.931	-0.353	- 2.63	0.011
	$u_{2,t-3}$	- 0.046	0.099	- 0.46	0.649
	$u_{1,t-4}$	- 0.095	0.278	- 0.34	0.735
	$u_{2,t-4}$	- 0.132	0.180	- 0.73	0.470

Equation	**Parameter**	**Estimate**	**Standard Error**	**t- value**	**Pr > |t|**
y_2	Constant	12.808	0.000		1
	$y_{1,t-1}$	- 1.428	1.303	- 1.10	0.279
	$y_{2,t-1}$	0.916	0.238	3.84	0.000
	$y_{1,t-2}$	0.737	0.601	1.23	0.226
	$y_{2,t-2}$	0.511	0.530	0.96	0.340
	$u_{1,t-1}$	- 0.385	1.410	- 0.27	0.786
	$u_{2,t-1}$	0.208	0.284	0.73	0.468
	$u_{1,t-2}$	- 0.990	1.310	- 0.76	0.454
	$u_{2,t-2}$	- 0.239	0.197	- 1.21	0.231
	$u_{1,t-3}$	- 1.454	1.456	- 1.00	0.323
	$u_{2,t-3}$	- 0.780	0.232	- 3.37	0.002
	$u_{1,t-4}$	- 0.547	1.760	- 0.31	0.757
	$u_{2,t-4}$	0.702	0.323	2.14	0.038

Table 8.13: Parameter estimates for the VARMA(2,4) model obtained from using the Quasi-Newton method

There does not appear to be evidence of an association between inflation and lagged values of itself and wage increases. This method has unfortunately not provided the t statistics in order to test for the significance of the parameters. We have thus decided to evaluate the equations by the use of the Newton-Raphson iterative procedure instead of Quasi-Newton iterations. We have used a large number of maximum iterations (50 000) and function calls (100 000). The results for the Newton-Raphson parameter estimates are shown below.

Equation	**Parameter**	**Estimate**	**Standard Error**	**t- value**	**Pr > |t|**
y_1	Constant	3.815	0.000		1
	$y_{1,t-1}$	0.598	0.353	1.69	0.097
	$y_{2,t-1}$	0.199	0.121	1.64	0.108
	$y_{1,t-2}$	- 0.216	0.243	- 0.89	0.379
	$y_{2,t-2}$	0.187	0.070	2.68	0.010
	$u_{1,t-1}$	- 0.863	0.412	- 2.09	0.042
	$u_{2,t-1}$	- 0.154	0.189	- 0.81	0.419
	$u_{1,t-2}$	- 0.740	0.422	- 1.76	0.086
	$u_{2,t-2}$	- 0.096	0.098	- 0.98	0.333
	$u_{1,t-3}$	0.931	-0.353	- 2.63	0.011
	$u_{2,t-3}$	- 0.046	0.099	- 0.46	0.649
	$u_{1,t-4}$	- 0.095	0.278	- 0.34	0.735
	$u_{2,t-4}$	- 0.132	0.180	- 0.73	0.470

Equation	**Parameter**	**Estimate**	**Standard Error**	**t- value**	**Pr > |t|**
y_2	Constant	12.808	0.000		1
	$y_{1,t-1}$	- 1.428	1.303	-1.10	0.279
	$y_{2,t-1}$	0.916	0.238	3.84	0.000
	$y_{1,t-2}$	0.737	0.601	1.23	0.226
	$y_{2,t-2}$	0.511	0.530	0.96	0.340
	$u_{1,t-1}$	- 0.385	1.410	- 0.27	0.786
	$u_{2,t-1}$	0.208	0.284	0.73	0.468
	$u_{1,t-2}$	- 0.990	1.310	- 0.76	0.454
	$u_{2,t-2}$	- 0.239	0.197	- 1.21	0.231
	$u_{1,t-3}$	- 1.454	1.456	- 1.00	0.323
	$u_{2,t-3}$	- 0.780	0.232	- 3.37	0.002
	$u_{1,t-4}$	- 0.547	1.760	- 0.31	0.757
	$u_{2,t-4}$	0.702	0.323	2.14	0.038

Table 8.14: Parameter estimates for the VARMA(2,4) model obtained from using the Newton-Raphson method

This procedure is computationally burdensome (It took approximately 1 and a half hours on SAS) however the test statistics for the significance of the parameters have been successfully computed. The current value of inflation (y_1) demonstrates a statistically significant relationship with the past value of wage increases at lag 2 (t $= 2.68$, $p = 0.010$). It also shows a significantly negative dependence on its own past shocks at time lags 1 and 3 in particular ($p =$ 0.042 and 0.011 at lags 1 and 3 respectively). There is very little evidence to suggest that inflation is dependent on the past innovations of wage increases at all time horizons ($p > 0.4$ at all time horizons).

The current value of wage increases shows a positive and very significant association with its previous value at lag 1 (t $= 3.84$, $p = 0.000$). There is very little association between any of the other lagged variables apart from its own shocks at time lags 3 and 4 ($p = 0.002$ and 0.038 at lags 3 and 4 respectively).

Next, a Granger-causality test under the null hypothesis of non-causality was conducted. Test 1 tests the alternative hypothesis that inflation is Granger-caused by wage increases and Test 2 tests the alternative that wage increases are Granger-caused by inflation.

Test	Degrees of Freedom	Chi-square	P-Value
1	2	10.89	0.0043
2	2	4.13	0.1267

Table 8.15: Granger-causality Wald tests for the VARMA(2,4) model

There is sufficient evidence to reject the null hypothesis that inflation is not Granger-caused by wage increases ($\chi^2(2) = 10.89$, $p = 0.0043$). Conversely, the null hypothesis that wage increases are not Granger-caused by inflation is accepted ($\chi^2(2) = 4.13$, $p = 0.1267$). It can be concluded therefore that inflation does not have a major impact on future values of wage increases. These findings are similar to that found when the VAR(4) model was tested for the direction of Granger-causality.

The final step of the model building process is diagnostic checking and I have decided to compare the goodness of fit between the VARMA$(2,4)$ and VAR(4) models. The first step is to compare the information criteria of the two models.

Information Criteria VAR(4)		Information Criteria VARMA(2,4)	
AICC	2.375	AICC	2.297
HQC	2.460	HQC	2.259
AIC	2.193	AIC	1.873
SBC	2.902	SBC	2.897
FPEC	9.052	FPEC	6.706

The values for all of the information criteria are lower for the VARMA(2,4) model than the VAR(4) which indicates that in theory, the VARMA(2,4) model should be selected instead of that of the VAR(4). The residual plot and Portmanteau statistics however reveal the following.

Schematic Representation of Cross Correlations of Residuals

Variable/ Lag	0	1	2	3	4	5	6	7	8	9	10	11	12
y1	++	.+	-.	--	+.	++	-.	--	+.	++	.+	--	.-
y2	++	-.	-.	+-	++	..	-.	.-	++	..	-.	.-	+.

+ is $> 2 *$ std error, − is $< -2 *$ std error, . is between

Up to lag	Degrees of Freedom	Chi-square	p-value
7	4	266.39	< .0001
8	8	304.37	< .0001
9	12	327.40	< .0001
10	16	368.60	< .0001
11	20	423.23	< .0001
12	24	446.22	< .0001

Table 8.16: Portmanteau statistics for the VARMA(2,4) model

The residual plots demonstrate that a large number of the residuals are outside the 2 standard error bounds which indicate that there is a strong possibility that the residuals are serially correlated. The presence of high Chi-square values results in a rejection of the null hypothesis that the model is adequate. This can be confirmed by $p < 0.0001$ up to all lags. It can thus be concluded that the VARMA(2,4) model is not a suitable model to analyse this data set.

Despite the VARMA(2,4) model not showing signs of adequacy, we have still decided to compute the forecasts in order to confirm the popular assertion that imprecise parameter estimates result in poor forecasts. The 3 step ahead forecasts as well as the predicted forecasts for observations 50 and 51 (based on the trends shown for observations 1 to 49) for inflation and wage changes are shown in Tables 8.17 and 8.18 respectively.

Variable	Obs	Forecast	Standard Error	95% Confidence Limits		Actual	Residual
y_1	50	16.518	0.677	15.192	17.845	13.404	- 3.114
	51	19.383	1.617	16.215	22.551	11.155	- 8.228
	52	24.596	2.780	19.147	30.045		
	53	27.780	3.924	20.089	35.470		
	54	30.957	4.910	21.335	40.580		

Table 8.17 : 3 step ahead forecasts for inflation generated from the use of the VARMA(2,4) model

Variable	Obs	Forecast	Standard Error	95% Confidence Limits		Actual	Residual
y_2	50	24.302	2.283	19.827	28.777	13.400	-10.902
	51	30.463	3.018	24.548	36.378	10.100	-20.363
	52	44.995	3.905	37.341	52.650		
	53	46.389	5.291	36.019	56.760		
	54	55.024	5.819	43.619	66.430		

Table 8.18: 3 step ahead forecasts for wage increases generated from the use of the VARMA(2,4) model

In sharp contrast to the VAR(4) model estimated earlier, there is a substantial difference between the actual values and forecasted values (residuals of -3.114 and -8.228) for observations 50 and 51. This combined with a large interval of the confidence limits for observations 53 and 54 (the 2nd and 3rd quarters of 2009) indicates that the forecasts are not accurate. This confirms the fact that a model which does not have accurate parameter estimates produces poor forecasts.

The results are even more spurious for the forecast of wage increases. The difference between the actual and the predicted values is more than 20 for the 51st observation (the 4th quarter of 2008). There is also expected to be a large increase in wages in contrast to the VAR(4) model.

For completeness, we also conducted separate tests under the assumption that the data is non-stationary. I have not reported these results but they confirmed that the VAR model fitted this dataset better than the VARMA model.

In conclusion, we ran 3 models, a VAR(4) model estimated using least squares estimation under the assumption that both the series are stationary, and two stationary VARMA$(2,4)$ models estimated using maximum likelihood with the nonlinear equations solved using both Quasi-Newton and Newton-Raphson methods. The VAR model fitted the data well and produced accurate forecasts. The VARMA model evaluated with Quasi-Newton estimation was unable to generate t statistics to test for the significance of the parameter estimates while the model that was evaluated using Newton-Raphson estimation produced estimates which did not fit the data well. This shows that a more complex model does not necessarily lead to more accurate results. The VAR model shows that there is strong evidence that wage increases occur independently of inflation and while inflation is also dependent on past values, there is evidence which points to an association between itself and wage changes at short time horizons. The values of both inflation and rate of wage increases were expected to decrease in the following 3 quarters.

CHAPTER 9

Conclusion

The main objective of this study was to describe and compare model building procedures and the forecasting performance of the VAR(p), VMA(q) and VARMA(p,q) models in an attempt to identify factors which could explain model selection.

The popularity of the VAR(p) model is largely due to the fact that it is easy to specify because only one lag order needs to be chosen. The estimation methods such as that of least squares estimation and Yule-Walker estimation are relatively easy to use and in addition, there is also the maximum likelihood estimation procedure which is more efficient. The VAR(p) model also produces reliable forecasts.

The VAR(p) model has the ability to determine the association between different variables. However, although the VAR(p) model can be used to determine the interdependence among two or more series, it does not take into account the effect of innovations or shocks at different time lags and neither is it parsimonious. A finite multivariate time series model which only takes into account the relationship between y_t and its various shocks at q time lags is known as the vector moving average model of order q (VMA(q)).

The VMA(q) model has received very little attention in the literature, probably because it cannot be used to determine the relationship between a variable and its own past values as well as the past values of all the other variables in the system. It is instead used to measure the effects of shocks/innovations in the system. Consequently, there has not been much literature on the model building process and the estimation procedure is limited to the maximum likelihood procedure.

The VARMA(p,q) model can be used to investigate the relationship between a variable and its past values and shocks of itself and other variables in the system. It has several advantages over the VAR(p) model. The main advantage is that it represents the data generating process in a more parsimonious manner due to its ability to summarise the high order autoregressive lags into low order lagged shocks. This parsimony improves the efficiency of the parameter estimates without taking away the associations amongst the variables. In addition, VARMA(p,q) models are also known for producing superior forecasts and impulse responses than VAR(p) models.

However, there are some distinct weaknesses in the use of the VARMA(p,q) model which may account for it receiving less attention. It has not been as popular as the VAR(p) model, because

more than one lag order is required to be chosen and the standard form of the model is not unique. This means that two different representations can lead to the same infinite moving average representation. As a result of this, restrictions have to be placed in order to ensure uniqueness. These restrictions make the specification and estimation procedures significantly more complicated.

The two most common approaches to model specification include the scalar component model (SCM) developed by Tiao and Tsay (1989) and the echelon form representation developed for the univariate case by Hannan and Rissanen (1982) and extended to the multivariate case by Lütkepohl and Poskitt (1996). The SCM and the echelon form representation have the same general appearance even though the specification procedures are not equivalent and do not normally give rise to the same model specification (Mélard, Roy & Saidi, 2004).

The main advantage of echelon forms is that the asymptotic inference is straightforward while there are problems in scalar component models if the transformation of the variables is not data dependent (Lütkepohl & Poskitt, 1996). The echelon form is also practical and feasible as the process can be fully automated unlike the scalar component model which can be partly automated but requires discretion and judgement from the analyst (Athanasopoulos et al., 2012).

The main advantage of scalar component models is that they are very flexible because the maximum autoregressive order does not have to be of the same order as that of the moving average component. They also produce better forecasts than the echelon form VARMA models (Athanasopoulos et al., 2012). Scalar component models however, may lead to computational difficulties due to the evaluation of a large number of eigenvalues (Dufour & Jouini, 2008).

The maximum likelihood estimation procedure is the most commonly used estimation procedure used to estimate VARMA models. It is not easy to maximise the likelihood function if the model is not identified properly and it usually requires nonlinear optimisation as well as iterative methods in order to evaluate the parameter estimates (Kascha, 2010). In addition, unlike that for the VAR model, the method of least squares estimation requires iterative procedures and has only recently begun to gain attention in the literature.

The VAR(p) model does not accurately capture the dynamics of the system if there is non-stationarity present in the data particularly if there is cointegration present i.e. the variables share a common trend. The model then needs to be written in VECM form which separates the cointegration and long run relations from the short term dynamics of the model. Gupta (2006) confirmed that the Bayesian VECM model produced the most accurate out of sample forecasts for chosen South African macroeconomic indicators. As is the case with stationary models, co integrated VARMA models have not gained much attention in the literature as compared to cointegrated VAR models even though they are more parsimonious and produce better forecasts.

In order to demonstrate the technique and to illustrate model selection, we analysed quarterly South African wage and inflation data from April 1996 to December 2008 using the program SAS version 9.2. The adequacy of the VAR(p) and VARMA (p, q) models was compared and their forecasting performance was evaluated.

In this application, three models were run, a VAR model and two VARMA models. The VAR model was estimated using the least squares method. Both VARMA models were estimated using maximum likelihood methods and solved using two optimisation techniques. The first model employed the Quasi-Newton method and the second, the Newton-Raphson procedure. The model with the best fit was the VAR model as the forecasts were reliable (there was not much difference between the observed and the predicted values), while the small values of the Portmanteau statistic indicated that there was little serial correlation in the residuals. The VARMA models in contrast, had large values of the Portmanteau statistic as well as unreliable forecasts and were thus were found not to fit the data well.

Analysis of this data shows the current value of inflation is dependent on its previous value as well as on the previous value of wage increases. The current value of wage increases is highly dependent on its own past values and there is evidence to suggest that higher inflation rates do not lead to an increase in wages.

This study has extended the understanding of multivariate time series models by describing the model building procedures and their more recent modifications in the literature. Modelling of multivariate time series is demonstrated by an application of the technique to local wage – inflation data which confirms that model selection is important to avoid obtaining spurious results.

In conclusion, multivariate time series analysis is a dynamic statistical procedure, which is used extensively to analyse the interrelationships between variables over a period of time and is supported by a large body of literature demonstrating its utility globally in the fields of economics.

Further research needs to be undertaken in order to simplify the model building procedures for the VARMA model. The echelon form models tend to show more signs of over parameterisation than scalar component models and hence research needs to be conducted to make them more parsimonious. There is also a need to develop methods for simplifying the second stage of the specification procedure.

Research incorporating more current wage and inflation data is required and the use of multivariate time series modelling techniques should be expanded in the natural and health sciences where it is underutilised.

References

Agénor, P.R. & Hoffmaister, A.W. (1997). Money, Wages and Inflation in Middle-Income Developing Countries, *Working Paper of the International Monetary Fund*.

Ahn, S. K. (1997). Inference of Vector Autoregressive Models with Cointegration and Scalar components, *Journal of the American Statistical Association* **92(437):** 350 - 356.

Ansari, M. I. (1996). Monetary and Fiscal Policy: Some Evidence from Vector Autoregression for India, *Journal of Asian Economics* **7(4):** 677-698.

Asghar, Z. (2008). Simulation Evidence of Granger Causality in presence of a confounding variable, *International Journal of Applied Econometrics and Quantitative Studies* **5(2)**: 71 – 86.

Athanasopoulos, G. & Vahid, F. (2006). A Complete VARMA Modelling Methodology Based on Scalar Components,Department of Econometrics and Business Statistics, Monash University, Working Paper **2(06),** http://www.buseco.edu.au/depts/ebs/pubs/wpapers.

Athanasopoulos, G. & Vahid, F. (2008). VARMA versus VAR for macroeconomic forecasting, *Journal of Business and Economic Statistics* **26:** 237 – 252.

Athanasopoulos, G., de Carvalho Guillén, O. T., Issler, J. V., Vahid, F. (2011). Model selection, estimation and forecasting in VAR models with short-run and long-run restrictions, *Journal of Econometrics* **164 :** 116 – 129.

Athanasopoulos, G., Poskitt, D. S. & Vahid, F. (2012). Two Canonical VARMA Forms: Scalar Component Models Vis-á-Vis the Echelon Form, *Econometric Reviews* **31(1):**60-83.

Bartel, H. & Lütkepohl, H. (1998). Estimating the Kronecker indices of cointegrated echelon form VARMA models, *Econometrics Journal* **1:** 676 – 699.

Boubacar Mainassara, Y. (2010). Selection of weak VARMA models by modified Akaike's information criteria, Unpublished Paper 32.

Boudjelabba, H., Dufour, J. M & Roy, R. (1992), Testing Causality Between Two Vectors in Multivariate Autoregressive Moving Average Models, *Journal of the American Statistical Association* **87:** 1082 – 1090.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, Hoboken, New Jersey.

Brandt, P. T & Williams, J. T. (2007). *Multiple Time Series Models*, Sage Publications Inc, Thousand Oaks, CA .

Brockwell, P. J. & Davis, R. A. (1996). *Introduction to Time Series and Forecasting*, Springer-Verlag, New York Inc.

Chao, J. C. & Phillips, P. C. B (1999). Model selection in partially non-stationary vector autoregressive processes with reduced rank structure, *Journal of Econometrics* **91:** 227 – 271.

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction* 6th edition, Chapman and Hall/CRC, Boca Raton, Florida.

Chen, C. & Lee, C.J. (1990). A VARMA Test on the Gibson Paradox, *The Review of Economics and Statistics* **72(1):** 96-107.

Chien, H., Lee, S. & Tsai, Y. (2006). The time series relation between monthly sales and stock prices, Atlantis Press, http://www.atlantis-press.com/php/download_paper.php?id=115.

Chigira, H. & Yamamoto, T. (2003). The Granger Non – Causality test in cointegrated vector autoregressions, Unpublised Paper, Department of Economics, Hitosuboshi University.

Chin, D.A. (1995). A scale model of multivariate rainfall time series, *Journal of Hydrology* **168:** 1-15.

Demetrescu, M., Lutkepohl, H. & Saikkonen, P. (2009). Testing for the Cointegrating Rank of a Vector Autoregressive Process with Uncertain Deterministic Trend Term , *Econometrics Journal* **12:** 414 – 435.

de Waele, S. & Broersen, P.M.T. (2003). Order Selection for Vector Autoregressive Models, *IEEE Transactions on Signal Processing* **51(2):** 427-433.

Dias, G. F. & Kapetanios, G. (2011). Forecasting Medium and Large Datasets with Vector Autoregressive Moving Average (VARMA) models, Unpublished Paper, Queen Mary University of London.

Dolado , J. J., Gonzalo, J. & Marmol, F. (1999). Cointegration, Department of Economics, Statistics and Econometrics,Universidad Carlos III de Madrid,C/Madrid 186,28903, Getate Madrid, SPAIN.

Du Frutos, R. F. & Serrano, G. R. (1997). A generalized least square estimation method for invertible vector moving average models, *Economics Letters*, **57:** 149 – 156.

Dufour, J. M. & Pelletier, D. (2002). Linear methods for estimating VARMA models with a macroeconomic application, Unpublished Paper, Département de sciences économics, Université de Montreal.

Dufour, J. M. & Pelletier, D. (2004). Linear Estimation of weak VARMA models with a macroeconomic application, Université de Montreal and North Carolina State University, Working Paper.

Dufour, J. M. & Pelletier, D. (2008). Practical methods for modelling weak VARMA processes: identification , estimation and specification with a macroeconomic application, Discussion Paper, McGill University, CIREQ and CIRANO.

Dufour, J. M. & Jouini, T. (2008). Simplified order selection and efficient linear estimation for VARMA models with a macroeconomic application, Université de Montreal.

Dufour, J. M. & Taamouti, A. (2010). Short and long run causality measures, theory and inference, *Journal of Econometrics* **154:** 42-58.

Enders, W. (2004). *Applied Econometric Time Series,* 2nd edition, Wiley, New York.

Enders, W. & Sandler, T. (1993). The Effectiveness of Antiterrorism Policies: A Vector Autoregressive Intervention Analysis, *The American Political Review* **87(4):** 829 -844.

Engle, R. F. & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing, *Econometrica* **55:** 251 – 276.

Engle, R. F. & Yoo, B. S. (1987). Forecasting and testing in Co-integrated systems, *Journal of Econometrics* **35:** 143-157.

Escanciano, J. C, Lobato, I. N & Zhu, L (2010), Automatic Diagnostic Checking for Vector autoregressions, Working Paper, econ.duke.edu/∼brossi/NBERNSF/Lobato.pdf.

Ewing, B.T., Kruse, J.B., Schroeder, J.L. & Smith, D. A. (2007). Time series analysis of wind speed using VAR and the generalized impulse response technique, *Journal of Wind Engineering and Industrial Aerodynamics* **95(3):** 209-219.

Ewing, B.T., Riggs, K. & Ewing, K. L. (2007). Time series analysis of a predator prey system: Application and generalized impulse response function, *Ecological economics* **60:** 605 – 612.

Fin24 .(2010). 'Wage demand won't affect inflation', [Online] Available at: http://www.fin24.com/Business/Wage-demand-wont-affect-inflation-20100812.

Fackler, J. & Krieger, S. (1986), An Application of Vector Time Series Techniques to Macroeconomic Forecasting, *Journal of Business and Economic Statistics* **4(1):** 71-80.

Friedman J., & Shachmurave, Y. (1997), Co-Movements of Major European Community Stock Markets: A Vector Autoregression Analysis, *Global Finance Journal* **8(2):** 257-277.

Galbraith J. W., Ullah, A. & Zinde-Walsh, V. (2002), Estimation of the Vector Moving Average Model by Vector Autoregression, *Econometric Reviews* **21(2):** 205-219.

Gan, J. (2006). Causality among wildfire, ENSO, timber harvest and urban sprawl: The vector autoregression approach, *Ecological Modelling* **191:** 304-314.

Geweke, J., Messe, R. & Dent, W. (1983). Comparing alternative tests of causality in temporal systems, Analytic results and experimental evidence, *Journal of Econometrics* **21:** 161 – 194.

Ghali, K. (1999). Wage Growth and the Inflation Process: A Multivariate Cointegration Analysis, *Journal of Money, Credit and Banking* **44(2):** 417-431.

Gospodinov, N. (2004). Asymptotic confidence intervals for impulse responses of near – integrated processes, *Econometrics Journal* **7:** 505 – 527.

Granger, C. W .J & Newbold, P. (1986). *Forecasting Economic Time Series*, 2nd edition, Academic Press Inc., San Diego, CA.

Grimaldi, S., Tallerini, C., Serinaldi, F.(2005). Multivariate linear parametric models applied to daily rainfall series, *Advances in Geosciences* **2:** 87-92.

Grubb, H. (1992). A Multivariate Time Series Analysis of some Flour Price Data, *Journal of the Royal Statistical Society Series C (Applied Statistics)* **41(1):** 95-107.

Gupta, R. (2006). Forecasting the South African Economy with VARs and VECMs, *South African Journal of Economics* **27(1):** 315-323.

Gupta, R., Jurgilas, M. & Kabundi, A. (2010). The effect of monetary policy on real house price growth in South Africa: A factor-augmented vector autoregression (FAVAR) approach, *Economic Modelling* **27(1):** 315-323.

Haden, K.L. & van Tassel, L. W. (1988).Application of Vector Autoregression to dynamic relationships within the U.S Dairy Sector, *North Central Journal of Agricultural Economics* **10(2)**: 209 – 216.

Hagnell, M. (1991). A Multivariate Time Series Analysis of Fertility, Adult Mortality, Nuptiality and Real Wages in Sweden 1751 – 1850: a Comparison of Two Different Approaches *, Journal of Official Statistics* **7(4):** 437-455.

Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Hannan, E. J. & Kavaliers, L. (1984). Multivariate linear time series models*, Advances in Applied Probability* **16:** 492 – 561.

Hannan, E. J. & Rissanen, J. (1982).Recursive estimation of mixed autoregressive moving average order, *Biometrika* **69:** 81 -94.

Hanson, M. S. (2006). Varying monetary policy regimes: A vector autoregressive investigation, *Journal of Business and Economic Statistics* **58:** 407-427.

Harris, R. I. J. (1995). *Using Cointegration analysis in econometric modelling*, Prentice Hall/Harvester Wheatsheaf, Hertfordshire, UK .

Hatemi, J. A. & Hacker, R. S. (2009). Can the LR test be helpful in choosing the optimal lag order in the VAR model when information criteria suggest different lag orders?, *Applied Economics*, **41(9):** 1121 – 1125.

Hess, G. & Schweitzer, M.E. (2000). Does Wage Inflation Cause Price Inflation, Federal Reserve Bank of Cleveland Policy, Discussion Paper No 10.

Hillmer, S. C. & Tiao, G. C. (1979). Likelihood Function of Stationary Multiple Autoregressive Moving Average Models, *Journal of the American Statistical Association* **74 :** 652 – 660.

Hosking, J. R. M. (1980). The Multivariate Portmanteau Statistic, *Journal of the American Statistical Association* **75:** 602-605.

Hurvich, C. M. & Tsai, C-L. (1989). Regression and Time Series Model Selection in Small Samples, *Biometrika* **76(2):** 297-307.

James, C., Koreisha, S. & Partch, M. (1985). A VARMA Analysis of the Causal Relations Among Stock Returns, Real Output and Nominal Interest Rates, *Journal of Finance* **40(5):** 1375 – 1384.

Johansen, G. (1988). Statistical analysis of co-integration vectors, *Journal of Economic Dynamics and Control* **12**: 231 – 254.

Johansen, S. & Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration with applications to the demand for money, *Oxford Bulletin of Economics and Statistics* **52:** 169 – 210.

Jonsson, M. & Palmqvist, S. (2004). Do higher wages cause inflation?, Sveriges Riksbank Working Paper Series No. 159.

Kargbo, J. M. (2007). Forecasting agricultural exports and imports in South Africa*, Applied Economics*, **39:** 2069-2084.

Karimi, M. (2011), Order selection criteria for vector autoregressive models, *Signal Processes* **91 :** 955 – 969.

Kascha, C. (2010). A Comparison of Estimation Methods for Vector Autoregressive Moving Average Models, Unpublished Paper, Norges Bank.

Kascha, C. & Trenkler, C. (2011). Forecasting US interest rates with Cointegrated VARMA Models, Working Paper No 33, Department of Economics, University of Zurich.

Kilian, L. (1998). Small- sample confidence intervals for impulse response functions, *Review of Economics and Statistics* **80:** 218 – 230.

Kulshreshtha, M. & Parikh, J. K. (2000). Modelling demand for coal in India : vector autoregression models with cointegrated variables, *Energy* **25:** 149 -168.

Kumar, S. & Webber, D.J. & Perry, G. (2009). Real wages, inflation and labour productivity in Australia, Discussion Papers 0921, University of the West of England, Department of Economics.

Lanne, M. and Saikkonen, P. (2009). Noncausal vector autoregression, Bank of Finland Research Discussion Papers 18.

Lu, M.(2001). Vector Autoregression (VAR) An Approach to Dynamic Analysis of Geographic Processes, *Geografiska Annaler: Series B, Human Geography* **83(2):** 67-75.

Lütkepohl, H & Claessen, H. (1997). Analysis of cointegrated VARMA processes, *Journal of Econometrics* **80:** 223 -239.

Lütkepohl, H & Poskitt, D. S. (1996). Specification of Echelon-Form VARMA Models *, Journal of Business and Economic Statistics* **14 :** 69 – 79.

Lütkepohl, H. & Saikonnen, P. (1999). A lag augmentation test for the cointegrating order of a VAR process, *Economics Letters* **63:** 23 – 27.

Lütkepohl, H. (2004). Forecasting Cointegrated VARMA Processes, *A Companion to Economic Forecasting*, Blackwell Publishing, Oxford,UK.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, Springer, Berlin.

Ma, C. (1997). On the Exact Likelihood Function of a Multivariate Autoregressive Moving average Model, *Biometrika* **84:** 957 – 984.

Maddala, G.S. & Kim,I.-M. (1999). *Unit Roots, Cointegration and Structural Change*, Cambridge, UK.

Marcucci, J & Quagliarello, M. (2008). Is bank portfolio riskiness procyclical: Evidence from Italy using a vector autoregression, *Journal of International Financial Markets, Institutions & Money,* **18:** 46-63.

Mauricio, J. A. (2006). Exact maximum likelihood estimation of partially nonstationary vector ARMA models, *Computational Statistics and Data Analysis* **50:** 2644 – 2662.

Mehra, Y.P. (1993). Unit Labour Costs and the Price Level, *Economic Review* **79(4):** 35-52.

Melard, G., Roy, R. & Saidi, A. (2004). Exact Maximum Likelihood Estimation of Structured or Unit Root Multivariate Time Series Models. Technical report CRM – 312Q, Centre de recherché mathématiques, Université de Montreal.

Montgomery, A.L & Moe, W. W. (2002). Should Music Labels Pay for Radio Airplay? Investigating the Relationship Between Album Sales and Radio Airplay, August 2002.

Morin, N. (2010). Likelihood Ratio Tests on Cointegrating Vectors, Disequilibrium Adjustment Vectors and Their Orthogonal Complements, *European Journal of Pure and Applied Mathematics* **3(3):** 541-571.

Nattrass, N. (2011). The new growth path: Game changing vision or cop-out , *South African Journal of Science* **107:** 8 pages.

Nielsen, B. (2006). Order determination in general vector autoregressions, *IMS Lecture Notes – Monograph Series Time Series and Related Topics* **52:** 93 -112.

Osborn, D. R. (1977). Exact and Approximate Maximum Likelihood estimates for Vector Moving Average Processes, *Journal of the Royal Statistical Society Series B (Methodological)* **39:** 114 – 118.

Oxley, L. & Greasley, D. (1998). Vector autoregression, cointegration and causality: testing for causes of the British industrial revolution, *Applied Economics* **30:** 1387-1397.

Papaikonomou, D. & Pires, J. (2006). Are US output expectations unbiased? A cointegrated VAR analysis in real time, *Economics Letters* **92:** 440 – 446.

Pesavento, E. & Rossi, B. (2006). Small sample Confidence Intervals for Multivariate Impulse Response Functions at Long Horizons, *Journal of Applied Econometrics* **21:** 1135 – 1155.

Pfaff, B. (2008). VAR, SVAR and SVEC Models: Implementation Within R Package vars., *Journal of Statistical Software*, **27(4):** http://www.jstatsoft.org/v27/i04/.

Qu, Z. & Perron, P. (2007). A Modified Information Criterion for Cointegration Tests based on a VAR approximation, *Economic Theory* **23:**638-685.

Raghavan, M. Athanasopoulos, G & Silvapulle, P. (2009). VARMA models for Malaysian Monetary Policy Analysis, Monash Econometrics and Business Statistics Working Papers.

Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis,* 2nd edition, Springer, New York.

Roy, A., Fuller, W. A. & Zhu, Y. (2009). A likelihood based estimator for vector autoregressive processes, *Statistical Methodology* **6:** 304 – 319.

Saidi, A. (2007). Consistent Testing of Non – Correlation of Two Cointegrated ARMA Time Series, *The Canadian Journal of Statistics* **35(1)** 169 – 188.

Saikonnen, P. & Lütkepohl, H. (1996). Infinite-order Cointegrated Vector Autoregressive Processes : Estimation and inference, *Economic Theory* **12(5):** 814 -844.

Simkins, S. (1995). Forecasting with vector autoregressive (VAR) models subject to business cycle restrictions, *International Journal of Forecasting* **11(4)**: 569-583.

Sims, C. A. (1980). Macroeconomics and Reality, *Econometrica* **48(1):** 1-48.

Sims, C. A. & Zha, T. (1999). Error bands for impulse responses, *Econometrica* **67:** 1113 -1155.

Sims, C. A., Stock, J. H. & Watson, M. W. (1990). Inference in linear time series models with some unit roots, *Econometric*a **58:** 113 – 146.

Singh, N., Yadavalli, V.S.S. & Peiris, M.S. (2002).A Note on the Modelling and Analysis of Vector Arma Processes with Nonstationary Innovations, *Mathematical and Computer Modelling* **36:** 1409-1424.

Stergiou, K. I. & Christou, E. D. (1996), Modelling and forecasting annual fisheries catches: comparison of regression, univariate and multivariate time series methods, *Fisheries Research* **25:**105 – 138.

Steyn L. Wild cat strikes tear into economy's growth. Mail and Guardian , 5 October 2012. http:// www.mg.co.za/article/2012-10-05-00. (Accessed 12 November 2012)

Stock, J. H. & Watson, M. W. (2001). Vector Autoregressions, *Journal of Economic Perspectives,* **15(4):**101-115.

Tahir, R. & Ghani, A. A. (2004). Relationship Between Exchange Rates and Stock Prices: Empirical Evidence from Bahrain's Financial Markets*, EcoMod2004 International Conference on Policy, Modeling* ,[Online] Available at: http://www.cepii.fr/anglaisgraph/meetings/2004/3006020704.htm

Tiao, G.C. & Box, G. E. P. (1981). Modeling Multiple Time Series with Applications, *Journal of the American Statistical Association,* **76:** 802 – 816.

Tiao, G. C. & Tsay, R. S. (1983). Multiple Time Series Modeling and Extended Sample Cross-Correlations, *Journal of Business and Economic Statistics* **1:** 43 – 56.

Tiao, G. C. & Tsay, R. S. (1989). Model Specification in Multivariate Time Series(with discussion), *Journal of the Royal Statistical Society Series B (Methodological)* **51:** 157-213.

Toda, H. Y. & Phillips, P. C. B.(1993). Vector autoregression and causality, *Econometrica* **61:** 1367 – 1393.

Toda, H. Y. & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes, *Journal of Econometrics* **66:** 225 – 250.

Todani, K.R., (2006). A partial VAR analysis of wages and price formation in South Africa, Research Department- South African Reserve Bank.

Tsay, R. S. (1989). Parsimonious Parameterization of Vector Autoregressive Moving Average Models, *Journal of Business and Economic Statistics* **7:** 327-341.

Tsay, R. S.(2005). *Analysis of Financial Time Series*, 2nd edition, John Wiley & Sons,Inc,New York.

Veenstra, A. W & Haralambides, H. E. (2001) Multivariate autoregressive models for forecasting seaborne trade flows, *Transportation research Part E* **37:**311-319.

Wei, W. W. S. (1990). *Time Series Analysis, Univariate and Multivariate Methods,* Addison-Wesley, Redwood City, CA.

Wei, W. W. S. (2006). *Time Series Analysis, Univariate and Multivariate Methods: Second Edition,* Addison-Wesley, Redwood City, CA.

Yap, S. F. & Reinsel, G. C. (1995). Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model, *Journal of the American Statistical Association* **90:** 253 -267.

Appendix

Computation of the Maximum likelihood estimates

As discussed earlier, it is not always easy to solve the maximum likelihood estimates.If the normal equations are nonlinear in the parameters, then as noted by Dufour & Jouini (2008), maximizing the likelihood function can be computationally burdensome to determine the order of p and q. The equations thus need to be evaluated by using optimisation techniques, i.e. finding the coefficients which minimise the likelihood function that is specified. These methods are based on those described in Lütkepohl (2005).

In order to proceed, it is necessary to make an assumption that the objective is to minimise a log likelihood function which can be partially differentiated twice say $L(\boldsymbol{\kappa})$. If for a given vector $\boldsymbol{\kappa}_i$, in order to minimise the likelihood function $L(\boldsymbol{\kappa})$, it is important to determine the direction d in which the objective function declines as well as a step of length say s which is performed in that particular direction. Hence in mathematical terms, the objective is to find suitable values of d and s such that the relation $L(\boldsymbol{\kappa}_i + sd) < L(\boldsymbol{\kappa}_i)$ is satisfied

The objective function will always decrease if the step length s is small and the direction d is downward. Thus for a decreasing function, the partial derivative of $L(\boldsymbol{\kappa}_i + sd)$ with respect to the step length s is

$$\frac{dL(\boldsymbol{\kappa}_i + \mathrm{sd})}{ds} = \nabla(\boldsymbol{\kappa}_i)[\frac{dL(\boldsymbol{\kappa}_i + \mathrm{sd})}{ds} | s = 0] = \nabla(\boldsymbol{\kappa}_i)d, \qquad \text{(A.1)}$$

where $\nabla(\boldsymbol{\kappa}_i) = \frac{dL(\boldsymbol{\kappa})}{d\kappa}$ is the gradient of $L(\boldsymbol{\kappa})$ evaluated at point $\boldsymbol{\kappa}_i$.

Suppose d is related to the gradient matrix $\nabla(\boldsymbol{\kappa}_i)$ matrix from the relation $d = -\boldsymbol{D}_i\nabla(\boldsymbol{\kappa}_i)$ where $\boldsymbol{D}_i$ is any positive definite matrix. If the gradient vector of $L(\boldsymbol{\kappa})$ evaluated at $\boldsymbol{\kappa}_i$ is nonzero, then $-\nabla(\boldsymbol{\kappa}_i)'\boldsymbol{D}_i\nabla(\boldsymbol{\kappa}_i) < 0$.

The local minimum occurs when the gradient vector is zero ($\nabla(\boldsymbol{\kappa}_i) = 0$) and from this it follows that the direction d is also 0.

The next step is to evaluate the gradient vector which evaluated at the $i + 1$ th iteration is

$$\boldsymbol{\kappa}_{i+1} = \boldsymbol{\kappa}_i - s_i\boldsymbol{D}_i\nabla(\boldsymbol{\kappa}_i) .$$

This method is known as the steepest descent method. s_i is the step length in the ith direction and $\boldsymbol{D}_i$ as defined earlier is a positive definite matrix. The main limitation of this method is that a large number of iterations are needed. This can be overcome using a modification known as the Newton-Raphson procedure.

In this procedure, note that the objective function $L(\boldsymbol{\kappa}_i)$ can be extended in a Taylor Series around $\boldsymbol{\kappa}_i$

$$L(\boldsymbol{\kappa}) \approx L(\boldsymbol{\kappa}_i) + \nabla(\boldsymbol{\kappa}_i)(\boldsymbol{\kappa} - \boldsymbol{\kappa}_i) + \frac{1}{2}(\boldsymbol{\kappa} - \boldsymbol{\kappa}_i)'\boldsymbol{H}_i(\boldsymbol{\kappa} - \boldsymbol{\kappa}_i) . \quad \text{(A.2)}$$

$\boldsymbol{H}_i = \frac{d^2 L(\boldsymbol{\kappa}_i)}{d\boldsymbol{\kappa} d\boldsymbol{\kappa}'}$ is known as the Hessian matrix and defined as the matrix of partial derivatives of second order with respect to $L(\boldsymbol{\kappa})$ which are evaluated at $\boldsymbol{\kappa}_i$. This Hessian matrix can altenatively be computed by updating the successive gradient matrices. If the Hessian matrix is computed in this way, then the optimisation method used will be known as the Quasi-Newton optimisation method.

The aim of the Newton-Raphson method is to choose the value of $\boldsymbol{\kappa}$ in order to minimise $L(\boldsymbol{\kappa})$. If $L(\kappa)$ is a quadratic function, then the right hand side of (A.2) is equal to $L(\kappa)$ and thus in order for a minimum to be located, the right hand side needs to be partially differentiated once and set equal to 0.

$$\nabla(\boldsymbol{\kappa}_i) + \boldsymbol{H}_i(\boldsymbol{\kappa} - \boldsymbol{\kappa}_i)' = 0 . \quad \text{(A.3)}$$

(A.3) can be used to obtain κ iteratively by starting with an initial guess and then updating using the representation,

$$\boldsymbol{\kappa}_{i+1} = \boldsymbol{\kappa}_i - \boldsymbol{H}_i^{-1}\nabla(\boldsymbol{\kappa}_i) .$$

This is a suitable approximation because κ_i should converge to $\boldsymbol{\kappa}$ as $i \to \infty$.

Thus a suitable choice of the positive definite matrix $\boldsymbol{D}_i$ is $-\boldsymbol{H}_i^{-1}$ .The minimum of $\boldsymbol{D}_i$ can be reached in a step of length s_i = 1 and should converge at a quicker rate than for the previous method. Hamilton (1994) noted that if $L(\boldsymbol{\kappa})$ is not a quadratic function, then the Newton-Raphson procedure is not very powerful, however Lütkepohl (2005) noted that even if $L(\boldsymbol{\kappa})$ is not a quadratic function, a suitable choice of $\boldsymbol{D}_i$ will still be $-\boldsymbol{H}_i^{-1}$.

Since it can become complicated to find the first and second order partial derivatives, the information matrix can be used to estimate the values. The information matrix is defined as

$$\mathrm{Inf}(\boldsymbol{\kappa}) = E\left(\frac{d^2 - lnL(\boldsymbol{\kappa}_i)}{d\boldsymbol{\kappa} d\boldsymbol{\kappa}'}\right) .$$

In most cases, the value of $\mathrm{Inf}(\kappa)$ is estimated using an iteration, $\widehat{\mathrm{Inf}}(\boldsymbol{\kappa}_\mathrm{i})$ and the update $\boldsymbol{\kappa}_{i+1}$ is calculated from an algorithm which is known as the scoring algorithm from

$$\boldsymbol{\kappa}_{i+1} = \boldsymbol{\kappa}_i + s_i\nabla(\boldsymbol{\kappa}_i) .$$

Despite this simplified form, there are still some conditions required for this algorithm to be effective. The first is that the value for the initial iteration $\boldsymbol{\kappa}_1$ needs to be a number close to the minimising vector. Secondly, the step length s_i has to be decided, however for the purposes of simplicity, the value of s_i chosen is usually chosen to be equal to one.

THE KRONECKER PRODUCT

Suppose there are two matrices A and B where

$$\text{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \text{ and B} = \begin{bmatrix} b_{11} & \cdots & b_{1q} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pq} \end{bmatrix}$$

The Kronecker product, A $\otimes$B is defined as

$$\text{A} \otimes \text{B} = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

SAS CODES

```
title'Analysis of South African wage and inflation data';
data inflation;
date = intnx( 'qtr', '01apr96'd, _n_-1 );
format date yyq. ;
input y1 y2  ;

label y1 = 'quarterly inflation rate'
      y2 = 'quarterly wage increases';
datalines;
6.042741341 11.9
7.675438596 11.9
9.130434783 11.9
9.709425939 12
9.381514941 11.6
8.621860149 8.4
7.038512616 4
5.426356589 8.9
5.146124524 10.6
7.6875      8.3
9.05707196  14.7
8.455882353 6.2
7.311178248 5.5
3.308183401 6.4
1.934015927 2.5
2.824858757 5.1
4.898648649 5.5
6.685393258 5.2
7.03125     7.2
7.417582418 7
6.441223833 5.5
```

```
4.739336493 6.7
4.223149114 9
5.677749361 8.3
7.715582451 8.8
10.40723982 15.2
12.75637819 8.4
10.64859632 10.3
7.771535581 10.4
4.690346084 5.4
0.798580302 8.5
0.437445319 10.6
0.651607298 11
1.304915181 14.8
3.125       10.7
2.87456446  12.3
3.236944325 12.5
3.821382568 6.83
3.627827571 11.8
3.767993226 8.63
4.013377926 10.8
5.252274607 11.3
5.560131796 8.05
5.956752346 10.5
6.953376206 10.1
6.99410609  9.7
8.427623878 14.7
9.857527917 15.7
11.68733559 15.5
13.40433346 13.4
11.15509176 10.1
;

proc univariate data = inflation plots;
var y1;
run;

proc univariate data = inflation plots;
var y2;
run;

proc arima data=inflation;
            identify var=y1
            ;
run;

proc arima data=inflation;
            identify var=y2
            ;
run;


proc varmax data=inflation;
            model y1 y2 / lagmax=7 print=(pcorr(12)) print=(covy(5))
print=(corry(5)) print=(parcoef(5))
            minic=(p=7 q=7) dftest;
```

```
                                run;

proc varmax data=inflation;
model y1 y2 / p=4 print=(decompose(15))
printform=univariate;
causal group1=(y1) group2=(y2);
causal group1=(y2) group2=(y1);
output lead=5 back=2;
run;

proc varmax data=inflation;
model y1 y2 / p=2 q=4 print=(decompose(15))
printform=univariate;
causal group1=(y1) group2=(y2);
causal group1=(y2) group2=(y1);
output lead=5 back=2;
run;

proc varmax data=inflation;
nloptions maxiter=50000 maxfunc=100000;
model y1 y2 / p=2 q=4 print=(decompose(15))
printform=univariate;
causal group1=(y1) group2=(y2);
causal group1=(y2) group2=(y1);
output lead=5 back=2;
run;
```

SAS OUTPUTS

The UNIVARIATE Procedure
Variable: y1 (quarterly inflation rate)

Moments

N	51	Sum Weights	51
Mean	6.32869223	Sum Observations	322.763304
Std Deviation	3.13104977	Variance	9.80347266
Skewness	0.13614801	Kurtosis	-0.4320656
Uncorrected SS	2532.84324	Corrected SS	490.173633
Coeff Variation	49.4738827	Std Error Mean	0.43843466

Basic Statistical Measures

Location		Variability	
Mean	6.328692	Std Deviation	3.13105
Median	6.441224	Variance	9.80347
Mode	.	Range	12.96689
		Interquartile Range	4.63450

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	14.43474	Pr > |t|	<.0001
Sign	M	25.5	Pr >= |M|	<.0001
Signed Rank	S	663	Pr >= |S|	<.0001

The UNIVARIATE Procedure
Variable: y2 (quarterly wage increases)

Moments

N	51	Sum Weights	51
Mean	9.61392157	Sum Observations	490.31
Std Deviation	3.18479643	Variance	10.1429283
Skewness	0.01319907	Kurtosis	-0.5354843
Uncorrected SS	5220.9483	Corrected SS	507.146416
Coeff Variation	33.1269234	Std Error Mean	0.4459607

Basic Statistical Measures

Location		Variability	
Mean	9.61392	Std Deviation	3.18480
Median	10.10000	Variance	10.14293
Mode	5.50000	Range	13.20000
		Interquartile Range	4.90000

Simple Summary Statistics

Variable	Type	N	Mean	Standard Deviation	Min	Max
y1	Dependent	51	6.32869	3.13105	0.43745	13.4043
y2	Dependent	51	9.61392	3.18480	2.50000	15.70000

Simple Summary Statistics

Variable	Label
y1	quarterly inflation rate
y2	quarterly wage increases

Dickey-Fuller Unit Root Tests

Variable	Type	Rho	Pr < Rho	Tau	Pr < Tau
y1	Zero Mean	-6.12	0.0829	-1.71	0.0827
	Single Mean	-49.27	0.0004	-4.42	0.0008
	Trend	-50.76	<.0001	-4.36	0.0058
y2	Zero Mean	-1.14	0.4462	-0.77	0.3778
	Single Mean	-15.74	0.0225	-2.74	0.0748
	Trend	-21.28	0.0297	-3.33	0.0736

Lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6
AR 0	4.4497394	4.4867851	3.983289	3.1432471	3.0464515	2.9582079	3.180371
AR 1	2.8663686	3.1129495	2.956916	2.4259054	2.3728138	2.2660139	2.5721951
AR 2	2.2074759	2.540936	2.5304398	2.0613891	2.0589693	2.2445299	2.6022467
AR 3	2.1090981	2.2924446	2.3495812	2.2287159	2.3135218	2.5742089	3.0074063

AR 4	2.052167	2.3811605	2.3935697	2.3867772	2.7058879	3.0049428	3.4461994
AR 5	2.2036022	2.2035825	2.402616	2.6649135	3.0764695	3.2534078	3.8206427
AR 6	2.2659342	2.6409342	2.891706	3.2162643	3.7329068	4.0556353	4.6388246
AR 7	2.3175016	2.7670546	3.2885361	3.8310036	4.4958937	5.0486499	5.8833136

Schematic Representation
of Partial Autoregression

Variable/ Lag	1	2	3	4	5
y1	++	--	..	..	..
y2	.+	..	..	.-	..

Cross Correlations of Dependent Series

Lag	Variable	y1	y2
0	y1	1.00000	0.28485
	y2	0.28485	1.00000
1	y1	0.84276	0.08733
	y2	0.44766	0.49537
2	y1	0.51968	-0.10352
	y2	0.37562	0.40790
3	y1	0.19518	-0.21137
	y2	0.15565	0.34206
4	y1	-0.02642	-0.15330
	y2	-0.05428	-0.05553
5	y1	-0.07850	-0.05494
	y2	-0.19474	0.15917

The VARMAX Procedure

Model Parameter Estimates

Equation	Parameter	Estimate	Standard Error	t Value	Pr > |t|	Variable
y1	CONST1	1.39137	0.78768	1.77	0.0854	1
	AR1_1_1	1.39767	0.17216	8.12	0.0001	y1(t-1)
	AR1_1_2	0.21729	0.06567	3.31	0.0021	y2(t-1)
	AR2_1_1	-0.58608	0.29791	-1.97	0.0565	y1(t-2)
	AR2_1_2	-0.08602	0.06829	-1.26	0.2155	y2(t-2)
	AR3_1_1	-0.07334	0.29494	-0.25	0.8050	y1(t-3)
	AR3_1_2	-0.12947	0.06755	-1.92	0.0628	y2(t-3)
	AR4_1_1	0.10414	0.15903	0.65	0.5165	y1(t-4)
	AR4_1_2	-0.05022	0.07716	-0.65	0.5190	y2(t-4)
y2	CONST2	3.79718	1.87147	2.03	0.0495	1
	AR1_2_1	0.11414	0.40903	0.28	0.7817	y1(t-1)
	AR1_2_2	0.37584	0.15603	2.41	0.0210	y2(t-1)
	AR2_2_1	-0.29076	0.70781	-0.41	0.6835	y1(t-2)
	AR2_2_2	0.31155	0.16226	1.92	0.0624	y2(t-2)
	AR3_2_1	-0.41308	0.70076	-0.59	0.5590	y1(t-3)
	AR3_2_2	0.35861	0.16049	2.23	0.0314	y2(t-3)
	AR4_2_1	0.50622	0.37785	1.34	0.1883	y1(t-4)
	AR4_2_2	-0.39704	0.18333	-2.17	0.0367	y2(t-4)

Granger-Causality Wald Test

Test	DF	Chi-Square	Pr > ChiSq
1	4	14.83	0.0051
2	4	6.55	0.1617

Test 1: Group 1 Variables: y1
Group 2 Variables: y2

Test 2: Group 1 Variables: y2
Group 2 Variables: y1

Portmanteau Test for Cross Correlations of Residuals

Up To Lag	DF	Chi-Square	Pr > ChiSq
5	4	14.44	0.0060
6	8	15.93	0.0434
7	12	16.11	0.1862
8	16	16.97	0.3875
9	20	17.80	0.6004
10	24	22.32	0.5602
11	28	25.46	0.6025
12	32	28.89	0.6249

Forecasts

Variable	Obs	Forecast	Standard Error	95% Confidence Limits		Actual	Residual
y1	50	11.68644	1.05676	9.61522	13.75765	13.40433	1.71790
	51	10.06815	2.02256	6.10400	14.03230	11.15509	1.08694
	52	7.34850	2.73846	1.98122	12.71577		
	53	4.62249	3.06628	-1.38730	10.63229		
	54	2.91102	3.13883	-3.24097	9.06300		

Variable	Obs	Forecast	Error	95% Confidence Limits		Actual	Residual
y2	50	14.46145	2.51080	9.54038	19.38252	13.40000	-1.06145
	51	11.98510	2.69797	6.69718	17.27303	10.10000	-1.88510
	52	10.04560	2.94229	4.27883	15.81238		
	53	9.33889	3.30776	2.85581	15.82198		
	54	9.14098	3.37140	2.53316	15.74880		

Proportions of Prediction Error Covariances by Variable

Variable	Lead	y1	y2
y1	1	1.00000	0.00000
	2	0.93410	0.06590
	3	0.89583	0.10417
	4	0.88394	0.11606
	5	0.87939	0.12061
	6	0.87490	0.12510
	7	0.84260	0.15740
	8	0.79929	0.20071
	9	0.75826	0.24174
	10	0.74161	0.25839
	11	0.74271	0.25729
	12	0.74771	0.25229
	13	0.75032	0.24968
	14	0.75056	0.24944

	15	0.74943	0.25057
y2	1	0.09428	0.90572
	2	0.10479	0.89521
	3	0.09684	0.90316
	4	0.08076	0.91924
	5	0.11473	0.88527
	6	0.13590	0.86410
	7	0.14870	0.85130
	8	0.14860	0.85140
	9	0.15120	0.84880
	10	0.15844	0.84156
	11	0.16644	0.83356
	12	0.16655	0.83345
	13	0.16646	0.83354
	14	0.16739	0.83261
	15	0.16930	0.83070

The VARMAX Procedure

Type of Model	VARMA(2,4)
Estimation Method	Maximum Likelihood Estimation

Constant Estimates

Variable	Constant
y1	3.73696
y2	12.74667

AR Coefficient Estimates

Lag	Variable	y1	y2
1	y1	1.15791	0.01765
	y2	-0.27098	0.46715
2	y1	-0.39890	0.08994
	y2	0.04077	0.60299

MA Coefficient Estimates

Lag	Variable	e1	e2
1	y1	-0.38776	-0.31232
	y2	0.32227	-0.43611
2	y1	-0.19622	-0.25146
	y2	-0.54982	-0.15947
3	y1	-0.63903	-0.11136
	y2	-1.04289	-0.66730
4	y1	0.16698	-0.20811
	y2	-0.06508	0.06941

Schematic Representation of Parameter Estimates

Variable/ Lag	C	AR1	AR2	MA1	MA2	MA3	MA4
y1	*	**	**	**	**	*-	+*
y2	*	**	**	*-	**	**	-*

+ is > 2*std error, - is < -2*std error, . is between, * is N/A

The VARMAX Procedure

Model Parameter Estimates

Equation	Parameter	Estimate	Standard Error	t Value	Pr > |t|	Variable
y1	CONST1	3.73696	0.00000			1
	AR1_1_1	1.15791	0.00000			y1(t-1)
	AR1_1_2	0.01765	0.00000			y2(t-1)
	AR2_1_1	-0.39890	0.00000			y1(t-2)
	AR2_1_2	0.08994	0.00000			y2(t-2)
	MA1_1_1	-0.38776	0.00000			e1(t-1)
	MA1_1_2	-0.31232	0.00000			e2(t-1)
	MA2_1_1	-0.19622	0.00000			e1(t-2)
	MA2_1_2	-0.25146	0.00000			e2(t-2)
	MA3_1_1	-0.63903	0.00000			e1(t-3)
	MA3_1_2	-0.11136	0.00000	-999.00	0.0001	e2(t-3)
	MA4_1_1	0.16698	0.00000	999.00	0.0001	e1(t-4)
	MA4_1_2	-0.20811	0.00000			e2(t-4)
y2	CONST2	12.74667	0.00000			1
	AR1_2_1	-0.27098	0.00000			y1(t-1)
	AR1_2_2	0.46715	0.00000			y2(t-1)
	AR2_2_1	0.04077	0.00000			y1(t-2)
	AR2_2_2	0.60299	0.00000			y2(t-2)
	MA1_2_1	0.32227	0.00000			e1(t-1)
	MA1_2_2	-0.43611	0.00000	-999.00	0.0001	e2(t-1)
	MA2_2_1	-0.54982	0.00000			e1(t-2)
	MA2_2_2	-0.15947	0.00000			e2(t-2)
	MA3_2_1	-1.04289	0.00000			e1(t-3)
	MA3_2_2	-0.66730	0.00000			e2(t-3)
	MA4_2_1	-0.06508	0.00000	-999.00	0.0001	e1(t-4)
	MA4_2_2	0.06941	0.00000			e2(t-4)

The VARMAX Procedure

Type of Model	VARMA(2,4)
Estimation Method	Maximum Likelihood Estimation

Constant Estimates

Variable	Constant
y1	3.81465
y2	12.80784

AR Coefficient Estimates

Lag	Variable	y1	y2
1	y1	0.59802	0.19855
	y2	-1.42826	0.91630
2	y1	-0.21608	0.18711
	y2	0.73688	0.51096

MA Coefficient Estimates

Lag	Variable	e1	e2
1	y1	-0.86262	-0.15374
	y2	-0.38548	0.20832
2	y1	-0.74015	-0.09590
	y2	-0.99028	-0.23890

3	y1	-0.93062	-0.04554
	y2	-1.45365	-0.78026
4	y1	-0.09462	-0.13181
	y2	-0.54736	0.70188

Schematic Representation of Parameter Estimates

Variable/ Lag	C	AR1	AR2	MA1	MA2	MA3	MA4
y1	*	..	.+	-.	..	-.	..
y2	*	.+	..	..	..	.-	.+

+ is > 2*std error, - is < -2*std error, . is between, * is N/A

The VARMAX Procedure

Model Parameter Estimates

Equation	Parameter	Estimate	Standard Error	t Value	Pr > |t|	Variable
y1	CONST1	3.81465	0.00000			1
	AR1_1_1	0.59802	0.35338	1.69	0.0972	y1(t-1)
	AR1_1_2	0.19855	0.12121	1.64	0.1081	y2(t-1)
	AR2_1_1	-0.21608	0.24323	-0.89	0.3789	y1(t-2)
	AR2_1_2	0.18711	0.06976	2.68	0.0101	y2(t-2)
	MA1_1_1	-0.86262	0.41240	-2.09	0.0419	e1(t-1)
	MA1_1_2	-0.15374	0.18866	-0.81	0.4192	e2(t-1)
	MA2_1_1	-0.74015	0.42172	-1.76	0.0858	e1(t-2)
	MA2_1_2	-0.09590	0.09796	-0.98	0.3326	e2(t-2)
	MA3_1_1	-0.93062	0.35355	-2.63	0.0114	e1(t-3)
	MA3_1_2	-0.04554	0.09936	-0.46	0.6488	e2(t-3)
	MA4_1_1	-0.09462	0.27833	-0.34	0.7354	e1(t-4)
	MA4_1_2	-0.13181	0.18078	-0.73	0.4696	e2(t-4)
y2	CONST2	12.80784	0.00000			1
	AR1_2_1	-1.42826	1.30266	-1.10	0.2785	y1(t-1)
	AR1_2_2	0.91630	0.23837	3.84	0.0004	y2(t-1)
	AR2_2_1	0.73688	0.60088	1.23	0.2262	y1(t-2)
	AR2_2_2	0.51096	0.52952	0.96	0.3395	y2(t-2)
	MA1_2_1	-0.38548	1.40968	-0.27	0.7857	e1(t-1)
	MA1_2_2	0.20832	0.28449	0.73	0.4676	e2(t-1)
	MA2_2_1	-0.99028	1.31007	-0.76	0.4535	e1(t-2)
	MA2_2_2	-0.23890	0.19693	-1.21	0.2312	e2(t-2)
	MA3_2_1	-1.45365	1.45597	-1.00	0.3232	e1(t-3)
	MA3_2_2	-0.78026	0.23179	-3.37	0.0015	e2(t-3)
	MA4_2_1	-0.54736	1.75971	-0.31	0.7571	e1(t-4)
	MA4_2_2	0.70188	0.32836	2.14	0.0378	e2(t-4)

Granger-Causality Wald Test

Test	DF	Chi-Square	Pr > ChiSq
1	2	10.89	0.0043
2	2	4.13	0.1267

Test 1: Group 1 Variables: y1
Group 2 Variables: y2

Test 2: Group 1 Variables: y2
Group 2 Variables: y1

Portmanteau Test for Cross Correlations of Residuals

Up To Lag	DF	Chi-Square	Pr > ChiSq
7	4	266.39	<.0001
8	8	304.37	<.0001
9	12	327.40	<.0001
10	16	368.60	<.0001
11	20	423.23	<.0001
12	24	446.22	<.0001

Forecasts

Variable	Obs	Forecast	Standard Error	95% Confidence Limits		Actual	Residual
y1	50	16.51844	0.67678	15.19198	17.84490	13.40433	-3.11411
	51	19.38295	1.61659	16.21450	22.55140	11.15509	-8.22786
	52	24.59603	2.78026	19.14683	30.04524		
	53	27.77922	3.92403	20.08826	35.47018		
	54	30.95718	4.90962	21.33451	40.57985		

Variable	Obs	Forecast	Error	95% Confidence Limits		Actual	Residual
y2	50	24.30200	2.28313	19.82714	28.77685	13.40000	-10.90200
	51	30.46294	3.01800	24.54776	36.37812	10.10000	-20.36294
	52	44.99528	3.90537	37.34089	52.64967		
	53	46.38927	5.29124	36.01863	56.75991		
	54	55.02422	5.81945	43.61831	66.43013		