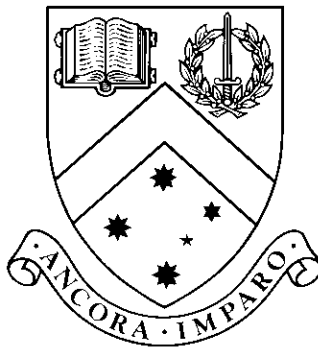


# Minimum Message Length Model Selection for Autoregressive Conditional Heteroskedastic Time Series

by

Samuel Gundry, BCS



**Thesis**

Submitted by Samuel Gundry

in partial fulfillment of the Requirements for the Degree of  
**Bachelor of Computer Science with Honours (1608)**

Supervisor: Assoc. Prof. David L. Dowe

**Clayton School of Information Technology  
Monash University**

November, 2006

© Copyright

by

Samuel Gundry

2006

# Contents

<b>List of Tables</b> . . . . .	<b>vi</b>
<b>List of Figures</b> . . . . .	<b>vii</b>
<b>Abstract</b> . . . . .	<b>viii</b>
<b>Acknowledgments</b> . . . . .	<b>x</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
<b>2 Time Series</b> . . . . .	<b>4</b>
2.1 Overview . . . . .	4
2.2 Definition . . . . .	4
2.3 Discrete and Continuous . . . . .	5
2.4 Deterministic and Stochastic . . . . .	5
2.5 Features . . . . .	6
2.5.1 Trends . . . . .	6
2.5.2 Seasonality . . . . .	6
2.5.3 Irregularity . . . . .	6
2.5.4 Conditional Heteroskedasticity . . . . .	7
2.5.5 Non-Linearity . . . . .	8
2.6 Analysis . . . . .	8
2.7 Summary . . . . .	9
<b>3 Time Series Models</b> . . . . .	<b>10</b>
3.1 Overview . . . . .	10
3.2 Concepts . . . . .	11
3.2.1 Mean and Expectation . . . . .	11
3.2.2 Variance . . . . .	11
3.2.3 Autocovariance . . . . .	11
3.2.4 Autocorrelation . . . . .	12
3.2.5 Stationarity . . . . .	12

3.2.6	Disturbances . . . . .	12
3.2.7	Model Order . . . . .	13
3.3	Homoskedastic Models . . . . .	13
3.3.1	Moving Average (MA) . . . . .	13
3.3.2	Autoregressive (AR) . . . . .	13
3.3.3	Autoregressive Moving Average (ARMA) . . . . .	14
3.3.4	Autoregressive Integrated Moving Average (ARIMA) . . . . .	15
3.4	Heteroskedastic Models . . . . .	16
3.4.1	Autoregressive Conditional Heteroskedastic (ARCH) . . . . .	16
3.4.2	Generalised ARCH (GARCH) . . . . .	18
3.5	Summary . . . . .	19
<b>4</b>	<b>Model Selection . . . . .</b>	<b>20</b>
4.1	Overview . . . . .	20
4.2	Existing Criteria . . . . .	21
4.2.1	Akaike's Information Criterion (AIC) . . . . .	21
4.2.2	Corrected Akaike's Information Criterion ( $AIC_c$ ) . . . . .	21
4.2.3	Bayesian Information Criterion (BIC) . . . . .	22
4.2.4	Hannan-Quinn Criterion (HQ) . . . . .	22
4.2.5	Minimum Description Length (MDL) . . . . .	22
4.3	Minimum Message Length . . . . .	23
4.4	Summary . . . . .	24
<b>5</b>	<b>Minimum Message Length . . . . .</b>	<b>25</b>
5.1	Introduction . . . . .	25
5.2	Encoding the Message . . . . .	26
5.3	Strict Minimum Message Length (SMML) . . . . .	27
5.4	MML87 . . . . .	28
5.5	MMLD (IID) . . . . .	29
5.6	Summary . . . . .	30
<b>6</b>	<b>MML87 Model Selection Formulation . . . . .</b>	<b>31</b>
6.1	Likelihood . . . . .	32
6.2	Bayesian Priors . . . . .	33
6.3	Lattice Constants . . . . .	34
6.4	Fisher Information . . . . .	34
6.5	Small Sample Approximation . . . . .	35
6.6	Summary . . . . .	35

<b>7</b>	<b>MMLD (IID) Model Selection Formulation</b>	<b>36</b>
7.1	Message from Monte Carlo	37
7.2	Sampling from the Posterior	38
7.3	An Envelope Distribution	38
7.4	Approximating the Optimal Uncertainty Region	39
7.5	Choosing the Point Estimate	40
7.6	Summary	40
<b>8</b>	<b>Evaluations</b>	<b>41</b>
8.1	Performance Measures	41
8.2	Simulations	42
8.3	Results	43
8.3.1	MML87 vs Existing Criteria	43
8.3.2	MMLD Results	46
8.4	Discussion of Results	46
8.4.1	MMLD Discussion	47
8.5	Summary	48
<b>9</b>	<b>Conclusion</b>	<b>50</b>
9.1	Future Work	51
	<b>References</b>	<b>53</b>
	<b>Appendix A ARCH(p) First and Second Derivatives</b>	<b>58</b>
	<b>Appendix B Score Algorithm</b>	<b>60</b>
	<b>Appendix C ARCH(p) Fisher Information</b>	<b>61</b>
	<b>Appendix D Complete Results</b>	<b>63</b>

# List of Tables

8.1	Results for Average Correct Model Order Selection . . . . .	44
8.2	Results for Average Mean Square Prediction Error . . . . .	44
8.3	Results for Average Negative Log Likelihood . . . . .	44
8.4	Aggregate results for Criteria under/correctly/over selecting the true model . . . . .	44
8.5	Results for MMLD for $T = \{40, 70\}$ and $p = \{1, 2, 3\}$ . . . . .	46
D.1	Results for $T = 40$ and $p = 1$ . . . . .	63
D.2	Results for $T = 40$ and $p = 2$ . . . . .	64
D.3	Results for $T = 40$ and $p = 3$ . . . . .	64
D.4	Results for $T = 40$ and $p = 4$ . . . . .	64
D.5	Results for $T = 70$ and $p = 1$ . . . . .	65
D.6	Results for $T = 70$ and $p = 2$ . . . . .	65
D.7	Results for $T = 70$ and $p = 3$ . . . . .	65
D.8	Results for $T = 70$ and $p = 4$ . . . . .	66
D.9	Results for $T = 100$ and $p = 1$ . . . . .	66
D.10	Results for $T = 100$ and $p = 2$ . . . . .	66
D.11	Results for $T = 100$ and $p = 3$ . . . . .	67
D.12	Results for $T = 100$ and $p = 4$ . . . . .	67
D.13	Results for $T = 200$ and $p = 1$ . . . . .	67
D.14	Results for $T = 200$ and $p = 2$ . . . . .	68
D.15	Results for $T = 200$ and $p = 3$ . . . . .	68
D.16	Results for $T = 200$ and $p = 4$ . . . . .	68

# List of Figures

1.1	Time Series of Melbourne's average monthly price for unleaded petrol[3].	2
1.2	Time Series of the global monthly average temperature anomalies[41].	3
2.1	Example of a time series exhibiting upward trend and seasonal features[13].	5
2.2	Example of a time series exhibiting changing trend[14]. . . . .	7
2.3	Example of a time series exhibiting an irregularity[15]. . . . .	8
2.4	Example of a time series exhibiting volatility[16]. . . . .	9
3.1	Time series plot of an MA(1) model with $\beta_1 = 0.2$ . . . . .	14
3.2	Time series plot of an AR(1) model with $\alpha_1 = 0.9$ . . . . .	15
3.3	Time series plot of a zero mean ARCH(1) process with $\alpha = (0.2, 0.9)$ .	17
3.4	Time series plot of an AR(1) process with ARCH(1) errors. . . . .	18
8.1	Results for Average Correct Model Order Selection (%) . . . . .	45
8.2	Results for Average Mean Square Prediction Error . . . . .	45
8.3	Results for Average Negative Log Likelihood . . . . .	46

# Minimum Message Length Model Selection for Autoregressive Conditional Heteroskedastic Time Series

Samuel Gundry, BCS  
sdgun1@student.monash.edu.au  
Monash University, 2006

Supervisor: Assoc. Prof. David L. Dowe  
David.Dowe@infotech.monash.edu.au

## Abstract

Time series exist in a broad range of disciplines and are of particular interest in econometrics and finance. Mathematical processes may be used to model time series data, facilitating a better understanding and providing an ability to forecast future behaviour. A time series exhibiting volatility is more difficult to model and, consequently, predict. If the variance changes over time then it is referred to heteroskedastic.

Traditional time series models assume the variance is constant over time and do not allow for heteroskedasticity. Engle introduced his Autoregressive Conditional Heteroskedastic (ARCH) time series model to avoid this assumption and, in doing so, provides a process to model the variance as a function of time.

Model selection aims to identify which model fits best given some time series data. Choosing the ‘best’ model will improve our understanding and increase future behaviour of the underlying phenomenon being studied. Currently, existing model selection methods such as Akaike’s Information Criterion (AIC) and Schwarz’ Bayesian IC (BIC) are used. However, their statistical properties as ARCH model selection is mostly unknown.

This thesis provides two new Minimum Message Length (MML) selection criteria for ARCH time series models. MML is an information-theoretic framework for statistical and inductive inference, and has been successfully applied as model selection to AR and MA time series.

We present ARCH formulations for two MML approximations, MML87 and MMLD. These are empirically compared in Monte Carlo simulations against current existing criteria AIC, corrected AIC, BIC and HQ.

Performance results from these simulations are quantitatively assessed, some insights given and possible future work for our ARCH MML estimators are suggested.

# Minimum Message Length Model Selection for Autoregressive Conditional Heteroskedastic Time Series

## Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

---

Samuel Gundry  
November 20, 2006

# Acknowledgments

I am most grateful to my supervisor, Assoc. Prof. David Dowe, not only for endless hours of support, advice and guidance, but also for the wondrous tales, bemusing analogies and most importantly, for his unwavering encouragement over the last eight months. Thank you.

To my girlfriend, Katie, thank you for putting up with a stressed boyfriend and countless late nights. You helped tremendously through the rough periods and provided a sanctuary away from thesis thoughts.

To my family (and Anna), who have shown interest for the entire year, thank you for your continued support.

Samuel Gundry

*Monash University*

*November 2006*

# Chapter 1

## Introduction

Time series are sequences of observations recorded in time of a particular phenomenon. Any ‘thing’ which has information available over time may be modelled using time series and, as such, are studied extensively in research and have widespread applications to a diverse range of fields.

Local petrol prices and global temperatures are two interesting phenomena to observe of late. Figure 1.1 is a time series of the average (unleaded) petrol prices in Melbourne for the last six months[3]. Figure 1.2 is a time series of the global average monthly temperature anomalies<sup>1</sup> since 1860. Examples of other time series include yearly population rates, average runs per innings, heart rate per second, average monthly rainfall and daily stock prices. Clearly, from the above examples, time series exist in many areas such as Economics and Finance, Meteorology, Statistics, Sport, Medicine and Health and the many sciences.

Often, time series are modelled to assist understanding and forecast future behaviour of the phenomenon being studied. Both deterministic and stochastic methods are employed to model the different kinds of time series. A broad range of model classes use randomness to simulate phenomena irregularities. Consequently, with a large range of models to choose from, it is possible (and likely) that more than one may be appropriate to a particular phenomenon.

The practise of selecting the most appropriate model from a set of candidate models is known as *Model Selection*. Obviously, the ability to select the model which fits best is desirable. However, what is less clear, is that fitting a model ‘too well’ can lead to an inability to accurately forecast future behaviour and generalise to other, similar, time series. Model selection methods try avoid over-fitting whilst still choosing a model which fits adequately.

---

<sup>1</sup> Anomalies in temperature are the departures from the time period average. According to the Bureau of Meteorology[41], “anomalies tend to be more consistent throughout wide areas than actual temperatures.”

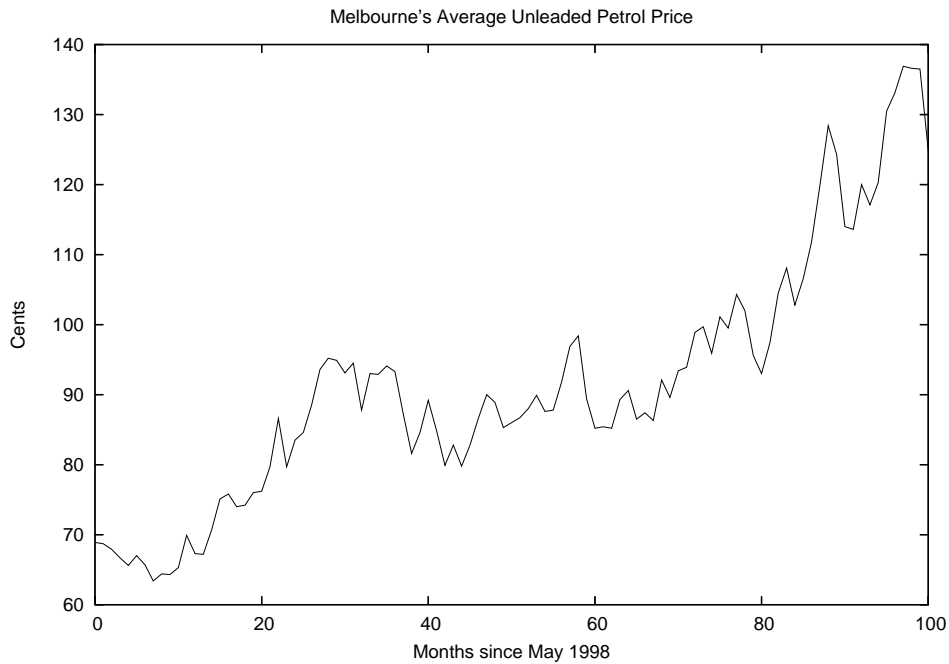


Figure 1.1: Time Series of Melbourne's average monthly price for unleaded petrol[3].

Many time series, especially in Econometrics and Finance, exhibit high volatility. Highly volatile time series occur when a phenomenon's behaviour, such as the petrol price (Figure 1.1), is dramatically different over time periods. Often, this variance changes over time and is referred to as *heteroskedasticity*. That is, there may be a high volatility during one period of time while during the subsequent time periods, there is a much lower variance. Obviously, highly volatile time series, particularly with changing variance, are more difficult to model and predict. Consequently, a class of models were designed specifically to simulate a changing variance over time. These are referred to as Autoregressive Conditional Heteroskedastic, or ARCH, models[21].

Model selection specific to ARCH models, however, has not yet been thoroughly investigated. Currently, existing methods are used which were originally developed for different model classes and, as such, their properties in the ARCH context are unknown[20, page 135].

The Minimum Message Length (MML) principle[52, 48] may provide a method for ARCH model selection and possibly improve upon existing methods in this regard. MML has been successfully used for model selection to a number of problems. Specifically to time series, MML performed well when compared to other existing methods during empirical simulations for Autoregressive models[28] and Moving Average models[44].

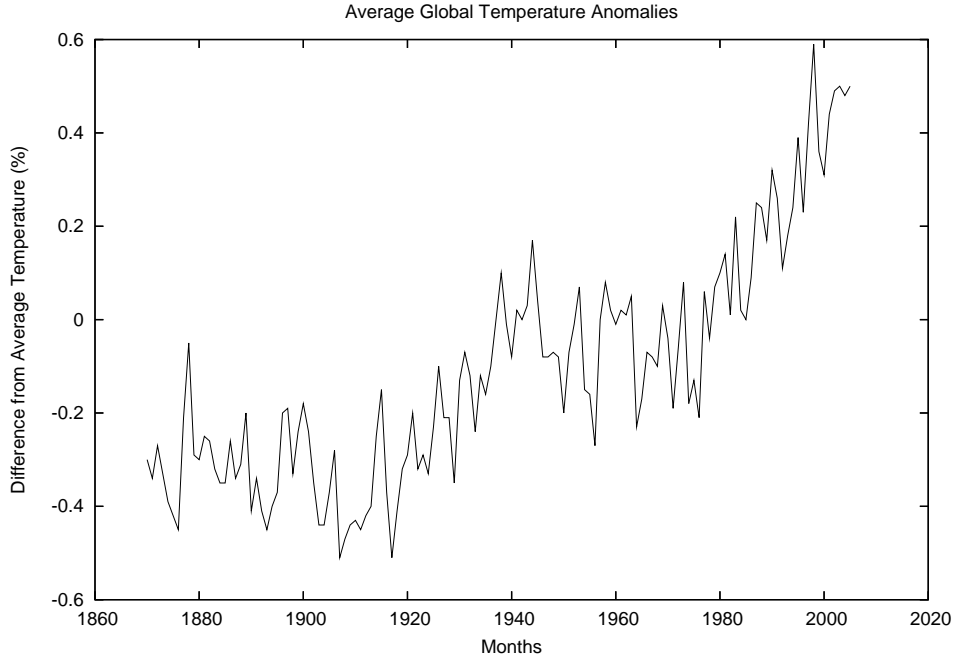


Figure 1.2: Time Series of the global monthly average temperature anomalies[41].

Our thesis investigated MML as model selection for volatile time series, specifically to the ARCH model family.

We begin with an overview of time series and their applications in Chapter 2, where we also discuss their main features and provide examples. Time series models, their underlying concepts, two main classes (homoskedastic and heteroskedastic) and their specific models are introduced in Chapter 3. The ARCH model is also formally defined and an example given in this chapter.

Chapter 4 provides an overview of model selection and its importance, additional to some background information and definitions of existing methods. MML as model selection is also discussed before we introduce you formally to the MML principle in the following chapter. The two MML approximations we investigated, MML87[52] and MMLD[37, 24, 26, 27], are introduced in this chapter before their formulations as ARCH model selection are presented in Chapters 6 and 7, respectively.

Our evaluation methodologies are discussed and results from empirical simulations (comparing MML87 and MMLD to other existing methods) are presented in Chapter 8. The following chapter discusses these results before concluding in Chapter 10 with our research limitations and possible future work.

# Chapter 2

## Time Series

### 2.1 Overview

Time series are studied and applied extensively throughout many disciplines such as engineering, meteorology, social sciences and particularly in finance and economics. Monthly rainfall, daily temperature, annual unemployment rates, daily share prices, and yearly populations are a small example of the vast applications of time series.

*Time plots* are an intuitive, natural and effective representation method for visualising time series and may show important features. Figure 2.1 is a time plot for residential sales of a gas and electric company. The upward trend and regular yearly seasonal affect are two obvious features, possibly attributed to an increasing client base (i.e. population) and seasonal temperature changes (i.e. Summer and Winter). Chatfield[10] considers the time plot as the most important step in time series analysis.

The ability to analyse and predict time series is possible through time series models. According to Mansfield[39, page 559], “business executives and economists pore over time series”, hence, the ability to model such data to allow analysis for a better understanding of the underlying processes and more refined capabilities to forecast or predict future behaviour is of great interest.

### 2.2 Definition

Formally, a time series is a series of observations,  $y_t$ , recorded<sup>1</sup> sequentially and uniformly in time,  $t$ , of a particular phenomenon. The initial observation is considered recorded at  $t = 1$  and the final observation at  $t = T$ , consequently, observations range from  $y_1$  to  $y_T$ . The units of measurement depend on the phenomenon being observed and can be any unit of time (e.g. decade, year, hour, minute, second, etc).

---

<sup>1</sup>Throughout this thesis; recorded, taken, measured and observed all refer to the collection or recording of data or ‘observations’ and used interchangeably.

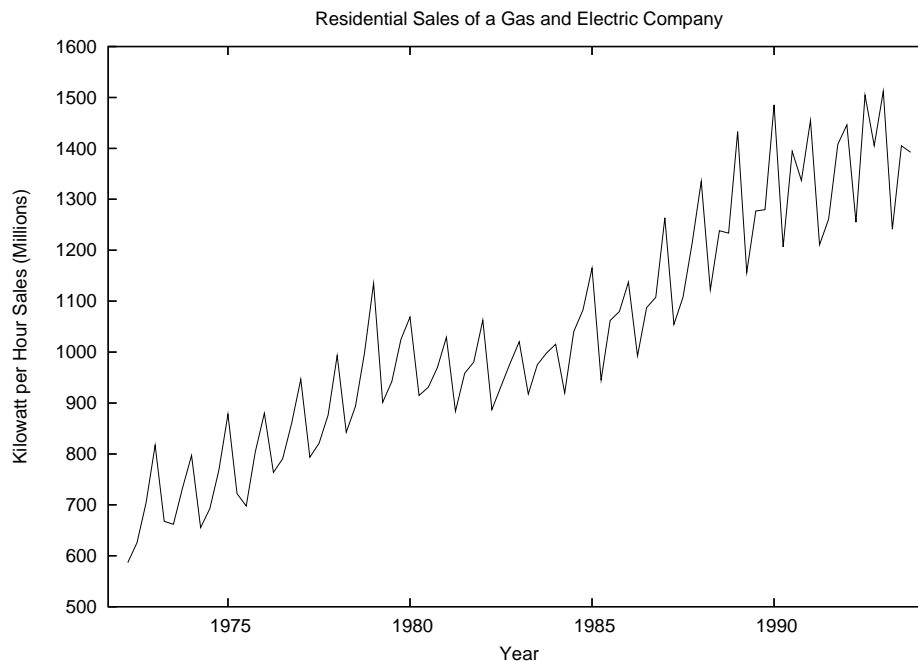


Figure 2.1: Example of a time series exhibiting upward trend and seasonal features[13].

## 2.3 Discrete and Continuous

Time series can be either *discrete* or *continuous*. A continuous time series occurs when the observed data,  $y_t$ , has values continuously through time,  $t$ . Whereas, discrete time series occur when the observed data,  $y_t$ , has values only at fixed time intervals,  $\tau$ , such that  $\{y_t : t = 1, 1 + \tau, \dots, T - \tau, T\}$ . A discrete time series may arise when either continuous data are recorded discretely (e.g. temperature measured once per day), accumulative values of observations are of interest (e.g. rainfall over a week), or the phenomenon observed is actually discrete (e.g. length of a queue for a day)[7, 10].

## 2.4 Deterministic and Stochastic

An important feature of time series data is the dependence on time order. A time series observation at  $y_t$  may partially or fully depend on previous observations,  $y_{t-1}$ ,  $y_{t-2}$ , etc. Fully dependent time series are referred to as *deterministic* as their future behaviour can be predicted exactly using only past observations. However, in practise, these are unlikely to occur as observations also partially depend on exogenous factors[47, page 2921]. Such time series - only partially dependent on past observations - are referred to as *stochastic* since the exogenous factors appear to behave randomly.

## 2.5 Features

Additional to the broad categories of time series previously introduced, all time series - either discrete or continuous and either deterministic or stochastic - may share common patterns or features. These features are sometimes obvious, such as the upward trend and seasonal components in figure 2.1. Franses[29] lists both *trend* and *seasonality*, along with *aberrant observations* (also referred to as irregularities, disturbances, random observations or even noise), *conditional heteroskedasticity* (changing variance) and *non-linearity* as the five key features of time series. A time series may, and frequently does, display more than one key feature. These five key features will now be discussed.

### 2.5.1 Trends

A trend is the general behaviour of a series over a given length of time. They can be, for example, either upward or downward, steep or slight, exponential or linear, and may also change over time. Figure 2.2 shows the consumption of cigarettes per adult in Turkey between 1960 and 1990. There is a slight upward trend during the first ten years before a steep upward trend - corresponding to rapid increase in consumption - during the seventies. A rapid decrease in consumption leads to a steep downward trend in the final decade. Of course, the period of time considered will alter a trend. For instance, in the long term, the trend in Figure 2.2 may be considered only slightly upward. That is, in 1990, the consumption is only marginally higher than in 1960.

### 2.5.2 Seasonality

A time series exhibits seasonal behaviour when observations at regular time periods are significantly different from those otherwise. Seasonal affects, often referred to as cyclical, may change over time. Figure 2.1 shows a clearly discernible yearly seasonal component with regular cyclical peaks and troughs during each year. These are most likely attributed to the cold and warm seasons. Another example, which may be familiar to the reader, is the increased sales in retail each year before Christmas.

### 2.5.3 Irregularity

Time series are often affected by events which seem to occur randomly, such as, strikes, government decisions, natural disasters, war. Economic and financial time series, according to Chatfield[10], are especially affected. A stochastic or random event at time  $t$  is usually evident from a significant change in observations, and may not only effect the current observation,  $y_t$ , but also subsequent observations,  $y_{t+1}$ ,

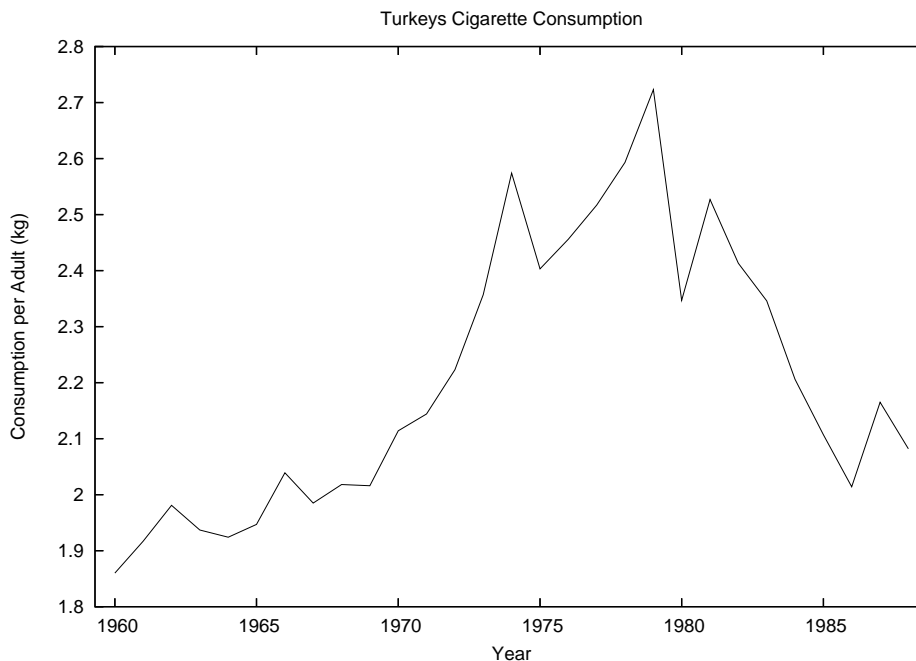


Figure 2.2: Example of a time series exhibiting changing trend[14].

$y_{t+2}$ , etc. Franses[29] refers to such features as “aberrant observations”, Chatfield[10] as “irregular fluctuations” and according to Enders[20], they are “irregular components”.

Figure 2.3 contains an example of random behaviour. The sharp drop in total expenditure by the U.S. military between 1945 and 1947 corresponds to the end of World War II - an event which you would not expect to occur regularly.

### 2.5.4 Conditional Heteroskedasticity

Often, particularly in econometric and financial time series[29, 10], irregularities cluster together and successive observations exhibit sharp changes. These are generally referred to as “volatility clusters”[29]. Frequently, sequences of highly volatile observations are followed by sequences of low volatility, or conversely, observations with low volatility are followed by highly volatile sequences. Figure 2.4 shows such behaviour. The time series is the monthly returns in percentage for a stock listed on the New York Stock Exchange. The successive observations during 1993 to 1995 have relatively low volatility, followed by observations with relatively high volatility. Clearly, periods with higher volatility are more difficult to predict[8, 20].



Figure 2.3: Example of a time series exhibiting an irregularity[15].

### 2.5.5 Non-Linearity

A different average change between successive observations in a time series usually indicates non-linear behaviour. That is, a series may either increase at a faster rate than it falls or, alternatively, fall more rapidly than increase. Enders suggests a number of economic time series should display non-linearity[20]. Quite often, these will either fall sharply and rise slowly (e.g. industrial output) or rise sharply and fall slowly (e.g. unemployment rate)[29].

For the purpose of this research, linear time series only will be considered. Non-linear time series imply more complicated mathematical models[10] and are considered beyond the scope of this research. An interested reader is invited to consult Chatfield[10, Ch 11], Enders[20, Ch 7] and Franses[29, Ch 8] for a good introduction.

## 2.6 Analysis

The process of formally identifying these five key features and understanding the factors responsible for time series behaviour is referred to as *time series analysis*. According to Chatfield[10], there are four objectives of time series analysis, *description*, *explanation*, *prediction* and *control*.

Occasionally, simple descriptive techniques, such as the time plot, may identify obvious features and describe a time series sufficiently[10]. Such series are often modelled adequately using simple processes which decompose a series into a trend,

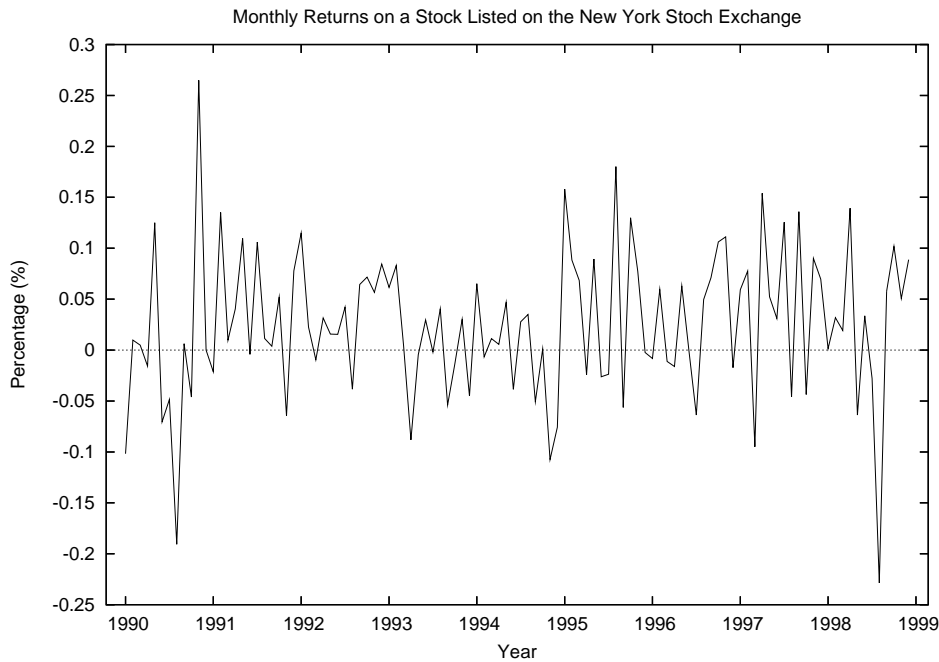


Figure 2.4: Example of a time series exhibiting volatility[16].

seasonal or cyclical components and possibly an irregular component. However, for most time series, more sophisticated models are required to provide an adequate description and more accurate predictions for future behaviour.

## 2.7 Summary

Time series were defined and their importance to many disciplines, such as engineering, meteorology, medicine and particularly in finance and economics, were briefly discussed. The two main types of time series currently studied and their five key features (trends, seasonality, irregularities, heteroskedasticity and non-linearity) were introduced and examples provided. The objectives of time series analysis, description, explanation, prediction and control were briefly explained and various approaches, including modelling, briefly discussed. The concepts of time series modelling, specific models and their underlying foundations will now be introduced.

# Chapter 3

## Time Series Models

### 3.1 Overview

Mathematical processes have long been used to model and describe the behaviour of physical phenomena. The corresponding model is capable of representing observed data concisely, facilitate a better understanding of the underlying processes and, consequently, forecast future behaviour.

If the future behaviour of a particular phenomenon can be predicted exactly, the phenomenon is called *deterministic*. However, Box and Jenkins[7] suggest that no phenomena are fully deterministic, and that unknown factors imply models for purely deterministic phenomena need not apply. Chatfield partially agrees and believes that ‘most’ phenomena contain a ‘random element’[10, page 33].

In such circumstances, stochastic or statistical processes are used to determine a phenomenon’s behaviour probabilistically. When applied to sequences of time dependent observations (time series), these stochastic processes are referred to as *stochastic time series models*<sup>1</sup>.

Thus, a time series is considered one particular realisation of the infinite set, or so called *ensemble*, of the infinite possible series generated by a specific stochastic process or time series model.

Time series models are a convenient method for conceptualising a time series[38]. A model assists with interpretation and understanding of the underlying processes, predicting or forecasting future observations and, subsequently, helping make informed decisions.

There are a vast array of models available for time series (see, for example, [33, 7, 20]), however, many of these extend either the ARIMA models, pioneered by Box and Jenkins[7] or the ARCH[21] and GARCH[5] model family. Certain models

---

<sup>1</sup>Usually, the stochastic element is assumed and such processes are simply referred to as *time series models*.

are more appropriate to different time series, and can be categorised according to their assumptions, namely, stationarity, linearity and volatility.

Volatility (§2.5.4) and linearity (§2.5.5) have already been qualitatively discussed, however, both these and other important time series concepts, such as expectation, correlations and stationarity, need to be introduced and defined more formally to assist understanding of the models introduced in Sections 3.3 and 3.4.

## 3.2 Concepts

### 3.2.1 Mean and Expectation

The mean for a particular observation,  $y_t$ , at time  $t$  is given by

$$\mu(t) = E(y_t)$$

where  $t = 1, \dots, T$  and  $E(y_t)$  is the expectation function which calculates the expected value of observation  $y_t$ .

### 3.2.2 Variance

The variance for a particular observation  $y_t$  at time  $t$  is given by

$$\sigma^2(t) = E[(y_t - \mu(t))^2]$$

where  $\mu(t)$  is the mean at time  $t$  and  $E(\cdot)$  is the expectation function (§3.2.1). If the variance is constant for all  $t$ , then it will be given by  $\sigma^2$ .

### 3.2.3 Autocovariance

The autocovariance function<sup>2</sup> (abbreviated acf) measures the dependency between two variables generated from the same stochastic process. Consider the observations  $y_t$  and  $y_{t+\tau}$  generated by a stochastic process and separated by  $\tau$  time intervals. The acf is then given by

$$\gamma(\tau) = COV(y_t, y_{t+\tau}) = E[(y_t - \mu(t))(y_{t+\tau} - \mu(t + \tau))]$$

where  $\mu(\cdot)$  is the mean, from above.

---

<sup>2</sup>The time series covariance defined here measures the dependence between values generated by the same process - hence the prefix 'auto' - as against the traditional use of covariance when measuring dependence between two values generated by different processes.

The acf will equal zero if  $y_t$  and  $y_{t+\tau}$  are independent, otherwise positive or negative depending if high  $y_t$  values go with high or low  $y_{t+\tau}$  values. Interpretation of the autocovariance function is made more difficult since the size depends on the units of measurement of  $y$ , thus it is useful to standardise it[10].

### 3.2.4 Autocorrelation

The autocorrelation function standardises the autocovariance (§3.2.3) and is given by

$$\begin{aligned}\rho(\tau) &= \frac{E[(y_t - \mu(t))(y_{t+\tau} - \mu(t + \tau))]}{\sqrt{E[(y_t - \mu(t))^2]E[(y_{t+\tau} - \mu(t + \tau))^2]}} \\ &= \frac{\gamma(\tau)}{\sqrt{\sigma^2(t)\sigma^2(t + \tau)}}\end{aligned}$$

where, from above,  $\gamma(\tau)$  is the autocovariance at  $\tau$  and  $\sigma^2(t)$  and  $\sigma^2(t + \tau)$  are the variances at time  $t$  and  $t + \tau$  respectively.

### 3.2.5 Stationarity

A process,  $y_t$ , is said to be *strictly stationary*, also referred to as *strongly stationary*, if and only if the joint distribution of  $(y_{t+t_1}, \dots, y_{t+t_\tau})$  is the same as the joint distribution of  $(y_{t_1}, \dots, y_{t_\tau})$  for every  $t, t_1, \tau, t_n$ [33].

However, often in practise a weaker definition for stationarity is used and is defined

$$\begin{aligned}\mu_t &= \mu \quad \forall t \\ COV[y_t, y_{t+\tau}] &= \gamma(\tau) \quad \forall t\end{aligned}$$

That is, the mean does not depend on time and the autocovariance depends only on the time interval. Such processes are referred to as *weakly stationary* or *second-order stationary*.

### 3.2.6 Disturbances

A phenomenon's irregularities may be incorporated as *disturbances*, or *errors*, into a time series model via a random process,  $\epsilon_t$ . If  $\epsilon_t$  has zero mean and constant variance:

$$\begin{aligned}E(\epsilon_t) &= 0 \\ E(\epsilon_t^2) &= \sigma^2\end{aligned}$$

and is independent across time:

$$E(\epsilon_t \epsilon_\tau) = 0 \quad \forall t \neq \tau$$

then they are referred to as *white noise* or simply *noise*.

Often,  $\epsilon_t$  is assumed to be generated from a Normal distribution, denoted by

$$\epsilon_t \sim N(\mu, \sigma^2)$$

where  $\mu$  is the mean and  $\sigma^2$  is the variance. Generally, it is assumed  $\mu = 0$

### 3.2.7 Model Order

The model order corresponds to the number of model parameters. A higher order model has more parameters than a lower order model.

## 3.3 Homoskedastic Models

An important class of models assume constant variance and are referred to as “homoskedastic” models. That is, the variance of the noise component does not depend on time. The moving average (MA), autoregressive (AR), autoregressive moving average (ARMA) and integrated ARMA (ARIMA) are all families of this class. These will now be briefly introduced.

### 3.3.1 Moving Average (MA)

The moving average model,  $y_t$  of order  $q$  (abbreviated to MA( $q$ )) is given by

$$y_t = \beta_0 e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

where  $\beta_i$  are constants and  $e_t \sim N(0, \sigma^2)$  is the error term (§3.2.6).

According to Chatfield[10], MA models are appropriate for modelling econometric time series data, which often display irregularities (§2.5.3).

Figure 3.1 is a time series plot of an MA(1) model  $\beta_1 = 0.2$ .

### 3.3.2 Autoregressive (AR)

An autoregressive model  $y_t$  of order  $p$  (abbreviated to AR( $p$ )) is similar to an MA model, however,  $y_t$  is regressed onto past  $y_t$  values.

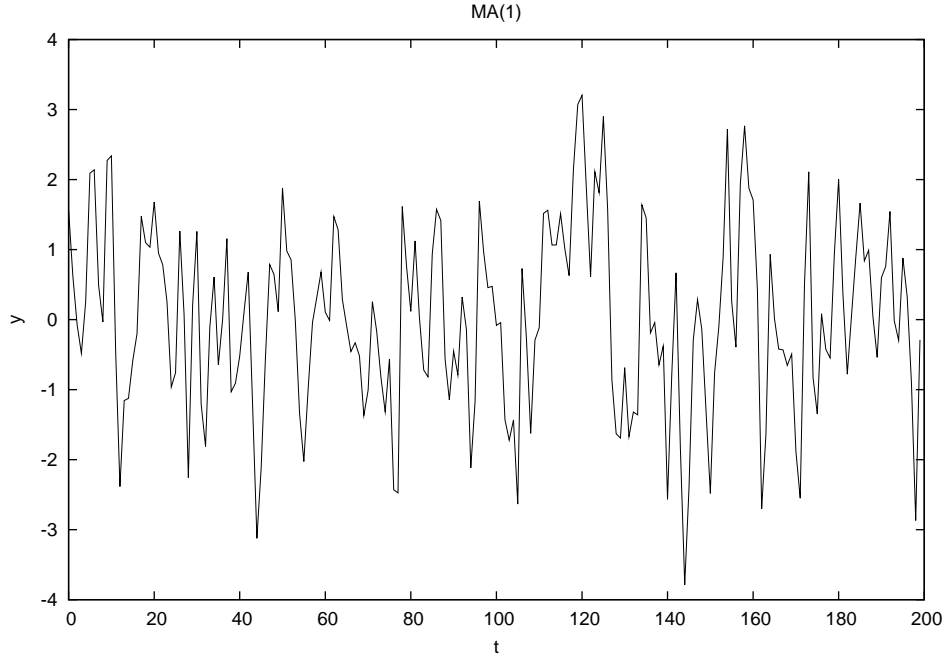


Figure 3.1: Time series plot of an MA(1) model with  $\beta_1 = 0.2$ .

A *first order* autoregressive model (AR(1)) is defined as

$$y_t = \alpha_1 y_{t-1} + e_t$$

where  $\alpha_i$  are constants and  $e_t \sim N(0, \sigma^2)$  and is uncorrelated with  $y_j \forall j \leq t$ .

The first order model can be generalised to a  $p$ th order model (AR( $p$ )) by

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + e_t$$

where  $\alpha_i$  and  $e_t$  as above.

Box and Jenkins[7] advocate that the AR process is an “extremely useful” [7, page 9] model for representing practical time series.

Figure 3.2 is a time series plot of an AR(1) model  $\alpha_1 = 0.9$ .

### 3.3.3 Autoregressive Moving Average (ARMA)

An ARMA model,  $y_t$  of order  $(p, q)$  (abbreviated to ARMA( $p, q$ )) is given by a combination of AR( $p$ ) and MA( $q$ ) models and is defined as

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

where  $\alpha_i$ ,  $\beta_i$  and  $e_t$  as above.

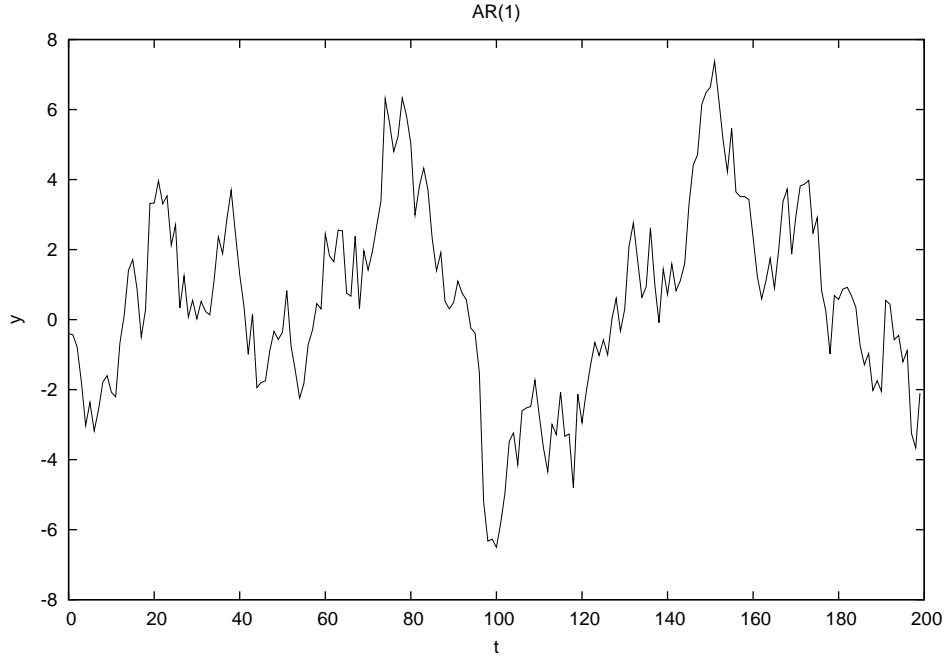


Figure 3.2: Time series plot of an AR(1) model with  $\alpha_1 = 0.9$ .

The ARMA models have experienced success due to their tractability and ease of implementation and parameter estimation (using linear least squares)[33]. Moreover, the ARMA model has the ability to adequately model stationary time series more parsimoniously than an AR or MA model by itself[10].

### 3.3.4 Autoregressive Integrated Moving Average (ARIMA)

Many time series, particularly financial, exhibit homogeneous non-stationary behaviour and demonstrate fluctuations at different levels at different times. Box and Jenkins[7] introduced the Autoregressive Integrated Moving Average class which, by comparison to the ARMA models, does not assume stationarity and is useful for modelling such time series.

An ARIMA process of order  $(p,d,q)$  is given by

$$\nabla^d X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$$

where  $\alpha_i$ ,  $\beta_i$  and  $Z_t$  as above and  $\Delta^d X_t$  is the  $d$ th *backward difference*[7] of the process defined by

$$\nabla^d X_t = X_t - X_{t-d}$$

## 3.4 Heteroskedastic Models

The previous models have all assumed a constant variance over time. However, many economic time series display periods of high volatility followed by periods of relative stability (§2.5.4). The conditional variance in such series (i.e. variance at  $x_t$  given the variance at  $x_{t-1}$ ) is considered non-constant and will vary over time. The ability to model changing conditional variance series is capable through the so called *heteroskedastic* time series models. The fundamental heteroskedastic models are the Autoregressive Conditional Heteroskedastic (ARCH) and Generalised ARCH (GARCH).

### 3.4.1 Autoregressive Conditional Heteroskedastic (ARCH)

The Autoregressive Conditional Heteroskedastic (ARCH) model family was introduced in 1982 - as the first heteroskedastic model - by 2003 Economics Nobel Laureate Robert Engle[21]. At the time, the “traditional”[21] models only considered changes in the mean of  $y_t$  and did not consider changes in the variance.

According to Engle[21], McNee suggests that “large and small errors tend to cluster together (in contiguous time periods)” for financial and econometric time series and that different forecast periods “vary widely over time due to inherent uncertainty and randomness”. Consequently, the more traditional models were found to be implausible for such time series. Engle recognised the usefulness of the ARCH model for these time series, where the variance may change over time and is predicted by previous forecast errors.

An ARCH process of order  $p$  (abbreviated to ARCH( $p$ )) facilitates modelling a changing variance by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i}^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2 \quad (3.1)$$

where  $\alpha_i \geq 0 \forall i$ . Notice that  $\sigma^2$ , is a function of time and past realised values of  $y$ . The ARCH process, with a noise process,  $\epsilon_t$ , may then form a constant mean ARCH model with changing variance by

$$y_t = \sigma_t \epsilon_t$$

where  $\sigma_t$  is the the local conditional variance at time  $t$ , given by the squared-root of the ARCH process (equation 3.1).

If  $\epsilon_t$  has zero mean and unit variance, then  $y_t$  may be referred to as a *zero mean ARCH model*:

$$y_t \sim N(0, \sigma^2)$$

Figure 3.3 is a time series plot of zero mean ARCH(1) process with  $\alpha_0 = 0.2$  and  $\alpha_1 = 0.9$ . Notice the jump in variance during periods 70 to 100 from relatively low to extremely high in the subsequent time periods. This indicates the primary purpose of the ARCH process: modelling a changing variance.

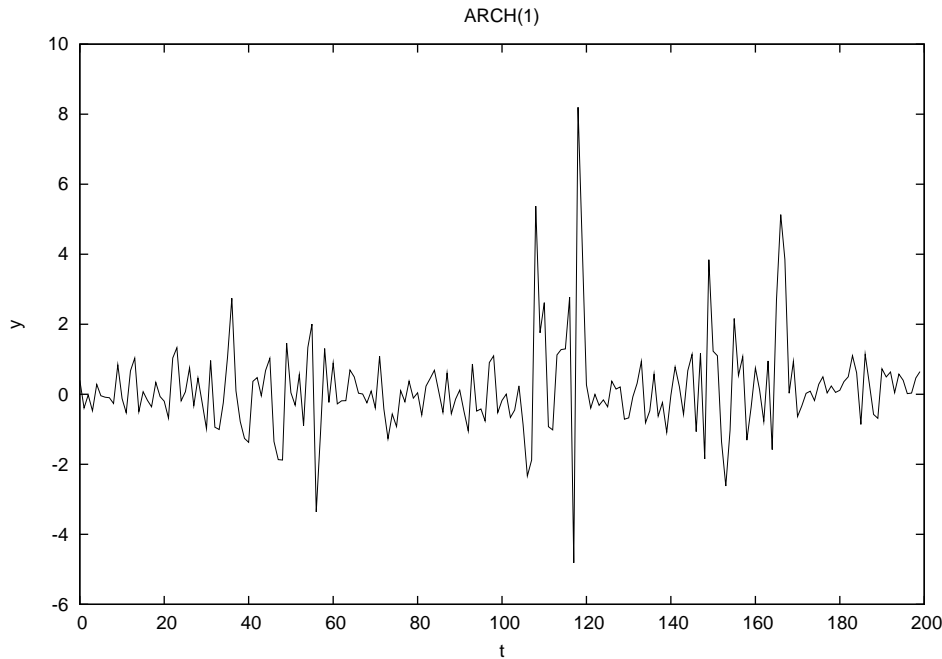


Figure 3.3: Time series plot of a zero mean ARCH(1) process with  $\alpha = (0.2, 0.9)$ .

A zero mean ARCH model is stationary under the following two assumptions[34]:

$$\alpha_i \geq 0 \quad \forall i$$

$$\sum_{i=1}^p \alpha_i < 1$$

An ARCH process may be used as the variance, or errors, to other models, such as the AR and MA process, described in Sections 3.3.2 and 3.3.1, respectively. By doing so, although the combined model's complexity is increased, the ability to model more sophisticated phenomena is improved.

Figure 3.4 is an example of an AR(1) process with ARCH(1) errors, occasionally denoted by AR(1)-ARCH(1). The dashed line represents the ARCH process while the solid line are the realised values of  $y$ . The corresponding parameter values are  $a_1 = 0.9$  for the AR process and  $\alpha = (0.2, 0.9)$  for the ARCH process.

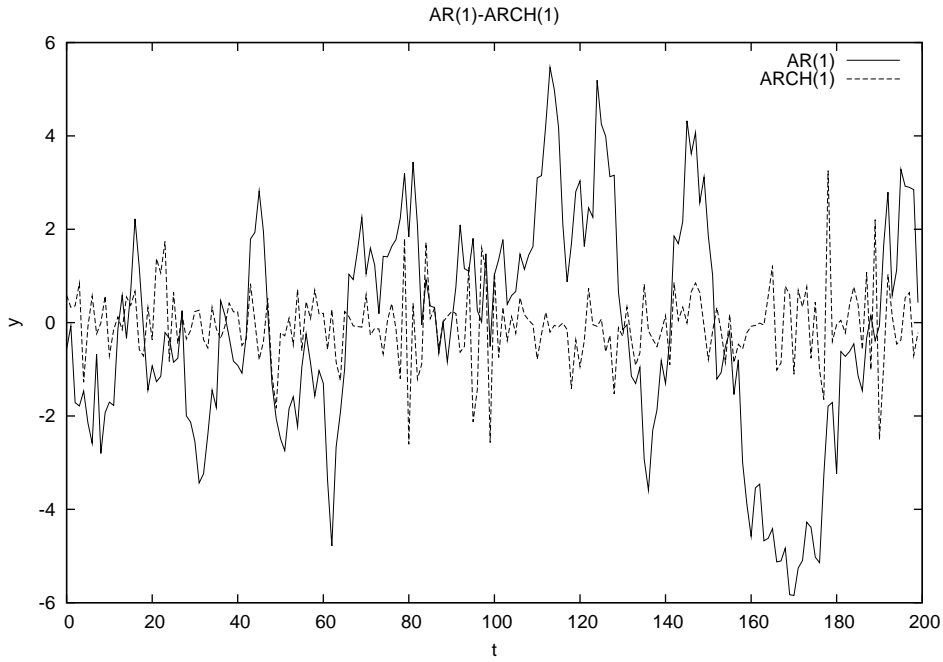


Figure 3.4: Time series plot of an AR(1) process with ARCH(1) errors.

### 3.4.2 Generalised ARCH (GARCH)

Bollerslev[5] introduced the Generalised ARCH process in 1986 as an extension to Engle's ARCH model, such that the process depends on past  $\sigma_t^2$  values additional to past  $y_t$  values. Consequently, GARCH allows for both a longer memory and more flexible lag structure[5, page 308].

A GARCH process with  $p$  regressive  $y_t$  terms and  $q$  autoregressive  $\sigma_t^2$  terms is abbreviated to GARCH( $p, q$ ) and given by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

where  $\alpha_i, \beta_j \geq 0 \forall i, j$ .

The GARCH(0,0) equivalent to white noise and the GARCH( $p, 0$ ) model is equivalent to an ARCH( $p$ ) model. Analogous to the ARMA class, the GARCH model has the ability to adequately model a time series more parsimoniously than an ARCH model[5].

There are an astounding variety of ARCH and GARCH models - many with convenient abbreviations such as threshold-ARCH (TARCH), structural-ARCH (STARARCH), exponential-GARCH (EGARCH) and integrated-GARCH (IGARCH). The reader is referred to Bollerslev et al.[6] (and references within) for further information.

## 3.5 Summary

Time series modelling allows greater understanding and better predictions for the modelled phenomena. The important concepts in time series modelling, in particular stationarity and volatility, have been introduced and discussed. Both, homoskedastic and heteroskedastic, models and their assumptions for the various types of series have been defined. While our research concentrated only on ARCH models, the AR and MA models were included for background knowledge and to assist with understanding of the ARCH concepts. The GARCH model is a natural extension of the ARCH model and was included here for completeness.

Once we have decided on the particular model family to use, we need to choose which model ‘best’ fits the observed (and potential future) data. This choice occurs during model selection and is considered a critical step in the model building process. The next section provides a brief overview to model selection and formally defines existing popular criteria, such as Akaike’s Information Criterion, before Minimum Message Length model selection, specifically to time series, is presented.

# Chapter 4

## Model Selection

### 4.1 Overview

The possible model orders and parameter estimates for an observed time series are formulated in the *model estimation* stage during the *model building process*. However, it can be difficult to know which particular model class and order is appropriate.

Possible candidates are identified during the *model formulation* or *model specification* stage. Chatfield[10] believes model specification is more important than model estimation and provides general rules of thumb for identifying models which may suit the data[10, pages 70-71]. Occasionally, more than one model, possibly from varying classes, are identified as suitable.

In such situations, choosing the “best” model between competing models is known as *model selection*[53]. Simple residual analysis statistics, such as  $R^2$ , can be used to compare models and goodness of fit and will generally choose high order models. The fit measurement for noise variances errors will generally decrease for higher order models, however, will often lead to over-fitting.

Over-fitting is not desirable, particularly when forecasting, as the fit measurement when forecasting will depend on the parameter estimate errors and will increase for higher order models[8, 10]. Linhart and Zucchini agree and believe over-fitting is a disadvantage due to greater model instability:

“Repeated samples collected under comparable conditions lead to widely different models, each of which conveys more about the particularities of its corresponding data set than about the underlying phenomena.”[38, page 2]

Hence, the general idea for applying the *Principle of Parsimony* to model selection. That is, a model is penalised proportional to the number of parameters it uses, consequently, a model with less parameters, which fits the data adequately, is

preferred over a higher order model. For example, if both AR(1) and AR(2) models fit the data adequately, then the more *parsimonious* model, AR(1), will be chosen.

## 4.2 Existing Criteria

Commonly used model selection criteria, such as Akaike's Information Criterion-akaike74, corrected AIC[36], Schwarz's Bayesian Information Criterion[45] and Hannan-Quinn[35], employ parameter penalties and will try to select the most parsimonious model. These criteria were mostly developed for ARMA time series and their behaviour for such time series are well understood. Extensive research has compared criterion performances (see, for example, [40, 53, 8, 10, 20]).

These criteria will now be introduced.

### 4.2.1 Akaike's Information Criterion (AIC)

Originally proposed by Akaike in 1974[2], Akaike's Information Criteria, abbreviated to AIC, is the more popular and widely used criterion[20, 40]. It is given by

$$AIC = -2 \log(\sigma^2) + 2r$$

where  $\sigma^2$  is the maximum likelihood estimate of the variance and  $r$  denotes the number of independent parameters, including the variance  $\sigma^2$ . For example,  $r = p + q + 1$  for an  $ARMA(p, q)$  model.

Although AIC performs adequately, and may outperform BIC (§4.2.3) for small sample size to parameter ratios[20], it is known to bias over-parameterised models[8]. AIC is also inconsistent, that is, it does not lead to the true model with probability one for large sample sizes[36].

Akaike also developed the final error prediction criterion[20, page 106-7], however, this requires a constant variance and is subsequently not suitable for ARCH model selection[40].

### 4.2.2 Corrected Akaike's Information Criterion ( $AIC_c$ )

Hurvich and Tsai[36] derived a bias correcting version for AIC in 1989 called corrected AIC (abbreviated in the literature as either AICC, AICc or  $AIC_c$ ). It is given by

$$AIC_c = -2 \log(\sigma^2) + 2 \frac{rn}{(n - r - 1)}$$

or equivalently

$$AIC_c = AIC + \frac{2r(r+1)}{(T-r-1)}$$

where  $T$  is the sample size and  $r, \sigma^2$  are as above.

$AIC_c$  will perform similarly to  $AIC$  and will tend to select the same model if the sample size to parameter ratio,  $T/r$ , is sufficiently large[9]. Hence, Burnham and Anderson[9] advocate the use of  $AIC_c$  when  $T/r \leq \approx 40$ [9, page 66] since in other situations, when  $T/r \geq \approx 40$ , the bias-correcting term is negligible.

### 4.2.3 Bayesian Information Criterion (BIC)

Schwarz's derived his criterion in a Bayesian context and subsequently became known as Bayesian Information Criterion (BIC). It is defined as

$$BIC = -2 \log(\sigma^2) + 2r \log T$$

where  $T, r$  and  $\sigma^2$  are as above.

BIC's performance is superior for large sample sizes and will always select a more parsimonious model than  $AIC$ [20].

### 4.2.4 Hannan-Quinn Criterion (HQ)

Hannan-Quinn introduced another consistent criteria, known as HQ, in 1979 and is given by

$$HQ = -\log(\sigma^2) + 2r \log(\log(T))$$

where  $T, r$  and  $\sigma^2$  are as above.

HQ performed poorly for Fitzgibbon et al.[28] for all sample sizes and Sak et al.[44] found HQ's performance to vary across their assessment criteria. However, Mitchell and McKenzie[40] suggest HQ performs similarly to  $AIC$  for most "practical data sets" but performed "remarkably well in selecting an optimal ARCH model" during their statistical analysis.

### 4.2.5 Minimum Description Length (MDL)

Rissanen developed a version of his minimum description length criterion in 1978, sometimes referred to as MDL78. It is given by

$$MDL78 = \log(\sigma^2) + \frac{r \log(T)}{T}$$

where  $T, r$  and  $\sigma^2$  are as above.

Hastie et al. showed, noted by Mitchell and McKenzie[40], that MDL78, is in fact equivalent to BIC. However, Mitchell and McKenzie suggest Rissanen's 1987 version[42] performs favourably for ARCH model selection against popular model selection criteria, such as AIC,  $AIC_c$  and BIC.

The behaviour of the aforementioned criteria for ARMA model selection is well understood. However, as noted by Enders[20, page 135], Bollerslev et al.[6] argue that "their statistical properties in the ARCH context are largely unknown.". Additionally, according to Mitchell and McKenzie[40], popular model selection criteria, such as AIC,  $AIC_c$  and BIC, are potentially inappropriate when used for ARCH model selection, of which information theoretic procedures (such as MDL87 and MML) provide a better theoretical basis[40].

### 4.3 Minimum Message Length

Minimum Message Length principle, introduced formally in Section 5, is a method for statistical and inductive inference and has been successfully used as model selection to a variety of problems. Wallace and Boulton introduced strict minimum message length (strict MML or SMML) in 1975[50, 52], however, SMML proved intractable for all but the simplest problems[50, 52, 22, 23, 26, 24].

As such, a number of approximations to SMML have been developed. The seemingly most popular is Wallace and Freeman's approximation introduced in 1987[52] (abbreviated to MML87). MML87 has been successfully applied as time series model selection criteria for MA models by Sak et al.[44] and AR models by Fitzgibbon et al.[28] and Schmidt (private correspondence) and also utilised by Collie et al.[11] for their AR agent-based stock market simulation.

MML87 outperformed AIC,  $AIC_c$ , BIC and HQ for the best average mean squared prediction error[28], best average log likelihood[44] and chose the best model order[28]. Schmidt (private correspondence) also found MML performs well competing against other criteria for AR and non-linear time series. Collie et al.[11] found that the agents using MML model selection (in conjunction with MML parameter inference), particularly the MML agent which assumed non-stationarity, significantly outperformed agents using different techniques.

Specifically to ARCH models, although only preliminary results, Sak et al.[44] found MML87 model selection to perform well for small sample sizes, coming second behind AIC.

Dowe's MML (abbreviated to MMLD) is another approximation to SMML originally suggested by D. L. Dowe in 1999. MMLD is a more recent approximation than

the popular MML87 consequently, less work has been published. Thus far; Fitzgibbon et al.[26] empirically compared MMLD's performance in model selection and parameter estimation to change-point<sup>1</sup> problems against SMML and another MML estimator; Fitzgibbon et al.[27] investigated MMLD inference to univariate polynomial approximation to noisy data points; and Agusta and Dowe[1] used MMLD as approximations for multivariate Gaussian mixture modelling.

MML87 and MMLD were the approximations of interest throughout our research and are discussed in more detail in the following sections.

## 4.4 Summary

Model selection is an important stage in time series analysis. We have briefly discussed the motivation for model selection and introduced its primary concerns; over-fitting and the Principle of Parsimony. Some existing criteria were formally defined whose statistical properties specific to ARCH models, according to Bollerslev et al[6], are largely unknown.

The Minimum Message Length principle and two approximations, MML87 and MMLD were introduced. MML87, as time series model selection, has been found (for AR and MA models) to perform more than competitively against the previous mentioned criteria.

We introduce MML in detail in the following chapter. Two approximations, MML87 and MMLD, are also described before their formulations as ARCH model selection criteria are presented in Chapters 6 and 7, respectively.

---

<sup>1</sup>Change-points occur when data needs to be partitioned into contiguous groups which may be modelled distinctly[26, page 1]. The parameters relate to where data is thought to have been generated from a different model, i.e, the change points.

# Chapter 5

## Minimum Message Length

### 5.1 Introduction

Originally developed in 1968 by Wallace and Boulton[49], Minimum Message Length (MML) is an information-theoretic framework for statistical and inductive inference from data. It has wide-spread applications, particularly to machine learning, parameter estimation, model selection and “data mining”.

MML seeks to explain some observed data by a hypothesis, or set of best hypotheses, by encoding and transmitting a two-part message between a theoretical sender and receiver. The first part of the message contains the optimally encoded hypothesis whilst the second part contains the data encoded by this hypothesis. Upon receiving the message, the receiver retrieves the originally observed data by decoding the second part (i.e, the transmitted data) using the hypothesis from the first part.

The sender may choose any hypothesis from the set of all possible hypotheses, or so called “codebook”[24], to encode the data. Clearly, if the hypothesis chosen contains no information, then its encoded length will be zero but the second part of the message will need to encode the entire observed data. Conversely, a complex hypothesis may be chosen such that very little, or even no, data is encoded in the second part. However, the increased complexity of the hypothesis will lead to a longer encoding of the first part.

This allows a natural trade-off between model complexity and goodness-of-fit. If an overly complex hypothesis, or model, is chosen, then it is likely to fit the observed data well but will not generalise to other, similar data. This is known as *over-fitting* (§4.1). Alternatively, if a less complex, more *parsimonious* model is chosen, then it may not over-fit but will, consequently, require more data in the second part of the message. The sender must then aim to choose the optimal model - the model which minimises the length of the encoded message - hence Minimum Message Length.

In this sense, MML can be thought of as a formal expression of *Occam's Razor* theory. That is, a message should be represented by as small a hypothesis as necessary for the entire message to be understood, as C. S. Wallace eloquently put it:

“The best explanation of the facts is the shortest”[48, page 1]

The MML principle and its foundations have been briefly introduced. Although, once we observe some data and decide on the hypotheses space,  $\Theta$ , we still need to encode the message to obtain its length.

## 5.2 Encoding the Message

Shannon[46] showed that an event with probability  $p$  can be optimally encoded into a string with length

$$-\log_b p$$

where  $b$  is the number of distinct symbols in the code alphabet,  $A$ . For instance,  $b = 2$  for the binary code,  $A = \{0, 1\}$ , and the code length is measured in ‘bits’. (Often for mathematical convenience, it is assumed data is encoded using  $b = e = 2.718\dots$  (the natural log) such that the code length is then measured in ‘nits’[48, page 78].)

We may then encode a series of  $n$  events, each with probability  $\{p_i; i = 1, \dots, n\}$ , by concatenating their encoded strings to form a complete word with minimum expected length

$$-\sum_{i=1}^n p_i \log_b p_i$$

If we observe some data<sup>1</sup>  $x$ , and consider it a string of symbols from a finite alphabet, say, for example  $A = \{0, \dots, 9, +, -, *, \dots\}$ , then we can encode this data using Shannon’s theory. Moreover, if there are deterministic or stochastic patterns in  $x$ , that is, the data is not purely random, then we may encode  $x$  using a hypothesis,  $\theta$ , chosen from the set of all possible hypotheses,  $\Theta$ .

Once the sender has chosen a particular hypothesis  $\hat{\theta}$ ,  $x$  may be encoded over the set of all possible data values,  $X$ , according to how probable it is given  $\hat{\theta}$ , denoted  $f(x|\hat{\theta})$  and referred to as the likelihood function.

The sender and receiver agree on  $\Theta$ ,  $X$ ,  $f(x|\theta)$  and a Bayesian prior distribution,  $h(\theta)$ , before transmission[48, page 153].

---

<sup>1</sup>We refer to data as  $x$  throughout this chapter instead of  $y$ , like the previous chapters, to maintain consistency with the literature.

Thus, the encoded message consists of two parts; the prior probability distribution of the hypothesis,  $h(\theta)$  and the likelihood of the data given the hypothesis,  $f(x|\theta)$ , and has length[48, page 116]<sup>2</sup>

$$\begin{aligned} \text{MsgLen}(x, \theta) &= -\log h(\theta) - \log f(x|\theta) \\ &= -\log (h(\theta)f(x|\theta)) \end{aligned}$$

Therefore, minimising the message length is not dissimilar to maximising  $h(\theta)f(x|\theta)$ , the posterior distribution of  $x$ . Wallace and Freeman[52, Page 245], noted by Fitzgibbon[24, Page 15], indicate the difference:

“MML looks for broader peaks of the posterior distribution and essentially chooses the local posterior mode with the greatest probability content rather than simply the highest one.”

Indeed, if a message has length less than the length of  $x$  encoded using no hypothesis, then we have compressed the data and consequently reduced the length of the transmission.

Difficulties arise when constructing such codes. For instance, we have assumed that both the sender and receiver know and can choose the ideal hypothesis which minimises the message length. However, the number of possible hypotheses is infinite for many problems [24, page 14] and, consequently, each hypothesis cannot be considered. Therefore, we select the region in  $\Theta$  that has the minimum expected message length on average[48].

### 5.3 Strict Minimum Message Length (SMML)

Wallace and Boulton introduced strict minimum message length (strict MML or SMML) in 1975[50, 52]. SMML partitions the data space,  $X$  into separate data groups, one for each hypothesis in the selected subset,  $\Theta^*$ , of  $\Theta$  and requires calculation of the marginal distribution  $r(x)$  over  $X$  (See [48, Page 222] why).

Consequently, SMML has been found, in practise, intractable, difficult to construct and computationally infeasible for all but the simplest problems[48, page 143].

---

<sup>2</sup>This assumes both the sender and receiver have agreed on the hypothesis space,  $\Theta$ , before transmission.

As such, several approximations have been developed. Our work involved Wallace and Freeman's MML87[52] and Dowe's MMLD[48, 37, 28, ch 4.10] approximations. Both of these are statistically invariant and have been shown to be statistically consistent on all problems considered[48, pages 218-219].

MML87 and MMLD shall now be introduced before formulation of their ARCH model selection equations are presented in Chapters 6 and 7, respectively.

## 5.4 MML87

Wallace and Freeman's popular 1987 approximation to MML, referred to as MML87[52], approximates the log likelihood quadratically around  $\theta$ . Whereas SMML partitioned the data space  $X$ , MML87 partitions the parameter space  $\Theta$ .

The approximation for  $n$  parameters  $\theta = (\theta_1, \dots, \theta_n)$  is given by[48, page 235]

$$\text{MML87 } \text{MsgLen}(\theta, x) = \left[ -\log \frac{h(\theta)}{\sqrt{F(\theta)\kappa_n}} \right] - \log f(x|\theta) + \frac{n}{2} \quad (5.1)$$

where  $x$  is the observed data,  $h(\theta)$  is the Bayesian prior over the  $n$  parameter values,  $\kappa_n$  is the lattice constant relating to the optimum method for tiling an  $n$ -dimensional space[48, page 178],  $f(x|\theta)$  is the likelihood function, and  $F(\theta)$  is the determinant of the expected Fisher Information matrix, given by the expected values for the second partial derivatives of the negative log likelihood function:

$$F(\theta) = \det \left[ -E \left( \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right) \right] \quad (5.2)$$

The first term of equation 5.1 gives the encoded hypothesis length while the second and third terms give the length of the encoded data given the hypothesis.

MML87 has been applied successfully to large variety problems, such as parameter estimation (see, for example,[52, 18, 19, 51]), univariate polynomial inference[11], inference of segmented points of ordered data (time series)[25], and, as indicated in Section 4.3, to model selection (see, for example,[52, 28, 44]), among others.

The formulation of MML87 as model selection specifically to ARCH time series is given in Chapter 6.

## 5.5 MMLD (I1D)

Similarly to MML87, Dowe’s Minimum Message Length (referred to as either MMLD or I1D) attempts to choose a region of the parameter space containing all models which would result from an SMML code[48, page 211]. Suggested by D. L. Dowe, and developed further by Lam[37], Fitzgibbon[24] and Fitzgibbon, Dowe and Allison[26, 27], MMLD was partly motivated as a better approximation to the uncertainty region than MML87’s quadratic approximation[24, page 69]. Like MML87, MMLD is also invariant under re-parameterisation[24, 48].

Given some observed data  $x$ , a Bayesian prior over the  $n$  parameter values  $h(\theta)$ , the likelihood function  $f(x|\theta)$  and the uncertainty region  $R$ , MMLD approximates the message length by

$$\text{MMLD } MsgLen(R, x) = -\log \left( \int_R h(\theta) d\theta \right) - \frac{\int_R h(\theta) \log f(x|\theta) d\theta}{\int_R h(\theta) d\theta} \quad (5.3)$$

The first term in Equation (5.3) gives the volume of the prior over  $R$ , approximating the first part of the message, while the second term gives the average prior-weighted negative log likelihood over  $R$ , approximating the second part of the message.

The uncertainty region,  $R$ , and, consequently, the minimum length of Equation 5.3, is determined according to the Boundary Rule[48, §4.10.2], such that

$$\theta \in R \text{ iff } -\log f(x|\theta) \leq -\frac{\int_R h(\theta) \log f(x|\theta) d\theta}{\int_R h(\theta) d\theta} + 1 \text{ nit} \quad (5.4)$$

In other words,  $R$  contains all  $\theta$  that have negative log likelihoods less than the average prior-weighted negative log likelihood of  $R$  plus one nit.

Equation 5.3 provides no method for choosing which parameter to use, that is, which point estimate  $\hat{\theta} \in R$  to encode the data. Wallace presents a “crude”[48, page 210] way to choose  $\hat{\theta}$ , which he termed “Random Coding of Estimates”, by generating a random sequence of values<sup>3</sup>

$$\{\theta_i : i = 1, 2, \dots\}$$

from the prior density  $h(\theta)$  and choosing the first value, say  $\theta_n$ , such that  $\theta_n \in R$ .

Dowe advocates (personal correspondance) using the minimum posterior-weighted Expected Kullback-Leibler (EKL) estimate over  $R$  as the point estimate, given by

---

<sup>3</sup>The sender and receiver agree on the random number generator and seed before transmission to ensure efficient coding.

Fitzgibbon[24, Equation 5.12]:

$$\hat{\theta} = \min \left( \frac{\int_{\mathcal{R}} h(\theta) KL(\theta, \hat{\theta}) d\theta}{\int_{\mathcal{R}} h(\theta) d\theta} \right) \quad (5.5)$$

where  $KL(\theta, \hat{\theta})$  is the Kullback-Leibler distance, a measurement of the difference between the true model,  $\theta$ , and the inferred model,  $\hat{\theta}$ , given by[48, §4.6]:

$$KL(a, b) = \sum_{x \in X} f(x|\theta) \log \frac{f(x|\theta)}{f(x|\hat{\theta})}$$

According to Fitzgibbon[24, page 72], Wallace also suggests using EKL, weighting, however, by the prior instead of the posterior.

The formulation of MMLD as model selection specifically to ARCH time series is given in Chapter 7.

## 5.6 Summary

The Minimum Message Length principle provides a formal specification for Occam's Razor, incorporating elements of information theory and Bayesian analysis with statistics. This chapter described MML and two SMML approximations, MML87 and MMLD, in some detail. MML87 has been widely and successfully used to a number of problems including time series model selection. MMLD is a newer approximation than MML87 and has experienced some success[27, 26, 24]. These approximations will now be formulated as ARCH model selection criteria in the following chapters.

# Chapter 6

## MML87 Model Selection Formulation

Recall from Section 3.4.1 that a  $p$ th order constant mean ARCH process is given by

$$y_t = v_t \sigma_t \tag{6.1}$$

where

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i}^2 \tag{6.2}$$

Also, recall from Equation 5.1, that the MML87 message length for some observed data<sup>1</sup>  $y$  is given by

$$\text{MML87 } \textit{MsgLen}(\theta, x) = \left[ -\log \frac{h(\theta)}{\sqrt{F(\theta)\kappa_n}} \right] - \log f(y|\theta) + \frac{n}{2} \tag{6.3}$$

where  $h(\theta)$  is the Bayesian prior over the  $n$  parameter values,  $\kappa_n$  is the lattice constant,  $f(y|\theta)$  is the likelihood function, and  $F(\theta)$  is the determinant of the expected Fisher Information given by Equation 5.2.

In this chapter, we discuss the necessary formulations to calculate the MML87 message length for zero mean ARCH models. These are, in section order, calculating the likelihood and estimating the parameters, deciding on an appropriate Bayesian prior to represent our beliefs, choosing the lattice constant and calculating the (observed) Fisher Information matrix.

---

<sup>1</sup>We have denoted the data as  $y$  in this, and subsequent, chapters, opposed to  $x$  previously, to remain consistent with the usual practise for time series

## 6.1 Likelihood

Assuming  $v_t \sim N(0, 1)$  and is independent of  $y_t$ , then Equation 6.1 is normally distributed conditional on the information set available at  $t$ ,  $\Psi_{t-1} = (y_0, y_1, \dots, y_{t-1})$  and the parameters  $\theta = (\alpha_0, \alpha_1, \dots, \alpha_p)$ :

$$y_t | \Psi_{t-1} \sim N(0, \sigma_t^2)$$

The conditional distribution of  $y_t$  is then:

$$f(y_t | \Psi_{t-1}; \theta) = f(y_t | \theta, y_1, y_2, \dots, y_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(\frac{-y_t^2}{2\sigma_t^2}\right) \quad (6.4)$$

and the conditional log likelihood at observation  $t$  is

$$\log f_t(y | \Psi_{t-1}; \theta) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_t^2 - \frac{1}{2} \frac{y_t^2}{\sigma_t^2} \quad (6.5)$$

From Equation 6.5, the full conditional log likelihood for all  $T$  observations has the form

$$\begin{aligned} \log f(y | \Psi_{t-1}; \theta) &= \sum_{t=1}^T \log f(y_t | \Psi_{t-1}; \theta) \\ &= \sum_{t=1}^T \left( -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_t^2 - \frac{1}{2} \frac{y_t^2}{\sigma_t^2} \right) \\ &= -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^T \log(\sigma_t^2) - \frac{1}{2} \sum_{t=1}^T \left( \frac{y_t^2}{\sigma_t^2} \right) \end{aligned} \quad (6.6)$$

Using the log likelihood (Equation 6.6), we are able to estimate that parameters,  $\hat{\theta}$ , which maximise the probability of  $y$  occurring given  $\hat{\theta}$ . We do this by maximising Equation 6.6, or, equivalently, minimising the negative log likelihood. This method is referred to as, unsurprisingly, Maximum Likelihood[34].

In his 1982 paper introducing ARCH, Engle[21] advocates using the, so called, ‘Score’ algorithm to maximise the likelihood. To do so, the first and second partial derivatives of the (log) likelihood are required. These are given below; their derivations, along with a description of the Score algorithm, are presented in Appendices A and B, respectively.

The first derivatives of Equation 6.6 are

$$\begin{aligned}\frac{\partial L}{\partial \alpha_0} &= \sum_{t=1}^T \frac{1}{2\sigma_t^2} \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) \\ \frac{\partial L}{\partial \alpha_i} &= \sum_{t=1}^T \frac{1}{2\sigma_t^2} y_{t-i}^2 \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i > 0\end{aligned}$$

and the second derivatives are

$$\begin{aligned}\frac{\partial L}{\partial \alpha_0^2} &= -\frac{1}{2} \sum_{t=1}^T \frac{1}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right) \\ \frac{\partial L}{\partial \alpha_i^2} &= -\frac{1}{2} \sum_{t=1}^T \frac{y_{t-i}^4}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i > 0 \\ \frac{\partial L}{\partial \alpha_0 \partial \alpha_i} &= \frac{\partial L}{\partial \alpha_i \partial \alpha_0} = -\frac{1}{2} \sum_{t=1}^T \frac{y_{t-i}^2}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i > 0 \\ \frac{\partial L}{\partial \alpha_i \partial \alpha_j} &= \frac{\partial L}{\partial \alpha_j \partial \alpha_i} = -\frac{1}{2} \sum_{t=1}^T \frac{y_{t-i}^2 y_{t-j}^2}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i, j > 0 \wedge i \neq j\end{aligned}$$

## 6.2 Bayesian Priors

The Bayesian priors reflect any prior belief of the parameters before any data is observed. We place an uninformative prior on number of parameters:

$$h(p) \propto 1 \tag{6.7}$$

and, expecting  $y_t$  to be stationary, we follow Sak[43] and place a uniform prior on the  $p$  parameter values:

$$h(\theta, \sigma^2) \propto \frac{1}{H_p} \times \frac{1}{\sigma^2} = \frac{1}{H_p \sigma^2} \tag{6.8}$$

where  $H_p$  is calculated by  $\left(\frac{0.99}{p}\right)^p$ , representing the product of  $p$  numbers from  $(0, 1]$  whose sum is 0.99[43, Page 25].

Geweke[30] also presents a prior on the parameter values for a stationary ARCH process:

$$h(\theta) \propto \left( \frac{1 - \sum_{j=1}^p \alpha_j}{\alpha_0} \right)^{\frac{1}{2}} \tag{6.9}$$

Both priors (Equations 6.8 and 6.9) are used during our simulations (§8.3).

### 6.3 Lattice Constants

The lattice constants ( $\kappa_n$  in Equation 6.3) relate to the expected error in the log likelihood from tiling (or partitioning) an  $n$ -dimensional space[48, 28]. The exact values of  $\kappa_n$  are known for small  $n$  only. However, it is known that  $\kappa_n \rightarrow 1/(2\pi \exp)$  as  $n \rightarrow \infty$ [12].

Sufficient for our purposes, replacing  $n$  for the number of parameters  $p$ , we use the following values[28]:

p	$\kappa_p$
1	1/12
2	5/(36 $\sqrt{3}$ )
3	19/(192 * 2 <sup>1/3</sup> )
$\geq 4$	1/(2 $\pi \exp(1)$ )

### 6.4 Fisher Information

The Fisher Information matrix is the negative expectation of the second derivatives of the log likelihood (§5.4). It is used in the MML87 approximation since the parameter estimates must be transmitted before the data has been received (i.e, the encoding of the first term cannot depend on data)[48, Page 240].

The exact expected values of the second derivatives for the ARCH( $p$ ) model are difficult to calculate. These require integration over  $p$  dimensions, and the complexity of time dependent variables also contribute to this difficulty.

Instead, we may use the observed values[48, §5.3] to approximate the Fisher Information, which is then referred to as the empirical (or observed) Fisher Information.

Using the estimated expectations of the second derivatives of the negative log likelihood (calculated in Appendix C), the empirical Fisher Information matrix for an ARCH( $p$ ) process has the form:

$$F(\theta, y) = \frac{1}{2T} \sum_{t=1}^T \begin{bmatrix} \frac{1}{\sigma_t^4} & \frac{y_{t-1}^2}{\sigma_t^4} & \dots & \frac{y_{t-p}^2}{\sigma_t^4} \\ \frac{y_{t-1}^2}{\sigma_t^4} & \frac{y_{t-1}^4}{\sigma_t^4} & \dots & \frac{y_{t-1}^2 y_{t-p}^2}{\sigma_t^4} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{y_{t-p}^2}{\sigma_t^4} & \frac{y_{t-p}^2 y_{t-1}^2}{\sigma_t^4} & \dots & \frac{y_{t-p}^4}{\sigma_t^4} \end{bmatrix} \quad (6.10)$$

The estimated expectations and empirical Fisher Information are calculated in full in Appendix C.

Substituting Equation 6.10 and the findings from the above sections into Equation 6.3 (and restructuring), our MML87 approximation for model selection of an ARCH(p) process is

$$\begin{aligned} \text{MML87 } \textit{MsgLen}(\theta, y) = & -\log h(\theta) + \frac{1}{2} \log (|F(\theta, y)|) - \\ & \log f(y|\theta) + \frac{p}{2}(1 + \log \kappa_p) - \log h(p) \end{aligned} \quad (6.11)$$

## 6.5 Small Sample Approximation

Wallace[48, §5.2.9] presents an alternative but “crude” [48, Page 236] approximation to message lengths (Equation 6.3) for little amounts of observed data. Using the same calculations from the previous sections, the MML87 small sample size approximation for ARCH model selection is:

$$\begin{aligned} \text{MML87 (sml) } \textit{MsgLen} \approx & \frac{1}{2} \log \left( 1 + \frac{F(\hat{\theta})\kappa_p}{h(\hat{\theta})^2} \right) - \\ & \log f(y|\hat{\theta}) + \frac{p}{2} - \log h(p) \end{aligned} \quad (6.12)$$

## 6.6 Summary

We have formulated the MML87 approximation for ARCH model selection. We discussed calculating the likelihood, estimating the parameters (via Maximum Likelihood), our choice of Bayesian priors, the implications of the lattice constant and finding the observed Fisher Information. In light of some tests (see Results, §8.3), we discovered that MML87 was underperforming to our expectations (see possibly why in discussion of results §8.4), hence, our investigation into the small sample size approximation (Equation 6.12) and Dowe’s MML approximation to SMML. The formulation of which, is discussed in the following chapter.

# Chapter 7

## MMLD (I1D) Model Selection Formulation

Recall from Section 3.4.1 that a  $p$ th order constant mean ARCH process is denoted by

$$y_t = v_t \sigma_t$$

where

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i}^2$$

Assuming  $v_t \sim N(0, 1)$ , then, from Section 6.1, the conditional likelihood and the condition log likelihood for the entire observed data,  $\{y_t : t = 1, \dots, T\}$ , are

$$f(y|\Psi_{T-1}; \theta) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(\frac{-y_t}{2\sigma_t^2}\right) \quad (7.1)$$

$$\log f(y|\Psi_{T-1}; \theta) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^T \log(\sigma_t^2) - \frac{1}{2} \sum_{t=1}^T \left(\frac{y_t^2}{\sigma_t^2}\right) \quad (7.2)$$

where  $\theta = (\alpha_0, \alpha_1, \dots, \alpha_p)$  and  $\Psi_t = (y_0, y_1, \dots, y_{t-1})$ .

Also, recall from Section 5.5, that, for some observed data  $y$ , the MMLD message length (Equation 5.3) has the form

$$\text{MMLD } MsgLen = -\log\left(\int_R h(\theta) d\theta\right) - \frac{\int_R h(\theta) \log f(y|\theta) d\theta}{\int_R h(\theta) d\theta} \quad (7.3)$$

where  $h(\theta)$  is a prior distribution over the  $p$  parameter values. For our MMLD implementation, we have used the prior given by Equation 6.8.

There are several challenges when calculating Equation 7.3, particularly in higher dimensions when the number of parameters is large. Firstly, we need to find the uncertainty region  $R$ , then evaluate two integrals (the integral in the first term can be reused as the denominator in the second term) and finally, since MMLD does not directly yield a point estimate,  $\hat{\theta} \in R$  (see §5.5), find  $\hat{\theta}$ .

Fortunately, Fitzgibbon[24] with Dowe and Allison[27] have developed an algorithm termed “Message From Monte Carlo” to approximate MMLD message lengths.

## 7.1 Message from Monte Carlo

Message from Monte Carlo (MMC) is based on posterior sampling and Monte Carlo integration (see [32] and various chapters within for a good overview of sampling and Monte Carlo methods) to find the uncertainty region,  $R$ , in Equation 7.3.

It does so by drawing a sample  $S = \{\theta_i : i = 1, \dots, N\}$  from the posterior distribution of the parameters,  $\theta$ :

$$\text{posterior}(\theta|y) = \frac{h(\theta)f(y|\theta)}{\int_{\theta \in \Theta} h(\theta)f(y|\theta)} \quad (7.4)$$

We then choose a subset,  $Q \subseteq S$ , where  $Q$  is considered a subset of the true uncertainty region,  $R \supseteq Q$ , and use the (finite) set  $Q$  and Monte Carlo integration to approximate the uncertainty region,  $R$  in Equation 7.3[24].

We could use the prior as the sampling distribution<sup>1</sup> to approximate  $R$ [24]. However, as Fitzgibbon notes[24, Page 84], the likelihood function becomes more concentrated as the number of observed data,  $T$ , increases. Thus, in general, sampling from the prior is not efficient, hence, MMC’s use of the posterior (7.4).

Therefore, the MMC message length approximation for MMLD message lengths (Equation 7.3), given by Fitzgibbon[24, Equation 6.9], is

$$\text{MMLD} \approx -\log \left( \frac{\sum_{\theta \in Q} f(y|\theta)^{-1}}{\sum_{\theta \in S} f(y|\theta)^{-1}} \right) + \frac{\sum_{\theta \in Q} \frac{-\log f(y|\theta)}{f(y|\theta)}}{\sum_{\theta \in Q} f(y|\theta)^{-1}} \quad (7.5)$$

Equation 7.5 is the basis for our MMLD model selection formulation for ARCH time series. We now discuss the methods required to calculate MMC’s approximation to MMLD.

---

<sup>1</sup>Importance sampling is when another distribution,  $g(x)$ , is used to increase the probability of generating points from the desired distribution,  $f(x)$ . This may be used when it is inefficient to sample from  $f(x)$ . See Gilks et al.[32] for a good introduction.

## 7.2 Sampling from the Posterior

The MMC algorithm requires a sample,  $S$ , generated from the posterior. There are many different sampling methods available. Fitzgibbon suggests using either Markov Chain Monte Carlo (MCMC), Gibbs sampling or Metropolis-Hastings rejection sampling, depending on the dimensions of the posterior and whether there is a full conditional distribution available[24, Page 82].

We used the Acceptance-Rejection method (also referred to as simply Rejection or Acceptance methods) explained by Gilks[31]. To generate  $S$ , we sample  $N$  points from an *envelope* distribution, denoted  $g(y)$ , to approximate a sample from our desired distribution,  $f(y)$ , by the following algorithm (based on an algorithm from [31, §5.3.1])

```

for all  $i$  such that  $1 \leq i \leq N$  do
  repeat
    Sample a point  $Y$  from  $g(\cdot)$ 
    Sample a point  $U$  from  $Uniform(0, 1)$ 
    If  $U \leq f(Y)/g(Y)$  accept  $Y$ 
  until one  $Y$  is accepted
   $S_i \leftarrow Y$ 
end for

```

For the sample to be plausible, we require

$$g(y) \geq f(y) \quad \forall y \tag{7.6}$$

That is, we sample from a known distribution that is greater or equal to our desired distribution for all points. The tighter the envelope, the less points we need to generate to obtain  $N$  samples, hence, ensuring greater efficiency[32, Page 81].

## 7.3 An Envelope Distribution

After careful inspection of the ARCH posterior in one and two dimensions, we decided to use a multi-variable exponential distribution as our envelope distribution. This is given by

$$E(\vec{B}, \vec{\lambda}) = \prod_{i=1}^p (\lambda_i \exp(B_i - y) \lambda_i) + 0.9 * posterior(\hat{\theta}_{ml}|y) \tag{7.7}$$

where the vectors  $\vec{B}$  and  $\vec{\lambda}$  are  $p$  dimensional and are parameterised to tighten the envelope fit.

$\vec{B}$ , along with the second term in Equation 7.7, position the exponential distribution according to the posterior mode. The elements of  $\vec{B}$ ,  $\{B_i : i = 1, \dots, p\}$  were set to the maximum likelihood estimates. The second term is scaled by 0.9 to ensure Equation 7.6 holds<sup>2</sup>. In other words, since  $\hat{\theta}_{ml}$  are only estimates and may not be the exact maximum of the likelihood, then if  $E(\cdot)$  were positioned at  $\hat{\theta}_{ml}$  without the adjustment, the envelope may not cover the posterior completely.

For example, consider the ARCH(2) process with  $\hat{\theta}_{ml} = (0.1, 0.6)$ . Then  $B = (0.1, 0.6)$  and would position  $E(\cdot)$  at  $(\alpha_0, \alpha_1) = (0.1, 0.6)$ . The second term would then offset  $E(0)$  by  $0.75 * h(\hat{\theta}_{ml})f(y|\hat{\theta}_{ml})$ .

The elements of  $\vec{\lambda}$  adjust the shape of the distribution. Unfortunately, these were set to 1 due to time constraints, consequently, the envelope was not very tight, reducing efficiency. Further investigations into these values may provide a method to obtain heuristics about the posterior, and shape  $E(\cdot)$  accordingly. For example, for small  $T$ , the shape of posterior is broad - since there is less data to infer the probability of  $y$  given  $\theta$  - therefore, we could choose  $\vec{\lambda} \propto T$ . Though, this is merely conjecture at this point as we have not performed any kind of formal analysis.

Using our envelope,  $E(\cdot)$ , we can now generate our sample from the posterior,  $S = \{\theta_i : i = 1, \dots, N\}$ .

## 7.4 Approximating the Optimal Uncertainty Region

We choose  $Q \subseteq S$  to approximate the true uncertainty region,  $R \supseteq Q$ . This is possible through the Boundary rule (from Section 5.5, Equation 5.4), by assuming that one or more  $\theta \in S$  must lie on the boundary of  $R$ [24, page 85]. We then sort each  $\theta$  into descending order of likelihood,  $f(y|\theta)$ , to form a boundary index,  $S = \{\theta_0, \theta_1, \dots, \theta_N\}$ , such that  $-\log f(y|\theta_i) \leq -\log f(y|\theta_{i+1}) \forall \theta \in S$ .

The boundary rule may then be approximated by Monte Carlo integration using the weighted samples from  $Q$ [24, Equation 6.11], such that<sup>3</sup>

$$\theta \in Q \text{ iff } -\log f(y|\theta) \leq -\frac{\sum_{\theta \in Q} h(\theta) \log f(y|\theta) I(\theta)^{-1}}{\sum_{\theta \in Q} h(\theta) I(\theta)^{-1}} + 1 \text{ nit} \quad (7.8)$$

<sup>2</sup>During formulation, we experimented with a number of scalars, testing to how often Equation 7.6 did not hold.

<sup>3</sup>The integrals from Equation 5.4 have been replaced with summations in Equation 7.8 since the sampled  $\theta \in Q$  are discrete.

where  $I(\theta)$  is the importance sampling distribution given by Equation 7.4.

Using the sorted samples in  $S$ , we can easily find  $Q$  by starting from  $\theta_0$  and adding each  $\theta_i$  to  $Q$  until the next sample, say,  $\theta_{i+1}$ , does not satisfy Equation 7.8.

In other words, we start with the closest sample to the posterior mode (i.e,  $\theta_0$ ), and grow the volume of the uncertainty region,  $Q$ , until the next sample,  $\theta_i + 1$  has a negative log likelihood greater than the negative average log likelihood for all  $\theta$  currently in  $Q$ , plus one nit.

## 7.5 Choosing the Point Estimate

Recall from Section 5.5, that the MMLD message length does not provide an estimate of the parameters. In this section, we also mention three methods for finding  $\hat{\theta}$ : Random Coding and prior or posterior weighted minimum Expected Kullback-Leibler estimation (EKL). For our MMLD approximation, we used Random Coding.

Random Coding requires samples,  $\{\theta_1, \theta_2, \dots\}$ , randomly drawn from the prior distribution. Conveniently, we may use the same sampling procedure as used in Section 7.2. Although we sampled from the posterior, and Random coding requires samples from the prior, we can (from Equation 7.4) simply weight each  $\theta$  by  $f(y|\theta)^{-1}$  to transform to a prior sample.

To obtain our point estimate,  $\hat{\theta}$ , we choose the first  $\theta$  that falls within the region, specified by  $Q$ , according to Equation 7.8.

By using the methods described in this Chapter, we have formulated an MMC approximation (Equation 7.5) to the MMLD message length (Equation 7.3) for ARCH time series.

## 7.6 Summary

We have discussed in some detail the Message from Monte Carlo (MMC) approximation for Dowe's Minimum Message Length estimator. We presented the methods used to obtain an MMLD approximate for ARCH time series. Unfortunately, the simulations and sampling methods required by our approximation are quite CPU intensive. As such, we have only been able to compare MMLD against MML87 and other model selection criteria in limited evaluations. These, and more detailed results for our MML87 model selection formulation (§6), are presented in the following chapter.

# Chapter 8

## Evaluations

In this chapter, our MML87 and MMLD ARCH model selection criteria are compared in Monte Carlo simulations against existing model selection criteria AIC, AICc, BIC and HQ (see Section 4.2).

Each simulation consists of a predetermined number of runs. For each run, we generate a ‘true’ ARCH series using random parameters, which are estimated by Maximum Likelihood (§6.1). Each criterion then infers the ‘best’ model from eight different model candidates according to these estimates.

We evaluate each criterion according to the model it infers by correct model order selection, mean square prediction error and average negative log likelihood.

We now discuss these performance measures briefly before describing our simulations in section 8.2

### 8.1 Performance Measures

#### Correct Model Selection

The correct model order selection measures whether each criterion inferred either a lower, higher or correct model order according to the true model. We present these results in total number and percentage.

#### MSPE

The mean square prediction error (MSPE) is also used to measure performance:

$$MSPE = \frac{1}{T} \sum_{t=1}^T \left( y_t - (\hat{\theta}_1 y_{t-1} + \dots + \hat{\theta}_p y_{t-p}) \right)^2 \quad (8.1)$$

## Negative Log Likelihood

The negative log likelihood is evaluated at the maximum likelihood estimates for the model selected by each criteria:  $-\log f(y|\hat{\theta})$

## 8.2 Simulations

Each simulation was configured by number of settings:

- Sample Size:  $T = \{40, 70, 100, 200\}$
- True Model Order:  $p = \{1, 2, 3, 4\}$
- Model candidates: ARCH(1) to ARCH(8)

Each simulation consisted of 100 runs. Each run proceeded as follows:

1. Randomly generate the true parameters,  $\alpha$ , maintaining the stationarity constraint (§3.4.1):

$$\alpha_0 \leftarrow U(0, 1)$$

**for all**  $i$  such that  $1 \leq i \leq p$  **do**

$$u \leftarrow 1 - \sum_{j=1}^{i-1} \alpha_j$$

$$\alpha_i \leftarrow U(0, u)$$

**end for**

where  $U(a, b)$  generates a uniformly random number between  $a$  and  $b$ .

2. Generate the true model  $y = \text{ARCH}(\alpha)$  for  $2 \cdot T$  data. The first  $T$  data,  $\{y_t : t = -T + 1, -T + 2, \dots, -T + T - 1, 0\}$ , are used as a ‘burn in’ to avoid unnecessary assumptions when evaluating the likelihood
3. Estimate the parameters,  $\hat{\theta}$ , for model orders  $p = 1, \dots, 8$ , forming our set of model candidates.
4. For each criterion, record the model inferred.
5. For each criterion, gather statistics (§8.1) according to the model inferred.

Simulations for true models ARCH(1) to ARCH(4), for each set size of  $T = 40, 70, 100, 200$ , were conducted, resulting in a data set of 1600 ( $4 * 100 * 4$ ) ARCH series.

## 8.3 Results

We evaluated four different MML87 estimators: two using the standard MML87 approximation (Equation 6.11) and two using the small sample size approximation (Equation 6.12). We placed the same uniform prior on the number of parameters (Equation 6.7),  $h(p)$ , for each estimator. However, we place two different priors, presented in Section 6.2, on the parameter values,  $h(\theta)$ .

The four estimators are:

- MML87a: Used the standard MML87 approximation with Equation 6.8 as the prior.
- MML87b: Used the standard MML87 approximation with Equation 6.9 as the prior.
- MML87c: Used the small sample MML87 approximation with Equation 6.8 as the prior.
- MML87d: Used the small sample MML87 approximation with Equation 6.9 as the prior.

We also evaluated our MMLD approximation in several of the simulations. We could only provide partially complete results since (our) MMLD's running time is substantially long. We generated 2000 samples from the posterior distribution, using half to obtain our uncertainty region  $S$  (see §7.4) and the other half for our point estimates (after adjusting them accordingly, see §7.5). We placed the same uniform prior on the number of parameters as above, and placed Equation 6.8 as the prior on the parameter values.

### 8.3.1 MML87 vs Existing Criteria

In this section, we present a summary of the results comparing the four MML87 estimators against existing criteria. Full results are given in Appendix D. Limited results for the MMLD estimator is given in the following section.

Tables, and corresponding figures, for average Correct Model Order selection (Table 8.1, Figure 8.1), average MSPE (Table 8.2, Figure 8.2) and average Negative Log Likelihood (Table 8.3, Figure 8.3) are given. Table 8.4 also shows the total number of times each criterion inferred a lower, correct or higher model order.

Best scores are highlighted in bold.

T	AIC	AICc	BIC	HQ	MML87a	MML87b	MML87c	MML87d
40	28.50	<b>29.7</b>	29.25	23.75	10.25	20.00	23.50	21.5
70	26.25	26.25	27.75	22.50	11.50	25.50	<b>28.75</b>	27.00
100	33.00	33.00	34.00	29.00	12.25	36.50	<b>38.75</b>	34.75
200	34.50	34.50	34.50	23.25	15.75	41.50	<b>45.50</b>	35.00

Table 8.1: Results for Average Correct Model Order Selection

T	AIC	AICc	BIC	HQ	MML87a	MML87b	MML87c	MML87d
40	0.20999	0.20958	0.20992	0.20446	0.34362	1.08212	<b>0.15603</b>	0.19699
70	0.10400	0.10391	0.10171	0.15039	0.22866	1.13899	<b>0.09712</b>	0.09801
100	0.32543	0.32543	0.32465	1.37359	0.37442	0.83074	<b>0.28941</b>	0.31053
200	0.05547	0.05547	<b>0.05534</b>	0.15733	0.06737	0.47954	0.05592	0.05949

Table 8.2: Results for Average Mean Square Prediction Error

T	AIC	AICc	BIC	HQ	MML87a	MML87b	MML87c	MML87d
40	<b>53.6403</b>	53.6466	53.6507	54.9503	55.7143	55.9448	54.1092	54.1392
70	<b>97.5993</b>	97.6001	97.6040	100.1656	99.7584	103.2687	98.0440	98.0733
100	<b>143.9523</b>	143.9528	143.9552	246.9647	145.9489	149.6704	144.3202	144.3644
200	<b>271.6157</b>	271.6157	271.6175	278.1194	273.2158	285.6161	271.9470	272.4964

Table 8.3: Results for Average Negative Log Likelihood

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	%
AIC	457	491	652	30.688
AICc	462	496	642	31.000
BIC	482	502	612	31.375
HQ	648	394	558	24.625
MML87a	130	251	1319	15.656
MML87b	927	494	179	30.875
MML87c	893	<b>546</b>	161	<b>34.125</b>
MML87d	785	493	322	30.813

Table 8.4: Aggregate results for Criteria under/correctly/over selecting the true model

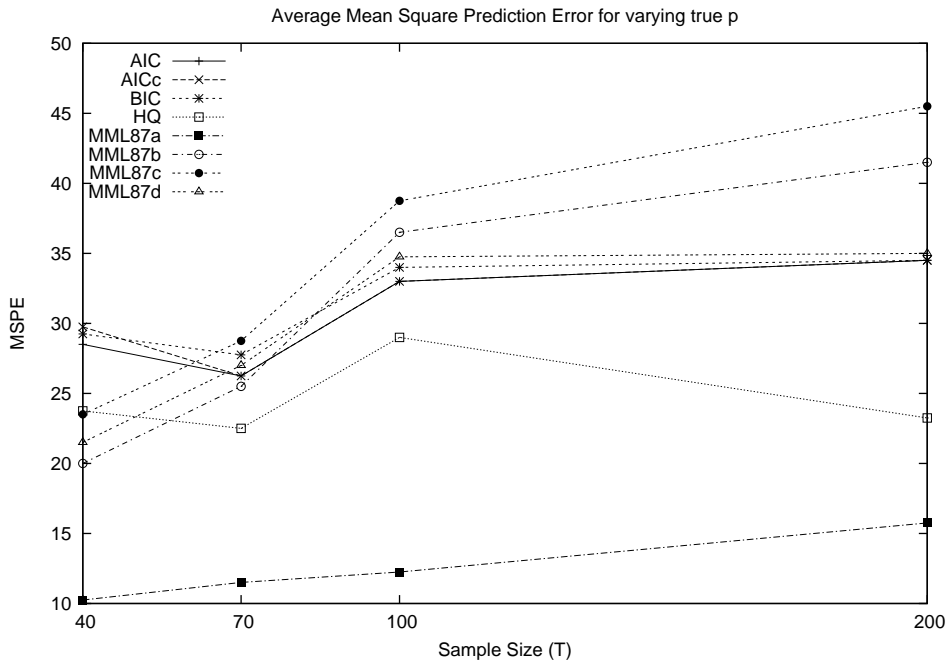


Figure 8.1: Results for Average Correct Model Order Selection (%)

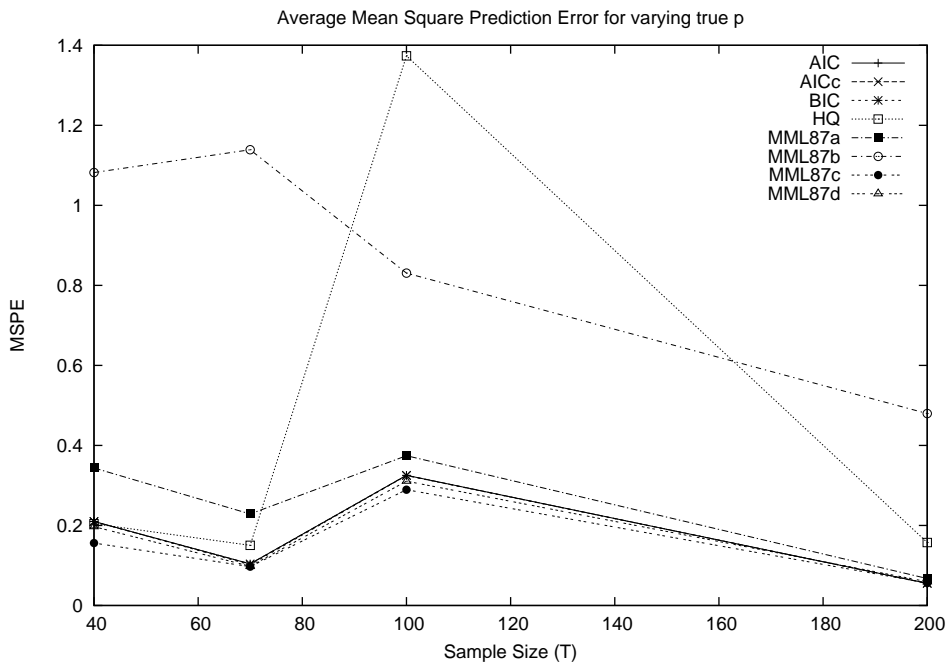


Figure 8.2: Results for Average Mean Square Prediction Error

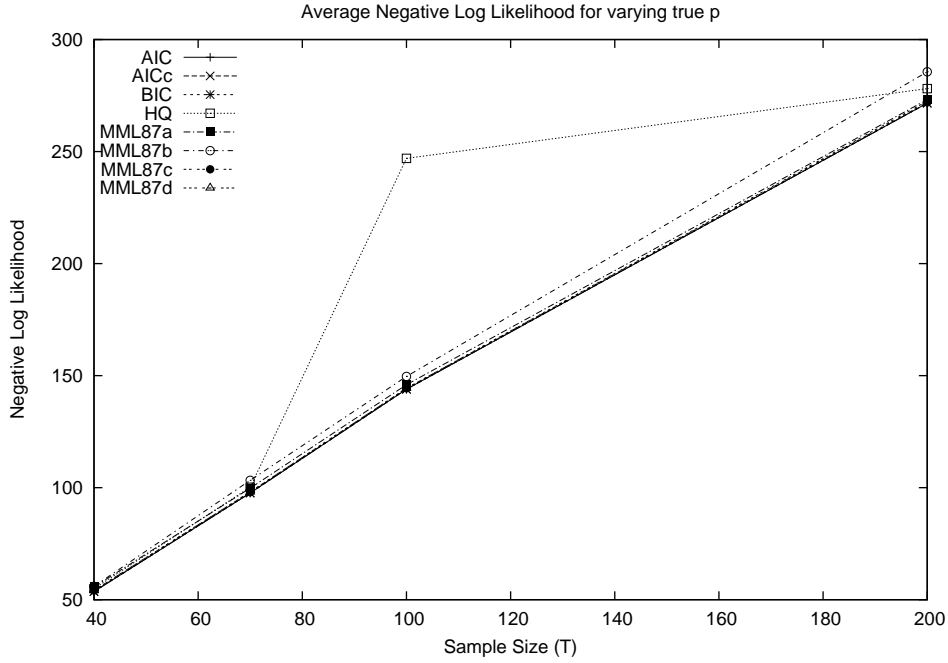


Figure 8.3: Results for Average Negative Log Likelihood

### 8.3.2 MMLD Results

Our approximations to the MMLD estimator have long running times, consequently, we only obtained results for data sample sizes 40 and 70 for true models ARCH(1), ARCH(2) and ARCH(3). The results for MMLD use the same sample data generated from the previous section. The results are presented in Table 8.5.

T	$\hat{p} < p$	$\hat{p} = p$ (%)	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
40	10	44 (14.67)	246	52.7831	0.12160
70	28	53 (17.67)	219	93.9317	0.07957

Table 8.5: Results for MMLD for  $T = \{40, 70\}$  and  $p = \{1, 2, 3\}$

## 8.4 Discussion of Results

Two of the four MML87 estimators, MMLc and MMLd, performed competitively according to all three measures for sample sizes 70, 100 and 200.

MML87c stood above the rest, outperforming all other criteria in overall correct model selection, correctly choosing the true model 546 times ( $\sim 34\%$ ), compared to the next best, BIC (502). It also scored the best average MSPE for sample sizes 40, 70 and 100.

MML87c selected the true model order in 45% of the runs for sample size 200 (for all  $p$ ), 4% ahead of MML87b and more than 10% ahead of the AIC, AICc, BIC and HQ. MML87c also selected the highest percentage of true models for sample sizes 70 and 100, although, the margin ahead of its competitors aren't as significant. For a sample size of 40, its performance is somewhat ( $\sim 6\%$ ) worse than the best criteria, AICc (29.7%).

These findings are reflected by MML87c's average MSPE results. It has the lowest average MSPE for sample sizes 40, 70 and 100 and is only marginal less than the best criteria, BIC, for sample size of 200. Its average MSPE is approximately 75% better than the next best criterion, MML87d for the smallest sample size simulated.

MML87d performed notably for average MSPE, coming second behind MML87c for sample sizes 40, 70 and 100 and only marginally less ( $\sim 0.004$ ) against AIC, AICc and BIC (whose performances improve for larger sample sets) for sample size 200. Its model selection accuracy is comparable to the existing criteria except for sample size 40.

MML87a clearly performed the worst when choosing the true model, selecting the correct model order in only 15% of runs. It is evident from Table 8.4, MML87a is bias towards higher order models, inferring a higher order model 4 runs in 5 (1319 times).

However, MML87a's average MSPE and negative log likelihood do not reflect this high inaccuracy. Its negative log likelihood is only slightly higher for all sample sizes, and outperforms HQ and MM87b in MSPE for sample size 200.

Our MML87 estimators did not perform as expected<sup>1</sup>, being only competitive for most simulations. This could possibly be explained by the (sometimes) inaccurate parameter estimates (see Appendix B). Although every other criteria (except MMLD) was measured on their performance using these estimates, the inaccuracies may be compounded to the MML87 estimators since we used the observed Fisher Information<sup>2</sup>.

Ideally, the exact Fisher Information could be found and used in conjunction with MML parameter estimates for ARCH time series (instead of using Maximum Likelihood). This, and other possible future work is discussed in the next chapter.

### 8.4.1 MMLD Discussion

Although only few simulations were conducted and, consequently, limited results obtained, the MMLD estimator showed some potential.

---

<sup>1</sup>Based on previous results for MML as time series model selection. See [28, 44].

<sup>2</sup>Recall from Section 6.4 that the parameter estimates are used when calculating Equation 6.10.

MMLD only estimated the true model in about 15% of runs. However, recall from Section 7.5, that the MMLD estimator infers parameters (using the point estimates), instead of relying on the Maximum Likelihood estimates<sup>3</sup>. However, if the inferred model's higher order co-efficients are insignificant, the inability to select the 'true' model order will not be a good indicator of performance[17, See Footnote 29, Page 68].

For example, if a 4th order model is chosen instead of the true second order model, then as long as the 3rd and 4th order co-efficients are (very) small, it is still a reasonable choice.

As such, it is important to consider MMLD's parameter estimates. In the tests conducted, an indirect measure is possible through its MSPE and negative log likelihood results. Both of these are comparable to the other criteria for samples sizes 40 and 70.

Ignoring the results for model order 4, MML87c's average MSPE and negative log likelihood for sample size 40 were 0.0910 and 51.8597, respectively. While the best performing criterion between AIC, AICc, BIC and HQ for average MSPE was AIC (51.4083) and AICc (0.10491). From Table 8.5, we see that MMLD's corresponding results are only marginally higher.

Obviously, these results are only partially complete. The approximations to MMLD may be parameterised much differently. Additionally, only 1000 points were used for the posterior sample due to time constraints. Ideally, the number of points should be as high as possible to improve the approximations as much as possible. These suggestions and others are discussed in Future Work in the following chapter.

## 8.5 Summary

The empirical simulations conducted and three performance measures were described in this chapter. Five MML model selection criteria, four MML87 estimators and our MMLD estimator, were evaluated against AIC, AICc, BIC and HQ. Although, overall MML87 under-performed to our expectations, two MML87 were found to perform competitively. Possible reasons were allured to and further discussion of improvements to these are presented in the next chapter.

Our MMLD estimator under limitations showed potential. Although its correct model order selection was low, we demonstrated that this is not the only, and sometimes, possibly incorrect, performance measure. MMLD infers its own parameter

---

<sup>3</sup>This is not entirely true. We used the ML estimates as an indicator of the posterior mode. However, the actually used parameters will vary from these.

estimates and appeared to bias higher order models. The MSPE and average negative log likelihood results showed that MMLD as an ARCH model selection criterion has much potential.

The next chapter concludes this thesis with a brief overview of what has been achieved and possible directions for future work.

# Chapter 9

## Conclusion

This thesis introduced time series, described their main features, and provided a small example of their applications in a diverse range of fields. Time series models provide a method for better understanding and an ability to predict future behaviour. The two main model types, homoskedastic and heteroskedastic were introduced, along with their specific models. Of particular interest, was the ARCH process for modelling a changing variance.

Model selection is a method for selecting the ‘best’ model between competing models. We briefly discussed model selection, it’s objectives and several existing popular criteria were introduced. Namely, AIC, corrected AIC, BIC and HQ.

The Minimum Message Length (MML) principle was described in detail, before introducing two approximations, Wallace and Freeman’s MML87 and Dowe’s MMLD. Four new MML87 and one MMLD model selection criteria for ARCH time series were formulated.

These were compared against existing criteria in empirical simulations and quantitative results presented. Two MML87 criteria proved competitive, while our MMLD criterion showed potential.

The limitations of our research were briefly stated during discussion of our results. Specifically, the use of the observed Fisher Information instead of the expected for the MML87 estimators and the lack of efficiency for the MMLD estimator.

The research contributed to the field of model selection by providing two new, theoretically different, criteria for ARCH models, extended upon Sak’s preliminary work[43]. Additional to this, the MMLD formulation appears to be the first, to my knowledge, to provide (although preliminary) results for Wallace’s Random Coding method.

Our MMLD estimator implemented the Message from Monte Carlo (MMC) algorithm. MMC uses stochastic simulation to approximate several difficult problems in Dowe’s minimum message length. Our research provided an initial method for

applying MMC to ARCH time series, however, the complexities ensure there is room for much improvement.

## 9.1 Future Work

We now suggest some possible improvements to both the MMLD estimator and our formulation methods for MML87.

Our MML87 formulation utilised the Maximum Likelihood (ML) estimates to calculate the likelihoods and observed Fisher Information. Optimisation of the ML estimates using the Score algorithm is time consuming, particularly for high dimensions. Additionally, the Score algorithm did not always provide plausible values, such as negative estimates, and conformity checks were required. Alternative optimisation techniques, such as the Berndt, Hall, Hall and Hausman (BHHH)[4] method, could be investigated.

MML may also be used for parameter estimation (see Chapter 5). Using MML instead of ML estimates could prove beneficial for our MML87 estimators, providing an advantage over the existing criteria.

Our MML87 formulation used the observed instead of the expected Fisher Information matrix. As noted in the results, the observed may have been inaccurate for some series. Use of the exact expectations would avoid this and is another area of possible future research.

There is much potential for improvement to our MMC implementation for the MMLD estimator. We employed the Acceptance sampling method, however, as indicated in Section 7.2 there are numerous methods available. Future work could look at using these different methods.

The envelope function used when sampling from the posterior is an area of further investigation. Improvements to the fit of the exponential function to increase efficiency could be looked into, as could altogether, alternative envelope functions.

Additional to the above sampling improvements, future research may also investigate the two other methods of point estimation; minimum Expected Kullback Estimation weighting by either the posterior or prior.

Apart from these suggested improvements, further work could look at other heteroskedastic time series models. There have been a large number introduced since ARCH's inception in 1982. Generalised ARCH(GARCH)[5] is the obvious next model to investigate for MML model selection. After GARCH, future research could

also look into other ARCH-type models, such as STARCH, EGARCH, TARCH, IGARCH.

# References

- [1] Y Agusta and D L Dowe. Unsupervised learning of correlated multivariate Gaussian mixture models using MML. In T D Gedeon and L C Fung, editors, *Proceedings of the Sixteenth Australian Joint Conference on Artificial Intelligence*, pages 477–489, Berlin, Germany, December 2003. Springer-Verlag.
- [2] H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [3] Australian Automobile Association. Victorian petrol prices. Data available from <http://www.aaa.asn.au/>.
- [4] E Berndt, B Hall, R Hall, and J Hausman. Estimation and inference in nonlinear structural models. *Annals of Social Measurement*, 3:653–665, 1974.
- [5] T Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.
- [6] T Bollerslev, R F Engle, and D B Nelson. *Handbook of Econometrics*, volume IV, chapter 49: ARCH Models, pages 2959–3038. Elsevier Science, Amsterdam, The Netherlands, 1994.
- [7] G E P Box and G M Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised (second) edition, 1976.
- [8] P J Brockwell and R A Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag New York, Inc., New York, second edition, 2002.
- [9] K P Burnham and D R Anderson. *Model Selection and Multimodel Inference*. Springer, New York, 2002.
- [10] C Chatfield. *The Analysis of Time Series: An Introduction*. Chapman and Hall, New York, sixth edition, 2004.

- [11] M J Collie, D L Dowe, and L J Fitzgibbon. Trading rule search with autoregressive inference agents. Technical report, School of Computer Science and Software Engineering, Monash University, 2005. Obtained from <http://www.fitzgibbon.name/leigh/papers/>.
- [12] J H Conway and N J A Sloane. *Sphere Packings, Lattices and Groups*. Springer-Verlag, New York, 1999.
- [13] Allin Cottrell. gretl: GNU regression, econometric and time-series library. Website: <http://gretl.sourceforge.net/>.
- [14] Allin Cottrell. gretl: GNU regression, econometric and time-series library, 2004. Source: “Cigarette demand, health scares, and education in Turkey,” by Aysit Tansel, Applied Economics, 1993, pp. 521-529.
- [15] Allin Cottrell. gretl: GNU regression, econometric and time-series library, 2004. Source: Annual data on U.S. Military expenditures and their determinants Data compiled by Jan Blackburn from the History of American Statistics from Colonial Times, and Historical Tables: Budget of the U.S. Government.
- [16] Allin Cottrell. gretl: GNU regression, econometric and time-series library, 2004. Source: Monthly data from 1990.01 through 1998.12 for Mellon Bank stock listed on the NYSE, compiled by Sebastian Badali.
- [17] D L Dowe, S Gardner, and G Oppy. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science*, Accepted 29 June 2006. To Appear In.
- [18] D L Dowe, J J Oliver, and C S Wallace. MML estimation of the parameters of the spherical Fisher distribution. *Lecture Notes in Artificial Intelligence (LNAI)*, 1160:213–227, 1996.
- [19] D L Dowe and C S Wallace. Resolving the Neyman-Scott problem by minimum message length. *Computing Science and Statistics Series*, 28:614–618, 1997.
- [20] W Enders. *Applied Econometric Time Series*. Wiley, Hoboken, second edition, 2004.
- [21] R F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 1982.
- [22] G E Farr and C S Wallace. *Research on Combinatorial Algorithms*, chapter Algorithmic and combinatorial problems in strict minimum message length inference, pages 50–58. 1997.

- [23] G E Farr and C S Wallace. The complexity of strict minimum message length inference. *The Computer Journal*, 45(3):285–292, 2002.
- [24] L Fitzgibbon. *Message from Monte Carlo: A Framework for Minimum Message Length Inference using Markov Chain Monte Carlo Methods*. PhD thesis, Monash University, Clayton, Wellington Rd, Victoria, 3800, Australia, 2004.
- [25] L J Fitzgibbon, L Allison, and D L Dowe. Minimum message length grouping of ordered data. In H Arimura and S Jain, editors, *Proceedings of the Eleventh International Conference on Algorithmic Learning Theory, Lecture Notes in Artificial Intelligence (LNAI)*, pages 56–70, Berlin, December 2000. Springer-Verlag.
- [26] L J Fitzgibbon, D L Dowe, and L Allison. Change-point estimation using new minimum message length approximations. In M Ishizuka and A Sattar, editors, *Proceedings of the Seventh Pacific Rim International Conference on Artificial Intelligence*, pages 244–254, August 2002.
- [27] L J Fitzgibbon, D L Dowe, and L Allison. Univariate polynomial inference by Monte Carlo message length. In *Nineteenth International Conference on Machine Learning (ICML)*, pages 147–154, Sydney, Australia, 8-12 July 2002.
- [28] L J Fitzgibbon, D L Dowe, and F Vahid. Minimum message length autoregressive model order selection. In M Palanaswami, C Chandra Sekhar, G Kumar Venayagamoorthy, S Mohan, and M K Ghantasala, editors, *International Conference on Intelligent Sensing and Information Processing (ICISIP)*, pages 439–444, Chennai, India, 4-7 January 2004.
- [29] P H Franses. *Time Series Models for Business and Economic Forecasting*. Cambridge University Press, Cambridge, 1998.
- [30] J Geweke. Exact predictive densities for linear models with ARCH disturbances. *Journal of Econometrics*, 40:63–86, 1989.
- [31] W R Gilks. *Markov Chain Monte Carlo in Practise*, chapter 5: Full conditional distributions. Chapman and Hall, London, 1996.
- [32] W R Gilks, S Richardson, and D J Spiegelhalter. *Markov Chain Monte Carlo in Practise*. Chapman and Hall, London, 1996.
- [33] C Gouriéroux. *ARCH Models and Financial Applications*. Springer, New York, 1997.
- [34] J D Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, New Jersey, 1994.

- [35] E J Hannan and B G Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41(2):190–195, 1979.
- [36] C M Hurvich and C Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.
- [37] E Lam. Improved approximations in MML, 2000. Honours Thesis.
- [38] H Linhart and W Zucchini. *Model Selection*. John Wiley & Sons, New York, 1986.
- [39] E Mansfield. *Statistics for Business and Economics*. W. W. Norton & Company, New York, fifth edition, 1994.
- [40] H Mitchell and M D McKenzie. GARCH model selection criteria. *Quantitative Finance*, 3(4):262–284, 2003.
- [41] Bureau of Meteorology. Timeseries - global climate variability and change. Data available from <http://www.bom.gov.au>.
- [42] J Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society, Series B*, 49(3):223–239, 1987.
- [43] M Sak. MA and ARCH time series model inference using minimum message length. Honours Thesis (only preliminary results for ARCH), Monash University, Clayton, Wellington Rd, Victoria, 3800, Australia, 2004.
- [44] M Sak, D L Dowe, and S Ray. Minimum message length moving average time series data mining. In *Proceedings of the First International ICSC Symposium on Advanced Computing in Financial Markets*, Istanbul, Turkey, 15-17 December 2005.
- [45] G Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [46] C E Shannon. A mathematical theory of communication. *Bell System Tech Journal*, 27:379–423, 1948.
- [47] T Teriásvirta, D Tjøstheim, and C W J Granger. *Handbook of Econometrics*, volume IV, chapter 48: Aspects of Modelling Nonlinear Time Series, pages 2917–2957. Elsevier Science, Amsterdam, The Netherlands, 1994.
- [48] C S Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, New York, 2005.

- [49] C S Wallace and D M Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.
- [50] C S Wallace and D M Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.
- [51] C S Wallace and D L Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.
- [52] C S Wallace and P R Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society. Series B*, 49(3):240–265, 1987.
- [53] W Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44:41–61, 200.

# Appendix A

## ARCH(p) First and Second Derivatives

The first partial derivatives of zero mean ARCH(p) process have the general form

$$\begin{aligned}\frac{\partial l_t}{\partial \alpha_i} &= \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \alpha_i} - \frac{1}{2} \frac{\partial \sigma_t^2}{\partial \alpha_i} \frac{y_t^2}{\sigma_t^4} \\ &= \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \alpha_i} \left( \frac{y_t^2}{\sigma_t^2} - 1 \right)\end{aligned}\tag{A.1}$$

where

$$\frac{\partial \sigma_t^2}{\partial \alpha_0} = 1\tag{A.2}$$

$$\frac{\partial \sigma_t^2}{\partial \alpha_i} = y_{t-i}^2 \forall i > 0\tag{A.3}$$

Substituting (A.2) and (A.3) into (A.1), the first partial derivatives of (6.5) for an ARCH(p) process are

$$\begin{aligned}\frac{\partial l_t}{\partial \alpha_0} &= \frac{1}{2\sigma_t^2} \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) \\ \frac{\partial l_t}{\partial \alpha_i} &= \frac{1}{2\sigma_t^2} y_{t-i}^2 \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i > 0\end{aligned}$$

and follows that

$$\frac{\partial L}{\partial \alpha_0} = \sum_{t=1}^T \frac{1}{2\sigma_t^2} \left( \frac{y_t^2}{\sigma_t^2} - 1 \right)\tag{A.4}$$

$$\frac{\partial L}{\partial \alpha_i} = \sum_{t=1}^T \frac{1}{2\sigma_t^2} y_{t-i}^2 \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i > 0\tag{A.5}$$

From (A.1), the second partial derivatives have the general form

$$\begin{aligned}\frac{\partial l_t^2}{\partial \alpha_i \partial \alpha_j} &= \frac{\partial}{\partial \alpha_i} \left( \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \alpha_i} \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) \right) \\ &= \frac{\partial}{\partial \alpha_j} \left[ \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \alpha_i} \right] \left( \frac{y_t^2}{\sigma_t^2} - 1 \right) - \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \alpha_i} \frac{\partial \sigma_t^2}{\partial \alpha_j} \frac{y_t^2}{\sigma_t^4}\end{aligned}$$

The second partial derivatives of (6.5) for an ARCH(p) process are then

$$\begin{aligned}\frac{\partial^2 l_t}{\partial \alpha_0^2} &= \frac{1}{2\sigma_t^4} - \frac{y_t^2}{\sigma_t^6} \\ \frac{\partial^2 l_t}{\partial \alpha_i^2} &= \frac{y_{t-i}^4}{2\sigma_t^4} - \frac{y_t^2 y_{t-i}^4}{\sigma_t^6} \forall i > 0 \\ \frac{\partial^2 l_t}{\partial \alpha_0 \partial \alpha_i} &= \frac{\partial^2 l_t}{\partial \alpha_i \partial \alpha_0} = \frac{y_{t-i}^2}{2\sigma_t^4} - \frac{y_t^2 y_{t-i}^2}{\sigma_t^6} \forall i > 0 \\ \frac{\partial^2 l_t}{\partial \alpha_i \partial \alpha_j} &= \frac{\partial^2 l_t}{\partial \alpha_j \partial \alpha_i} = \frac{y_{t-i}^2 y_{t-j}^2}{2\sigma_t^4} - \frac{y_t^2 y_{t-i}^2 y_{t-j}^2}{\sigma_t^6} \forall i, j > 0 \wedge i \neq j\end{aligned}$$

then it follows

$$\begin{aligned}\frac{\partial L}{\partial \alpha_0^2} &= \sum_{t=1}^T \left( \frac{1}{2\sigma_t^4} - \frac{y_t^2}{\sigma_t^6} \right) \\ &= -\frac{1}{2} \sum_{t=1}^T \frac{1}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right)\end{aligned}\tag{A.6}$$

$$\begin{aligned}\frac{\partial L}{\partial \alpha_i^2} &= \sum_{t=1}^T \left( \frac{y_{t-i}^4}{2\sigma_t^4} - \frac{y_t^2 y_{t-i}^4}{\sigma_t^6} \right) \forall i > 0 \\ &= -\frac{1}{2} \sum_{t=1}^T \frac{y_{t-i}^4}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right)\end{aligned}\tag{A.7}$$

$$\begin{aligned}\frac{\partial L}{\partial \alpha_0 \partial \alpha_i} &= \frac{\partial L}{\partial \alpha_i \partial \alpha_0} = \sum_{t=1}^T \left( \frac{y_{t-i}^2}{2\sigma_t^4} - \frac{y_t^2 y_{t-i}^2}{\sigma_t^6} \right) \\ &= -\frac{1}{2} \sum_{t=1}^T \frac{y_{t-i}^2}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i > 0\end{aligned}\tag{A.8}$$

$$\begin{aligned}\frac{\partial L}{\partial \alpha_i \partial \alpha_j} &= \frac{\partial L}{\partial \alpha_j \partial \alpha_i} = \sum_{t=1}^T \left( \frac{y_{t-i}^2 y_{t-j}^2}{2\sigma_t^4} - \frac{y_t^2 y_{t-i}^2 y_{t-j}^2}{\sigma_t^6} \right) \\ &= -\frac{1}{2} \sum_{t=1}^T \frac{y_{t-i}^2 y_{t-j}^2}{\sigma_t^4} \left( 2 \frac{y_t^2}{\sigma_t^2} - 1 \right) \forall i, j > 0 \wedge i \neq j\end{aligned}\tag{A.9}$$

# Appendix B

## Score Algorithm

Engle[21] suggests the use of the ‘Score’ algorithm to numerically optimise the maximum likelihood function and, consequently, estimate the parameters, of an ARCH(p) process.

The Score algorithm is an iterative process, using the first and second derivatives of the function whose optimum we are seeking. Given initial starting values for our parameters,  $\theta_0$ , we estimated the maximum of the ARCH (log) likelihood by:

$$\theta_0 \leftarrow U(0, 1)$$

**repeat**

$$\theta_{i+1} = \theta_i + F(\theta_i)^{-1} \frac{1}{T} \sum_{t=1}^T \frac{\partial f(y_t|\theta_i)}{\partial \theta_i}$$

**until**  $\nabla \theta \leq \tau$

where  $f(y_t|\theta_i)$  is the likelihood function and  $F(\theta_i)$  is the Fisher Information, both evaluated with the  $i$ th parameter estimate.

We stopped the algorithm once the change in the parameter estimates,  $\nabla \theta = \theta_{i+1} - \theta_i$  is less than some threshold,  $\tau$ .

# Appendix C

## ARCH(p) Fisher Information

The Fisher Information is the matrix of expectations of the second partial derivatives of the log likelihood. The exact expectations are difficult to find analytically, hence, the observed expectations were used instead.

Since  $v_t = y_t/\sigma_t$  and  $E(v_t) = 0$  and  $Var(v_t) = E(v_t^2) = 1$  then

$$E\left(\frac{y_t^2}{\sigma_t^2}\right) = E(v_t^2) = 1$$

consequently

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_0^2}\right) &= -\frac{1}{2}E\left(\sum_{t=1}^T \frac{1}{\sigma_t^4} \left(2\frac{y_t^2}{\sigma_t^2} - 1\right)\right) \\ &= -\frac{1}{2}\sum_{t=1}^T E\left(\frac{1}{\sigma_t^4}\right) \end{aligned} \quad (C.1)$$

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_i^2}\right) &= -\frac{1}{2}E\left(\sum_{t=1}^T \frac{y_{t-i}^4}{\sigma_t^4} \left(2\frac{y_t^2}{\sigma_t^2} - 1\right)\right) \\ &= -\frac{1}{2}\sum_{t=1}^T E\left(\frac{y_{t-i}^4}{\sigma_t^4}\right) \forall i > 0 \end{aligned} \quad (C.2)$$

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_0 \partial \alpha_i}\right) &= E\left(\frac{\partial L}{\partial \alpha_i \partial \alpha_0}\right) = -\frac{1}{2}E\left(\sum_{t=1}^T \frac{y_{t-i}^2}{\sigma_t^4} \left(2\frac{y_t^2}{\sigma_t^2} - 1\right)\right) \\ &= -\frac{1}{2}\sum_{t=1}^T E\left(\frac{y_{t-i}^2}{\sigma_t^4}\right) \forall i > 0 \end{aligned} \quad (C.3)$$

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_i \partial \alpha_j}\right) &= E\left(\frac{\partial L}{\partial \alpha_j \partial \alpha_i}\right) = -\frac{1}{2}E\left(\sum_{t=1}^T \frac{y_{t-i}^2 y_{t-j}^2}{\sigma_t^4} \left(2\frac{y_t^2}{\sigma_t^2} - 1\right)\right) \\ &= -\frac{1}{2}\sum_{t=1}^T E\left(\frac{y_{t-i}^2 y_{t-j}^2}{\sigma_t^4}\right) \forall i, j > 0 \wedge i \neq j \end{aligned} \quad (C.4)$$

Equations (C.1,C.2,C.3,C.4) are estimated by

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_0^2}\right) &= \sum_{t=1}^T E\left(\frac{\partial l_t}{\partial \alpha_0^2}\right) \\ &= -\frac{1}{2T} \sum_{t=1}^T \frac{1}{\sigma_t^4} \end{aligned} \quad (\text{C.5})$$

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_i^2}\right) &= \sum_{t=1}^T E\left(\frac{\partial l_t}{\partial \alpha_i^2}\right) \\ &= -\frac{1}{2T} \sum_{t=1}^T \frac{y_{t-i}^4}{\sigma_t^4} \end{aligned} \quad (\text{C.6})$$

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_0 \partial \alpha_i}\right) &= E\left(\frac{\partial L}{\partial \alpha_i \partial \alpha_0}\right) = \sum_{t=1}^T E\left(\frac{\partial l_t}{\partial \alpha_0 \partial \alpha_i}\right) \\ &= -\frac{1}{2T} \sum_{t=1}^T \frac{y_{t-i}^2}{\sigma_t^4} \end{aligned} \quad (\text{C.7})$$

$$\begin{aligned} E\left(\frac{\partial L}{\partial \alpha_i \partial \alpha_j}\right) &= E\left(\frac{\partial L}{\partial \alpha_j \partial \alpha_i}\right) = \sum_{t=1}^T E\left(\frac{\partial l_t}{\partial \alpha_i \partial \alpha_j}\right) \\ &= -\frac{1}{2T} \sum_{t=1}^T \frac{y_{t-i}^2 y_{t-j}^2}{\sigma_t^4} \end{aligned} \quad (\text{C.8})$$

Forming the matrix from equations (C.5,C.6,C.7,C.8), the empirical Fisher Information matrix is given by

$$I_{\alpha_i \alpha_j} = \frac{1}{2T} \sum_{t=1}^T \begin{bmatrix} \frac{1}{\sigma_t^4} & \frac{y_{t-1}^2}{\sigma_t^4} & \dots & \frac{y_{t-p}^2}{\sigma_t^4} \\ \frac{y_{t-1}^2}{\sigma_t^4} & \frac{y_{t-1}^4}{\sigma_t^4} & \dots & \frac{y_{t-1}^2 y_{t-p}^2}{\sigma_t^4} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{y_{t-p}^2}{\sigma_t^4} & \frac{y_{t-p}^2 y_{t-1}^2}{\sigma_t^4} & \dots & \frac{y_{t-p}^4}{\sigma_t^4} \end{bmatrix} \quad (\text{C.9})$$

# Appendix D

## Complete Results

The complete results for all simulations are given below. Each table represents each simulation containing 100 runs. The sample size and true model order are given for each table.

Each criterion is listed along with corresponding statistics for under/correctly/over inferring the model (since these are out of 100 they are percentages too), average negative log likelihood and mean square prediction error. The best scores are highlighted in bold.

Results for MMLD were only obtained for  $T = \{40, 70\}$  and  $p = \{1, 2, 3\}$ .

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	17	44	39	<b>46.5967</b>	0.08150
AICc	17	<b>45</b>	38	46.5973	0.08148
BIC	18	44	38	46.5979	0.08151
HQ	26	34	40	47.1841	0.10352
MML87a		0	100	48.9763	0.23054
MML87b	42	43	15	47.3695	0.10162
MML87c	53	41	6	47.0150	0.04227
MML87d	52	35	13	46.9982	<b>0.04189</b>
MMLD	2	10	88	47.02518	0.07883

Table D.1: Results for  $T = 40$  and  $p = 1$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	28	31	41	<b>49.0236</b>	0.08478
AICc	29	<b>33</b>	38	49.0355	<b>0.08388</b>
BIC	32	32	36	49.0421	0.08459
HQ	34	25	41	49.8430	0.10869
MML87a	0	1	99	51.0868	0.26034
MML87b	69	21	10	51.9897	0.37069
MML87c	67	28	5	49.5346	0.09058
MML87d	58	23	19	49.5557	0.08471
MMLD	2	12	86	50.2917	0.08915

Table D.2: Results for  $T = 40$  and  $p = 2$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	48	22	30	<b>58.6046</b>	0.14999
AICc	50	22	28	58.6123	0.14936
BIC	50	<b>23</b>	27	58.6138	0.14954
HQ	52	22	26	60.0334	0.18939
MML87a	2	0	98	60.3830	0.27908
MML87b	85	9	6	61.0339	0.43658
MML87c	77	18	5	59.0296	<b>0.14052</b>
MML87d	71	18	11	59.1126	0.16614
MMLD	6	22	72	61.0317	0.19682

Table D.3: Results for  $T = 40$  and  $p = 3$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	61	17	22	<b>60.3363</b>	0.52369
AICc	61	<b>19</b>	20	60.3415	0.52356
BIC	64	18	18	60.3491	0.52404
HQ	74	14	12	62.7408	0.41624
MML87a	1	0	99	62.4112	0.60451
MML87b	85	7	8	63.3861	3.41959
MML87c	87	7	6	60.8576	<b>0.35073</b>
MML87d	80	10	10	60.8905	0.49522
MMLD	-	-	-	-	-

Table D.4: Results for  $T = 40$  and  $p = 4$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	12	28	60	<b>83.3167</b>	0.02598
AICc	12	28	60	83.3176	0.02592
BIC	14	32	54	83.3217	0.02591
HQ	21	27	52	84.8804	0.05899
MML87a	0	1	99	85.8962	0.073316
MML87b	36	<b>53</b>	11	85.8102	0.10592
MML87c	39	52	9	83.8086	<b>0.02507</b>
MML87d	39	39	22	83.7822	0.02740
MMLD	8	18	74	84.7481	0.04639

Table D.5: Results for  $T = 70$  and  $p = 1$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	26	26	48	<b>95.2712</b>	0.08583
AICc	26	26	48	95.2718	0.08582
BIC	27	28	45	95.2752	0.08570
HQ	37	29	34	99.3758	0.16002
MML87a	4	0	96	97.5358	0.22230
MML87b	59	22	19	100.7036	0.80718
MML87c	57	<b>30</b>	13	95.7683	0.07283
MML87d	46	24	30	95.6274	<b>0.07578</b>
MMLD	10	14	76	96.0184	0.08695

Table D.6: Results for  $T = 70$  and  $p = 2$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	36	24	40	<b>99.5113</b>	0.09936
AICc	36	24	40	99.5116	0.09936
BIC	38	25	37	99.5170	0.09948
HQ	48	20	32	101.2527	0.11381
MML87a	1	1	98	101.3842	0.25503
MML87b	74	12	14	105.6256	0.66634
MML87c	70	21	9	99.9395	<b>0.09209</b>
MML87d	58	<b>29</b>	13	100.0830	0.09609
MMLD	10	21	69	101.0286	0.10538

Table D.7: Results for  $T = 70$  and  $p = 3$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	52	<b>27</b>	21	<b>112.2979</b>	0.20483
AICc	53	<b>27</b>	20	112.2993	0.20450
BIC	55	26	19	112.3020	0.19575
HQ	63	14	23	115.1535	0.26870
MML87a	3	4	93	114.2172	0.36397
MML87b	81	15	4	120.9354	2.97652
MML87c	80	12	8	112.6597	0.19846
MML87d	72	16	12	112.8004	<b>0.19275</b>
MMLD	-	-	-	-	-

Table D.8: Results for  $T = 70$  and  $p = 4$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	6	46	48	<b>117.5582</b>	0.01774
AICc	6	46	48	117.5584	0.01774
BIC	6	49	45	117.5615	0.01768
HQ	13	49	38	118.7457	0.03560
MML87a	0	2	98	120.0962	0.06061
MML87b	22	68	10	121.1493	0.16803
MML87c	18	<b>75</b>	7	117.8976	<b>0.01417</b>
MML87d	19	61	20	117.9430	0.01887
MMLD	-	-	-	-	-

Table D.9: Results for  $T = 100$  and  $p = 1$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	20	37	43	<b>143.5067</b>	0.16451
AICc	20	37	43	<b>143.5067</b>	0.16451
BIC	20	<b>38</b>	42	143.5083	0.16443
HQ	32	26	42	145.8955	0.22319
MML87a	2	2	96	145.6107	0.21850
MML87b	52	36	12	148.6377	0.63548
MML87c	50	34	16	143.8714	<b>0.05625</b>
MML87d	42	27	31	143.8321	0.05856
MMLD	-	-	-	-	-

Table D.10: Results for  $T = 100$  and  $p = 2$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	24	<b>33</b>	43	<b>152.8973</b>	0.96608
AICc	24	<b>33</b>	43	152.8980	0.96611
BIC	25	<b>33</b>	42	152.8992	0.96622
HQ	43	21	36	556.5056	4.91710
MML87a	2	3	95	154.7048	0.92960
MML87b	57	27	16	158.9252	1.87498
MML87c	56	31	13	153.2429	<b>0.92899</b>
MML87d	39	37	24	153.3030	0.97832
MMLD	-	-	-	-	-

Table D.11: Results for  $T = 100$  and  $p = 3$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	41	16	43	<b>161.8468</b>	0.15335
AICc	42	16	42	161.8481	0.15335
BIC	44	16	40	161.8518	<b>0.15024</b>
HQ	54	<b>20</b>	26	166.7121	0.31841
MML87a	2	2	96	163.3839	0.28896
MML87b	74	15	11	169.9694	0.64444
MML87c	74	15	11	162.2691	0.15820
MML87d	65	14	21	162.3793	0.18633
MMLD	-	-	-	-	-

Table D.12: Results for  $T = 100$  and  $p = 4$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	6	48	46	<b>229.2366</b>	<b>0.00894</b>
AICc	6	48	46	<b>229.2366</b>	<b>0.00894</b>
BIC	7	47	46	229.2379	0.00883
HQ	19	35	46	234.2422	0.03095
MML87a	0	10	90	231.6111	0.02473
MML87b	14	<b>75</b>	11	234.7390	0.18634
MML87c	17	67	16	229.5720	0.00896
MML87d	19	55	26	229.6046	0.00984
MMLD	-	-	-	-	-

Table D.13: Results for  $T = 200$  and  $p = 1$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	14	32	54	<b>262.7395</b>	0.03021
AICc	14	32	54	<b>262.7395</b>	0.03021
BIC	17	31	52	262.7435	0.03042
HQ	38	23	39	267.4480	0.07540
MML87a	5	3	92	264.1398	0.03612
MML87b	45	46	9	279.4255	0.43104
MML87c	32	<b>54</b>	14	263.0718	0.02965
MML87d	30	45	25	263.5956	<b>0.02601</b>
MMLD	-	-	-	-	-

Table D.14: Results for  $T = 200$  and  $p = 2$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	23	29	48	<b>297.2035</b>	<b>0.11796</b>
AICc	23	29	48	<b>297.2035</b>	<b>0.11796</b>
BIC	24	28	48	297.2040	0.11815
HQ	46	17	37	303.6190	0.27367
MML87a	3	3	94	298.6619	0.12586
MML87b	58	31	11	316.4034	0.88872
MML87c	46	<b>39</b>	15	297.5178	0.11961
MML87d	36	35	29	297.6675	0.11803
MMLD	-	-	-	-	-

Table D.15: Results for  $T = 200$  and  $p = 3$ 

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Avg $-\log f(y \theta)$	MSPE
AIC	43	31	26	<b>297.2831</b>	0.06475
AICc	43	31	26	<b>297.2831</b>	0.06475
BIC	45	<b>32</b>	23	297.2844	<b>0.06391</b>
HQ	48	18	34	307.1683	0.24926
MML87a	8	7	85	298.4503	0.08274
MML87b	74	14	12	311.8965	0.41205
MML87c	70	22	8	297.6264	0.06544
MML87d	59	25	16	299.1179	0.08408
MMLD	-	-	-	-	-

Table D.16: Results for  $T = 200$  and  $p = 4$