MIXTURE AUTOREGRESSIVE MODELS: ASYMPTOTIC PROPERTIES AND APPLICATION TO FINANCIAL RISK

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2013

Mary Idowu Akinyemi
School of Mathematics

Contents

List of Tables

List of Figures

The University of Manchester

Mary Idowu Akinyemi
Doctor of Philosophy
Mixture Autoregressive Models: Asymptotic Properties and Application to Financial Risk
June 14, 2013

This thesis extensively studies the class of Mixture autoregressive (MAR) models in terms of its asymptotic properties and applications to financial risk evaluation.

We establish geometric ergodicity of the MAR models and by implication absolute regular and strong-mixing properties of the models. In addition, we also show the consistency and asymptotic normality of the maximum likelihood estimators of the MAR models.

We compare the estimates of Value at Risk (VaR) and Expected Shortfall (ES) based on the MAR models to estimates based on a number of other methods, for individual stocks, exchange rates and stock indices. We find that the MAR models consistently perform better than the other models. In addition, tail density forecast performance of individual stocks, stock indices and exchange rate, based on some popular GARCH models are compared to tail forecasts based on MAR models with both Gaussian and Student-t innovations. The MAR models mostly outperform the other models. Confirming the claim that MAR models are better suited to capture the kind of data dynamics present in financial data. All the data analysis are implemented in R.

The traditional residuals of the MAR model are computed as the difference between the observed values and their conditional means. We show that these residuals form a martingale difference sequence and that the unconditional variance of these residuals is strictly positive and bounded by the expected value of its conditional variance. We compare the class of MAR Models to the class of GARCH models and observed that both the GARCH type models and MAR models can be cast into the framework of random coefficient autoregressive models as well as generalized hidden markov models. We also show that for the MAR(2;1,1) model, the variance-covariance matrix is positive definite and the same for both the conditional least square and maximum likelihood penalty functions.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Head of the School of Mathematics.

Acknowledgements

I would like to thank the almighty God who gives all wisdom for seeing me through this work from start to finish.

I would like to thank my supervisor Dr. Georgi Boshnakov for his guidance, patience and the willingness to share his time and knowledge, his ideas and experience were indispensable to this thesis.

I would like to thank my husband for his prayerful support and understanding. I will also like to thank my two lovely little ladies (Miss Pearl and Miss Zoe-thanks for being my constant companions). I must not forget also my entire family for all the encouragement. I must also mention my friends and colleagues for their constant encouragement. I should also mention that my PhD was funded by the School of Mathematics, University of Manchester, for this I am most grateful.

Dedication

To the Lord most high, who holds all wisdom and bestows on all mortals that ask Him-Daniel 2:20–23,ESV Bible

Chapter 1

Introduction

Mixture Autoregressive models belong to the class of finite mixture models, this class of models have a number of interesting properties that make them viable models for several time series data in real life scenarios. Some of these properties include their ability to model both unimodal and multimodal conditional distribution as well as capture conditional heteroskedasticity. These properties have made the MAR models and its different variations popular in modelling non-linear time series. The application of this class of models can be found in finance, medicine, engineering among others.

This thesis focuses on the application of MAR model to finance, as the flexibility of the class of models has made them increasingly preferred candidates for capturing stylised properties of different financial time series.

1.1 Background

Time series models are useful for practical purposes only if they can be fitted to data and their parameters estimated. Computational procedures for determining parameters for the various model classes have been widely studied. Mixture distributions come into play when time series data is made up of undefined subgroups mixed in random proportions. Mixture distributions are made up of finite or infinite number of components that describe the different characteristics of the various subgroups of the data. In this work, we focus on finite mixture distributions with a finite number of components.

Finite Mixture Models and Financial Modelling Recent advances in financial modelling have ushered in a rapid expansion in modelling based on finite mixture distributions and Markov switching models.

The Markov switching models, also known as the regime switching models, involve multiple structures that are able to describe time series behaviors in different regimes. These class of models are able to capture more complex dynamic patterns by allowing switching between these structures, Another important feature of the Markov switching model is that the switching mechanism is controlled by an unobservable state variable that follows a first-order Markov chain. The Markovian property ensures that the current value of the state variable depends on its immediate past value, so that a structure might hold for a random period of time, after which it may be displaced by another structure when switching takes place (Hamilton (1989)), making it well suited for describing correlated data that exhibit distinct dynamic patterns during different time periods.

Finite mixture distributions arise in a natural way as marginal distribution for statistical models involving discrete latent variables such as clustering or latent class models. They are able to capture many specific properties of real data such as multi-modality, skewness, kurtosis, and unobserved heterogeneity. Their extension to Markov mixture models is able to deal with many features of practical time series, for example, spurious long-range dependence and conditional heteroscedasticity (Fruhirth-Schnatter (2006)).

Geometric Ergodicity, mixing conditions and Maximum likelihood (MLE) estimation The advent of faster and more sophisticated computational numerical methods have made the maximum likelihood method of parameter estimation a very popular approach to mixture density estimation problems.

A maximum likelihood estimate associated with a sample of observations is a choice of parameters that maximizes the probability density function of the sample, called in this context the likelihood function. MLE is of fundamental importance in the theory of inference and it forms the basis of many inferential techniques in statistics (Myung (2003)). There is vast literature on MLE and it's applications as well as it's properties available in literature among which is Wald (1949), Andersen (1970) and many others. When a model has a higher maximized loglikelihood value than other models, the model becomes more viable for further investigation than the other models.

Geometric ergodicity is very useful in establishing mixing conditions and central limit results for parameter estimates of a model, it also justifies the use of laws of large numbers and in essence form a basis for exploring asymptotic theory of the model. This further translates into examining the consistency and asymptotic normality of the parameter estimates of the model (Tjostheim (1990)). Meyn and Tweedie (1993), Stockis et al. (2010), Tweedie (1988),Bradley (2005) and many others give detailed discussions on geometric ergodicity and mixing conditions.

Risk Recent occurrences in the global financial industry and the substantial losses that companies and major financial houses have suffered in the past decade have made the concept of managing risk extremely vital to businesses, hence the popularity of Value at Risk (VaR) as a measure of risk.

A detailed review of Value at Risk (VaR) and its limitations as well as Expected Shortfall (ES) as a viable option had been looked into in literature (see Jorion (1997), Manganelli and Engle (2001), Tasche (2002),Tsay (1997) and many others).

Some of the arguments that make ES a more suitable alternative to VaR that can be found in literature are that VaR does not satisfy the subadditivity axiom which contradicts the framework of modern portfolio theory, that is, diversification should reduce risk (Jadhav et al. (2009)). It does not consider tail distribution beyond it's value, hence it disregards tail risk and permits the construction of proxies portfolios having low VaR resulting from a trade-off of heavy tail loss (Mamon and (Eds.) (2007)). In additon, Rational investors who wish to maximize expected utility can be misled by the information given by VaR (see Mamon and (Eds.) (2007) for details).

Some properties of ES that gives it an advantage over VaR include that ES is coherent, that is it is monotonous, sub-additive, positively homogeneous,and translative invariant. The coherence property is by far the most important property of ES. ES is such that $ES_\alpha(X) \geq VaR_\alpha(X)$ and ES_α is *law invariant.* Furthermore, for a real-valued random variable X with $E[X^-] < \infty$, $\alpha \to ES_\alpha$ is absolutely continuous on $(0, 1)$ and non-decreasing. This implies that ES_α is continuous with respect to α thus insensitive to the changes in the confidence level α, finally, ES_α is *comonotonic additive.*

Density Forecast Forecasts play a very significant role in economics and finance just as it does in any other science, hence, evaluating accurate/dependable forecasts is of primary concern. A large chunk of the existing forecast literature is focused on evaluating point forecasts then a smaller slice on interval forecasts and a much thiner slice on probability forecasts. Point forecast have been noted to be generally unsuitable for most models as forecasts based on quite a number of financial and economic models are not readily summarised by point forecasts (Berkowitz (2001)). Furthermore, a lot of important financial calculations are based on estimates which are not summarized by the point forecasts, examples of which are Value at Risk (VaR) and Expected shortfall (Jorion (1997)) as well as the Standard Portfolio Analysis of Risk (SPAN) system. Hence, density forecasts have received increasing attention over the past decade in both economics and finance related fields and most especially in the area of risk management (Diebold et al. (1998)).

This thesis, focuses on a class of finite mixture models introduced by Wong and Li (2000), the Mixture autoregressive model. This class of models and its extensions employ itself to many applications in various industries including finance (Saikkonen (2007)), neural networks (Martinetz et al. (1993)) and many others (Shao (2006)).

Boshnakov (2009) and Boshnakov (2011b) explore the predictive distributions of the MAR model as well as the conditions for first and second order stationarity of the model. Wong et al. (2009), Lanne and Saikkonen (2003), Ni and Yin (2009), Jin and Li (2006) study different extensions of the MAR model and apply them to financial modelling. These extensitons include the MAR model with ARCH innovations (MARCH model) (Wong and Li (2001)), Student t-mixture autoregressive (TMAR) model (Wong et al. (2009)), MAR model with GARCH innovations (MAR-GARCH model) (Lanne and Saikkonen (2003)), Self Organizing Mixture Autoregressive model (SOMAR) (Ni and Yin (2008)) and Mixture Autoregressive Panel (MARP) model (Jin and Li (2006)). We give details of these extensions and how they are applied in financial modelling.

1.2 Structure of Thesis

In this thesis, we study the traditional residuals of the MAR model and show that they form a martingale difference sequence and that the unconditional variance of the residuals is bounded by the expectation of the conditional variance. We compare the class of MAR models to the class of GARCH models, we observe that both the GARCH models and the MAR models can be cast into the framework of Random Coefficient Autoregressive (RCA) models as well as Hidden Markov Models (HMM). In addition, we show that the MAR model is geometrically ergodic and by implication satisfies the absolutely regular and strong mixing conditions. We then examine the asymptotic properties of the maximum-likelihood estimates of the model.

Furthermore, we propose the use of the MAR model in measuring VaR and ES and show that the class of MAR models perform comparably better than the other approaches in literature. Finally, we evaluate the tail density forecast of some financial time series based on the MAR model. The performance of the MAR model is compared to that of some popular GARCH models, we found that the MAR model better forecasts the tail density of the financial time series selected.

The rest of this thesis is structured as follows, In chapter 2 a detailed description of the Mixture Autoregressive model (MAR) is given, the properties of the traditional residuals of the model are explored, it is shown that these residuals form a martingale difference sequence and that the unconditional variance of the MAR model is strictly positive and bounded by the expected value of the conditional variance. The chapter then proceeds to outline some extensions of the MAR model and their applications. The chapter ends with a comparison of the class of MAR model to the class of GARCH model.

Chapter 3 extends the work done by Klimko and Nelson (1978) and Tjostheim (1986) on the consistency and asymptotic normality of parameter estimation based on conditional least squares and maximum likelihood estimator penalty functions. Expressions are given for the Variance-Covariance matrix of the MAR$(2; 1, 1)$ model and show that this Variance-covariance matrix is the same for both penalty functions.

The geometric ergodicity of the MAR model is proved in chapter 4 and as a consequence, the $\beta-$ mixing of the model is established. By implication, $\alpha-$ mixing is also established for the model. It is also shown that the model has a stationary distribution with finite second moments.

The asymptotic properties of the maximum likelihood estimator of the parameters of the MAR model are explored chapter 5. Consistency and asymptotic normality of the maximum likelihood estimator of the MAR model is proved, we leverage on the ideas in Douc et al. (2004) for the proofs.

In chapter 6 the concept of risk and risk management is examined, we discuss the various classes of risk measures viz; coherent, convex and spectral measures of risk and give some examples of risk measures that fall into these classes. The chapter then proceeds to discuss Value as Risk as the most popular risk measure in practise. The merits and demerits of Value at risk are mentioned. This is then followed by a detailed description of Expected shortfall as an alternative/complement to Value at Risk. Expected shortfall is defined in terms of some other risk measures found in literature. Detailed descriptions of the different methods for evaluating Value at Risk and Expected shortfall are given. The MAR model is then proposed as a viable underlying model for evaluating VaR and ES. The chapter ends with a description of backtesting methodology for assessing the performance of both VaR and ES evaluation methodologies. Chapter 7 applies a three component MAR model to computing VaR and ES and compares the results to some existing methods in literature. This is done by evaluating one-step ahead out of sample VaR and ES for daily returns of some financial time series and backtesting using the backtest methodology described in chapter 6.

Chapter 8 explores the tail forecast density of some financial time series based on the MAR model and apply the Berkowitz density test to check the fit of the MAR model to some financial time series data. The performance of the MAR model is compared to that of some popular GARCH models.

Finally, chapter 9 concludes with recommendations and opportunities for further research.

1.3 Contributions

The results of this thesis readily lends themselves to real life applications. We not only study the asymptotic properties of the class of MAR models, but also show examples of it application to risk management and find that the models do perform better than some popular models.

We are currently preparing Chapters 3, 4, 5, 6, 7 and 8 for publication. Chapter 8 has been submitted for the EURO-INFORMS 2013 Joint International Conference EURO XXVI.

Our contributions are:

- We establish that the MAR model is geometrically ergodic and by implication satisfies the absolutely regular ($\beta-$mixing) and strong (α-mixing) mixing conditions (see Chapter 4, Section 4.3).

- We show that the maximum likelihood estimators of the MAR model are both consistent and asymptotically normal (see Chapter 5).
- We show that the traditional residuals of the MAR model form a martingale difference sequence, a very useful property for establishing some asymptotic properties of the parameter estimates. We also show that the unconditional variance of these residuals strictly positive and bounded by the expected value of its conditional variance (see Chapter 2, Section 2.2.1).
- We apply the work done by Klimko and Nelson (1978) on an estimation procedure for stochastic processes to the Mixture Autoregressive model. We give an example for the MAR(2;1,1) model and show that for the model, the variance-covariance matrix is positive definite and identical for both the conditional least square and maximum likelihood penalty functions (see Chapter 3).
- We propose the use of the MAR model for evaluating VaR and ES. We show that the MAR models do perform comparatively better than the other approaches examined (see Chapter 7).
- We compare the tail density forecast of some financial time series based on the MAR models with Gaussian and student-t innovations to some popular GARCH models. We find that the MaR model better captures the distributional properties at the tails of financial time series (see Chapter 8).
- We compare the class of mixture autoregressive models to the class of GARCH models and observe that both the GARCH type models and MAR models

can be cast into the framework of both Random Coefficient Autoregressive (RCA) models and Generalized Hidden Markov (GHM) models (see Chapter 2, Section 2.4).

Chapter 2

Mixture Autoregressive Model

2.1 Finite Mixture models

Recent advances in financial modelling have ushered in a rapid expansion in modelling based on finite mixture distributions and Markov switching models. Some features of finite mixture distributions that render them useful in statistical modelling include:

1. Finite mixture distributions arise in a natural way as marginal distribution for statistical models involving discrete latent variables such as clustering or latent class models.

2. Statistical models which are based on finite mixture distributions capture many specific properties of real data such as multi-modality, skewness, kurtosis, and unobserved heterogeneity.

3. Their extension to Markov mixture models enables dealing with many features of practical time series, for example, spurious long-range dependence and conditional heteroscedasticity. (Fruhirth-Schnatter (2006)).

The focus here is on a class of finite mixture models introduced by Wong and Li (2000) the *Mixture Autoregressive Model*. An important property of the MAR model is that the shape of the conditional distribution of a forecast depends on the recent history of the process (Boshnakov (2009)). This property gives the MAR model the flexibility to model unimodal and multimodal time series. It also provides a suitable platform for capturing conditional heteroscedasticity, a property that occurs in most financial time series. In addition, the MAR model lends a flexible approach for capturing multiple regimes in financial data and hence, changes in volatility persistence. The residuals of the MAR model is computed as the difference between the observed values and their conditional means. These residuals are quite useful as they give information on how close the observed values are to the means of the corresponding predictive distribution. The properties of these residuals are examined and it is shown that the traditional residuals of the MAR model forms a martingale difference sequence, a very useful property for establishing some asymptotic properties of the parameter estimates. Furthermore, it is shown that the unconditional variance of the residuals is strictly positive and bounded by the expectation of the conditional variance.

Some extensions of the MAR model, as well as their application to financial modelling are also given. A comparison of the class of MAR models to the class of GARCH models is also discussed.

2.2 The Mixture Autoregressive Model

Definition 2.2.1. Mixture Autoregresssive (MAR) Model

A process $\{y_t\}$ is said to be a mixture autoregressive process if the conditional distribution function of y_t given past information is given by,

$$F_{t|t-1}(x) = \sum_{k=1}^{g} \pi_k F_k \left(\frac{x - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i}}{\sigma_k} \right), \tag{2.2.0.1}$$

where

1. $F_{t|t-1}(x) = F(y_t \mid \mathcal{F}_{t-1})$ is the conditional distribution of y_t given information up to and including time $t-1$;

2. $\mathcal{F}_t$ is the sigma field generated by the process $\{y_t\}$ up to and including time t;

3. g is a positive integer representing the number of components in the model;

4. $\pi_k > 0$, $k = 1, \ldots, g$, $\sum_{k=1}^{g} \pi_k = 1$, are probabilities and they define a discrete distribution π. π_k are referred to as mixing proportions and can be either time invariant or functions of observed variables (e.g. lagged observations);

5. $\sigma_k > 0$ is a scaling factor for the kth noise component;

6. $F_k(\cdot)$ is a (conditional) cumulative distribution function;

7. $\phi_{k,0}$ and $\phi_{k,i}$ $i = 1, \ldots, p_k$ are autoregressive coefficients and $\phi_{k,i} = 0$ for $i > p_k$;

8. p_k is the order of the kth autoregressive model and set $p = \max_{1 \le k \le g} p_k$;

9. We assume that the model is stationary (see Boshnakov (2009)).

Conditions for first and second order stationarity of the MAR model

For the MAR model represented as in Equation (2.2.0.8) below, let each $\epsilon_k(t)$ be jointly independent and also independent of past ys in the sense that for each t, the σ field generated by the set of random variables $\{\epsilon_k(t+n), n \geq 1, \leq k \leq g\}$ is independent of $\mathcal{F}_t$, furthermore, the choice of each component at any time t (i.e. z_t) does not depend on $\mathcal{F}_{t-1}$.

For $k = 1, \ldots, g$ define $\boldsymbol{A}_k$ by

$$\boldsymbol{A}_k = C[\phi_{k,1}, \ldots, \phi_{k,p}] \equiv \begin{pmatrix} \phi_{k,1} & \phi_{k,2} & \ldots & \phi_{k,p-1} & \phi_{k,p} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \tag{2.2.0.2}$$

. Denote the expected value of $\boldsymbol{A}_{z_t}$ by $\boldsymbol{A}$. Then,

$$A \equiv E(\boldsymbol{A}_{z_t}) = \sum_{k=1}^{g} \pi_k \boldsymbol{A}_k \tag{2.2.0.3}$$

So that the vector $\boldsymbol{Y}_t = (y_t, \ldots, y_{t-p})'$ is such that the vector process $\boldsymbol{Y}_t$ is a first order random coefficient autoregressive process, such that

$$\boldsymbol{Y}_t = \boldsymbol{c}_{z_t} + \boldsymbol{A}_{z_t} \boldsymbol{Y}_{t-1} + \boldsymbol{\epsilon}_{t,z_t} \tag{2.2.0.4}$$

where $\boldsymbol{\epsilon}_{t,z_t} = (\sigma_{z_t}\epsilon_{z_t}(t), 0, \ldots, 0)'$,

$$\boldsymbol{c}_{z_t} = \begin{pmatrix} \phi_{z_t,0} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \boldsymbol{c} = E\boldsymbol{c}_{z_t} = \begin{pmatrix} E\phi_{z_t,0} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{2.2.0.5}$$

The following results are due to Boshnakov (2009).

First order stationarity of the MAR process

Theorem 2.2.1. *The process $\{y_t\}$, $t = 1-p, \ldots, 0, 1, 2, \ldots$ is stationary in the mean if and only if $E\epsilon_{z_t}(t)$ exists and one of the following three cases holds:*

1. *$\boldsymbol{c} = \mathbf{0}$ and $\boldsymbol{\mu}_0 = \mathbf{0}$*

2. *$\boldsymbol{c} = \mathbf{0}$,$\mathbf{1}$ is an eigenvector of $\boldsymbol{A}$ associated with eigenvalue 1, and $\boldsymbol{\mu}_0 = \mu\mathbf{1}$ for some constant μ*

3. *$\boldsymbol{c} \neq \mathbf{0}$, 1 is not an eigenvalue of $\boldsymbol{A}$, and $\boldsymbol{\mu}_0 = \mu\mathbf{1}$ where $\mu = c/(1 - \sum \phi_i)$.*

where $\boldsymbol{\mu}_0$ is the mean of the initial vector $\boldsymbol{Y}_0$ and μ is a scalar constant and $\mathbf{1}$ is the identity matrix.

See Boshnakov (2009) for proof.

Second order stationarity of the MAR process

Theorem 2.2.2. *Let $\lambda(\boldsymbol{A} \otimes \boldsymbol{A} + E\{\boldsymbol{U}_{z_t} \otimes \boldsymbol{U}_{z_t}\}) < 1$ and $\Delta_1 \neq \mathbf{0}$. The process $\{y_t\}$, $t = 1-p, \ldots, 0, 1, 2, \ldots$ is second order stationary if and only if the initial vector $(y_0, y_{-1}, \ldots, y_{1-p})'$ has mean $\mu\mathbf{1}$, where μ is some scalar, covariance $C_{0,0}$ which is the solution of the equation*

$$C_{0,0} = \boldsymbol{A}C_{0,0}\boldsymbol{A}' + E\{\boldsymbol{U}_{z_t}C_{0,0}\boldsymbol{U}'_{z_t}\} + \Delta_1 \tag{2.2.0.6}$$

.

see Boshnakov (2009) for proof.

The MAR model is such that at each time t, one of g autoregressive-like equations is picked at random to generate y_t. The process y_t can be written in the following form (Boshnakov (2009)). Denote the past values of y_t as y_t' that is,

$$y_t' = (y_{t-1}, \ldots, y_{t-p}) \qquad (\text{where, } p = \max_{1\leq k\leq g} p_k). \tag{2.2.0.7}$$

$$\begin{aligned} y_t &= \mu_k(y_t') + \sigma_k\epsilon_k(t) \quad \text{and} \\ \mu_k(y_t') &= \phi_{k,0} + \sum_{i=1}^{p_k} \phi_{k,i}y_{t-i}. \end{aligned} \tag{2.2.0.8}$$

where each $\mu_k(y_t^{'})$ is an autoregressive model of order p_k and make up the components of the MAR model, $\epsilon_k(t)$ is the kth noise component, a strict white noise with distribution function $F_k(\cdot)$, corresponding density function and characteristic function f_k and φ_k respectively. We assume that the $\epsilon_k(t)$s are jointly independent and independent of past ys. Wong and Li (2001) show that $\sum_{k=1}^{g} \pi_k \mu_{k,t}^2 - (E(y_t \mid \mathcal{F}_{t-1})^2$ is non-negative and zero iff $\mu_{1,t} = \mu_{2,t} = \cdots = \mu_{g,t}$.

The model is referred to as the MAR$(k; p_1, \ldots, p_k)$ model i.e. a k-component MAR model, each component with corresponding order p_k. The conditional density of y_t given only the past values of y_t is,

$$f_\theta(y_t \mid y_t^{'}) = \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} f_k \left(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y(t-i)}{\sigma_k} \right), \tag{2.2.0.9}$$

where, $f_k(\cdot)$ and φ_k represent the conditional probability density function and the characteristic function whose distribution function is defined by $F_k(\cdot)$ for each $k = 1, \ldots, g$;

Let $\{z_t\}$ be an iid sequence of random variables with distribution π such that $Pr\{z_t = k\} = \pi_k, \quad k = 1, \ldots, g$, define a vector $Z_t = [Z_{t,1}, \ldots, Z_{t,g}]^{'}$ such that,

$$Z_{t,k} = \begin{cases} 1 & \text{if } z_t = k \\ 0 & \text{otherwise} \end{cases}$$

Then, the process y_t can be written as Boshnakov (2009),

$$y_t = \mu_{z_t}(y_t^{'}) + \sigma_{z_t} \epsilon_{z_t}(t) \tag{2.2.0.10}$$

where

$$\mu_{z_t}(y_t^{'}) = \phi_{z_t,0} + \sum_{i=1}^{p} (\phi_{z_t,i} y_{t-i}) \qquad (p = \max_{1 \le k \le g} p_k). \tag{2.2.0.11}$$

The conditional density of y_t given both the past values of y_t and z_t is,

$$f_\theta(y_t \mid y_t^{'}, z_t) = \frac{1}{\sigma_{z_t}} f_{z_t}\left(\frac{y_t - \phi_{z_t,0} - \sum_{i=1}^{p_{z_t}} \phi_{z_t,i} y(t-i)}{\sigma_{z_t}}\right), \tag{2.2.0.12}$$

$\{Z_t, t > 0\}$ is a simple case of a hidden Markov chain on a finite state space $S \in [0,1]$ with stationary k-step transition probability matrix. $\{Z_t, t > 0\}$ drives the dynamics of $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$, so that we can write a chain,

$$Q_t = (Z_t, Y_t), \tag{2.2.0.13}$$

where, Q_t is an aperiodic $S \times \mathbb{R}^p$-valued Markov chain.

Let A be a non -negative $g \times g$ matrix such that $A = (a_{ij})$ and $\sum_j a_{ij} = 1$. Let θ be the vector of all the free parameters of the model. We assume that θ belongs to a compact subset of R^d denoted by Θ.

The following assumptions are made on the chain Q_t.

Assumptions A

(i) The true parameter value which we represent by θ^0 lies in the interior of Θ.

(ii) For each $k \in \{1, \ldots, g\}$, $\{Z_{t,k} : t \geq 0\}$ is an irreducible, aperiodic Markov chain on a finite space S with probability distribution $\pi_1, \ldots, \pi_g$ and transition probability matrix $A = (a_{ij})$, so that $Z_{t,k}$ inherits the properties of $\{Z_t\}$.

(iii) The chain $\{Z_t\}$ is independent of the ϵ_t, also, for $\mathcal{F}_{t-1} = \sigma\{Y_r,\ r \leq t-1\}$ and all i, j,

$$P(z_t = j \mid z_{t-1} = i, \mathcal{F}_{t-1}) = P(z_t = j \mid z_{t-1} = i). \tag{2.2.0.14}$$

(iv) $\{\epsilon_t\}$ are jointly independent and are independent of past ys.

(v) $\{\epsilon_t\}$ has a probability density function that is continuous and positive everywhere.

(vi) $f_{z_t}(y)$ is non periodic and bounded on all compacts sets for all k and $z_t \in S$.

2.2.1 Residuals of the Mixture Autoregressive model

The linear predictor of the process y_t is given as,

$$\hat{y}_t = E(y_t \mid \mathcal{F}_{t-1}) = a_0 + \sum_{i=1}^{p} a_i y_{t-i}. \tag{2.2.1.1}$$

where

$$a_0 = \sum_{k=1}^{g} \pi_k \phi_{k,0} \qquad \text{and} \qquad a_i = \sum_{k=1}^{g} \pi_k \phi_{k-i}, \tag{2.2.1.2}$$

Wong and Chan give the conditional expectation of the MAR model defined in Equation (2.2.0.8) as,

$$E(y_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{g} \pi_k(\phi_{k,0} + \phi_{k,1} y_{t-1} \cdots + \phi_{k,p_k} y_{t-p_k}), \tag{2.2.1.3}$$

indicating that the conditional expectation of the process y_t is linear.

The traditional residuals of y_t can be obtained as the difference between y_t and $\hat{y}_t$, that is,

$$u_t = y_t - \hat{y}_t = y_t - a_0 - \sum_{i=1}^{p} a_i y_{t-i}. \tag{2.2.1.4}$$

$\{u_t\}$ is an uncorrelated but dependent sequence with the following properties:

1.

$$\begin{aligned} E(u_t \mid \mathcal{F}_{t-1}) &= E\left[y_t - \hat{y}_t \mid \mathcal{F}_{t-1}\right] \\ E(y_t \mid \mathcal{F}_{t-1}) - \hat{y}_t &= \hat{y}_t - \hat{y}_t = 0 \end{aligned} \tag{2.2.1.5}$$

2.

$$E(u_t) = E(E(u_t) \mid \mathcal{F}_{t-i}) = E(0) = 0 \tag{2.2.1.6}$$

3. Choose $s \leq v$ and $t > v$

$$E(u_t u_s \mid \mathcal{F}_v) = u_s E(u_t \mid \mathcal{F}_v) = 0 \tag{2.2.1.7}$$

Now let $t \geq s > v$

$$E(u_t u_s \mid \mathcal{F}_v) = E(E(u_t u_s \mid \mathcal{F}_s) \mid \mathcal{F}_v) = 0 \tag{2.2.1.8}$$

4.

$$\begin{aligned} E(u_t u_s) &= E[(y_t - \hat{y}_t)(y_s - \hat{y}_s)] && (2.2.1.9) \\ &= E(y_t y_s - y_t \hat{y}_s - y_s \hat{y}_t + \hat{y}_t \hat{y}_s) && (2.2.1.10) \\ &= E[E(u_t u_s \mid \mathcal{F}_v)] = E(0) = 0. && (2.2.1.11) \end{aligned}$$

5.

$$Var(u_t \mid \mathcal{F}_{t-1}) = E[(y_t - \hat{y}_t)^2 \mid \mathcal{F}_{t-1}]. \tag{2.2.1.12}$$

$$\begin{aligned} Var(u_t) &= E[u_t - (Eu_t)]^2 \\ &= E[E[u_t - (Eu_t)]^2 \mid \mathcal{F}_{t-1}] && (2.2.1.13) \\ &= E[Var(u_t) \mid \mathcal{F}_{t-1}]. \end{aligned}$$

We give a quick definition of a *martingale difference sequence*

Definition 2.2.2. Davidson (1997) A *Martingale Difference Sequence*(mds)$\{X\}_{-\infty}^{\infty}$ is an adapted sequence on $(\Omega, \mathcal{F}, P)$ satisfying the following properties:

1. $E|X| \leq \infty$,

2. $E(X \mid \mathcal{F}_{t-1}) = 0, \quad$ a.s., for all t

Propositon 2.2.1. The traditional residuals u_t, of the MAR model, given by equation 2.2.0.1 is a martingale difference sequence (MDS). Furthermore, the unconditional variance of u_t is strictly positive and bounded by the expectation of its conditional variance.

Proof. By Definition 2.2.2, and Equations (2.2.1.5) and (2.2.1.6) it follows that the traditional residuals u_t, of the MAR model is MDS.

We expand Equations (2.2.1.12) and (2.2.1.13) as,

$$Var(y_t \mid \mathcal{F}_{t-1}) = E(y_t^2 \mid \mathcal{F}_{t-1}) - (E(y_t \mid \mathcal{F}_{t-1}))^2 \tag{2.2.1.14}$$

$$= \sum_{k=1}^{g}[\pi_k \sigma_k^2] + \sum_{k=1}^{g} \pi_k[\phi_{k,0} + \sum_{i=1}^{p_i} \phi_{k,i} y_{t-i}]^2 - (E(y_t \mid \mathcal{F}_{t-1}))^2 \tag{2.2.1.15}$$

$$0 < \sum_{k=1}^{g} \pi_k \sigma_k^2 \leq Var(y_t \mid \mathcal{F}_{t-1}) \tag{2.2.1.16}$$

Thus,

$$E(\sum_{k=1} \pi_k \sigma_k^2) \leq Var(u_t) < \infty$$

and $Var(u_t) \leq E(Var(u_t \mid \mathcal{F}_{t-1}))$ and is strictly positive as required. $\square$

2.3 Extensions of the Mixture Autoregressive model and their applications in financial modeling

2.3.1 The MAR model and Financial modeling

The flexibility of the MAR models have made them preferred candidates for capturing stylized properties of different financial time series. Some of these applications are discussed in this section.

A. **Market Returns and Stock Index**

(a) Wong and Chan model the IBM stock daily closing price from May 17 1961 to November 2, 1962. They considered 2-component and 3-component MAR models for the return series they also computed the one-step and two-step ahead predictive distributions. They compared this models to other existing models in literature viz: the ARIMA model (see Box G. E. P. and C. (1994)), the SETAR model (see Tong (1990)) and the GMTD model (see Le et al. (1996)). They found that the MAR model out performed both the ARIMA and the SETAR model. They also found that although the empirical coverage of the prediction intervals for both the GMTD model and the MAR model are similar, the BIC values reveal the MAR model as the preferred model.

(b) Wong et al. (2009) extended the MAR model to include ARCH innovations, they call this class of models the MARCH model.

A time series y_t is said to follow a MARCH$(K, p_1, \ldots, p_K; q_1, \ldots, q_K)$ process if the cumulative distribution function of y_t given past information is given by;

$$F(y_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{K} \alpha_k G\left(\frac{e_{k,t}}{\sqrt{h_{k,t}}}\right), \tag{2.3.1.1}$$

where

$$\begin{aligned} e_{k,t} &= y_t - \mu_{k,t}, \qquad (2.3.1.2)\\ \mu_{k,t} &= \phi_{k,0} + \phi_{k,1}y_{t-1} + \cdots + \phi_{k,p_k}y_{t-p_k},\\ h_{k,t} &= \beta_{k,0} + \beta_{k,1}e^2_{k,t-1} + \cdots + \beta_{k,q_k}e^2_{k,t-q_k}. \end{aligned}$$

$\mathcal{F}_{t-1}$ is the information set up to time $t-1$.; $G(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\alpha_k > 0$ is such that $\alpha_1 + \cdots + \alpha_K = 1$ are the mixing proportions. The ex-ante conditional probability of a MARCH model is defined as:

$$\pi_{k,t} = E[Z_{k,t} \mid \mathcal{F}_{t-1}] \tag{2.3.1.3}$$

where $Z_{k,t}$ is the unobservable random vector such that $Z_{k,t} = 1$ if y_t comes from the kth component of the conditional distribution function and is zero otherwise.

Here the conditional mean of y_t follows an AR process while its conditional variance follows an ARCH process.

Wong and Chan apply this model to the monthly returns of the TSE 300 index with dividends reinvested from January 1956 to December 1999 as well as monthly S&P 500 total return series from 1956 to 1999. They compare the performance of the MARCH model to the ILN model, the 2-regime RSLN (RSLN2) model and the 2-point mixture of independent normal distributions (MIND2) model. These comparisons were done by comparing each of the model characteristics such as moments, percentiles, autocorrelations and squared autocorrelations to those of the observed data. They find that for both data sets, the MARCH model captures crucial characteristics of the observed data such as kurtosis (i.e. thickness of tails) and extreme observations (e.g the October 1987 crash). In addition, they found that the MARCH model is able to flexibly model volatility clustering in the data.

(c) In order to accommodate the the excess kurtosis exhibited by most financial time series (Wong et al. (2009)) introduced the Student t-mixture autoregressive (TMAR) model. The model consists of a mixture of g autoregressive components with Student-t distributed innovations.

The conditional cumulative distribution function of the TMAR $(g; p_1, \ldots, p_g)$ model is defined by the following conditional cumulative distribution function;

$$F(y_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{g} \alpha_k F_{v_k} \left(\frac{y_t - \phi_{k,0} - \phi_{k,1} y_{t-1} - \cdots - \phi_{k,p_k} y_{t-p_k}}{\sigma_k} \right) \tag{2.3.1.4}$$

$\mathcal{F}_{t-1}$ is the information set up to time t; $F_{v_k}(\cdot)$ is the commutative distribution function of the standardized student t-distribution with v_k degrees of freedom; $\alpha_1 + \cdots + \alpha_g = 1$; and $\alpha_k > 0$, for $k = 1, \ldots, g$.

The TMAR $(g; p_1, \ldots, p_g)$ model can also be expressed as follows;

$$y_t = \phi_{k,0} + \phi_{k,1} y_{t-1} + \cdots + \phi_{k,p_k} y_{t-p_k} + \sigma_k \varepsilon_t \tag{2.3.1.5}$$

with probability α_k for $k = 1, \ldots, g$. ε_t are iid standardised student t-distributions with degrees of freedom v_k.

The TMAR model is reduced to the MAR model when the degrees of freedom of all the components of the of the TMAR model tend to infinity hence the MAR model can be refereed to as a limiting case of the TMAR model.

The TMAR model is flexible in modelling the tails of the conditional distribution as well as capture leptokutosis in financial data.

Wong et al. (2009) apply the TMAR model to the daily returns series of the Hong Kong Hang Seng Index over the January 2, 1996 to December 30, 2005 period. They compare the performance of the TMAR model to the AR, MA, RW, AR-GARCH,GARCH-t,AR-EGARCH,AR-TGARCH. The comparison is based on their empirical coverage of the in-sample one-step ahead prediction intervals as well as the empirical coverage of the one sided lower prediction intervals of the data.

They find that the the TMAR model generally out performed all the models considered, as the empirical coverage based on the TMAR model is closer to the nominal coverage than the all the other models while for the lower prediction interval the TMAR and the GARCH-t models are similar.

B. **Interest Rates**

i. Wong et al. (2009) model 3-year, 5-year and 10-year interest rate swap spread series in Australia using the MARCH model defined in Equation (2.3.1.1). They find that the MARCH model is able to capture volatility persistence and dependence of volatility on the level of data as well as allow for regime switches in the swap spreads.

ii. Lanne and Saikkonen (2003) extend the MAR model to have GARCH innovations, they call it the MAR-GARCH model.

The model is defined as follows;

$$y_t = \sum_{i=1}^{m}(v_i + b_{i,t}y_{t-1} + \cdots + b_{i,p}y_{t-p} + \sigma_{i,t}\varepsilon_t)I(c+\eta_t \leq y_{t-d} < c+\eta_t) \tag{2.3.1.6}$$

with conditional density

$$f_{t,t-1}(y_t) = \sum_{i=1}^{m}\frac{1}{\sigma_{i,t}}\phi((y_t - v_i - b_{i,t}y_{t-1} - \cdots - b_{i,p}y_{t-p})/\sigma_{i,t})\pi_{i,t-d} \tag{2.3.1.7}$$

where $\sigma_{i,t}$ is obtained from the GARCH (r,q) process viz;

$$\sigma_{i,t}^2 = \sigma_i^2 + \beta_{i,1}\sigma_{i,t-1}^2 + \cdots + \beta_{i,r}\sigma_{i,t-r}^2 + \alpha_{i,1}u_{i,t-1}^2 + \cdots + u_{i,q}^2 \tag{2.3.1.8}$$

and $$u_{i,t} = y_t - v_i - b_{i,1}y_{t-1} - \cdots - b_{i,p}y_{t-p}. \tag{2.3.1.9}$$

Here, the assumed GARCH structure implies conditional heteroskedasticity even when the conditional variance is nearly constant i.e. when one of the mixing proportions is close to 1 and the others close to zero. Notice, that the parameters in Equation (2.3.1.8) is dependent on the index i indicating that conditional heteroskedasticity in this model is generally regime dependent making it a viable candidate for capturing multiple regimes in financial time series.

Lanne and Saikkonen (2003) apply this model to short term interest rate and Bond pricing. In particular, they apply it to weekly data on the US 3-month treasury bill. They also apply the model to estimating bond prices maturing at 1,2 and 3years; They find that the realizations generated from the MAR-GARCH model is stable and that the properties are similar to those of the observed series. Furthermore, the forecasting performance of the model was found to be better than those existing in literature. In addition, the implied drift and diffusion functions given by the MAR-GARCH model is consistent. They obtain an estimate of market price of interest rate risk by complementing the model for the short term interest rate with with a model for a longer term interest rate and also computed bond prices by applying the risk-neutral valuation principle (see Hull (2000)) to market price of interest rate computed. They find that the model is able to produce term structure patterns that agree with those historically observed. They noted that the method can also be applied to pricing derivative securities.

C. **Forex Rate** Ni and Yin (2008) introduced the Self Organizing Mixture Autoregressive model (SOMAR) this model is an improvement on the Self Organizing Autoregressive model (SOAR) of Lampinen and Oja (1989). The SOMAR model extends the SOM based local regression model (Strickert and Hammer (2005)) and the SOAR model.The model combines topological clustering and linear regression to provide better temporal modeling capability of non-stationary time series. They observe that this joint estimation effectively reduces computational cost.

The model consists of a number of topologically ordered mixture of local regressive models.

The SOMAR model measures the competence of a local model by the autocorrelation of the error instead of the error itself, it expects that for a model following a correct path, the modeling error should be gradually close to white noise.The more white an error series is the smaller the sum of it's autocorrelation coefficient and vise versa. This autocorrelation based similarity measure makes the network more effective and robust in identifying correct local models given input segments compared to the error based measures.

Given a consecutive set of p modeling errors, $\{e(1), \cdots, e(p)\}$, with mean μ and variance σ^2, $p > 1$. The winning local model is said to be the model that generates the smallest sum of autocorrelation coefficients (SAC) of the modeling errors,

$$R_i(k) = \frac{1}{(p-k)\sigma^2} \sum_{t=1}^{p-k} (e_i(t) - \mu_i)(e_i(t+k) - \mu_i) \tag{2.3.1.10}$$

they define the winning local model as,

$$v = \operatorname*{argmin}_i \Big(\sum_{j=-k}^{k} |R_i(k)| \Big) \qquad i = 1, \ldots, N \tag{2.3.1.11}$$

k is the number of lags and i is the index of a local regressive model and N is the number of local models.

The winning local AR model and its neighboring models update their model parameters by the ordinary recursive least-mean square method according to the updating rule:

$$w_i(t) = w_i(t-1) + h(\lambda, k_i)\eta(t)e_i(t)x(t) \tag{2.3.1.12}$$

where, $e_i(t) = x(t) - x(t)^T w_i$, $h(\lambda, k_i)$ is the neighborhood function and $\eta(t)$ is the adaptation strength.

The Neural-gas (NG) algorithm (Martinetz et al. (1993)) is then used to find the optimal representation.

Ni and Yin (2009) apply the SOMAR model to Forex data, they consider 15years of daily exchange rates from the PACIFIC Exchange Rate Service provided by W.Antwiler of the UBC's Saunder school of Business. They examine both the predicted return and the predicted rate for the data. They compare the results of the SOMAR network for the predicted FX rate and the predicted FX price to that of the Vector Self Organizing Map (VSOM), Self organizing Autoregressive (SOAR) network , Recurrent Self Organizing Map (RSOM), Recursive Self Organizing Map (RecSOM), Neural Gas (NG), Self Organizing Map with Support Vector Machine Regression (SOM-SVM), GARCH model and ARIMA model. They find that the SOMAR more efficiently accommodates the non-stationarity of the FX prices, they also find that the SOMAR performs better that other SOM based methods in modeling and predicting non-stationary FX rates.

D. **Panel Time Series** Jin and Li (2006) introduce the Mixture Autoregressive Panel (MARP) model. The model enlarges the stationarity region of the traditional AR model and is able to capture multimodality in some panel data sets.

The conditional cumulative distribution function for a $N-$component finite mixture autoregressive model for a panel time series X_{jt}, $j = 1, \dots M$, $t = 1, \dots, T_j$ is given by:

$$\begin{aligned} F(X_{jt} \mid \mathcal{F}_{t-1}) = \sum_{i=1}^{N-1} \pi_i \Phi\big(\frac{X_{jt} - \phi_{0ij} - \sum_{k=1}^{p} \phi_{kij} X_{j,t-k}}{\sigma_{ij}}\big) \\ + \pi_N \Phi\big(\frac{X_{jt} - \phi_{0N} - \sum_{k=1}^{p} \phi_{kN} X_{j,t-k}}{\sigma_{Nj}}\big) \end{aligned} \tag{2.3.1.13}$$

where $i = 1, \dots N$ is the index of the components and $j = 1, \dots, N$ is the index of the series and $t = 1, \dots, T_j$ is the time index. k is the index of the of the lags and p is the autoregressive order. Jin and Li (2006) assumed that for the model 2.3.1.13 the order of each component in each series is the same. If the orders $p^{'}_{ij}$ are not the same, set $p = \max_{ij}\{p^{'}_{ij}\}$ and those $\phi_{kij} = 0$ when $k > p^{'}_{ij}$.

The random variable X_{jt} is evaluated at x_{jt} given past information up to time $t-1$ and is taken from the ith component with probability π_i. Where $\sum_{i=1}^{N}\pi_i = 1$, $\pi_N = 1 - \pi_1 - \ldots \pi_{N-1}$ and $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_{N-1}$. $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. It is assumed that the error terms $\epsilon_{ijt} \sim N(0, \sigma^2_{ij})$ and mutually independent.

Note that although it is assumed that the order p are the same, the length T_j are not assumed to be the same.

Jin and Li (2006) assumed that the last component in each of the series have the same coefficient ϕ_{kN}, $k = 0, \ldots, p$. The noise of each of the components are assumed to be Gaussian, however, the composition noise of each series by finite mixture are non-Gaussian and hence give rise to more flexible modelling.

Write

$$\begin{aligned}
\boldsymbol{\theta}_{ij} &= (\phi_{0ij}, \ldots, \phi_{pij}, \sigma^2_{ij})^T, \\
\boldsymbol{\theta}_{j} &= (\boldsymbol{\theta}^T_{1j} \ldots \boldsymbol{\theta}^T_{N-1,j})^T, \\
\boldsymbol{\theta}_{NL} &= (\phi_{0N}, \ldots, \phi_{pN}, \sigma^2_{N1}, \ldots, \sigma^2_{NM})^T \text{ and} \\
\boldsymbol{\theta} &= (\boldsymbol{\theta}^T_{1} \ldots \boldsymbol{\theta}^T_{M}, \boldsymbol{\theta}^T_{NL})^T, \text{(the subscript } L \text{ represents "Last" component)}
\end{aligned}$$

Now for the vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$, $\boldsymbol{Z}_{jt} = (Z_{1jt}, Z_{2jt}, \ldots, Z_{Njt})^T$, $\boldsymbol{Z}_j = (Z^T_{1j}, Z^T_{2j}, \ldots, Z^T_{NT_j})^T$ and $\boldsymbol{Z} = (Z^T_1, Z^T_2, \ldots, Z^T_M)^T$. The vector $\boldsymbol{Z}_{jt}$ contains the unobservable random variable Z_{ijt} where $Z_{ijt} = 1$ when at time t, in the jth series, the X_{jt} comes from the ith component of the conditional distribution function and $Z_{ijt} = 0$ otherwise. So that the parameters are naturally divided into two groups. Model 2.3.1.13 can thus be rewritten as,

$$X_{jt} = \sum_{i=1}^{N} Z_{ijt}(\phi_{0ij} + \sum_{k=1}^{p} \phi_{kij} X_{j,t-k} + \varepsilon_{ijt}) + Z_{Njt}(\phi_{0N} + \sum_{k=1}^{p} \phi_{kN} X_{j,t-k} + \varepsilon_{N_j t}) \tag{2.3.1.14}$$

where ϵ_{ijt} is the white noise process corresponding to the component of the jth series. The $\{\epsilon_{ijt}\}$ are also independent for $i = 1, \ldots, N$, $j = 1, \ldots, M$ and all t.

Representing the MARP in the form of Equation (2.3.1.14) drives home the fact that the MARP model is actually a mixture of N Gaussian AR models.

The conditional mean of the jth series is given as:

$$E(X_{jt} \mid \mathcal{F}_{t-1}) = \sum_{i=1}^{N-1} \pi_i(\phi_{0ij} + \sum_{k=1}^{p} \phi_{kij} x_{j,t-k}) + \pi_N(\phi_{0N} + \sum_{k=1}^{p} \phi_{kN} x_{j,t-k}) \tag{2.3.1.15}$$

Since the conditional mean depends on past information of the time series, the shape of the conditional distributions will change from time to time and can be uni-modal or multi-modal.

In the same literature, Jin and Li (2006) relax the assumption that the parameters π_i are common to every series. They give an illustration with a two-component and order one model. So that Model 2.3.1.13 is modified as:

$$\begin{aligned} F(X_{jt} \mid \mathcal{F}_{t-1}) = {} & \pi_{1j}\Phi\big(\frac{X_{jt} - \phi_{01j} - \phi_{11j} X_{j,t-1}}{\sigma_{1j}}\big) \\ & + \pi_{2j}\Phi\big(\frac{X_{jt} - \phi_{02} - \phi_{12} X_{j,t-1}}{\sigma_{2j}}\big) \end{aligned} \tag{2.3.1.16}$$

where $\pi_{1j} + \pi_{2j} = 1$ for each j, $\qquad j = 1, \ldots, M$.

They apply this model to the gray-sided voles data (Hsiao (1986)) and (Hjellvik and Tjostheim (1999)). They compare the MARP model to the method applied by Hjellvik and Tjostheim (1999). The BIC value shows that the MARP model gives a better fit for the data.

2.4 Mixture Autoregressive Model versus the Generalized Autoregressive Hetetoskedastic (GARCH) Model

Here the the class of Mixture Autoregressive Models (MAR) is compared to the class of Generalised Autoregressive (GARCH) models. The similarities of this class of models to the GARCH models are examined.

2.4.1 The Generalised Autoregressive (GARCH) Model

The motivation for models that put more weight on recent information can be traced to the following reasons:

1. Research has shown that although longer periods increase the precision of volatility estimate, they could miss underlying variation in volatility.

2. Traditional estimates of VaR are based on the assumption that the standard deviation in returns does not change over time (i.e. they are homoskedatic), however, Manganelli and Engle (2001) argue that better estimates can be achieved by adopting models that explicitly allow the standard deviation to change over time (i.e. heteroskedasticity).

The Generalized Autoregressive Heteroskedastic (GARCH) model which is part of the family of Autoregressive Heteroskedastic (ARCH) models introduced by

Engle (1982)) and (Bollerslev (1986) is one of such models. These models have so far successfully lent themselves to financial data (see Bollerslev et al. (1986) for a survey). Bollerslev (1986) in his paper introduced a more general extension of the ARCH model of Engle (1982), he called this generalization the GARCH (Generalized Autoregressive Conditional Heteroskedastic) processes. These class of models allow for a much more flexible lag structure. The extension of the ARCH process to the GARCH process bears much resemblance to the extension of the standard time series AR process to the general ARMA process (Bollerslev (1986)).

The GARCH(p,q) model is defined as follows,

Definition 2.4.1. A process is called a GARCH(p,q) process if its first 2 conditional moments exist and satisfy:

(i) The first moment is given by: $E(y_t \mid y_s, s < t) = 0, \qquad t \in \mathbb{Z}$

(ii) There exists constants $\omega, \alpha_i, i = 1, \ldots, q$ and $\beta_j, j = 1, \ldots, p$ such that

$$\sigma_t^2 = Var(y_t \mid y_s, s < t) = \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{i=1}^{p} \beta_j \sigma_{t-j}^2, \qquad t \in \mathbb{Z} \quad (2.4.1.1)$$

Francq and Zakoian (2010) refers to this class of GARCH processes as semi-strong processes. They also introduce the notion of "Strong GARCH(p,q)" process defined as follows,

Definition 2.4.2 (Francq and Zakoian (2010))**.** Let (η) be an iid sequence with distribution η. The process (y_t) is called a strong GARCH(p,q) (with respect to the sequence (η_t)) if

$$\begin{cases} y_t = \sigma_t \eta_t \\ \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{j=1}^{2} \beta_j \sigma_{t-j}^2 \end{cases} \quad (2.4.1.2)$$

where α_i and β_j are nonnegative constants and ω is a (strictly) positive constant.

$\{\eta\}$ is a sequence of iid random variables with mean 0 and unit variance and can be assumed to be standard normal or standardized Student -t distribution or generalized error distribution (Tsay (1997)). $\alpha_i \geq 0$, $\beta_j \geq 0$ are referred to as the ARCH and GARCH parameters and $\sum_{i=1}^{\max(m,s)}(\alpha_i + \beta_i) < 1$, such that $\alpha_i = 0$ for $i > m$ and $\beta_j = 0$ for $j > s$.

The constraint imposed on $\alpha_i + \beta_i$ implies that the unconditional variance of y_t is finite, while it's conditional variance σ_t^2 evolves over time.
Equation (2.4.1.2)above can be further represented as follows, by letting $\epsilon_t = y_t^2 - \sigma_t^2$ so that $\sigma_t^2 = a_t^2 - \epsilon_t$ and then putting $\sigma_{t-i}^2 = a_{t-i}^2 - \epsilon_{t-i}$ $(i = 0, \ldots, s)$ into Equation (2.4.1.2) and hence we rewrite the GARCH model as follows,

$$a_t = \alpha_0 + \sum_{i=1}^{\max m,s} (\alpha_i + \beta_i) a_{t-i}^2 + \eta_t - \sum_{j=1}^{s} \beta_j \eta_{t-j} \tag{2.4.1.3}$$

Now substituting y_{t-i} by $\sigma_{t-i}\eta_{t-i}$ in 2.4.1.1 gives:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2 \eta_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{2.4.1.4}$$

which can be written as

$$\sigma_t^2 = \omega + \sum_{i=1}^{r} a_i(\eta_{t-i}) \sigma_{t-i}^2 \tag{2.4.1.5}$$

where $a_i(z) = \alpha_i z^2 + \beta_i, \qquad i = 1, \ldots, r$. Francq and Zakoian (2010) show that this representation shows that the volatility process of a strong GARCH process is the solution of an autoregressive equation with random coefficients. We discuss below the the IGARCH model a class of GARCH models that is popular in financial modelling especially in financial risk management.

Integrated GARCH Models

A GARCH(p,q) process is called an IGARCH(p,q) process when,

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j = 1. \tag{2.4.1.6}$$

The integrated GARCH process are originally developed to cater for data that exhibit persistent changes in volatility. They can be either non stationary or stationary with infinite variance hence making the class of models well suited to heavy tailed data. The class of IGARCH models are sometimes referred to as the unit root GARCH models (S.Tsay (1997)). Another key characteristic of the IGARCH models is that the impact of past squared shocks $\eta_{t-i} = a_{t-i}^2 - \sigma_{t-i}^2$ for $i > 0$ on a_t^2 is persistent.

An IGARCH$(1, 1)$ model can be written as follows:

$$a_t = \sigma_t \epsilon_t, \;\; \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + (1 - \beta_1) a_{t-1}^2, \tag{2.4.1.7}$$

where $1 > \beta_1 > 0$ and $\{\epsilon\}$ is a sequence of iid random variables with mean 0 and unit variance and can be assumed to be either standard normal, standardized student-t distribution or generalized error distribution . The volatility process σ_t^2 is martingale and under certain conditions it is strictly stationary, but not weakly stationary (Francq and Zakoian (2010) and Tsay (1997)).

A case of specific interest in the study of the IGARCH $(1, 1)$ model is that of $\alpha_0 = 0$. Here the volatility forecasts denoted by $\sigma_h^2(\ell)$ for all forecast horizons, given as: $(\sigma_h^2(\ell) = \sigma_h^2(1) + (\ell - 1)\alpha_0, \; \ell \geq 1)$. This special IGARCH $(1, 1)$ model is the volatility model used in the RiskMetrics method for calculating VaR, it is also an *exponential smoothing* model for $\{a_t^2\}$. Rewrite Equation (2.4.1.7) for $\alpha = 0$ as follows:

$$\begin{aligned}\sigma_t^2 &= (1-\beta_1)a_{t-1}^2 + \beta_1\sigma_{t-1}^2 \quad (2.4.1.8)\\ &= (1-\beta_1)a_{t-1}^2 + \beta_1[(1-\beta_1)a_{t-2}^2 + \beta_1\sigma_{t-2}^2]\\ &= (1-\beta_1)a_{t-1}^2 + (1-\beta_1)\beta_1 a_{t-2}^2 + \beta_1^2\sigma_{t-2}^2\end{aligned}$$

and by repeated substitution for $\sigma_{t-2}^2, \sigma_{t-3}^2, \ldots$ we have,

$$\sigma_t^2 = (1-\beta_1)(a_{t-1}^2 + \beta_1 a_{t-2}^2 + \beta_1^2 a_{t-3}^3 + \cdot), \quad (2.4.1.9)$$

This is known as the *exponential smoothing formation* with β_1 being the discounting factor. Therefore, exponential smoothing methods can be used to estimate an IGARCH$(1, 1)$ model represented by Equation (2.4.1.9).

2.4.2 Comparison of the MAR and GARCH model based on 1st and 2nd Conditional moments

Consider the MAR model given in Equation (2.2.0.1) above, y_t has the following conditional 1st and 2nd moments,

$$E(y_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{g} \pi_k \mu_{k,t}(y') \quad (2.4.2.1)$$

$$\begin{aligned}Var(y_t \mid \mathcal{F}_{t-1}) &= E(y_t^2 \mid \mathcal{F}_{t-1}) - [E(y_t \mid \mathcal{F}_{t-1})]^2 \quad (2.4.2.2)\\ &= \sum_{k=1}^{g}\pi_k\sigma_k^2 + \sum_{k=1}^{g}\pi_k\mu_{t,k}^2(y') - \left[\sum_{k=1}^{g}\pi_k\mu_{t,k}(y')\right]^2 \quad (2.4.2.3)\end{aligned}$$

The 1st and 2nd conditional moments for the the GARCH(p,q) model defined in Equation (2.4.1.1) is as follows,

$$E(y_t \mid \mathcal{F}_{t-1}) = 0 \quad (2.4.2.4)$$

$$Var(y_t \mid \mathcal{F}_{t-1}) \equiv \sigma_t^2 = \omega + \sum_{j=1}^{q}\alpha_j y_{t-j}^2 + \sum_{j=1}^{p}\beta_j\sigma_{t-j}^2 \quad (2.4.2.5)$$

Notice that the the 1st conditional moment of the MAR model depends on past values of the time series, thus it is able to capture changes in the shape of the conditional distributions of the of the series over time, making it a more interesting alternative for modeling financial time series than the GARCH class of models as the conditional distribution of financial time series can be changed from short-tailed to fat-tailed, from symmetric to asymmetric, unimodal to bimodal or even multimodal in some cases.

Furthermore, a closer examination show that that the the conditional variances in both models have the same linear structure and are both dependent on squared past values of the time series. However, the conditional variance of the MAR model depends on the conditional mean and is able to capture changes in conditional variance. The last 2 terms of Equation (2.4.2.1) (conditional variance equation), that is, $\sum_{k=1}^{g} \pi_k \mu_{t,k}^2(y^{'}) - [\sum_{k=1}^{g} \pi_k \mu_{t,k}(y^{'})]^2$ is non-negative and zero only if $\mu_{t,1}(y^{'}) = \mu_{t,2}(y^{'}) = \cdots = \mu_{t,g}(y^{'})$. The conditional variance is large when the $\mu_{t,k}(y^{'})s$ are very different and smallest when they are all the same, this is refereed to in literature as the smallest possible conditional variance, and the baseline conditional variance or volatility is given as $\sum_{k=1}^{g} \pi_K \sigma_k^2$ (see Francq and Roussignol (1998)).

It is also worth noting that if we set the RHS of Equation (2.4.2.1) to zero i.e. $\sum_{k=1}^{g} \pi_k \mu_{k,t}(y^{'}) = 0$ we have a special case of the semi-strong GARCH processes.

2.4.3 MAR and GARCH models as forms of Vector RCA class models

Both the GARCH models and MAR models can be cast into the framework of Random coefficient Autoregressive models.

The Random Coefficient Autoregressive (RCA) model is defined as,

Definition 2.4.3 (The Random coefficient Autoregressive (RCA) Model)**.** The process $\{X_t\}$ is generated by a random coefficient autoregressive model if,

$$X_t = (\theta + \boldsymbol{A}_t)X_{t-1} + \varepsilon_t \tag{2.4.3.1}$$

where θ is a $p \times p$ non-random matrix, $\{\boldsymbol{A}_t\}$ is a random sequence, ε_t is sequence of iid random variables.

The following assumptions are made on Model 2.4.3.1,

1. The sequence $\{\boldsymbol{A}_t\}$ and ε_t are iid and also independent of each other
2. ε_t is sequence of iid random variables. with zero mean and common positive definite covariance matrix.

Quinn and Nicholls (1981) gives conditions for the stability of this class of model and (Feigin and Tweedie (1985)) gives conditions for stationarity and finiteness of moments.

2.4.4 The MAR model as an RCA model

Consider the Equation (2.2.0.10) representation of the MAR model. For $k = 1, \ldots, g$, define $\boldsymbol{A}_k$ by the column matrix,

$$\boldsymbol{A}_k = C[\phi_{k,1}, \ldots, \phi_{k,p}] \tag{2.4.4.1}$$

so that

$$\boldsymbol{A} = E(\boldsymbol{A}_{z_t}) = \sum_{k=1}^{g} \pi_k \boldsymbol{A}_k \tag{2.4.4.2}$$

Hence, for $t \geq 0$, let $\boldsymbol{Y}_t = (y_t, \ldots, y_{t+1-p})'$ be a vector of p values of the time series $\{y_t\}$. Then the vector process $\{\boldsymbol{Y}_t\}$ is a first order random coefficient autoregressive process defined as (Boshnakov (2009)),

$$\boldsymbol{Y}_t = \boldsymbol{c}_{z_t} + \boldsymbol{A}_{z_t}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_{t,z_t} \tag{2.4.4.3}$$

where

$$\begin{aligned} \boldsymbol{\varepsilon}_{t,z_t} &= (\sigma_{z_t}\varepsilon_t(t), 0, \ldots, 0)^T, \\ \boldsymbol{c}_{z_t} &= (\phi_{z_t,0}, 0, \ldots, 0)^T, \\ \boldsymbol{c} &= (E(\phi_{z_t,0}), 0, \ldots, 0)^T = (c, 0, \ldots, 0)^T. \end{aligned} \tag{2.4.4.4}$$

Boshnakov (2009) gives conditions for first and second order stationarity of the model.

2.4.5 The GARCH model as an RCA model

Now consider the augmented GARCH(1,1) process introduced by Duan (1997) defined as follows,

$$\begin{aligned} \varepsilon_t &= \sigma_t\eta_t, \quad t = 0, 1, \ldots, \\ \boldsymbol{\sigma}_t^2 &= \boldsymbol{c}_{e_t}\boldsymbol{\sigma}_{t-1}^2 + \boldsymbol{f}_{e_t} \end{aligned} \tag{2.4.5.1}$$

where the process $\{\boldsymbol{\sigma}_t^2\}$ is a real valued process. This process is synonymous to the strong GARCH process defined in Equation (2.4.2) above. The sequence e_t is some measurable function of η_t and is such that the sequence $\{\eta_t\}$ is a sequence of iid real valued sequence random variables with mean 0 and variance 1 and is independent of σ_0^2. $\boldsymbol{c}_{e_t}$ is an $m \times m$ matrix valued polynomial function and $\boldsymbol{f}_{e_t}$ is an $m \times 1$ vector valued polynomial function. The process $\{\boldsymbol{\sigma}_t^2\}$ is a generalized polynomial random coefficient vector autoregressive model (Carrasco and Chen (2002)). Duan (1997) gives sufficient conditions for the strict stationarity of the model.

2.4.6 MAR and GARCH models as forms of Generalized Hidden Markov models

Both the MAR and GARCH models can be given Makovian structure and hence viewed as generalized hidden Markov models (GHMM). The Generalized hidden Markov process is defined as,

Definition 2.4.4 (Generalized Hidden Markov Model)**.** [Carrasco and Chen (2002)] A process $\{Y_t, t \geq 0\}$ with state space $(S, \mathcal{B}(S))$ follows a generalized hidden Markov model with a hidden chain $\{X_t, t \geq 0\}$ if the following hold true:

1. $\{X_t, t \geq 0\}$ is a strictly unobserved strictly stationary Markov chain with state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.
2. For all $t \geq 1$, the conditional distribution of Y_t given $(X_t, Y_{t-1}, X_{t-1}, \ldots, Y_0, X_0)$ depends only on X_t.

2.4.7 The MAR model as a Generalized Hidden Markov model

Consider the MAR process as defined in Equation (2.4.4.3) assume that $\{\epsilon_t\}$ are jointly independent and are independent of past ys and has a probability density function that is continuous and positive everywhere. Furthermore, let $\{Z_{t,k} : t \geq 0\}$ with $k = \{1, \ldots, g\}$ be an irreducible, aperiodic Markov chain on a finite space S with probability distribution $\pi_1, \ldots, \pi_g$ and transition probability matrix $A =$

$(a_{ij}) \in S$. So that $Z_{t,k}$ inherits the properties of $\{Z_t\}$. Assume further that the chain (Z_1) is independent of the noise term ϵ_t and for $\mathcal{F}_{t-1} = \sigma\{Y_r,\ r \leq t-1\}$,

$$P(z_t = j \mid z_{t-1} = i, \mathcal{F}_{t-1}) = P(z_t = j \mid z_{t-1} = i) \qquad \forall i, j. \tag{2.4.7.1}$$

The process defined by:

$$Y_t = (y_t, \ldots, y_{t-p+1})' \text{ and } \quad Q_t = (Z_t, Y_t) \tag{2.4.7.2}$$

is a generalised hidden Markov model with hidden chain $\{Z_t\}$.

2.4.8 The GARCH model as a Generalized Hidden Markov model

The augmented GARCH model defined in Equation (2.4.5.1), η_t has a positive continuous density with respect to the lebesgue measure on a real line, hence $\{\varepsilon_t\}$ is a generalized hidden markov model with hidden chain $\{\sigma_t^2\}$.

2.4.9 Comparison Based on Persistence

Persistence in variance of a random variable evolving through time and volatility clustering have a lot to do with the dynamics of the properties of the conditional variance. The degree to which the conditional variance of a random variable is persistent in financial data is a major cause of economic concern. A number of research have gone into the study of the extent to which volatility affects financial data.

Poterba and Summers (1986) show that for stock prices, the extent to which the volatility of stock returns affect the prices of the corresponding stocks depends largely on the persistence of shock to variance. Furthermore, an understanding of persistence of shocks to variance is critical to the pricing of contingent claims like options.

The GARCH model defined in Definition 2.4.1 is said to be a useful way of empirically capturing momentum in conditional variance. However, Lamoureux and Lastrapes (1990) show that the persistence of the volatility of shocks depends on the sum of the GARCH parameters , they find that as the sum of the parameters of the model approaches one from below, the effects of past shocks on current variance become stronger. Bollerslev and Wooldridge (1988) find that when the GARCH model is applied to high frequency data like daily asset prices, shocks to variance are very persistent that is the sum of the parameters of the model is very close to one. Lamoureux and Lastrapes (1990) further show that ignoring simple structural shifts in unconditional volatility can lead to spurious appearance of extremely strong persistence in variance when using GARCH models.

Recent results shown that most financial data exhibit the presence of multiple regimes (Lanne and Saikkonen (2003)), hence, models in the regime switching framework have been suggested to more effectively capture this changes in volatility persistence. The MAR model has a lot of properties that fit into this framework, making it better suited to capturing persistence and volatility clustering in financial data.

2.5 Summary

The traditional residuals of the MAR model, computed as the difference between the observed values and their conditional means is quite useful as they give information on how close the observed values are to the means of the corresponding predictive distribution.

We have shown here that the traditional residuals of the MAR model form a martingale difference sequence, a very useful property for establishing some asymptotic properties of the parameter estimates.

We have also been able to establish that the unconditional variance of the traditional residuals are strictly positive and bounded by the expectation of its conditional variance.

In addition, we compared the MAR Model to the class of GARCH models. We observed that both the GARCH type models and MAR models can be cast into the framework of random coefficient autoregressive model as well as generalized hidden markov models.

Chapter 3

Conditional Least Squares vs Maximum Likelihood Type Penalty function for MAR models

This chapter is based on the work of Klimko and Nelson (1978), Tjostheim (1986) and Masanobu Taniguchi (2000). Klimko and Nelson (1978) developed an estimation procedure for stochastic processes based on the minimization of a sum of squared deviations about conditional expectations. They studied stationary ergodic processes as well as Markov processes which are asymptotically stationary and ergodic and worked out a detailed example for a subcritical branching process with immigration.

Tjostheim (1986) extended the work by developing a general framework for analyzing estimates in nonlinear time series models. He further derived general conditions for strong consistency and asymptotic normality for both conditional least squares and maximum likelihood type penalty function estimates. He outlined examples for the Exponential AR model, RCA and Threshold AR models.

We show here that these techniques can be applied to the Mixture Autoregressive model. In particular, we give an example for the MAR(2;1,1) model and show that for the model, the variance-covariance matrix is for both the conditional least square and maximum likelihood type penalty functions are positive definite and the same.

3.1 Conditional Least Squares

The conditional least squares (CLS) procedure provides an integrated means of handling estimation problems for commonly used stochastic models. This method stemmed out of the assumption that normally distributed error terms in autoregressive models make maximum likelihood estimation similar to the minimization of a sum of squares. The conditional least squares method is motivated by the interpretation of conditional expectation as an orthogonal projection in L^2.

Klimko and Nelson (1978) showed under a variety of conditions that the CLS estimators are strongly consistent and asymptotically normally jointly distributed with rate of convergence $(\frac{\log\log n}{n})^{\frac{1}{2}}$.
The following is some notation used in this chapter:

1. Let $y_t,\ t = 1, 2, \ldots$ be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, P_\theta)$, with parameter vector $\theta = (\theta_1, \ldots, \theta_g)'$, which we assume lies in some open set, A of a Euclidean $p-$space. The true value of θ will be denoted

by $\theta^o = (\theta_1^o, \ldots, \theta_g^o)'$. The parameter θ includes

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_g \end{pmatrix}, \boldsymbol{\sigma} = \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_g \end{pmatrix}, \boldsymbol{\phi}_i = \begin{pmatrix} \phi_{1,i} \\ \vdots \\ \phi_{g,i} \end{pmatrix}, \qquad i = 1, \ldots, g \tag{3.1.0.1}$$

and each $\boldsymbol{\phi}_{k,i} = (\phi_{k,1}, \ldots, \phi_{k,p})'$ are the autoregressive coefficients for each of the components of the model.

2. Let $\| \cdot \|$,$E_\theta(\cdot)$ and $E_\theta(\cdot \mid \cdot)$ denote the Euclidean norm, the expectation, conditional expectation under P_θ respectively. So that, $\|\theta\| = (\theta'\theta)^{1/2}$.

3. For $\delta > 0$, define $N_\delta = \{\theta : \|\theta - \theta^o\| < \delta\}$.

4. Let $\mathcal{Q}_n = \mathcal{Q}_n(\theta) = \mathcal{Q}_n(y_1, \ldots, y_n; \theta)$ be a general real valued penalty function that is almost surely twice continuously differentiable in a neighbourhood S of θ^o and $N_\delta \subset S$. Moreover, let $(\partial \mathcal{Q}_n/\partial\theta)$ be the column vector defined by $(\partial \mathcal{Q}_n/\partial\theta_i),\ i = 1, \ldots g$, likewise, let $(\partial^2 \mathcal{Q}_n/\partial\theta^2)$ be the $g \times g$ matrix defined by $(\partial^2 \mathcal{Q}_n)/\partial\theta_i\theta_j),\ i, j = 1, \ldots, g$.

The Taylor's expansion for the penalty function $\mathcal{Q}_n$ around θ^o is

$$\begin{aligned}
\mathcal{Q}_n(\theta) &= \mathcal{Q}_n(\theta^o) + (\theta - \theta^o)' \frac{\partial \mathcal{Q}_n(\theta^o)}{\partial\theta} \\
&+ \frac{1}{2}(\theta - \theta^o)' \frac{\partial^2 \mathcal{Q}_n(\theta^o)}{\partial\theta\partial\theta'}(\theta - \theta^o) \\
&+ \frac{1}{2}(\theta - \theta^o)' \left\{ \frac{\partial^2 \mathcal{Q}_n(\theta^*)}{\partial\theta\partial\theta'} - \frac{\partial^2 \mathcal{Q}_n(\theta^o)}{\partial\theta\partial\theta'} \right\} (\theta - \theta^o) \\
&= \mathcal{Q}_n(\theta^o) + (\theta - \theta^o)' \frac{\partial \mathcal{Q}_n(\theta^o)}{\partial\theta} \\
&+ \frac{1}{2}(\theta - \theta^o)' V_n(\theta - \theta^o) \\
&+ \frac{1}{2}(\theta - \theta^o)' T_n(\theta^*)(\theta - \theta^o),
\end{aligned} \tag{3.1.0.2}$$

where θ^* is appropriately chosen as an intermediate point between θ and θ^o such that $0 < \|\theta^o - \theta^*\| < \delta$.

$$V_n = \left(\frac{\partial^2 \mathcal{Q}_n(\theta^o)}{\partial\theta_i \partial\theta_j}\right)_{i,j=1,\ldots,g} \tag{3.1.0.3}$$

and

$$T_n^{g\times g}(\theta^*) = \left(\frac{\partial^2 \mathcal{Q}_n(\theta^*)}{\partial\theta_i \theta_j} - V_n\right)_{i,j=1,\ldots,g} \tag{3.1.0.4}$$

The following theorem is due to Klimko and Nelson (1978).

Theorem 3.1.1. *Assume that $\{y_t\}$ and $\mathcal{Q}_n$ are such that, as $n \to \infty$,*

B1:

$$n^{-1}\left(\frac{\partial \mathcal{Q}_n(\theta^o)}{\partial\theta_i}\right) \xrightarrow{a.s.} 0, \qquad a.e.\ for\ i \leq g, \tag{3.1.0.5}$$

B2: $(2n)^{-1}V_n \xrightarrow{a.e.} V$, where V is a $g \times g$ positive definite (symmetric) matrix of constants,

B3:

$$\lim_{n\to\infty} sup_{\delta\to 0}[(n\delta)^{-1}|T_n(\theta^*)_{ij}|] < \infty\ a.e., \qquad i,j \leq g. \tag{3.1.0.6}$$

Then there exists a sequence of estimators $\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{ng})'$ such that $\hat{\theta}_n \xrightarrow{a.s.} \theta^o$ as $n \to \infty$, and such that for $\epsilon > 0$, there is an E event in $(\Omega, \mathcal{F}, P)$ with $P(E) > 1 - \epsilon$ and an n_0, such that on E and for $n > n_0, \partial\mathcal{Q}_n(\hat{\theta}_n)/\partial\theta_i = 0$, for $i = 1, \ldots g$, and $\mathcal{Q}_n$ attains a relative minimum at $\hat{\theta}_n$.

We show that the MAR model satisfies the conditions of Theorem 3.1.1 in Theorem 3.1.2. For ease of notation, all through this chapter, we denote the conditional expectation of the MAR model defined in Equation (2.2.0.1) as $g(\theta, \mathcal{F}_t)$, that is,

$$g(\theta, \mathcal{F}_t) = E(y_t \mid \mathcal{F}_{t-1}) = a_0 + \sum_{i=1}^{p} a_i y_{t-i}, \tag{3.1.0.7}$$

where a_0 and a_i are given in Equation (2.2.1.2).

The conditional least squares penalty function Q_n for the model is,

$$\mathcal{Q}_n(\theta) = \sum_{t=0}^{n-1} [y_t - g(\theta, \mathcal{F}_t)]^2 \tag{3.1.0.8}$$

Theorem 3.1.2. *Let y_t be an MAR process given by Equation (2.2.0.8) and $\mathcal{Q}_n(\theta) = \sum_{t=0}^{n-1}[y_{t+1} - g(\theta, \mathcal{F}_t)]^2$. If Assumptions A hold, then there exists a sequence of estimators $\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{ng})'$ minimizing the penalty function Q_n such that $\hat{\theta}_n \xrightarrow{a.s.} \theta^o$ as $n \to \infty$. For $\epsilon > 0$, there is an E event in $(\Omega, \mathcal{F}, P)$ with $P(E) > 1 - \epsilon$ and an n_0, such that on E and for $n > n_0, \partial \mathcal{Q}_n(\hat{\theta}_n)/\partial \theta_i = 0$, for $i = 1, \ldots g$, and $\mathcal{Q}_n$ attains a relative minimum at $\hat{\theta}_n$.*

Proof. Let, $u_t = y_t - g(\theta, \mathcal{F}_t)$. u_t is a martingale difference sequence.

The first derivate of the penalty function Q_n is,

$$\begin{aligned}\frac{\partial \mathcal{Q}_n(\theta)}{\partial \theta_i}|_{\theta=\theta^0} &= 2 \sum (y_{t+1} - g(\theta, \mathcal{F}_t)) \frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}|_{\theta=\theta^0} \\ &= 2 \sum u_{t+1}(\theta^o) \frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}\ |_{\theta=\theta^0}\end{aligned} \tag{3.1.0.9}$$

Differentiating again, we obtain the second derivative,

$$\begin{aligned}\frac{(\partial^2 \mathcal{Q}_n(\theta))}{\partial \theta_i \partial \theta_j} &= 2 \left[-\sum_{t=0}^{n-1} \left(\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i} \cdot \frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_j} \right) + \sum_{t=0}^{n-1} (y_{t+1} - g(\theta, \mathcal{F}_t)) \left(\frac{\partial^2 g(\theta, \mathcal{F}_t)}{\partial \theta_i \partial \theta_j} \right) \right] \\ &= 2 \left[-\sum_{t=0}^{n-1} \left(\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i} \cdot \frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_j} \right) + \sum_{t=0}^{n-1} \left(u_{t+1} \frac{\partial^2 g(\theta, \mathcal{F}_t)}{\partial \theta_i \partial \theta_j} \right) \right]\end{aligned} \tag{3.1.0.10}$$

B1 From Equation (3.1.0.9),

$$n^{-1}\frac{\partial \mathcal{Q}_n(\theta^o)}{\partial \theta_i} = n^{-1}2\sum u_{t+1}(\theta^o)\frac{\partial g(\theta^0, \mathcal{F}_t)}{\partial \theta_i}$$

We have shown in Section 2.2.1 that u_{t+1} is martingale difference sequence. Furthermore, we have also shown in Section 2.2.1 that $Var(u_t) < E(Var(u_t \mid \mathcal{F}_{t-1})) \leq \infty$ and is positive. Now, let $s_n^2 = \sum_{t=1}^n E(u_t^2 \mid \mathcal{F}_{t-1}))$ and $v_n = (2\log\log s_n^2)^{\frac{1}{2}}$, then $\limsup \frac{\sum_{t=1}^n u_t}{(s_n v_n)} = 1$ *a.s.* (see Klimko and Nelson (1978) for proof). This together with the strong law of martingales implies that

$$n^{-1}2\sum u_{t+1}(\theta^o)\frac{\partial g(\theta^0, \mathcal{F}_t)}{\partial \theta_i} \to 0 \qquad \text{as required.}$$

B2 By Equations (3.1.0.10) and (3.1.0.3) we can write,

$$\frac{1}{2}V_n = \sum_{t=0}^{n-1}\left(\frac{\partial g(\theta^0, \mathcal{F}_t)}{\partial \theta_i}\cdot\frac{\partial g(\theta^0, \mathcal{F}_t)}{\partial \theta_j}\right) - \sum_{t=0}^{n-1}\left(u_{t+1}\frac{\partial^2 g(\theta^0, \mathcal{F}_t)}{\partial \theta_i \partial \theta_j}\right) \tag{3.1.0.11}$$

which after dividing by n gives

$$\frac{1}{2n}V_n = \frac{1}{n}\sum_{t=0}^{n-1}\left(\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}\cdot\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_j}\right) - \frac{1}{n}\sum_{t=0}^{n-1}\left(u_{t+1}\frac{\partial^2 g(\theta, \mathcal{F}_t)}{\partial \theta_i \partial \theta_j}\right).$$

By a martingale strong law (see appendix (4)) it follows that as n increases to infinity, $\frac{1}{n}\sum_{t=0}^n u_{t+1}(\theta^o) \to 0$. This together with the boundedness of the second derivative gives

$$\left(\frac{1}{n}\sum_{t=0}^{n-1}\left(u_{t+1}\frac{\partial^2 g(\theta, \mathcal{F}_t)}{\partial \theta_i \partial \theta_j}\right)\right)^{g\times g} \to 0^{g\times g}, \text{ a.e as } n\to\infty$$

and

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{t=0}^{n-1}\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}\cdot\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_j}\right)^{g\times g} \to V^{g\times g} \text{ a.e and } V^{g\times g} \text{ is positive definite,}$$

so that condition B2 is satisfied.

B3 By Equation (3.1.0.4), we can write,

$$\lim_{n\to\infty}\sup_{\delta\to 0}\frac{(|T_n(\theta^*)_{ij}|)}{n\delta} = \lim_{n\to\infty}\sup_{\delta\to 0}\frac{\left|\frac{\partial^2 \mathcal{Q}_n(\theta^*)}{\partial\theta_i\partial\theta_j} - \frac{\partial^2 \mathcal{Q}_n(\theta^*)}{\partial\theta_i\partial\theta_j}\right|}{n\delta} < \infty \quad \text{a.e.,}$$

for $i, j \leq g$ and $0 < \|\theta^o - \theta^*\| < \delta$.

Now write $V_n(\theta^0)_{ij} = \frac{\partial^2 \mathcal{Q}(\theta^0)}{\partial\theta_i\partial\theta_j}$ and $V_n(\theta^*)_{ij} = \frac{\partial^2 \mathcal{Q}(\theta^*)}{\partial\theta_i\partial\theta_j}$, so that,

$$\lim_{n\to\infty}\sup_{\delta\to 0}\frac{(|T_n(\theta^*)_{ij}|)}{n\delta} = \lim_{n\to\infty}\sup_{\delta\to 0}\frac{1}{n\delta}|V_n(\theta^*)_{ij} - V_n(\theta^0)_{ij}|, \tag{3.1.0.12}$$

From B2 we have that $\frac{1}{2n}V_n(\theta^0) \to V$, it follows that for any intermediate point between θ and θ^0, we have,

$$\frac{1}{2n}V_n(\theta^*) \to V \tag{3.1.0.13}$$

further more, notice that as $n \to \infty$ and $\delta \to 0$ the distance $\|\theta^0 - \theta^*\| \to 0$ so that, $T_n(\theta^*) \to 0 < \infty$ thus,

$$\lim_{n\to\infty}\sup_{\delta\to 0}\frac{1}{n\delta}|V_n(\theta^*)_{ij} - V_n(\theta^0)_{ij}| \to 0 \tag{3.1.0.14}$$

so that

$$\lim_{n\to\infty}\sup_{\delta\to 0}\frac{(|T_n(\theta^*)_{ij}|)}{n\delta} < \infty \tag{3.1.0.15}$$

as required.

Now, by assumption B1-B3 of Theorem 3.1.1 and Egoroff's theorem (see Loeve (1977)), choose $\epsilon > 0$, such that for an event E with $P(E) > 1 - \epsilon$, a positive $\delta^* < \delta$, an $M > 0$, and an n_0 such that on E, for any $n > n_0$ and $\theta \in N_{\delta^*} \equiv \{\theta : \|\theta - \theta^o\| < \delta^*\}$, the following three conditions hold:

i $|(\theta - \theta^o)'\partial\mathcal{Q}(\theta^o)/\partial\theta| < n\delta^3$, ii the minimum eigenvalue of $(2n)^{-1}V_n$ is greater than some $\Delta > 0$ (bearing in mind that $\lim_{n\to\infty}(2n)^{-1}V_n = V^{g\times g}$ is positive definite),

iii $1/2(\theta - \theta^o)' T_n(\theta^*)(\theta - \theta^o) < nM\delta^3$

Hence, using the Taylor's expansion in Equation (3.1.0.2), for θ on the boundary of N_{δ^*}, we have the following:

$$\begin{aligned} \mathcal{Q}_n(\theta) &\geq \mathcal{Q}_n(\theta^o) + n(-\delta^3 + \delta^2\Delta - M\delta^3) \\ &= \mathcal{Q}_n(\theta^o) + n\delta^2(\Delta - \delta - M\delta) \end{aligned} \tag{3.1.0.16}$$

Since $\Delta - \delta - M\delta$ can be made positive by initially choosing δ sufficiently small, $Q_n(\theta)$ must attain a minimum at some $\hat{\theta}_n = (\hat{\theta}_{n1}\hat{\theta}_{n2}\ldots\hat{\theta}_{ng})'$ in N_{δ^*}, at which point the least squares equations $(\partial\mathcal{Q}_n(\hat{\alpha}_n)/\partial\alpha) = 0$ must be satisfied on E for any $n > n_0$.

Next replacing ε by $\varepsilon_k = 2^{-k}$ and $\delta_k = k^{-1}, k = 1, 2, \ldots$ to determine a sequence of events $\{E_k\}$ and an increasing sequence $\{n_k\}$ such that $\partial\mathcal{Q}_n(\theta)/\partial\theta = 0$ has a solution on E_k for any $n > n_k$.

For $n_k < n \leq n_{k+1}$ define $\hat{\theta}_n$ on E_k to be a root of $\partial\mathcal{Q}_n(\theta)/\partial\theta = 0$ within δ of θ^o at which $\mathcal{Q}_n$ attains a relative minimum.

Then $\hat{\theta}_n \to \theta^0$ on $\liminf_{k\to\infty} E_k$ and $P(\liminf_{k\to\infty} E_k) = 1$ since

$$\begin{aligned} 1 - P(\liminf_{k\to\infty} E_k) &= P(\limsup_{k\to\infty} E_k^c) = \lim_{k\to\infty} P(\bigcup_{j=k}^{\infty})E_j^c) \\ &\leq \lim_{k\to\infty}\sum_{j=k}^{\infty} P(E_k^c) \leq \lim_{k\to\infty}\sum_{j=k}^{\infty} 2^{-j} = 0. \end{aligned} \tag{3.1.0.17}$$

$\square$

Theorem 3.1.3. *In addition to the conditions in Theorem 3.1.1, assume further that*

$B1^i$

$$\left(\frac{1}{2}\right) n^{-\frac{1}{2}} \partial \mathcal{Q}_n(\theta^o)/\partial\theta \xrightarrow{\mathscr{L}} MVN(0^{g\times 1}, W), \tag{3.1.0.18}$$

where $W^{g\times g}$ is a positive definite matrix.

Then,

$$n^{\frac{1}{2}}(\hat{\theta}_n - \theta^o) \xrightarrow{\mathscr{L}} MVN(0^{g\times 1}, V^{-1}WV^{-1}). \tag{3.1.0.19}$$

Proof. Using the results of theorem 3.1.2, we can choose $\{\hat{\theta}_n\}$ so that $\partial\mathcal{Q}_n(\theta)/\partial\theta = 0$. Expanding the vector $n^{-1/2}(\partial\mathcal{Q}_n(\theta^o)/\partial\theta)$ in a Taylor series about θ^o we obtain:

$$\begin{aligned} 0^{p\times g} &= n^{-1/2}\left\{\frac{\partial\mathcal{Q}_n(\hat{\theta}_n)}{\partial\theta}\right\} \qquad (3.1.0.20) \\ &= n^{-1/2}\left\{\frac{\partial\mathcal{Q}_n(\theta^o)}{\partial\theta}\right\} + n^{-1}\{V_n + T_n(\theta^*)\}\sqrt{n}(\hat{\theta}_n - \theta^o). \end{aligned}$$

By Assumptions $B2$ and $B3$ in Theorem 3.1.2, it follows that $n^{-1}\{V_n+T_n(\theta^*)\} \xrightarrow{a.s} 2V$, so that the limiting distribution of $\sqrt{n}(\hat{\theta}_n-\theta^o)$ is the same as $(2V)^{-1}n^{-1/2}\{\frac{\partial\mathcal{Q}_n(\hat{\theta}^o)}{\partial\theta}\}$ and thus together with condition $B1^i$ yields the desired result. □

3.2 Conditional Least Squares penalty function for Stationary Ergodic Processes

Consider a stationary ergodic sequence of integrable random variables $\{y_t\}_{t=0}^{\infty}$ and for an arbitrary positive integer m, $\mathcal{F}_t = \sigma(y_t, y_{t-1}, \cdots, y_{t-m+1})$, $t = m-1, m, \ldots$. Assume that the second moments of $\{y_t\}$ exist and $\{y_t\}$ is second order stationary. Define the function $g(\theta, \mathcal{F}_t) = E_\theta(y_t \mid \mathcal{F} - \infty_t)$. The conditional least squares penalty function is defined as in Equation (3.1.0.8). The conditional Least squares estimates are obtained by minimizing $\mathcal{Q}_n(\theta)$. The following theorem is due to

Tjostheim (1986) and Masanobu Taniguchi (2000), however, the arguments are due to Klimko and Nelson (1978).

3.2.1 Consistency

Theorem 3.2.1. *Assume that $\{y_t\}$ is a d-dimensional strictly stationary ergodic process with $E(|y_t|^2) < \infty$ and $g(\theta, \mathcal{F}_t) = E_\theta\{y_t \mid \mathcal{F}_{t-1}\}$ is almost surely three times continuously differentiable in an open set B containing θ^o. Moreover, suppose that:*

$C1$:

$$E\left\{\left|\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}(\theta^o)\right|^2\right\} < \infty \text{ and } E\left\{\left|\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i \partial \theta_j}(\theta^o)\right|^2\right\} < \infty$$

for $i, j = 1, \ldots, g$.

$C2$: The vectors $\partial g(\theta^o, \mathcal{F}_t)/\partial \theta_i$, $i = 1, \ldots, g$, are linearly independent in the sense that if $a_1, \ldots a_g$ are arbitrary real numbers, such that

$$E\left\{\left|\sum_{i=1}^{r} a_i \frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}(\theta^o)\right|^2\right\} = 0,$$

then $a_1 = a_2 = \cdots = a_g = 0$.

$C3$: For $\theta \in B$, there exist functions $G_{i-1}^{ijk}(y_1, \ldots, y_{t-1})$ and $H_{i-1}^{ijk}(y_1, \ldots, y_{t-1})$ such that

$$\left|\frac{\partial g(\theta, \mathcal{F}_t)}{\partial \theta_i}(\theta)\frac{\partial^2 g(\theta, \mathcal{F}_t)}{\partial \theta_i \theta_j}(\theta)\right| \leq G_{i-1}^{ijk},\ E(G_{i-1}^{ijk}) < \infty,$$
$$\left|\{y_t - g(\theta, \mathcal{F}_t)\}\frac{\partial^3 g(\theta, \mathcal{F}_t)}{\partial \theta_i \theta_j \theta_k}(\theta)\right| \leq H_{i-1}^{ijk},\ E(H_{i-1}^{ijk}) < \infty,$$

for $i, j, k = 1, \ldots, g$.

Then there exists a sequence of estimators $\hat{\theta}_n$ minimizing the penalty function $\mathcal{Q}_n$, such that then there exists a sequence of estimators $\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{ng})'$ minimizing the penalty function Q_n such that $\hat{\theta}_n \xrightarrow{a.s.} \theta^o$ as $n \to \infty$. For $\epsilon > 0$, there is an E event in $(\Omega, \mathcal{F}, P)$ with $P(E) > 1 - \epsilon$ and an n_0, such that on E and for $n > n_0, \partial\mathcal{Q}_n(\hat{\theta}_n)/\partial\theta_i = 0$, for $i = 1, \ldots g$, and $\mathcal{Q}_n$ attains a relative minimum at $\hat{\theta}_n$.

.

Proof. By the Ergodic theorem, strict stationarity and Assumption $C1$, together with Equations (3.1.0.9) and (3.1.0.10) we have that,

$$n^{-1}\frac{\partial}{\partial\theta}\mathcal{Q}_n(\theta) \xrightarrow{a.s} -2E[\frac{\partial}{\partial\theta}g(\theta^0, \mathcal{F}_t)'\{Y_t - g(\theta^0, \mathcal{F}_t)\}] = 0, \tag{3.2.1.1}$$

So that Assumption $B1$ of Theorem 3.1.1 is satisfied. Furthermore, by Assumption $C1$ and the ergodic theorem, we can write

$$n^{-1}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{Q}_n(\theta^0) \xrightarrow{a.s} 2V. \tag{3.2.1.2}$$

Assumption $C2$ ensures the positive definiteness of V. So that Assumption $B2$ of Theorem 3.1.1 is satisfied. Finally, by the mean value theorem and the ergodic theorem, Assumption $C3$ is satisfied and thus completes the proof. □

3.2.2 Asymptotic normality

Theorem 3.2.2. *Define the $g \times g$ matrix (W_{ij}) by*

$$W_{ij} = E\left(u_m^2(\theta^o)\frac{\partial g(\theta^o, \mathcal{F}_{m-1})}{\partial a_i} \cdot \frac{\partial g(\theta^o, \mathcal{F}_{m-1})}{\partial a_j}\right). \tag{3.2.2.1}$$

Assume,

D1: $E(y_t \mid y_{t-1}, \ldots, y_0) = E(y_t \mid y_{t-1}, \ldots, y_{t-m})$ a.e., $t \geq m$, and $W_{ij} < \infty$ $i, j \leq g$, where $u_m(\theta^o) = y_m - g(\theta^0, \mathcal{F}_{m-1})$.

Let $\{\hat{\theta}_n\}$ be the consistent sequence of estimators obtained in Theorem 3.2.1. Then

$$n^{\frac{1}{2}}(\hat{\theta}_n - \theta^o) \rightarrow MVN(0, V^{-1}WV^{-1}). \tag{3.2.2.2}$$

Proof. Expand the vector $\frac{\partial \mathcal{Q}_n(\hat{\theta})}{\partial \theta}$ in a Taylor series expansion about θ^o and multiply through by $n^{-\frac{1}{2}}$.

$$\begin{aligned} 0^{p\times 1} &= n^{-\frac{1}{2}} \frac{\partial \mathcal{Q}_n(\hat{\theta})}{\partial \theta} \\ &= n^{-\frac{1}{2}} \frac{\partial \mathcal{Q}_n(\theta^o)}{\partial \theta} + n^{-1}(V_n + U_n) n^{\frac{1}{2}}(\hat{\theta}_n - \theta^o). \end{aligned} \tag{3.2.2.3}$$

where $U_n^{g\times g} = (2^{-1} \sum_{k=1}^{p} (\hat{\theta}_{nk} - \theta_k^o)(\partial^3 \mathcal{Q}_n(\theta^*))/(\partial\theta_i \partial\theta_j \partial\theta_k))_{i\leq g; j\leq g}$ and assume that $\{\hat{\theta}_n\}$ satisfies the least squares equation and θ^* is an appropriate intermediate point. We have $n^{-1}U_n \rightarrow 0^{p\times p}$ a.e. Billingsley (961b) central limit theorem may be applied to the martingale

$$\sum_{t=m}^{n} \sum_{i=1}^{p} \left(c_i \frac{\partial g(\theta^o, \mathcal{F}_{t-1})}{\partial \theta} \right) u_t(\theta^o),$$

where c_i are non-zero constants so that the conditions in Theorem 3.2.1 are satisfied. □

Note that Klimko and Nelson (1978), give the relationship between the positive definiteness of V and W as follows, if

$$E([y_m - g(y_{m-1}, \cdots, y_0; \theta^0)]^2 \mid \mathcal{F}_{m-1}) > 0 \text{ a.e.}, \tag{3.2.2.4}$$

then the positive definiteness of V implies the same as that of W. (see Klimko and Nelson (1978) for details as well as for an example of a case where V is positive definite and W is semi-positive definite).

3.3 Maximum Likelihood Type Penalty Function

Let,

$$f_\theta(t, t-1) = E[(y_t - g(\theta, \mathcal{F}_t))(y_t - g(\theta, \mathcal{F}_t))^{'}] \tag{3.3.0.5}$$

and

$$\phi_t = [\ln\{\det f_\theta(t, t-1)\} + (y_t - g(\theta, \mathcal{F}_t)^{'}(f_\theta(t, t-1))^{-1}(y_t - g(\theta, \mathcal{F}_t)\}] \tag{3.3.0.6}$$

The likelihood type penalty function of Tjostheim (1986) is,

$$L_n = \sum_{t=m+1}^{n} \phi_t \tag{3.3.0.7}$$

Note that if $\{Y_t\}$ is a conditional Gaussian process, then L_n coincides with the log likelihood function with the exception of a multiplicative constant (Tjostheim (1986)). Here, L_n is treated as a general penalty function as it has the martingale property. The following theorem is due to Tjostheim (1986).

3.3.1 Consistency

Theorem 3.3.1 (Tjostheim (1986)). *Assume that $\{y_t\}$ is an m-dimensional strictly stationary and ergodic process with $E(|y_t|^2) < \infty$. Assume also that $g(\theta, \mathcal{F}_t)$ and $f_\theta(t, t-1)$ are almost surely three times continously differentiable in an open set A containing the true θ^o. If ϕ_t is defined as in Equation* (3.3.0.6), *and the following conditions hold:*

E1:

$$E\left(\left|\frac{\partial \phi_t}{\partial \theta_i}(\theta^o)\right|\right) < \infty \text{ and } E\left(\left|\frac{\partial^2 \phi_t}{\partial \theta_i \theta_j}(\theta^o)\right|\right) < \infty \tag{3.3.1.1}$$

for $i, j = 1, \ldots, g$.

E2: For arbitrary real numbers $a_1, \ldots, a_g$ *such that , for* $\theta = \theta^o$,

$$E\left(\left|f_\theta^{1/2}(t,t-1)\sum_{i=1}^{s} a_i \frac{\partial g(\theta, \mathcal{F}_{t-1})}{\partial \theta_i})\right|^2\right)$$
$$+ E\left[\left|f_\theta^{1/2}(t,t-1)\otimes f_\theta^{1/2}(t,t-1)\sum_{i=1}^{g} a_i \frac{\partial\{\vec{(}f_\theta(t,t-1))\}}{\partial \theta_i})\right|^2\right]$$
$$= 0, \quad (3.3.1.2)$$

then we have $a_1 = a_2 = \cdots = a_g = 0$.

E3: The vectors

$$f_\theta(t,t-1)^{1/2}(\frac{\partial g(\theta, \mathcal{F}_{t-1})}{\partial \theta_i}), \quad i = 1, \ldots, g \text{ are linearly independent} \quad (3.3.1.3)$$

E4: for $\theta \in A$, *there exists a function* $H_t^{ijk}(Y_1, \ldots, Y_t)$ *such that*

$$\left|\frac{\partial^3 \phi_i}{\partial\theta_i\partial\theta_j\partial\theta_k}(\theta)\right| \leq H_t^{ijk} \qquad and \qquad E(H_t^{ijk}) < \infty \quad (3.3.1.4)$$

for $i, j, k = 1, \ldots, g$

E5:

$$V = E\left\{\frac{\partial\phi_t(\theta^o)}{\partial\theta}\frac{\partial\phi_t\theta'}{\partial\theta}\right\} \leq \infty. \quad (3.3.1.5)$$

Then there exists a sequence of estimators $\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{ng})'$ *minimizing the penalty function* L_n *such that* $\hat{\theta}_n \xrightarrow{a.s.} \theta^o$ *as* $n \to \infty$. *For* $\epsilon > 0$, *there is an* E *event in* $(\Omega, \mathcal{F}, P)$ *with* $P(E) > 1 - \epsilon$ *and an* n_0, *such that on* E *and for* $n > n_0, \partial L_n(\hat{\theta}_n)/\partial\theta_i = 0$, *for* $i = 1, \ldots g$, *and* L_n *attains a relative minimum at* $\hat{\theta}_n$. *Furthermore,*

$$\sqrt{n}(\hat{\theta}_n - \theta^o) \xrightarrow{d} N(0, U^{-1}VU^{-1}), \quad (3.3.1.6)$$

where

$$U = E\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\phi_t(\theta^o)\right\} \tag{3.3.1.7}$$

We show that the MAR model satisfies the conditions of Theorem 3.3.1 in Theorem 3.3.2.

Theorem 3.3.2. *Let y_t be an MAR process given by Equation* (2.2.0.8). *If Assumptions 2.2 hold, then there exists a sequence of estimators $\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{ng})'$ minimizing the penalty function L_n such that $\hat{\theta}_n \xrightarrow{a.s.} \theta^o$ as $n \to \infty$. For $\epsilon > 0$, there is an E event in $(\Omega, \mathcal{F}, P)$ with $P(E) > 1 - \epsilon$ and an n_0, such that on E and for $n > n_0, \partial L_n(\hat{\theta}_n)/\partial\theta_i = 0$, for $i = 1, \ldots g$, and L_n attains a relative minimum at $\hat{\theta}_n$. Furthermore,*

$$\sqrt{n}(\hat{\theta}_n - \theta^o) \xrightarrow{d} N(0, U^{-1}VU^{-1}) \tag{3.3.1.8}$$

Proof. By the stationarity and ergodicity of the process as well as Assumption $(E1)$, we have that $n^{-1}\partial L_n(\theta^o)/\partial\theta_i \xrightarrow{a.s} E\{\partial\phi_t(\theta^o)/\partial\theta_i\}$ as $n \to \infty$.

However, because of the martingale increment property of $\{\partial\phi_t(\theta^o)/\partial\theta_i\}$ we have $E(\{\partial\phi_t(\theta^o)/\partial\theta_i\}) = E[E\{\partial\phi_t(\theta^o)/\partial\theta_i \mid \mathcal{F}^y_{t-1}\}] = 0$ satisfying Assumption $B1$ of Therorem 3.1.1. Similarly, Assumption $B3$ follows from $E3$, the mean value theorem and the ergodic theorem. From the second part of $E1$ and the ergodic theorem, we have

$$n^{-1}\frac{\partial^2 L_n}{\partial\theta_i\partial\theta_j}(\theta^o) \xrightarrow{a.s.} E\left[E\left\{\frac{\partial^2\phi_t}{\partial\theta_i\partial\theta_j} \mid \mathcal{F}^y_{t-1}\right\}\right] \triangleq V'_{ij} \tag{3.3.1.9}$$

It is now left to show that $E2$ implies the matrix $V' = V'_{ij}$ is positive definite

(Tjostheim (1986)).

$$E\left(\frac{\partial^2\phi_t}{\partial\theta_i\partial\theta_j} \mid \mathcal{F}^y_{t-1}\right) = Tr\left(f_\theta^{-1}(t,t-1)\frac{\partial f_\theta(t,t-1)}{\partial\theta_t}f_\theta^{-1}(t,t-1)\frac{\partial f_\theta(t,t-1)}{\partial\theta_j}\right) \tag{3.3.1.10}$$

$$+\,2\frac{\partial g(\theta,\mathcal{F}_{t-1})}{\partial\theta_i}f_\theta^{-1}(t,t-1)\frac{\partial g(\theta,\mathcal{F}_{t-1})}{\partial\theta_j}$$

So that for $\theta=\theta^o$ and arbitrary $a_1,\cdots,a_g$, we have:

$$\sum_{i=1}^{s}\sum_{j=1}^{s}a_ia_jE\left\{E\left(\frac{\partial^2\phi_t}{\partial\theta_i\partial\theta_j} \mid \mathcal{F}^y_{t-1}\right)\right\} = 2E\left(\left|(f_\theta^{-\frac{1}{2}}(t,t-1)\sum_{i=1}^{s}a_i\frac{\partial g(\theta,\mathcal{F}_{t-1})}{\partial\theta_i}\right|^2\right)$$

$$+\,E\left(\left|f_\theta^{-\frac{1}{2}}(t,t-1)\otimes f_\theta^{-\frac{1}{2}}(t,t-1)\sum_{i=1}^{s}a_ivec\left(\frac{\partial f_\theta(t,t-1)}{\partial\theta_i}\right)\right|^2\right)\geq 0 \tag{3.3.1.11}$$

Hence the matrix $(V)_{i,j}$ is non-negative definite, and since $f_\theta(t,t-1)$ is positive definite, it follows from Equation (3.3.1.11) and assumption $E2$ that V is also positive definite. This concludes the proof. □

3.3.2 Asymptotic normality

Tjostheim (1986) explore the asymptotic normality of the estimator $\hat{\theta}_n$ in the following theorem.

Theorem 3.3.3. *Assume that the conditions of Theorem 3.3.1 are fulfilled and that for $\theta=\theta^o$ and*

$F1$*:*

$$S_{ij}\triangleq\frac{1}{4}E\left\{\frac{1}{f_\theta^4(t,t-1)}\left(\frac{\partial f_\theta(t,t-1)}{\partial\theta_i}\frac{\partial f_\theta(t,t-1)}{\partial\theta_j}\left[E\{(y_t-g(\theta,\mathcal{F}_{t-1}))^4 \mid \mathcal{F}^y_{t-1}\}\right.\right.\right.$$

$$\left.-3f_\theta^2(t,t-1)\right]+2E\{(y_t-g(\theta,\mathcal{F}_{t-1}))^3 \mid \mathcal{F}^y_{t-1}\}f_\theta(t,t-1)]$$

$$\left.\left.\times\left(\frac{\partial g(\theta,\mathcal{F}_{t-1})}{\partial\theta_i}\frac{\partial f_\theta(t,t-1)}{\partial\theta_j}+\frac{\partial g(\theta,\mathcal{F}_{t-1})}{\partial\theta_j}\frac{\partial f_\theta(t,t-1)}{\partial\theta_i}\right)\right)\right\}<\infty \tag{3.3.2.1}$$

Let $S = S_{ij}$, and let $\{\hat{\theta}_n\}$ be the estimators obtained in Theorem 3.3.1. Then we have

$$S_{ij} = \frac{1}{4}E\left(\frac{\partial\phi_t}{\theta_i}\frac{\partial\phi_t}{\theta_j}\right) - U^{'}_{ij} \tag{3.3.2.2}$$

and

$$n^{1/2}(\hat{\theta}_n - \theta^o) \xrightarrow{d} \mathcal{N}(0, (U^{'})^{-1} + ((U^{'})^{-1}S(U^{'})^{-1}). \tag{3.3.2.3}$$

where

$$U^{'}_{ij} = E\left\{\frac{1}{f^2_\theta(t, t-1)}\left(f_\theta(t, t-1)\frac{\partial g(\theta, \mathcal{F}_{t-1})}{\partial\theta_i}\frac{\partial g(\theta, \mathcal{F}_{t-1})}{\partial\theta_j} + \frac{1}{2}\frac{\partial f_\theta(t, t-1)}{\partial\theta_i}\frac{\partial f_\theta(t, t-1)}{\partial\theta_j}\right)\right\} \tag{3.3.2.4}$$

Proof. Applying the *martingales central limit theorems* for the strictly stationary ergodic case together with a Cramer-Wald argument, we find that if the limiting covariance of $n^{-1/2}\partial L_n(\theta^0)/\partial\theta$ exists, then it follows that it has a multivariate normal distribution as its limiting distribution. By Theorem 3.1.1, the estimator $\hat{\theta}_n$ is asymptotically normal. Since $\{\partial L_n(\theta^0)/\partial\theta_i, \mathcal{F}_n\}$ is a martingale, we have that,

$$\begin{aligned} n^{-1}E\left\{\frac{\partial L_n}{\partial\theta_i}(\theta^0)\frac{\partial L_n}{\partial\theta_j}(\theta^0)\right\} &= n^{-1}\sum_{t=1}^{n}\left\{\frac{\partial\phi_t}{\partial\theta_i}(\theta^0)\frac{\partial\phi_t}{\partial\theta_j}(\theta^0)\right\} \\ &= E\left[E\left\{\frac{\partial\phi_t}{\partial\theta_i}(\theta^0)\frac{\partial\phi_t}{\partial\theta_j}(\theta^0) \mid \mathcal{F}_{t-1}\right\}\right] \end{aligned} \tag{3.3.2.5}$$

So that by the definition of ϕ_t (Equation (3.3.0.7)), it follows that $\theta = \theta^0$ and

$$E\left\{E\left(\frac{\partial\phi_t}{\partial\theta_i}\cdot\frac{\partial\phi_t}{\partial\theta_j} \mid \mathcal{F}_{t-1}\right)\right\} = 4(S_{ij} + U^{'}_{ij}). \tag{3.3.2.6}$$

By Assumptions $E1$ and $F1$

$$E\left\{\frac{n^{-1/2}\partial L_n(\theta^0)}{\partial\theta_i}\cdot\frac{n^{-1/2}\partial L_n(\theta^0)}{\partial\theta_j}\right\} < \infty \tag{3.3.2.7}$$

The covariance matrix in Equation (3.3.2.3) follows from Theorem 3.1.3 as well as the definition of S and $U^{'}$ □

In situations where $f_\theta(t, t-1)$ does not depend on θ, that is $\partial f_\theta(t, t-1)/\partial\theta = 0$ hence, $S = 0$ and in addition,

$$U^{'} = E\left[\frac{\partial g^{'}(\theta^o, \mathcal{F}_{t-1})}{\theta}\{E(f_\theta(t, t-1))\}^{-1}\frac{\partial g(\theta, \mathcal{F}_{t-1})}{\theta}\right] \tag{3.3.2.8}$$

so that we then have,

$$n^{1/2}(\hat{\theta}_n - \theta^o) \xrightarrow{d} \mathcal{N}(0, (U^{'})^{-1}). \tag{3.3.2.9}$$

and estimation using L_n of Equation (3.3.0.7) and Q_n of Theorem 3.1.1 would basically yield similar results.

3.4 An MAR (2; 1, 1) example

We extend the results of the paper by Tjostheim (1986) to the MAR model described in Chapter 2. In particular, we explore the MAR (2; 1, 1) model. An MAR model with two AR components each of order one, that is, $p_1 = p_2 = 1$ and $k = 2$.

The MAR(2,1,1) is such that,

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1}y_{t-1} + \sigma_1\epsilon_1(t) & \text{with probability } \pi_1 \\ \phi_{2,0} + \phi_{2,1}y_{t-1} + \sigma_2\epsilon_2(t) & \text{with probability } \pi_2, \end{cases}$$

with conditional distribution

$$F_{t|t-1}(x) = \pi_1 F_1\left(\frac{y_t - \phi_{11}y_{t-1}}{\sigma_1}\right) + \pi_2 F_2\left(\frac{y_t - \phi_{12}y_{t-1}}{\sigma_2}\right). \tag{3.4.0.10}$$

$$g(a, \mathcal{F}_t) = E(y_t \mid \mathcal{F}_{t-1}) = a_0 + a_1 y_{t-1}, \tag{3.4.0.11}$$

$$\text{where} \quad a_0 = \sum_{k=1}^{g}\pi_k\phi_{k,0} \quad \text{and} \quad a_1 = \sum_{k=1}^{g}\pi_k\phi_{k,1}. \tag{3.4.0.12}$$

The process y_t can also be written as:

$$y_t = f_{z_t}(t) + \sigma_{z_t}\epsilon_{z_t}(t) \tag{3.4.0.13}$$

where $f_{z_t}(t) = \phi_{z_t,0} + \phi_{z_t,1}(y(t-1))$ and z_t is an i.i.d sequence of random variables with distribution π, such that $Pr\{z_t = k\} = \pi_k,\ k = 1, 2$ (see Boshnakov (2009)).

3.4.1 The Conditional Least Square Type Penalty

Theorem 3.4.1. *Let* $\{y_t\}$ *be defined as in Equation* (2.2.0.8). *Assume that* $E(y_t^4) < \infty$. *Then there exists a unique distribution for the initial variable* y_1 *such that* $\{Y_t,\ t \geq 1\}$ *is strictly stationary and ergodic.*

Moreover, there exists a strongly consistent sequence of estimators $\{\hat{a}_n\}$ *minimizing the penalty function* Q_n *defined in Equation* (3.1.0.8) *in the manner of Theorem 3.2.1 such that* $\hat{a}_n \xrightarrow{a.s.} a$ *and* $\hat{a}_n$ *is asymptotically normal with variance-covariance matrix defined by* $V^{-1}WV^{-1}$, V *and* W *are as in Equations* (3.1.0.3) *and* (3.2.2.1) *respectively.*

Proof. Wong and Li (2000) and Boshnakov (2009) gives conditions for the stationary and ergodicity of the MAR model. The first and second derivative of Equation (3.1.0.7) is evaluated as,

$$\frac{\partial g(a, \mathcal{F}_t)}{\partial a_i} = \begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix},\ \frac{\partial^2 g(a, \mathcal{F}_t)}{\partial a_i a_j} = \begin{pmatrix} 1 & y_{t-1} \\ y_{t-1} & y_{t-1}^2 \end{pmatrix} \text{ and } \frac{\partial^3 g(a, \mathcal{F}_t)}{\partial a_i a_j a_k} = 0.$$

With corresponding expected values of the first and second derivative of Equation (3.1.0.7) evaluated as,

$$E\left[\frac{\partial g(a, \mathcal{F}_t)}{\partial a_i}\right] = E\begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix} = \begin{pmatrix} 1 \\ E(y_{t-1}) \end{pmatrix} < \infty$$

also

$$E\left[\frac{\partial^2 g(a,\mathcal{F}_t)}{\partial a_i a_j}\right] = E\begin{pmatrix} 1 & y_{t-1} \\ y_{t-1} & y_{t-1}^2 \end{pmatrix} = E\begin{pmatrix} 1 & E(y_{t-1}) \\ E(y_{t-1}) & E(y_{t-1}^2) \end{pmatrix} < \infty.$$

Since the expected values of both the first and second derivatives of Equation (3.4.0.11) are finite, Condition $C1$ of Theorem 3.2.1 is satisfied.

Now notice that,

$$\frac{\partial g(a,\mathcal{F}_t)}{\partial a_0} = 1$$
$$\frac{\partial g(a,\mathcal{F}_t)}{\partial a_1} = y_{t-1}$$

so that , for an arbitrary set of real numbers b_1, b_2,

$$E\left(\left|b_1\frac{\partial g(a,\mathcal{F}_t)}{\partial a_0} + b_2\frac{\partial g(a,\mathcal{F}_t)}{\partial a_1}\right|\right)$$

so that $E[b_1 + b_2 y_{t-1}] = 0$ if and only if $b_1 = 0$ implies $b_2 = 0$ and vice versa so that Condition $C2$ of Theorem 3.2.1 is also satisfied.

In addition,

$$E\left|\frac{\partial' g(a,\mathcal{F}_t)}{\partial a_i}\frac{\partial^2 g(a,\mathcal{F}_t)}{\partial a_i a_j}\right|$$
$$= E\left|\begin{pmatrix} 1 & y_{t-1} \end{pmatrix}\begin{pmatrix} 1 & y_{t-1} \\ y_{t-1} & y_{t-1}^2 \end{pmatrix}\right|$$
$$E\left|\begin{pmatrix} 1 + y_{t-1}^2 & y_{t-1} + y_{t-1}^3 \end{pmatrix}\right| < \infty$$

and

$$E\left|\{y_t - g(a,\mathcal{F}_t)\}'\frac{\partial^3 g(a,\mathcal{F}_t)}{\partial a_i a_j a_k}\right| = 0$$

shows that Condition $C3$ of Theorem 3.2.1 is satisfied.

Furthermore,

$$E\left\{\frac{\partial g(a,\mathcal{F}_t)}{\partial a} f_a(t,t-1)\frac{\partial g(a,\mathcal{F}_t)}{\partial a}\right\} =$$

$$E\left|\begin{pmatrix}1 & y_{t-1}\end{pmatrix}\sum_{k=1}^{g}\pi_k\sigma_k^2\begin{pmatrix}1\\ y_{t-1}\end{pmatrix}\right|$$

$$= E\sum_{k=1}^{g}\pi_k\sigma_k^2[1+y_{t-1}^2] < \infty$$

hence, Condition $D1$ of Theorem 3.2.2 is also satisfied.

Then from Theorem 3.2.1 and Theorem 3.2.2, there is an $\hat{a}_n$, such that $\hat{a}_n \xrightarrow{a.s.} a$ with variance-covariance matrix defined by $V^{-1}WV^{-1}$, V as required. □

3.4.2 The Maximum Likelihood Type Penalty

Theorem 3.4.2. *Let* $\{y_t\}$ *be defined by Equation* (2.2.0.8)*. Assume that* $E(y_t^4) < \infty$*. Then there exists a unique distribution for the initial variable* y_1 *such that* $\{Y_t,\ t \geq 1\}$ *is strictly stationary and ergodic.Moreover, there exists a strongly consistent sequence of estimators* $\{\hat{a}_n\}$ *minimizing the penalty function* L_n *defined in Equation* (3.3.0.7) *in the manner of theorem 3.3.1 such that* $\hat{a}_n \xrightarrow{a.s.} a$ *and* $\hat{a}_n$ *is asymptotically normal with variance-covariance matrix simply defined by* $(U')^{-1}$*,* U *is as in Equation* (3.3.1.7)*.*

Proof. Write,

$$\begin{aligned} f_a(t,t-1) &= E\{[(y_t - g(a,\mathcal{F}_t) \mid \mathcal{F}_{t-1})(y_t - g(a,\mathcal{F}_t))' \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\} \quad (3.4.2.1)\\ &= \sum_{k=1}^{g}\pi_k\sigma_k^2 \times \mathbf{I}. \end{aligned}$$

where

$$\mathbf{I} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

By Equation (3.4.0.11), the first derivate of $f_a(t, t-1)$ is evaluated as

$$\frac{\partial f_a(t,t-1)}{\partial a} = \frac{\sum_{k=1}^{g} \pi_k \sigma_k^2}{\partial a} = 0 \tag{3.4.2.2}$$

and the derivatives of $g(a, \mathcal{F}_t) = E(y_t \mid \mathcal{F}_{t-1})$ is as in the proof of theorem 3.4.1 above. So that the first partial derivates of equation (3.3.0.7) with respect to the parameters a_i is evaluated as,

$$\frac{\partial \phi_t}{\partial a_i} = \begin{pmatrix} \frac{2(y_t - g(a,\mathcal{F}_{t-1})}{E(y_t - g(a,\mathcal{F}_{t-1})^2} \\ \frac{2y_{t-1}(y_t - g(a,\mathcal{F}_{t-1})}{E(y_t - g(a,\mathcal{F}_{t-1})^2}. \end{pmatrix} \tag{3.4.2.3}$$

With the corresponding expected value given as,

$$E\left[\frac{\partial \phi_t}{\partial a_i}\right] = E\begin{pmatrix} \frac{2(y_t - g(\alpha,\mathcal{F}_{t-1})}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} \\ \frac{2y_{t-1}(y_t - g(\alpha,\mathcal{F}_{t-1})}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} \end{pmatrix} = 0 \text{ and hence } E\left[\frac{\partial \phi_t}{\partial a_i}\right] < \infty. \tag{3.4.2.4}$$

Similarly, the second partial derivates of Equation (3.3.0.7) with respect to the parameters a_i is evaluated as,

$$\frac{\partial^2 \phi_t}{\partial a_i \partial a_j} = \begin{pmatrix} \frac{2}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} & \frac{2y_{t-1}}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} \\ \frac{2y_{t-1}}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} & \frac{2y_{t-1}^2}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2}. \end{pmatrix} \tag{3.4.2.5}$$

With corresponding expectation given as,

$$E\left[\frac{\partial^2 \phi_t}{\partial a_i \partial a_j}\right] = E\begin{pmatrix} \frac{2}{(y_t - g(\alpha,\mathcal{F}_{t-1})^2} & \frac{2y_{t-1}}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} \\ \frac{2y_{t-1}}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} & \frac{2y_{t-1}^2}{E(y_t - g(\alpha,\mathcal{F}_{t-1})^2} \end{pmatrix} = \tag{3.4.2.6}$$

$$\frac{2}{\sum_{k=1}^{g} \pi_k \sigma_k^2} E\begin{pmatrix} 1 & y_{t-1} \\ y_{t-1} & y_{t-1}^2 \end{pmatrix} =$$

$$E\frac{2}{E(y_t - g(\alpha, \mathcal{F}_{t-1})^2}\begin{pmatrix} 1 & y_{t-1} \\ y_{t-1} & y_{t-1}^2 \end{pmatrix} < \infty.$$

Since the expected values of both the first and second derivatives of Equation (3.3.0.7) are finite, then condition $E1$ of Theorem 3.3.1 is satisfied.

From equation (3.4.2.2), the second part of the equation in Condition $E2$ equates to zero so that, Condition $D2$ follows from the proof of Condition $C2$ of equation 3.4.1.

Similarly Condition $E3$ follows from the fact that

$$\left| \frac{\partial^3 \phi_t}{\partial a_i \partial a_j \partial a_k} \right| = 0. \tag{3.4.2.7}$$

To show the asymptotic normality of the estimators $\hat{a}_n$, note that the 2nd term in the RHS of Equation (3.3.2.4) and the 1st and 3rd terms in the RHS of Equation (3.3.2.2) are both equal to zero since $|\frac{\partial^3 \phi_t}{\partial a_i \partial a_j \partial a_k}| = 0$ and $\frac{\partial f_a(t,t-1)}{\partial a_i} = 0$. So that

$$U_ij = E\{\frac{1}{f_a(t,t-1)} \frac{\partial g(a, \mathcal{F}_t)}{\partial a_i} \frac{\partial g(a, \mathcal{F}_t)}{\partial a_j}\} \tag{3.4.2.8}$$

and $S_{ij} \triangleq 0 < \infty$. Hence, the Variance-Covariance matrix is given as (U_{ij}^{-1}) as required. $\square$

3.4.3 Variance-Covariance (V-C) Matrix

The V-C matrix is derived using using both the conditional least squares and the maximum likelihood penalty function. We adapt the following notation:

- $V - C_{cls_i}$ and $V - C_{cls_{ii}}$ are the Variance-Covariance matrix based on the conditional least squares penalty function for scenario i and ii respectively
- $V - C_{ml_i}$ and $V - C_{ml_{ii}}$ are the Variance-Covariance matrix based on the maximum likelihood penalty function for scenario i and ii respectively.

The two scenarios explored are:

1. Klimko and Nelson (1978)

I

$$\begin{aligned} V - C_{cls_i} &= V^{-1}WV^{-1} \qquad \text{where} \\ V &= E[\partial^2 g(\alpha, \mathcal{F}_t)/\partial\alpha_i\alpha_j] \qquad \text{and} \\ W &= E[(y_t - E(y_t \mid \mathcal{F}_{t-1})^2 \partial^2 g(\alpha, \mathcal{F}_t)/\partial\alpha_i\alpha_j] \end{aligned} \tag{3.4.3.1}$$

II

$$\begin{aligned} V - C_{ml_i} &= (U^{'})^{-1} + (U^{'})^{-1}S(U^{'})^{-1} = (U^{'})^{-1} \qquad \text{where} \\ U^{'} &= E[\partial^2 g(\alpha, \mathcal{F}_t)/\partial\alpha_i\alpha_j\{E(f_\alpha(t, t-1))\}^{-1}] \end{aligned} \tag{3.4.3.2}$$

2. Tjostheim (1986)

I

$$\begin{aligned} V - C_{cls_{ii}} &= U^{-1}RU^{-1} \qquad \text{where} \\ U &= E[\frac{\partial g(\alpha, \mathcal{F}_t)}{\partial\alpha_i^{'}}\frac{\partial g(\alpha, \mathcal{F}_t)}{\partial\alpha_j}] \qquad \text{and} \\ R &= E[\frac{\partial g(\alpha, \mathcal{F}_t)}{\partial\alpha_i^{'}} f_\alpha(t, t-1)\frac{\partial g(\alpha, \mathcal{F}_t)}{\partial\alpha_j}] \end{aligned} \tag{3.4.3.3}$$

II

$$\begin{aligned} V - C_{ml_{ii}} &= (U^{'})^{-1} + (U^{'})^{-1}S(U^{'})^{-1} = (U^{'})^{-1} \qquad \text{where} \\ U^{'} &= E[\frac{\partial g(\alpha, \mathcal{F}_t)}{\partial\alpha_i^{'}}\{E(f_\alpha(t, t-1))\}^{-1}\frac{\partial g(\alpha, \mathcal{F}_t)}{\partial\alpha_j}]. \end{aligned} \tag{3.4.3.4}$$

We show that in both cases the Variance-Covariance matrix is equal.

Scenario I:$V = \frac{\partial^2 g(\alpha, \mathcal{F}_t)}{\partial\alpha_i\alpha_j}$

Variance-Covariance Matrix for the Conditional Least Squares Penalty Method

The Variance-Covariance Matrix for the Conditional Least Squares Penalty Method, as defined in Equation (1) is derived.

Recall from Theorem 3.2.2 that the estimators $\hat{\theta}_n$ are not only consistent but also asymptotically normal with the properties defined in Equation (3.2.2.2). In this section, we compute $V^{-1}WV^{-1}$ for the MAR $(2,1,1)$ model as follows:

$$V = \begin{pmatrix} 1 & Ey_{t-1} \\ Ey_{t-1} & Ey_{t-1}^2 \end{pmatrix} \tag{3.4.3.5}$$

and

$$V^{-1} = \frac{1}{Var(y_{t-1})} \begin{pmatrix} Ey_{t-1}^2 & -Ey_{t-1} \\ -Ey_{t-1} & 1 \end{pmatrix} \tag{3.4.3.6}$$

also

$$W = \sum_{k=1} \pi_k \sigma_k^2 \begin{pmatrix} 1 & Ey_{t-1} \\ Ey_{t-1} & Ey_{t-1}^2 \end{pmatrix} \tag{3.4.3.7}$$

Putting Equations (3.4.3.6) and (3.4.3.7) together we have

$$V^{-1}W = \frac{\sum_{k=1} \pi_k \sigma_k^2}{Var(y_{t-1})} \begin{pmatrix} Var(y_t) & 0 \\ 0 & Var(y_t) \end{pmatrix} = \sum_{k=1} \pi_k \sigma_k^2 \times \mathbf{I}$$

where

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

so that

$$V^{-1}WV^{-1} = \frac{\sum_{k=1} \pi_k \sigma_k^4}{Var(y_{t-1})} \begin{pmatrix} Ey_{t-1}^2 & -Ey_{t-1} \\ -Ey_{t-1} & 1 \end{pmatrix} = V - C_{cls_i} \tag{3.4.3.8}$$

Variance-Covariance Matrix for the Maximum Likelihood Penalty Method

The Variance-Covariance Matrix for the Maximum Likelihood Penalty Method as defined in Equation (1) is derived.

Recall from Theorem 3.3.3 that the estimators $\hat{\theta}_n$ are not only consistent but also asymptotically normal with the properties defined in Equation (3.3.2.3).

In this section, we compute $(U')^{-1}+((U')^{-1}S(U')^{-1})$ for the MAR $2;1,1$ model as follows:

Recall from the proof of Theorem 3.4.2 that $S_{ij} \triangleq 0$ since $f_a(t,t-1)$ is independent of a so that the variance covariance matrix is simply given by $(U')^{-1}$ computed as follows:

$$U' = E\frac{\partial^2 g(\alpha,\mathcal{F}_{t-1})}{\partial\alpha_i\partial\alpha_j}\{E(f_\alpha(t,t-1))\}^{-1} \tag{3.4.3.9}$$
$$= \frac{1}{\sum_{k=1}\pi_k\sigma_k^2}\begin{pmatrix} 1 & Ey_{t-1} \\ Ey_{t-1} & Ey_{t-1}^2 \end{pmatrix}$$

so that

$$(U')^{-1} = \frac{\sum_{k=1}\pi_k\sigma_k^2}{Var(y_{t-1})}\begin{pmatrix} Ey_{t-1}^2 & -Ey_{t-1} \\ -Ey_{t-1} & 1 \end{pmatrix} = V - C_{ml_i} \tag{3.4.3.10}$$

Equation (3.4.3.8)= Equation (3.4.3.10) implying that the V-C matrix for both the conditional least squares and maximum likelihood penalty functions are equal.

Scenario II:$V = \frac{\partial g(\alpha,\mathcal{F}_t)}{\partial\alpha'}\frac{\partial g(\alpha,\mathcal{F}_t)}{\partial\alpha_j}$

Variance-Covariance Matrix for the Conditional Least Squares Penalty Method

The Variance-Covariance Matrix for the Conditional Least Squares Penalty Method, based on Equation (2) is derived.

$$U = E\left[\begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix}\begin{pmatrix} 1 & y_{t-1} \end{pmatrix}\right] = \begin{pmatrix} 1 & Ey_{t-1} \\ Ey_{t-1} & Ey_{t-1}^2 \end{pmatrix} \tag{3.4.3.11}$$

and

$$(U')^{-1} = \frac{1}{Var(y_{t-1})}\begin{pmatrix} Ey_{t-1}^2 & -Ey_{t-1} \\ -Ey_{t-1} & 1 \end{pmatrix} \tag{3.4.3.12}$$

also

$$R = E\sum_{k=1} \pi_k \sigma_k^2 \left[\begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix} \begin{pmatrix} 1 & y_{t-1} \end{pmatrix} \right] = \sum_{k=1} \pi_k \sigma_k^2 \begin{pmatrix} 1 & Ey_{t-1} \\ Ey_{t-1} & Ey_{t-1}^2 \end{pmatrix} \tag{3.4.3.13}$$

Putting Equation 3.4.3.12 and Equation 3.4.3.13 together we have

$$U^{-1}R = \frac{\sum_{k=1} \pi_k \sigma_k^2}{Var(y_{t-1})} \begin{pmatrix} Var(y_t) & 0 \\ 0 & Var(y_t) \end{pmatrix} = \sum_{k=1} \pi_k \sigma_k^2 \times \mathbf{I}$$

where

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

so that

$$U^{-1}RU^{-1} = \frac{\sum_{k=1} \pi_k \sigma_k^4}{Var(y_{t-1})} \begin{pmatrix} Ey_{t-1}^2 & -Ey_{t-1} \\ -Ey_{t-1} & 1 \end{pmatrix} \tag{3.4.3.14}$$

Variance-Covariance Matrix for the Maximum Likelihood Penalty Method

The Variance-Covariance Matrix for the Maximum Likelihood Penalty Method as defined in Equation (2) is derived.

As before, recall from the proof of Theorem 3.4.2 that $S_{ij} \triangleq 0$ since $f_a(t, t-1)$ is independent of a so that the variance covariance matrix is simply given by$(U^{'})^{-1}$ computed as follows,

$$\begin{aligned} U^{'} &= E\frac{\partial^2 g(\alpha, \mathcal{F}_t)}{\partial \alpha_i \partial \alpha_j} \{E(f_\alpha(t, t-1))\}^{-1} \\ &= \frac{1}{\sum_{k=1} \pi_k \sigma_k^2} \begin{pmatrix} 1 & Ey_{t-1} \\ Ey_{t-1} & Ey_{t-1}^2 \end{pmatrix} \end{aligned} \tag{3.4.3.15}$$

so that

$$(U^{'})^{-1} = \frac{\sum_{k=1} \pi_k \sigma_k^2}{Var(y_{t-1})} \begin{pmatrix} Ey_{t-1}^2 & -Ey_{t-1} \\ -Ey_{t-1} & 1 \end{pmatrix} = V - C_{ml_{ii}} \tag{3.4.3.16}$$

Equation (3.4.3.14) equals equation 3.4.3.16 implying that the Variance-Covariance matrix for the two penalty functions are also equal, regardless of the scenarios used.

3.5 Summary

The conditional least squares procedure provides an integrated means of handling estimation problems for commonly used stochastic models. This method stemmed out of the assumption that normally distributed error terms in autoregressive models makes maximum likelihood estimation similar to the minimization of a sum of squares. Klimko and Nelson (1978) developed an estimation procedure for stochastic processes based on the minimization of a sum of squared deviations about conditional expectations. Tjostheim (1986) extended the work by developing a general framework for analyzing estimates in nonlinear time series models. He further derived general conditions for strong consistency and asymptotic normality for both conditional least squares and maximum likelihood type penalty function estimates. We apply these techniques to the Mixture Autoregressive model. In particular, we have given an example for the MAR(2;1,1) model and have shown that for the model, the variance-covariance matrix is positive definite and identical for both the conditional least square and maximum likelihood penalty functions.

Chapter 4

Geometric Ergodicity of the Mixture Autoregressive Model

Geometric ergodicity is very useful in establishing mixing conditions and central limit results for parameter estimates of a model. It justifies the use of laws of large numbers and forms part of the basis for exploring the asymptotic theory of the model (Tjostheim (1990)).

The aim here is to show that the MAR model is geometrically ergodic and by implication satisfies the absolutely regular and strong mixing conditions.

We use the following notation.

Denote the lebesgue measure on the Borel σ- Field, $B(\mathbb{R}^p)$, of $\mathbb{R}^p$ by φ and let $\|\cdot\|$ be any vector norm on $\mathbb{R}^p$, $p > 1$.

Also, for any function f, we write

$$E_z[f(z_t)] = E[f(z_t) \mid Z_{t-1} = z] \tag{4.0.0.1}$$

4.1 Definitions of some useful concepts related to Markov chains and geometric ergodicity

Consider a state space S and a σ-field $\mathcal{F}$. Let (Y_t) be a homogenous Markov chain evolving on S, i.e. for all set $A \in \mathcal{F}$ and all $s, t \in \mathbb{N}$, the transition probability $P^t(y, A)$ is defined as,

$$P^t(y, A) := \mathbb{P}(Y_{s+t} \in A \mid Y_r, r < s; Y_s = y). \tag{4.1.0.2}$$

The *Markov Property* implies that $P^t(y, A)$ does not depend on $Y_r, r < s$, given Y_s. *Time homogeneity* refers to the fact that the transition probability does not depend on s.

A *Transition Kernel* is a function $P : S \times \mathcal{F} \to [0, 1]$ with the following properties,

1. For all $A \in \mathcal{F}$, the transition probability, $P(\cdot, A)$ is measurable;

2. For every $y \in S, P(y, \cdot)$ is a probability measure on $(S, \mathcal{F})$.

The law of each $\{Y_t\}$ is determined by the initial probability measure π and the transition kernel P. Set

$$\begin{aligned}&\mathbb{P}_\pi(Y_0 \in A_0, \cdots, Y_t \in A_t) = \\ &\qquad \int_{y_0 \in A_0} \cdots \int_{y_{t-1} \in A_{t-1}} \pi(dy_0) P(y_0, dy_1) \cdots P(y_0, dy_{t-1}) P(y_{t-1}, A_t)\end{aligned}$$

The Markov chain Y_t is said to be $\varphi-$*irreducible* for a nontrivial (that is not identically equal to 0) measure φ on $(S, \mathcal{F})$, if

$$\forall A \in \mathcal{F},\ \varphi(A) > 0 \ \Rightarrow\ \forall y \in S,\ \exists t > 0,\ P^t(y, A) > 0. \tag{4.1.0.3}$$

If Y_t is $\varphi-$irreducible for some φ, Y_t is simply called *irreducible* and φ is called an *irreducibility measure* for Y_t.

Denote the average time that the chain Y_t spends in A when it starts at y by $U(y, A) = S \sum_{t=1}^{\infty} \mathbb{I}_A(Y_t)$, where $\mathbb{I}_A(\dot{})$ is an indicator function.

$M(A)$ is a *maximal irreducible measure* i.e. $M(\cdot)$ is such that all the other irreducible measures on Y_t are absolutely continuous with respect to $M(\cdot)$.

A $\varphi-$irreducible chain is called *recurrent* if

$$U(y, A) := \sum_{t=1}^{\infty} P^t(y, A) = +\infty,\ \forall y \in S,\ \forall A \in \mathcal{F}^+, \tag{4.1.0.4}$$

where

$$\mathcal{F}^+ = \{A \in \mathcal{F} \mid M(A) > 0\}, \tag{4.1.0.5}$$

that is the chain Y_t spends infinite time in A when it starts at y.

A $\varphi-$irreducible chain is called *transient* if

$$\exists (A_j),\ \mathcal{F} = \bigcup_j A_j,\ U(y, A_j) \le M_j < \infty,\ \forall y \in S \tag{4.1.0.6}$$

Furthermore, Y_t is said to be *positive recurrent* if

$$\limsup_{t \to \infty} P^t(y, A) > 0, \qquad \forall y \in S, \forall A \in \mathcal{F}^+. \tag{4.1.0.7}$$

This property is equivalent to the existence of a unique *invariant probability measure* i.e. a probability measure π such that

$$\forall A \in \mathcal{F},\ \pi(A) = \int P(y, A)\pi(dy). \tag{4.1.0.8}$$

A $\varphi-$irreducible Markov chain Y_t is said to be *Harris Recurrent* (see Meyn and Tweedie (1993)) if for all $A \in \mathcal{F}$, $\varphi(A) > 0$ and $\forall y \in S$ such that, $P(L_A \mid Y_0 =$

$y) = 1$, where $L_A = \inf\{n \geq 0 : Y_n \in A\}$. That is, given that the process starts in y it will eventually reach A in a finite number of steps with probability 1.

If a Markov chain is Harris recurrent and positive recurrent, then it is called *positive Harris recurrent.*

A non-null set $C \in \mathcal{F}$ is *small* if there exists a positive integer m, a constant $n > 0$ and a nontrivial probability measure $v(\cdot)$ on A, such that

$$P^n(y, A) \geq mv(A) \qquad \forall y \in C,\ A \in \mathcal{F} \tag{4.1.0.9}$$

A set $C \in \mathcal{B}(Y)$ is called *petite (v-petite)* if the chain satisfies the bound $K(y, A) \geq v(A)$, for all $y \in C$ and $A \in \mathcal{B}(Y)$ and $K(y, A) = \sum_{n=0}^{\infty} P^n(y, A)$ (see Meyn and Tweedie (1993)).

For a small set C let

$$I(C) = \{n \in N,\ P^n(y, A) \geq mv(A)\} \qquad \forall y \in C,\ B \in \mathcal{F} \tag{4.1.0.10}$$

Let $d(C)$ be the greatest common divisor of $I(C)$. For all small sets C, if $d(C) = 1$, then the Markov chain is called *aperiodic*; otherwise, it is called periodic with period $d = d(C)$ (see Masanobu Taniguchi (2000)).

A Markov chain is called ***ergodic*** if it is irreducible, aperiodic and positive Harris recurrent. That is, there exists a probability measure π on $S, \mathcal{F}$, such that,

$$\lim_{t \to +\infty} \|P^t(y, \cdot) - \pi(\cdot)\| = 0,\ y \in S, \tag{4.1.0.11}$$

where $\|\cdot\|$ here is the total variation norm (see Meyn and Tweedie (1993)).

Ergodicity is however not a sufficient condition for establishing $\beta-$ mixing properties. We discuss an additional condition for $\beta-$mixing below.

4.2 Geometric Ergodicity

The chain (Y_t) is called *geometrically ergodic* if there exists a positive constant $\rho < 1$ such that,

$$\lim_{t \to +\infty} \rho^{-t} \| P^t(y, \cdot) - \pi(\cdot) \| = 0, \; \forall y \in S \tag{4.2.0.12}$$

Recurrence, existence of an invariant probability measure and $\varphi-$irreduciblity properties are not generally easily verified for all models. Tjostheim (1990) and Meyn and Tweedie (1993) suggest exploring the use of the *drift condition* for proving geometric ergodicity.

The following theorem is due to Tjostheim (1990) and Meyn and Tweedie (1993).

Theorem 4.2.1. ***(Geometric ergodicity)*** *Suppose that the Markov process Y_t is aperiodic and $\varphi-$irreducible, suppose also that there exists a petite set A, positive constants $0 < \rho < 1$, $\varepsilon > 0$, $M < \infty$ and a non-negative measurable function $V \geq 1$ such that:*

$$E[V(Y_t) \mid y_{t-1} = y] \leq \rho V(y) - \varepsilon, \; y \in A^c \tag{4.2.0.13}$$

and

$$E[V(Y_t) \mid y_{t-1} = y] \leq M, \; y \in A. \tag{4.2.0.14}$$

Then Y_t is geometrically ergodic.

The function V is said to be a ***drift criterion***, it is also referred to as a *test function* (see Tweedie (1988)).

4.2.1 Geometric Ergodicity and mixing Conditions

Mixing coefficients/conditions are defined as follows,

(Davidson (1997)) consider a sequence $\{Y_t(\omega)\}_{-\infty}^{\infty}$, let $\mathcal{F}_a^b = \sigma(Y_t, a \leq t \leq b)$, $\mathcal{L}^2(\mathcal{F}_a^b)$ be a set of $\mathcal{F}_a^b-$ measurable random variables with finite and positive definite variance.

Consider σ-algebras of events separated by at least m time units,

$$\begin{aligned} \mathcal{F}_{t+m}^{\infty} &= \sigma(Y_{t+m}, Y_{t+m+1}, Y_{t+m+2}, \dots) \qquad \text{and} \\ \mathcal{F}_{-\infty}^{t} &= \sigma(\dots, Y_{t-2}, Y_{t-1}, Y_t) \end{aligned} \tag{4.2.1.1}$$

The following measures have been found useful in characterising the strength of dependendence between events in these σ-algebras:

$$\alpha(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{F}_{t+m}^{\infty}, B \in \mathcal{F}_{-\infty}^{t}\}$$

$$\phi(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) = \sup\{|\mathbb{P}(A \mid B) - \mathbb{P}(A)| : A \in \mathcal{F}_{t+m}^{\infty}, B \in \mathcal{F}_{-\infty}^{t}, \mathbb{P}(B) > 0\}$$

$$\beta(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) = \sup\{E(|\mathbb{P}(A \mid \mathcal{F}_{-\infty}^{t}) - \mathbb{P}(A)|) : A \in \mathcal{F}_{t+m}^{\infty}\}$$

$$\rho(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) = \sup\{cov(X, Y) : X \in \mathcal{L}^2(\mathcal{F}_{t+m}^{\infty}), Y \in \mathcal{L}^2(\mathcal{F}_{-\infty}^{t})\}$$

The following concepts make use of the σ-algebras listed above.

The sequence is said to be $\alpha-$*mixing* or *strong mixing* if $\lim_{m\to\infty} \alpha_m = 0$, where

$$\alpha_m = \sup_t \alpha(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}). \tag{4.2.1.2}$$

The sequence is said to be $\phi-$*mixing* or *uniform mixing* if $\lim_{m\to\infty} \phi_m = 0$, where

$$\phi_m = \sup_t \phi(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}). \tag{4.2.1.3}$$

The sequence is said to be $\beta-$*mixing* or *absolutely regular* if $\lim_{m\to\infty} \beta_m = 0$, where

$$\beta_m = \sup_t \beta(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}). \tag{4.2.1.4}$$

The sequence is said to be $\rho-$*mixing* or *completely regular* if

$$\rho_m = \sup_t \rho(\mathcal{F}^t_{-\infty}, \mathcal{F}^{\infty}_{t+m}) \to 0. \tag{4.2.1.5}$$

Davidson (1997) shows that $\phi-$mixing implies $\alpha-$mixing and $\beta-$mixing condition is intermediate between $\alpha-$mixing and $\phi-$mixing conditions.

In summary, $\phi-$mixing implies $\rho-$mixing and $\beta-$mixing; also, $\beta-$mixing and $\rho-$mixing imply $\alpha-$mixing (see, Dedecker et al. (2007)).

Geometric Egodicity and $\beta-$mixing conditions

Yu.A.Davydov (1973) and Bradley (2005) show that for an ergodic Markov chain Y_t, of invariant probability measure π,

$$\beta_Y(t) = \int \|P^t(y,.) - \pi\|\pi(dy). \tag{4.2.1.6}$$

The rate ρ in Equation (4.2.0.12) can be chosen independently of the initial point. If Equation (4.2.0.12) holds then it follows that $\beta_Y(t) = O(\rho^t)$. Then (Y_t) is a stationary and geometrically ergodic and hence $\beta-$mixing.

Since $\beta-$mixing implies $\alpha-$mixing then geometric ergodicity entails both $\alpha-$mixing and $\beta-$mixing. Hence, a major consequence of geometric ergodicity is that the Markov chain Y_t is absolutely regular that is, β-mixing and hence strong mixing that is α-mixing, at a geometric rate.

4.3 Geometric Ergodicity of the MAR model

Let y_t be an MAR process defined by Equation (2.2.0.8), we show here that $Y_t = (y_t, \ldots, y_{t-p+1})'$ is geometrically ergodic and by implication β-mixing.

Remark 4.3.1. We prove the geometric ergodicity of (Y_t) instead of the process y_t, however, if Y_t is geometrically ergodic, so is y_t.

The follow assumptions are made in addition to assumption A in Section (2.2).

Assumptions B

i For each $z \in S$, there exists $c_{i(z)}, d_{i(z)} \in \mathbb{R}^p$ and $c_{i(z)} \geq 0$, $d_{i(z)} \geq 0$, $i = 1, \ldots, p$ such that for $y = (y_1, \ldots, y_p)$

(a) $|f_{z_t}(y)| \leq \sum_{i=1}^{p} c_{i(z)}|y_i| + o(\|y\|)$ as $\|y\| \to \infty$ and

(b) $\sigma^2_{z_t}(y) \leq \sum_{i=1}^{p} d_{i(z)}|y_i^2| + o(\|y\|^2)$ as $\|y\| \to \infty$.

ii **Drift Condition:** The Foster-Lyapounov drift condition (Tjostheim (1990), Meyn and Tweedie (1993))

There exists a real valued measure function $V \geq 1$ such that for some constant $\varepsilon > 0$, $0 < \rho < 1$, a constant M_1 and a small set $A = \{y \in \mathbb{R} : \|y\| \leq M_1\}$:

$$E[V(Q_t) \mid Q_{t-1} = (q)] \leq \rho V(q) \qquad \text{for } y \in A^c \tag{4.3.0.7}$$

$$\sup_{x \in A} E[V(Q_t) \mid Q_{t-1} = q] < \infty \qquad \text{for } y \in A \tag{4.3.0.8}$$

We will use the following result by Meyn and Tweedie (1993) to prove the geometric ergodicity of the MAR model.

Lemma 4.3.1 (Meyn and Tweedie (1993)). For an aperiodic, $\varphi-$irreducible Markov chain, all petite sets are small sets.
(for proof see theorem 5.5.7 of Meyn and Tweedie (1993)).

Next we prove the following result for the chain $Q_t = (Z_t, Y_t)$.

Propositon 4.3.1. For the Markov chain $Q_t = (Z_t, Y_t)$, if for every $Z \in S$, $f_z(\cdot)$ is bounded on all compact sets, then Q_t is $v \times \varphi$-irreducible and for every compact set $C \in \mathbb{R}^p$, $S \times C$ is a small set.

Proof. Now, the density function f is continuous and positive everywhere. We have that if $\varphi(A) > 0$ and C is a compact subset of $\mathbb{R}^p$, then by Bhattacharya and Lee (1999, Lemma 1) and Stockis et al. (2010)), we can write

$$\int_A g(q, y \mid z) d\varphi(q) > 0 \tag{4.3.0.9}$$

and

$$\inf_{y \in C} \int_A g(q, y \mid z) d\varphi(q) > 0 \tag{4.3.0.10}$$

For any $S^{'} \subset S$,$z = (z_1, \ldots, z_t) \in S^{'}$ and $y = (y_1, \ldots, y_t) \in S$ and $t \geq p$, and the transition probabilities of moving between alternate states z_t is

$$p_{z_0 z_1}, \ldots, p_{z_{t-1} z_t} > 0 \tag{4.3.0.11}$$

Denote Y_t given $Z_1 = z_1, \ldots, Z_t = z_t$ by $Y_t^{(z)}$.

For $Q_0 = (z_0, y)$,

$$P(Q_t \in S^{'} \times A \mid Q_0) = \sum_{z_t \in S^{'}} \sum_{z_{t-1} \in S} \cdots \sum_{z_1 \in S} p_{z_0 z_1}, \ldots, p_{z_{t-1} z_t} P(Y_t^{(z)} \in A \mid Q_0) \tag{4.3.0.12}$$

from Equation (4.3.0.9)

$$P(Y_t^{(z)} \in A \mid Y_{t-p} = q) = \int_A g(q, y \mid z_{t-p+1}, \ldots, z_t) d\varphi(y) > 0, \ \forall q \tag{4.3.0.13}$$

combining Equations (4.3.0.9)-(4.3.0.13) we have,

$$P(Y_t^{(z)} \in A \mid Y_0 = y) > 0 \tag{4.3.0.14}$$

and

$$\inf_{y \in C} P(Y_t^{(z)} \in A \mid Y_0 = y) > 0 \tag{4.3.0.15}$$

Hence, Q_t is $\varphi-$irreducible.

Furthermore, for any compact set $C \in \mathbb{R}^p$

$$\inf_{(z_0,y) \in S \times C} \sum_{n=1}^{j} P^n((z_0, y), S' \times A) > 0 \tag{4.3.0.16}$$

and $S \times C$ is a small set. which completes the proof. □

To verify the geometric erogodicity of the MAR model, we need to:

1 Prove that the process $Q_t = (Z_t, Y_t)$ is $\varphi-$irreducible and aperiodic.

2 Show the existence of a test function $V(Q_t)$ satisfying the drift condition (Equation (4.3.0.7)) above.

The two steps are summarized in the following theorem.

Theorem 4.3.1. *Consider the aperiodic Markov Chain $Q_t = (Z_t, Y_t)$. For a small set A and the aperiodic and $\varphi-$irreducible process $\{Y_t; t \geq 0\}$ such that $Y_t = (y_t, \ldots, y_{t-p+1})'$. Each y_t is an MAR process defined by Equation* (2.2.0.10). *Suppose that Assumption (2.2) and Assumption (4.3) are satisfied and*

$$\sup_z E[\sum_{j=1}^{p} c_{i(z_t)} c_{j(z_t)} + E(\epsilon_{z_t}^2) d_{i(z_t)} \mid Z_{t-1} = z] < 1 \tag{4.3.0.17}$$

Then

1. $\{Y_t; t \geq 0\}$ *is geometrically ergodic with* $V(y) = 1 + \|y\|^2$

2. $\{Y_t; t \geq 0\}$ *has a stationary distribution with finite second moments i.e.* $E_{\pi_y}[y_t^2] < \infty$ *and*

3. $\{Y_t; t \geq 0\}$ *is* β*-mixing and hence strong mixing at geometric rate.*

where π *is unique invariant distribution of* Y_t *and* $\pi_y(A) = \pi(S \times A \times \mathbb{R}^{p-1})$, $A \in B(\mathbb{R})$

Proof. Step 1: To show that the drift condition 4.3.0.7 (Equation (4.3.0.7)) is satisfied,

let $\tau_i(z) = \sum_{j=1}^{p} c_{i(z)}c_{j(z)}$ and choose $\delta > 0$ so that $\sum_{i=1}^{p} \xi_i + \delta = 1$, where

$$\xi_i = \sup_z E[\sum_{j=1}^{p} c_{i(z)}c_{j(z)} + E(\epsilon_t^2)d_{i(z)} \mid Z_{t-1} = z] < 1 \tag{4.3.0.18}$$

Now define a test function $V : S \times R^p \to R$ by

$$V(z, y) = 1 + \|y^2\| \tag{4.3.0.19}$$

write

$$\xi_i \leq (1 - \frac{\delta}{p}) \qquad 1 \leq i \leq p - 1 \tag{4.3.0.20}$$

Hence for $y = y_1, \ldots, y_p$;

$$
\begin{aligned}
E[V(Q_t) \mid Q_{t-1} = q] &= E[V(Q_t) \mid Q_{t-1} = (z, y)] \qquad (4.3.0.21)\\
&= E[(f_{z_t}(y) + \sigma_{z_t}\epsilon_t)^2 \mid Z_{t-1} = z] + 1\\
&\leq E_z[(f_{z_t}(y) + \sigma_{z_t}\epsilon_t)^2] + \sum_{i=2}^{p} y_{i-1}^2 + 1\\
&\leq \sum_{i=1}^{p} E_z[\tau_{i(z)} + E\epsilon_{z_t}^2 d_{i(z)}]y_i^2 + \sum_{i=2}^{p} y_{i-1}^2\\
&+ E_z[(2o(\|y\|)(\sum_{i=1}^{p} c_{iz})|y_i|) + (o(\|y\|))^2 + E(\epsilon_t^2)o(\|y\|^2)]\\
&+ E(\epsilon_t^2)o(\|y\|^2) + 1\\
&\leq \sum_{i=1}^{p} \xi_i y_1^2 + \sum_{i=2}^{p} y_{i-1}^2 + E_z[L_{z_t}(y)] + 1\\
&\leq y_1(1 - \frac{\delta}{p}) + \sum_{i=2}^{p} y_{i-1}^2 + E_z[L_{z_t}(y)] + 1\\
&\leq \sum_{i=1}^{p} y_i^2 - \frac{\delta}{p}\sum_{i=1}^{p} y_i^2 + E_z[L_{z_t}(y)] + 1\\
&\leq \sum_{i=1}^{p} y_i^2 - \frac{\delta}{p}\sum_{i=1}^{p} y_i^2 + E_z[L_{z_t}(y)] + 1 + \frac{\delta}{p} - \frac{\delta}{p}\\
&= (1 + \sum_{i=1}^{p} y_i^2) - \frac{\delta}{p}(1 + \sum_{i=1}^{p} y_i^2) + E_z[L_{z_t}(y)] + \frac{\delta}{p}\\
&= V(z, y) - \frac{\delta}{p}(V(z, y)) + E_z[L_{z_t}(y)] + \frac{\delta}{p}\\
&= V(z, y)\left[1 - \frac{\delta}{p} + \frac{1}{V(z, y)}\left[E_z\left[L_{z_t}(y)\right] + \frac{\delta}{p}\right]\right]
\end{aligned}
$$

where

$$
L_{z_t}(y) = (2o(\|y\|)(\sum_{i=1}^{p} c_{i(z)})|y_i|) + (o(\|y\|))^2 + E(\epsilon_t^2)o(\|y\|^2). \qquad (4.3.0.22)
$$

$\frac{E_z[L_{z_t}(y)]}{V(z,y)} \to 0$ as $\|y\| \to \infty$, also,

$\frac{\delta/p}{V(z,y)} \to 0$ as $\|y\| \to \infty$ so that we have,

$$E[V(Y_t) \mid Y_{t-1} = z, y] \leq V(z,y)[1 - \frac{\delta}{p} + \frac{\delta/p}{V(z,y)}] = \\ V(z,y)(1 - \frac{\delta}{p}) \tag{4.3.0.23}$$

Now suppose that $y \in A^c$ and there exists $M_1 > 1$ such that $\|y\| > M_1$ so that $\frac{\delta}{p} < \varepsilon < 1$, ε is a strictly positive constant defined in Equation (4.3.0.18).

choose $1 - \frac{\delta}{p} < \rho < 1$ in Equation (4.3.0.23), it follows that the first part of Equation (4.3.0.7) holds. Furthermore, since $f_{z_t}(y)$ is locally bounded for $y \in A$, the second part of Equation (4.3.0.7) holds.

Thus,

$$E[V(Y_t) \mid Y_{t-1} = (z,y)] \leq \rho V(z,y) \text{ for } y \in A^c \tag{4.3.0.24}$$
$$\sup_{y \in A} E[V(Y_t) \mid Y_{t-1} = (z,y)] < \infty \text{ for } y \in A$$

Therefore, the geometric ergodicity and hence the strict stationarity and $\beta-$mixing property of Y_t and hence, y_t are established.

We prove Theorem 4.3.1(ii) as follows, by 4.3 we can write

$$\begin{aligned} y_t^2 &\leq [\sum_{i=1}^{p} c_{i(z)}|y_{i-1}| + o(\|y\|) + (\sum_{i=1}^{p} d_{i(z)}|y_{i-1}^2| + o\|y\|^2)^{\frac{1}{2}}\epsilon_{z_t}]^2 \\ &= [(\sum_{i=1}^{p} c_{i(z)}|y_{i-1}| + o(\|y\|))^2 + \sum_{i=1}^{p} d_{i(z)}|y_{i-1}^2| + o\|y\|^2)\epsilon_{z_t}^2] + 2f_{zt(y)}\sigma_{z_t}\epsilon_{z_t} \\ &= \sum_{i=1}^{p} c_{i(z)}|y_{i-1}|c_{j(z)}|y_{j-1}| + 2\sum_{i=1}^{p} c_{i(z)}y_{i-1}o(\|y\|) + (o\|y\|)^2 \sum_{i=1}^{p} d_{i(z)}|y^2| + o\|y\|^2)\epsilon_{z_t}^2] + 2f_{zt(y)}\sigma_{z_t}\epsilon_{z_t} \\ &= \sum_{i=1}^{p}(\tau_{i(z)} + \epsilon_{z_t}^2 d_{i(z)}y_{i-1}^2) + 2\sum_{i=1}^{p} c_{i(z)}y_{i-1}o(\|y\|) + (o\|y\|)^2 + o\|y\|^2)\epsilon_{z_t}^2] + 2f_{zt(y)}\sigma_{z_t}\epsilon_{z_t} \end{aligned} \tag{4.3.0.25}$$

Taking expectation and by the independence of y_{t-1} and ϵ_{z_t} as well as z_t and ϵ_{z_t} we have,

$$Ey_t^2 \leq \sum_{i=1}^{p}(\tau_{i(z)} + E\epsilon_{z_t}^2 d_{i(z)})Ey_{i-1}^2 + L_{z_t}(y) \tag{4.3.0.26}$$

$L_{z_t}(y)$ is the same as Equation (4.3.0.22) above.

$$EY_t^2 \leq \frac{L_{z_t}(y)}{1 - [\sum_{i=1}^{p}(\tau_{i(z)} + E\epsilon_{z_t}^2 d_{i(z)})]} \tag{4.3.0.27}$$

Now by the proof of theorem 4.3.1 i and ii above, the Foster criterion F1 and F2 of Tweedie (1988) hold. Hence, by Tweedie (1988, Theorem 2), there exists a finite invariant measure π and **?**, Theorem 1(iii)itettweedie88 holds. Hence, the RHS of Equation (4.3.0.27) is finite. and $E_\pi(y_t^2) < \infty$ as required. $\square$

4.4 Summary

Geometric ergodicity is very useful in establishing mixing conditions and central limit results for parameter estimates of a model. It also justifies the use of laws of large numbers and forms part of the basis for exploring the asymptotic theory of the model. A consequence of geometric ergodicity is $\beta-$mixing.

Since $\beta-$mixing implies $\alpha-$mixing we can say that geometric ergodicity entails both $\alpha-$mixing and $\beta-$mixing. So that the absolute regularity and hence strong mixing of the Markov chain Y_t is a major consequence of geometric ergodicity.

We have established the geometric ergodicity of the MAR model and by implication show that it satisfies the absolutely regular and strong mixing conditions. In addition, we show that the process $\{y_t\}$ has a stationary distribution with finite second moments.

Chapter 5

Asymptotic Properties of the Maximum Likelihood Estimator of the Mixture Autoregressive Model

A maximum likelihood estimate associated with a sample of observations is a choice of parameters that maximizes the probability density function of the sample, called in this context the likelihood function. In this chapter, we examine the asymptotic properties of the maximum-likelihood estimates of the MAR model and show that the MLE of the MAR model is both consistent and asymptotically normal.

Given $(Y_t, t \geq 0)$, $Y_t = (y_t, \ldots, y_{t-p+1})'$ each y_t is an MAR process defined by Equation (2.2.0.10) with conditional distribution function defined in Equation (2.2.0.1) Denote by θ^0 the true value of the parameters to be estimated and $\hat{\theta}$ the maximum likelihood estimate. Let $f_\theta(\cdot \mid y, k)$ denote the conditional density of y_t given $y_{t-1}, \ldots, y_{t-p}, Z_{t,k}$, this conditional density is defined in Equation (2.2.0.12). We write $\{Y_m, \ldots, Y_n\} = Y^{(m,n)}$. Also, by the markov property, the filtering distribution of the unknown state given past information is given by,

$$\begin{aligned}\mathbb{P}(z_t = k \mid z_s, Y_s, s = 0, \ldots, t-1) &= \mathbb{P}(z_t = k \mid Y_0, Y_s, s = 0, \ldots, t-1) \\ &= \mathbb{P}(z_t = k \mid z_{t-1}) \quad \text{for } k = 1, \ldots, g,.\end{aligned} \tag{5.0.0.1}$$

The conditional likelihood function of $Y^{(1,n)}$ given both Y_0 and $Z_0 = z_0$ is given as,

$$p_\theta(Y^{(1,n)} \mid Y_0, Z_0 = z_0) = \sum_{z_n=1}^{g} \cdots \sum_{z_1=1}^{g} \prod_{t=1}^{n} a_{z_{t-1},z_t} f_\theta(Y_t \mid Y_{t-1}, z_t) \tag{5.0.0.2}$$

where a_{ij} is the transition probability matrix such that $P(Y_t = i \mid Y_{t-1} = j)$. The corresponding conditional log-likelihood function is,

$$l_n(\theta, z_0) = \log p_\theta(Y^{(1,n)} \mid Y_0, Z_0 = z_0) = \sum_{t=1}^{n} \log p_\theta(Y_t \mid Y^{(0,t-1)} Y_0, Z_0 = z_0). \tag{5.0.0.3}$$

where

$$\begin{aligned} &p_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0) \\ &\quad = \sum_{z_{t-1}=1}^{g} \sum_{z_t=1}^{g} f_\theta(Y_t \mid Y_{t-1}, z_t) a_{z_{t-1},z_t} \mathbb{P}(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}, Z_0 = z_0) \end{aligned} \tag{5.0.0.4}$$

Similarly, the conditional log-likelihood function given Y_0 only is,

$$l_n(\theta) = \sum_{t=1}^{n} \log p_\theta(Y_t \mid Y^{(0,t-1)}) \tag{5.0.0.5}$$

where

$$p_\theta(Y_t \mid Y^{(0,t-1)}) = \sum_{z_{t-1}=1}^{g} \sum_{z_t=1}^{g} f_\theta(Y_t \mid Y_{t-1}, z_t) a_{z_{t-1},z_t} \mathbb{P}(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}) \tag{5.0.0.6}$$

Douc et al. (2004, Corollary 1) show that the total variation distance between the filtering probabilities $\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y_0)$ and $\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y_0, Z_0 = z_0)$ tends to zero exponentially fast as $t \to \infty$ uniformly with respect to z_0.

We will show that the following conditions hold for the MAR model.

Assumptions C

C1 $\{Y_t, t \geq 0\}$ is geometrically ergodic

C2 For all $y, y^{'} \in \mathbb{R}^p$, y' is a vector of past values of y.

$$\inf_{\theta} f_\theta(y \mid y^{'}) > 0, \qquad \sup_{\theta} f_\theta(y \mid y^{'}) < \infty \tag{5.0.0.7}$$

C3

$$b_+ = \sup_{\theta} \sup_{y,y^{'}} f_\theta(y \mid y^{'}, k) < \infty \tag{5.0.0.8}$$

and

$$E|\log \inf_{\theta} f_\theta(Y_1 \mid Y_0)| < \infty \tag{5.0.0.9}$$

Lemma 5.0.1. Let (y_t) be an MAR process and $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$. Then Assumption C holds.

Proof. 1. The geometric ergodicity of Y_t has been established in Section 4.3.

2. Assume

$$f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i})}{\sigma_k} \leq 1. \tag{5.0.0.10}$$

Choose a positive constant M such that for $\sigma_k^2 > 0$, let $\sigma_k^2 \geq M^2$. Then, by Equation 2.2.0.12 we can write,

$$\begin{aligned} f_\theta(y \mid y^{'}, k) &\leq \sum_{k=1}^{g} \frac{\pi_k}{M} \\ &= \frac{1}{M} \sum_{k=1}^{g} \pi_k = \frac{1}{M} \qquad (\text{since } \textstyle\sum_{k=1}^{g} \pi_k = 1) \\ &\text{this implies that } f_\theta(y \mid y^{'}, k) \leq \frac{1}{M} \\ &\text{which then implies that } f_\theta(y \mid y^{'}, k) \leq \frac{1}{\sigma_k} \end{aligned}$$

so that for all $y, y^{'} \in \mathbb{R}$, we can write $f_\theta(y, \mid y^{'}) \le \frac{1}{\sigma_k}$.

Furthermore by the compactness of Θ, we can choose $M > 0$ such that for $k = 1, \ldots, g$, $\phi_{k,0}^2, \phi_{k,i}^2, \sigma_k^2 \le M^2$. Then,

$$(y - \phi_{k,0}^2 - \phi_{k,i}^2 y^{'})^2 \le (|y + \phi_{k,0}^2 + \phi_{k,i}^2 y^{'})|^2 \le (|y| + M|y^{'}|)^2, \quad (5.0.0.11)$$

and

$$0 \le \sigma_k \le M(1 + |y^{'}|). \quad (5.0.0.12)$$

So that for all $\theta \in \Theta$,

$$\begin{aligned} f_\theta(y \mid y^{'}) &\ge \max_{k=1,\ldots,g} \frac{\pi_k}{\sigma_k} f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i})}{\sigma_k} \quad (5.0.0.13) \\ &\ge \max_{k=1,\ldots,g} \frac{\pi_k}{M(1 + |y^{'}|)} f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i})}{\sigma_k} \\ &\ge \frac{1}{g} \frac{1}{M(1 + |y^{'}|)} f_k(\frac{|y| + M|y^{'}|}{\sigma_k}) > 0. \end{aligned}$$

$\max_{k=1,\ldots,g} \pi_k \ge \frac{1}{g}$ and $\sum_{k=1}^{g} \pi_k = 1$ so that Assumption C2 follows.

3. By the definition of $f_\theta(\cdot)$ and proof of Assumption C2, Equation b_+ is trivially dominated by a positive constant thus the first part of Assumption C3 holds. To prove the second part,

$$\frac{1}{\sigma_k} \ge \inf_\theta f_\theta(Y_1 \mid Y_0) \ge \frac{1}{gM} \cdot \frac{1}{1 + |Y_0|} f_k(\frac{(|Y_1| + M|Y_0|}{\sigma_k}) > 0 \quad (5.0.0.14)$$

So that

$$\begin{aligned} \log(\inf_\theta f_\theta(Y_1 \mid Y_0)) &\ge \log|\frac{1}{gM}| + \log(\frac{1}{1 + |Y_0|}) + \log(f_k(\frac{|Y_1| + M|Y_0|}{\sigma_k})) > 0 \\ & \quad (5.0.0.15) \\ &= -\log|gM| - \log(1 + |Y_0|) + \log(f_k(\frac{|Y_1| + M|Y_0|}{\sigma_k})) \\ &\ge -\log|gM| - 0 + \log(f_k(\frac{|Y_1| + M|Y_0|}{\sigma_k})) > -\infty \end{aligned}$$

using the fact that $EY_t^2 < \infty$, $E\log(1+|Y_0|) \leq E|Y_0| < \infty$, hence the second part of Assumption C3 follows.

□

5.1 Consistency of the maximum likelihood estimator of the MAR model

Proving consistency of the maximum likelihood estimator of the MAR model involves checking that the limit of the normalized log-likelihood is only maximized at the true value of the parameter (θ^0) that is, $l(\theta) \leq l(\theta^0)$. We start by stating some useful lemmas and propositions. The following Lemma is due to Douc et al. (2004),

Lemma 5.1.1. Given assumption C1 and C2 above hold, for $(Y_t, t \geq 0)$, $Y_t = (y_t, \ldots, y_{t-p+1})'$ each y_t being an MAR process. Then the following holds for all $\theta \in \Theta$.

$$\sup_{\theta\in\Theta} |l_n(\theta, z_0) - l_n(\theta)| \leq \frac{1}{(1-\rho)^2} \qquad \text{a.s for some } 0 \leq \rho < 1 \tag{5.1.0.16}$$

where

$$\begin{aligned} l_n(\theta, z_0) &= \log P_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0) \qquad \text{and} \\ l_n(\theta) &= \log P_\theta(Y_t \mid Y^{(0,t-1)}), \end{aligned} \tag{5.1.0.17}$$

$$\rho = 1 - \frac{\mu_-}{\mu_+},$$

$$0 \leq \mu_- = \inf_\theta \inf_{i,j} a_{i,j} \quad \text{and}$$

$$\mu_+ = \sup_\theta \sup_{i,j} a_{i,j} < 1.$$

(a_{ij}) is the transition probability matrix and is such that $\sum_j a_{ij} = 1$

Proof. By Douc et al. (2004, Corollary 1), the total variation distance between the filtering predictions $\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y_0)$ and $\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y_0, Z_0 = z_0)$ is bounded by ρ^{t-1}. So that for $t \geq 1$ we have,

$$|P_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0) - P_\theta(Y_t \mid Y^{0,t-1}| = \sum_{z_{t-1}=1}^{K} \sum_{z_t=1}^{K} f_\theta(y_t \mid y_{t-1}) a_{z_{t-1}} a_{z_t}$$
$$\times [\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}, Z_0 = z_0) - \mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y^{0,t-1})]$$
$$\leq \rho^{t-1} \sup_{z_t-1} \sum f_\theta(y_t \mid y_{t-1}, z_t) a_{z_{t-1}} a_{z_t} \leq \rho^{t-1} \mu_+ \sum f_\theta(y_t \mid y_{t-1}, z_t) \quad (5.1.0.18)$$

Now,

$$P_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0) = \sum_{z_{t-1}=1}^{g} \sum_{z_t=1}^{g} f_\theta(y_t \mid y_{t-1}) a_{z_{t-1}} a_{z_t}$$
$$\times \mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}, Z_0 = z_0) \geq \mu_- \sum f_\theta(y_t \mid y_{t-1}, z) \quad (5.1.0.19)$$

Similarly,

$$P_\theta(Y_t \mid Y^{(0,t-1)}) = \sum_{z_{t-1}=1}^{K} \sum_{z_t=1}^{K} f_\theta(y_t \mid y_{t-1}) a_{z_{t-1}} a_{z_t}$$
$$\times \mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}) \geq \mu_- \sum f_\theta(y_t \mid y_{t-1}, z) \quad (5.1.0.20)$$

By the inequality, $|\log x - \log y| \leq \frac{|x-y|}{(x \wedge y)}$ where $x \wedge y = \min(x, y)$. We can write

$$|\log P_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0) - \log P_\theta(Y_t \mid Y^{(0,t-1)})| \leq \frac{\rho^{t-1}}{1-\rho} \quad (5.1.0.21)$$

summing over all t, the right hand side of Equation (5.1.0.21) gives

$$\sum_{t=1}^{n} \frac{\rho^{t-1}}{1-\rho} = \frac{1}{(1-\rho)^2} \quad (5.1.0.22)$$

so that we have,

$$\sup_{\theta \in \Theta} | \, l_n(\theta, z_0) - l_n(\theta) \, | \leq \frac{1}{(1-p)^2} \quad (5.1.0.23)$$

as required, which completes the proof. □

The next step is to show that $\frac{1}{n}l_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\log P_\theta(Y_t \mid Y^{(0,t-1)})$ can be approximated by $\frac{1}{n}\sum_{t=1}^{n}\log P_\theta(Y_t \mid Y^{(-\infty,t-1)})$, where $\frac{1}{n}\sum_{t=1}^{n}\log P_\theta(Y_t \mid Y^{(-\infty,t-1)})$ is the sample mean of observations from a two-sided stationary ergodic sequence of random variables in $L^{'}$. We summarize this in the following corollary.

Corollary 5.1.1. Given that the process Y_t satisfies Assumption C. Then for all z_0 and $\theta \in \Theta$, the following holds,

$$\lim_{n\to\infty}\frac{1}{n}l_n(\theta, z_0) = l(\theta). \tag{5.1.0.24}$$

Proof. We adapt the following notation from Douc et al. (2004),

$$\begin{aligned}\Delta_{t,m,z}(\theta) &= \log P_\theta(Y_t \mid Y^{(-m,t-1)}, Z_{-m} = z_{-m}) \qquad \text{and} \\ \Delta_{t,m}(\theta) &= \log P_\theta(Y_t \mid Y^{(-m,t-1)})\end{aligned} \tag{5.1.0.25}$$

so that

$$l_n(\theta) = \sum_{t=1}^{n}\Delta_{t,0}(\theta) \tag{5.1.0.26}$$

Douc et al. (2004, Lemma 3) show that $\Delta_{t,m,z}(\theta)$ and $\Delta_{t,m}(\theta)$ are uniform Cauchy sequence and converge uniformly with respect to θ a.s. They also show that they are uniformly bounded in L^1 for all m and that $\lim_{m\to\infty}\Delta_{t,m,z}(\theta) = \Delta_{t,\infty}(\theta)$. They say that the inequality does not depend on z and is a stationary ergodic process such that the following inequalities hold,

$$\begin{aligned}\sup_\theta\sup_z|\Delta_{t,m,z}(\theta) - \Delta_{t,m^{'},z^{'}}| &\le \frac{\rho^{t+(m\wedge m^{'})-1}}{1-\rho} \qquad \text{and} \\ \sup_\theta\sup_z|\Delta_{t,m,z}(\theta) - \Delta_{t,m}| &\le \frac{\rho^{t+m-1}}{1-\rho}\end{aligned} \tag{5.1.0.27}$$

Setting $m = 0$ and $m^{'} \to \infty$ in the system of Equations (5.1.0.27) gives

$$\begin{aligned}\sup_\theta|\Delta_{t,0,z}(\theta) - \Delta_{t,\infty}| &\le \frac{\rho^{t-1}}{1-\rho} \qquad \text{and} \\ \sup_\theta|\Delta_{t,0,z}(\theta) - \Delta_{t,0}| &\le \frac{\rho^{t-1}}{1-\rho}.\end{aligned} \tag{5.1.0.28}$$

Pulling them together and summing over all t we have,

$$\sum_{t=1}^{n} \sup_{\theta} |\Delta_{t,0}(\theta) - \Delta_{t,\infty}| \leq \frac{2}{(1-\rho)^2} \text{ a.s.} \tag{5.1.0.29}$$

Thus by Equation (5.1.0.29) $\frac{1}{n} l_n(\theta)$ can be approximated by the sample mean of a stationary ergodic sequence, uniformly with respect to $\theta \in \Theta$ (Douc et al. (2004)), so that by the ergodic theorem we can write,

$$\frac{1}{n} l_n(\theta) \to l(\theta) = \mathbb{E}\Delta_{0,\infty}(\theta) \quad \text{a.s.} \tag{5.1.0.30}$$

This together with Lemma 5.1.1 imply that for $\theta \in \Theta$,

$$\lim_{n\to\infty} \frac{1}{n} l_n(\theta, z_0) = l(\theta) \quad \text{a.s.} \tag{5.1.0.31}$$

$\square$

For the MAR process, at any initial point z_0, $\frac{1}{n}(l_n(\theta, z_\theta) - l_n(\theta)) \to 0$ uniformly with respect to $\theta \in \Theta$ due to the uniform forgetting of the conditional Markov chain (Douc et al. (2004)).

Hence, $\hat{\theta}_{n,z_0}$ and $\hat{\theta}_n$ are asymptotically equivalent and are the maximum of $l(\theta_{n,z_0})$ and $l(\theta_n)$ respectively. The following proposition summarizes this idea.

Propositon 5.1.1. Assume that Assumption C above holds for the MAR process y_t, then

$$\sup_{\theta} \sup_{1 \leq z_0 \leq k} \left| \frac{1}{n} l_n(\theta, z_0) - l(\theta) \right| \to 0 \quad \text{as } n \to \infty \tag{5.1.0.32}$$

Proof. The proof of this proposition consists of proving the existence of a deterministic function $l(\theta)$ such that asymptotically $\frac{1}{n} l_n(\theta, z_0) \to l(\theta)$ a.s. uniformly with respect to $\theta \in \Theta$. This is implied by Corollary 5.1.1.

We now show that Equation (5.1.0.32) holds.

By Lemma 5.1.1 and the compactness of Θ it is sufficient to prove that for all $\theta \in \Theta$

$$\limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{|\theta'-\theta|\le\delta}|\frac{1}{n}l_n(\theta^{'}) - l(\theta)| = 0, \text{ a.s. } \delta \ge 0, \tag{5.1.0.33}$$

which can be broken down into,

$$\begin{aligned}
&\limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{|\theta'-\theta|\le\delta}|\frac{1}{n}l_n(\theta^{'}) - l(\theta)| \qquad (5.1.0.34)\\
&= \limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{|\theta'-\theta|\le\delta}|\frac{1}{n}l_n(\theta^{'}) - \frac{1}{n}l_n(\theta)|\\
&\le \limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{|\theta'-\theta|\le\delta}\frac{1}{n}\sum_{t=1}^{n}|\Delta_{t,0}(\theta^{'}) - \Delta_{t,\infty}(\theta^{'})|\\
&+ \limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{|\theta'-\theta|\le\delta}\frac{1}{n}\sum_{t=1}^{n}|\Delta_{t,\infty}(\theta^{'}) - \Delta_{t,\infty}(\theta)|\\
&+ \limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}|\Delta_{t,\infty}(\theta) - \Delta_{t,0}(\theta)|
\end{aligned}$$

By Corollary 5.1.1 and Equation (5.1.0.29), the first and third terms can be equated to 0.

Also, by the ergodic theorem and the following equality from Douc et al. (2004, Lemma 4)

$$\lim_{\delta\to 0}\mathbb{E}_{\theta^*}\left[\sup_{|\theta'-\theta|\le\delta}|\Delta_{0,\infty}(\theta^{'}) - \Delta_{0,\infty}(\theta)|\right] = 0. \tag{5.1.0.35}$$

The second term proceeds as follows,

$$\begin{aligned}
&\limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{|\theta'-\theta|\le\delta}\frac{1}{n}\sum_{t=1}^{n}|\Delta_{t,\infty}(\theta^{'}) - \Delta_{t,\infty}(\theta)| \qquad (5.1.0.36)\\
&\le \limsup_{\delta\to 0}\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sup_{|\theta'-\theta|\le\delta}|\Delta_{t,\infty}(\theta^{'}) - \Delta_{t,\infty}(\theta)|\\
&= \limsup_{\delta\to 0}\mathbb{E}_{\theta^0}\left[\sup_{|\theta'-\theta|\le\delta}|\Delta_{0,\infty}(\theta^{'}) - \Delta_{0,\infty}(\theta)|\right] = 0 \qquad \text{a.s.}
\end{aligned}$$

which concludes the proof. $\square$

We now show that the stationary laws of the observed process associated with two different values of the parameters (say $\mathbb{P}^Y_\theta, \mathbb{P}^Y_{\theta^0}$) do not coincide unless the parameters do (Douc et al. (2004)). To do this, we will need the following assumptions:

Assumption D

(d)

$$\theta = \theta^0 \text{implies that} \quad E\left[\log \frac{P_\theta(y^{(1,p)} \mid y_0)}{P_{\theta^0}(y^{(1,p)} \mid y_0)}\right] = 0 \qquad \text{for all } p \geq 1 \quad (5.1.0.37)$$

$(d^{'})$

$$E\left[E_{\theta^0}\left[\log \frac{P_\theta(y^{(1,p)} \mid y_0)}{P_{\theta^0}(y^{(1,p)} \mid y_0)} \mid y_0\right]\right] = 0 \qquad \text{for all } p \geq 1 \qquad (5.1.0.38)$$

We summarize the idea in the following lemma.

Lemma 5.1.2. Let (y_t) be an MAR process and $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$. Assume that Assumption C holds, then Assumption D holds and

$$\mathbb{P}^Y_\theta = \mathbb{P}^Y_{\theta*} \quad \text{implies} \quad E\left[\log \frac{P_{\theta^0}(y^{(1,p)} \mid y_0)}{P_{\theta^0}(y^{(1,p)} \mid y_0)}\right] = 0 \qquad \text{for all } p \geq 1 \qquad (5.1.0.39)$$

Proof. The proof is the same as the proof of Douc et al. (2004, Lemma 6). For $p \geq 1$ and $m \geq 0$,

$$E_{\theta^0}\left[\log \frac{P_{\theta^0}(Y^{(1,p+m)} \mid Y_0)}{P_\theta(Y^{(1,p+m)} \mid Y_0)}\right] \qquad (5.1.0.40)$$

$$= E_{\theta^0}\left[\log \frac{P_{\theta^0}(Y^{(1,m)} \mid Y^{(m+1,p+m)}, Y_0)}{P_\theta(Y^{(1,m)} \mid^{(m+1,p+m)}, Y_0)}\right] + E_{\theta^0}\left[\log \frac{P_{\theta^0}(Y^{(m+1,p+m)} \mid Y_0)}{P_\theta(Y^{(m+1,p+m)} \mid Y_0)}\right]$$

The two terms on the RHS are non-negative as they are expectations of the *Kullback-Leibler* divergence functions (see Leroux (1992). Thus,

$$0 \geq E_{\theta^0}\left[\log \frac{P_{\theta^0}(Y^{(m+1,p+m)} \mid Y_0)}{P_{\theta^0}(Y^{(m+1,p+m)} \mid Y_0)}\right] = E_{\theta^0}\left[\log \frac{P_{\theta^0}(Y^{(1,p)} \mid Y_{-m})}{P_\theta(Y^{(1,p)} \mid Y_{-m})}\right] \qquad (5.1.0.41)$$

$$= E_{\theta^0}\left[\sum_{t=1}^{n} \log \frac{P_{\theta^0}(Y^{(1,p)} = y^{(1,p)} \mid Y_{-m})}{P_\theta(Y^{(1,p)} = y^{(1,p)} \mid Y_{-m})} P_{\theta^0}(Y^{(1,p)} = y^{(1,p)} \mid Y_{-m}) v^{\otimes p}\right]$$

So that, for all $m \geq 0$,

$$P_{\theta^0}(Y^{(1,p)} \mid Y_{-m}) = P_\theta(Y^{(1,p)} \mid Y_{-m}) \text{ , a.s.} \qquad (5.1.0.42)$$

Furthermore, by

$$\lim_{j\to\infty} \sup_{i\leq j} |p_\theta(Y^{(t,l)} \mid Y^{(i,j)} - p_\theta(Y^{(t,l)}| = 0 \text{ (see Douc et al. (2004) lemma 5 for proof).} \qquad (5.1.0.43)$$

write

$$\begin{aligned} &|P_{\theta^0}(Y^{(1,p)}) - P_\theta(Y^{(1,p)}) \\ &\qquad = \lim_{m\to\infty} |P_{\theta^0}(Y^{(1,p)} \mid Y_{-m}) - P_\theta(Y^{(1,p)} \mid Y_{-m}) = 0| \text{ a.s.} \end{aligned} \qquad (5.1.0.44)$$

and $P_{\theta^0}(Y^{(1,p)}) = P_\theta(Y^{(1,p)})$ as required. □

Propositon 5.1.2. Given Assumption C as well as Assumption D and d' above holds, then $l(\theta) = l(\theta^0)$ implies that $\theta = \theta^0$

Proof. By the dominated convergence theorem:

$$\begin{aligned} l(\theta) = E_{\theta^0}\left[\lim_{m\to\infty} \log p_\theta(Y_1 \mid Y^{(-m,0)})\right] &= \lim_{m\to\infty} E_{\theta^0}\left[\log p_\theta(Y_1 \mid Y^{(-m,0)})\right] \\ &= \lim_{m\to\infty} E_{\theta^0}\left[E_{\theta^0}\left[\log p_\theta(Y_1 \mid Y^{(-m,0)}) \mid Y^{(-m,0)}\right]\right] \end{aligned} \qquad (5.1.0.45)$$

So that $l(\theta) - l(\theta^0)$ is non-negative as the limit of the expectations of conditional *Kullback Leibler divergence functions* between θ and θ^0 (see Leroux (1992) and Francq and Roussignol (1998, Lemma 8)). θ^0 is a maximizer of the function $\theta \to l(\theta)$ (Douc et al. (2004)).

Now for all $t \geq 1$ and $m \geq 0$, we have that

$$E_{\theta^0}[\log p_\theta(Y^{(1,t)} \mid Y^{(-m,0)})] = \sum_{i=1}^{t} E_{\theta^0}[\log p_\theta(Y_1 \mid Y^{(-m-i+1,0)})], \qquad (5.1.0.46)$$

taking limits as $m \to \infty$ we have

$$\lim_{m\to\infty} \frac{1}{t} E_{\theta^0}[\log p_\theta(Y^{(1,t)} \mid Y^{(-m,0)})] = l(\theta). \tag{5.1.0.47}$$

So that

$$\begin{aligned} t(l(\theta^0) - l(\theta)) &= \lim_{m\to\infty} E_{\theta^0}\left[\log \frac{p_{\theta^0}(Y^{(1,t)} \mid Y^{(-m,0)})}{p_\theta(Y^{(1,t)} \mid Y^{(-m,0)})}\right] \\ &\geq \limsup_{m\to\infty} E_{\theta^0}\left[\log \frac{p_{\theta^0}(Y^{(t-p+1,t)} \mid Y_{t-p}, Y^{(-m,0)})}{p_\theta(Y^{(t-p+1,t)} \mid Y_{t-p}, Y^{(-m,0)})}\right] \\ &= \limsup_{m\to\infty} E_{\theta^0}\left[\log \frac{p_{\theta^0}(Y^{(1,p)} \mid Y_0, Y^{(p-t-m,p-t)})}{p_\theta(Y^{(1,p)} \mid Y_0, Y^{(p-t-m,p-t)})}\right] \end{aligned} \tag{5.1.0.48}$$

Note that for $p \geq 1$ and $\theta \in \Theta$, we have,

$$\begin{aligned} \lim_{t\to\infty} \sup_{m\geq t} \left| E_{\theta^0}\left[\log \frac{p_{\theta^0}(Y^{(1,p)} \mid Y_{-m}, Y^{(-m,-t)})}{p_\theta(Y^{(1,p)} \mid Y_{-m}, Y^{(-m,-t)})}\right] - E_{\theta^0}\left[\log \frac{p_{\theta^0}(Y^{(1,p)} \mid Y_0)}{p_\theta(Y^{(1,p)} \mid Y_0)}\right]\right| \\ = 0. \qquad \text{(see Douc et al. (2004) for proof)} \end{aligned} \tag{5.1.0.49}$$

Taking limits in Equation (5.1.0.48) as $t \to \infty$ and applying Equation (5.1.0.49) gives,

$$\lim_{t\to\infty} \limsup_{m\to\infty} E_{\theta^0}\left[\log \frac{p_{\theta^0}(Y_1, \cdots, Y_p \mid Y_0, Y_{p-t-m}, \cdots, Y_{p-t})}{p_\theta(Y_1, \cdots, Y_p \mid Y_0, Y_{p-t-m}, \cdots, Y_{p-t})}\right] = 0 \tag{5.1.0.50}$$

which concludes the proof. $\square$

The consistency of the MAR model is formally stated in the following theorem:

Theorem 5.1.1. *Let $Y_t = (y_t, \ldots, y_{t-p+1})'$, each y_t is an MAR model defined in Equation* (2.2.1)*. Given Assumptions 2.2, Assumption C and Assumption D hold. Then for any $z_0 \in 1, \ldots, g$*

$$\lim_{n\to\infty} \hat{\theta}_{n,z_0} = \theta^0 \ a.s., \tag{5.1.0.51}$$

where, $\hat{\theta}_{n,z_0} = arg \max_{\theta\in\Theta} l_n(\theta, z_0)$ is the maximum likelihood estimator of θ.

Proof. The proof of the above theorem like most of this chapter largely uses the results in Douc et al. (2004).Their assumptions/proofs are hinged on the paper by Wald (1949) which says that there exists a deterministic asymptotic criterion function $l(\theta)$ such that $n^{-1}l_n(\theta, z_0) \to l(\theta)$ a.s. uniformly with respect to $\theta \in \Theta$.

The conditional form of the log likelihood function that is $l_n(\theta, z_0)$ is considered instead of $l(\theta)$

For the MAR model, at any initial point z_0, $\frac{1}{n}(l_n(\theta, z_0) - l_n(\theta)) \to 0$ uniformly with respect to $\theta \in \Theta$ this follows from the proof of Proposition 5.1.0.32. The proposition also establishes the consistency of the conditional log-likelihood of the model.

Furthermore, the geometric ergodicity of the chain Y_t (and by implication the process y_t) establishes the β-mixing (see Section 4.3) property and hence absolute regularity of the process y_t so that Equation (5.1.0.49) is established. This together with Propositions 5.1.0.32 and Proposition5.1.2 as well as the identifiably condition established by Assumption D and Lemma 5.1.2 prove the consistency of the maximum likelihood estimators of the MAR model. □

5.2 Asymptotic normality of the maximum likelihood estimator of the MAR model

To prove asymptotic normality we need some further assumptions.

Assumptions E Let $\Theta^0 \subset \Theta$ denote an open neighbourhood of θ^0 contained in Θ.

For $\delta \geq 0$, let $\Theta^0 = \{\theta \in \Theta : |\theta - \theta^0| < \delta\}$, the following conditions hold:

(a) For all $k, l \in \{1, \ldots, g\}$ and $y, y' \in \mathbb{R}$, the function $\theta \to a_{kl}(\theta)$ and $\theta \to f_\theta(y \mid$

$y^{'}, k)$ are twice continuously differentiable on Θ^0.

(b)

$$\sup_{\theta\in\Theta^0}\sup_{k,l}\|\frac{\partial \log a_{kl}(\theta)}{\partial\theta}\| < \infty \quad \text{and} \tag{5.2.0.52}$$

$$\sup_{\theta\in\Theta^0}\sup_{k,l}\|\frac{\partial^2 \log a_{kl}(\theta)}{\partial\theta\partial\theta'}\| < \infty \tag{5.2.0.53}$$

(c)

$$E\{\sup_{\theta\in\Theta^0}\sup_{k}\|\frac{\partial \log f_\theta(Y_1 \mid Y_0, k)}{\partial\theta}\|^2 < \infty\} \quad \text{and} \tag{5.2.0.54}$$

$$E\{\sup_{\theta\in\Theta^0}\sup_{k}\|\frac{\partial^2 \log f_\theta(Y_1 \mid Y_0, k)}{\partial\theta\partial\theta'}\| < \infty\}. \tag{5.2.0.55}$$

Lemma 5.2.1. Let $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$, each y_t is an MAR model defined in Equation (2.2.1). Given that Assumptions 2.2 and C hold, then Assumptions E follows.

Proof. 1. The first parts of Assumption E (a) and (b) follow from the fact that the transition probabilities $a_{kl}(\theta)$ are parameters themselves and linear functions of the model parameters when $l = g$.

From Equation (2.2.0.9) both $\frac{\pi_k}{\sigma_{z_t}}$ and $f_k\left(\frac{y_t-\phi_{k,0}-\sum_{i=1}^{p_k}\phi_{k,i}y_{t-i}}{\sigma_k}\right)$ are continuous,measurable, finite, positive and differentiable functions. The function $f_\theta(y, y^{'}, k)$ is a continuous, measurable, finite, positive and differentiable function, so that the second parts of Assumption E (a) and (b) hold.

2. To validate Assumption E (c), it is sufficient to examine the first and second order partial derivatives with respect to each of the parameters that make

up θ. Write

$$H_k(\theta) = \log f_\theta(Y_t \mid Y_{t-1}, k) = \log \sum_{k=1}^{g} \frac{\pi_k}{\sigma_{z_t}} f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i}}{\sigma_k})$$
$$= \sum_{k=1}^{g} Z_{k,t} \log \pi_k - \sum_{k=1}^{g} Z_{k,t} \log \sigma_k + \sum_{k=1}^{g} Z_{k,t} \log f_k \left(\frac{e_{k,t}}{\sigma_k}\right) \quad (5.2.0.56)$$

where $e_{k,t} = y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i}$. The first order derivatives are as follows,

$$\frac{\partial H_k(\theta)}{\partial \pi_k} = \sum_{t=p+1}^{n} \left(\frac{Z_{k,t}}{\pi_k} - \frac{Z_{k,t}}{\pi_g}\right) \quad (5.2.0.57)$$
$$\frac{\partial H_k(\theta)}{\partial \phi_{k,0}} = \sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k^2} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)$$
$$\frac{\partial H_k(\theta)}{\partial \phi_{k,i}} = \sum_{t=p+1}^{n} \frac{Z_{k,t} y_{t-i}}{\sigma_k^2} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)$$
$$\frac{\partial H_k(\theta)}{\partial \sigma_k} = \sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)$$

and the second order derivatives are,

$$\frac{\partial^2 H_k(\theta)}{\partial \pi_k^2} = \frac{\partial}{\partial \pi_k} \left(\sum_{t=p+1}^{n} \left(\frac{Z_{k,t}}{\pi_k} - \frac{Z_{k,t}}{\pi_g}\right)\right) = -\sum_{t=p+1}^{n} \left(\frac{Z_{k,t}}{\pi_k^2} + \frac{Z_{k,t}}{\pi_g^2}\right) \quad (5.2.0.58)$$

$$\frac{\partial^2 H_k(\theta)}{\partial \pi_k \partial \pi_l} = \frac{\partial}{\partial \pi_l} \left(\sum_{t=p+1}^{n} \left(\frac{Z_{k,t}}{\pi_k} \frac{Z_{k,t}}{\pi_g}\right)\right) = -\sum_{t=p+1}^{n} \left(\frac{Z_{k,t}}{\pi_g^2}\right)$$
$$\frac{\partial^2 H_k(\theta)}{\partial \phi_{k,i}^2} = \frac{\partial}{\partial \phi_{k,i}} \left(\sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k^2} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)\right) = -\sum_{t=p+1}^{n} \frac{Z_{k,t} u(y_{t,i})^2}{\sigma_k^2}$$
$$\frac{\partial^2 H_k(\theta)}{\partial \phi_{k,i} \phi_{k,j}} = \frac{\partial}{\partial \phi_{k,j}} \left(\sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k^2} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)\right) = -\sum_{t=p+1}^{n} \frac{Z_{k,t} u(y_{t,i}) u(y_{t,j})}{\sigma_k^2}$$
$$\frac{\partial^2 H_k(\theta)}{\partial \sigma_k^2} = \frac{\partial}{\partial \sigma_k} \left(\sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)\right) = -\sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k^2} f_k^{''} \left(\frac{e_{k,t}}{\sigma_k}\right)$$
$$\frac{\partial^2 H_k(\theta)}{\partial \sigma_k \partial \phi_{k,i}} = \frac{\partial}{\partial \phi_{k,i}} \left(\sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k} f_k^{'} \left(\frac{e_{k,t}}{\sigma_k}\right)\right) = -\sum_{t=p+1}^{n} \frac{Z_{k,t}}{\sigma_k^3} e_{k,t} u(y_{t,i})$$

where,

$$u(y_t, i) = \begin{cases} 1 & i = 0 \\ y_{t-i} & i > 0 \end{cases} \tag{5.2.0.59}$$

Y_{t-1} and ϵ_{z_t} in Model 2.2.0.8 are independent and ϵ_{z_t} has finite variance. The assumption $E|Y_0| < \infty$ follows from the geometric ergodicity property.

□

Some additional regularity assumptions are further needed to set the table for the proof of the asymptotic normality of the MLE of the MAR model.

Assumption F

(a) There exists a function $g_0 : \mathbb{R}^2 \to \mathbb{R}^+$ satisfying $E[g_0(Y_t, Y_{t-1})] < \infty$, such that

$$\sup_{\theta \in \Theta^0} f_\theta(y \mid y^{'}, k) \leq g_0(y, y^{'}) \qquad \text{for all } y, y^{'} \in \mathbb{R} \tag{5.2.0.60}$$

(b) There exist functions $g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}^+$ satisfying $E[g_i(Y_t Y_{t-1})] < \infty,\ i = 1, 2,$ such that

$$||\frac{\partial f_\theta(y \mid y^{'}, k)}{\partial \theta}|| \leq g_1(y, y^{'}) \text{ and } ||\frac{\partial^2 f_\theta(y \mid y^{'}, k)}{\partial \theta \theta'}|| \leq g_2(y, y^{'}) \text{ for all } y, y^{'} \in \mathbb{R} \tag{5.2.0.61}$$

Lemma 5.2.2. Let $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$, each y_t is an MAR model defined in Equation (2.2.1). Given that Assumptions 2.2 and Assumption C hold, then Assumptions F follows.

Proof. From Lemma 5.0.1, we have that $0 \leq f_\theta(y \mid y^{'}, k) \leq \frac{1}{\sigma_k}$ so that Equation (5.2.0.60) and Assumption F(a) hold.

Also, Let d_k and h_k be arbitrary selections with replacement from $\phi_{k,i}$ and σ_k, so that, the first derivative of the conditional density of y_t with respect to d_k is,

$$\frac{\partial f_\theta(y \mid y^{'}, k)}{\partial d_k} = \frac{\partial H_k(\theta)}{\partial d_k} f_\theta(y \mid y^{'}, k) \tag{5.2.0.62}$$

and differentiating the conditional density of y_t with respect to both d_k and h_k gives

$$\frac{\partial^2 f_\theta(y \mid y^{'}, k)}{\partial d_k \partial h_k} = \left(\frac{\partial^2 H_k(\theta)}{\partial d_k \partial h_k} + \frac{\partial H_k(\theta)}{\partial d_k} \frac{\partial H_k(\theta)}{\partial h_k} \right) f_\theta(y \mid y^{'}, k) \tag{5.2.0.63}$$

$H_k(\theta)$ and the corresponding partial derivatives are detailed above. Assume further that all the first and second order partial derivatives of $f_\theta(y \mid y^{'}, k)$ with respect to θ are bounded by

$$f_\theta(y \mid y^{'}, k)(c_2(u(y_{t,i}))^2 + c_1) \leq \frac{1}{\sigma_k}(c_2(u(y_{t,i}))^2 + c_1) \tag{5.2.0.64}$$

for all $y, y^{'} \in \mathbb{R}$ and all $k = 1, \ldots, g$. Setting g_1, g_2 equal to the RHS of the above inequality and taking expectations, Assumption F(b) follows from the assumption that $E(Y_t^4) < \infty$. $\square$

Theorem 5.2.1. *Let $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$, each y_t is an MAR model defined in Equation* (2.2.1)*. Given that Assumptions 2.2, Assumption C and Assumption D hold. Assume that $E(\epsilon_t^4) < \infty$ and that the Fisher information matrix $(I(\theta^0))$ is positive definite, then for all $z_0 \in 1, \ldots, g$ we have,*

$$\frac{1}{n} \frac{\partial^2 l_n(\hat{\theta}_{n,z_0}, Z_0)}{\partial \theta \theta'} \rightarrow I(\theta^0) \tag{5.2.0.65}$$

and

$$\sqrt{n}(\hat{\theta}_{n,z_0} - \theta^0) \rightarrow \mathcal{N}(0, (I(\theta^0))^{-1}) \tag{5.2.0.66}$$

where

$$I(\theta^0) = -E_{\theta^0} \frac{\partial^2 \log p_{\theta^0}(Y_t \mid Y_{-\infty}, \dots, Y_{t-1})}{\partial\theta\partial\theta'}. \tag{5.2.0.67}$$

Proof. The proof of asymptotic normality makes use of the following,

1. A central limit theorem (CLT) for the Fisher score function $\frac{1}{\sqrt{n}} \frac{\partial l_n(\theta^0, z_0)}{\partial\theta}$

2. a local uniform law of large numbers for the observed Fisher information $\frac{1}{n} \frac{\partial^2 l_n(\theta^0, z_0)}{\partial\theta\partial\theta'}$ in the neighborhood of θ^0.

Douc et al. (2004) express the score function and the observed fisher information as functions of conditional expectations of the complete score function and the complete Fisher information.

5.2.1 A central Limit theorem for the score function

The method here for the Fisher identity is due to Louis (1982) (see also Tanner (1993)). The Louis *Missing Information Principle* says that,

Observed Information=Complete Information - Missing Information.

Now, for all z_0 and $\theta \in \Theta$,

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta^0, z_0)}{\partial\theta} &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log p_{\theta^0}(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0)}{\partial\theta} \\ &= \frac{1}{\sqrt{n}} \frac{\partial \sum_{t=1}^{n} \Delta_{t,0,z_0}(\theta^0)}{\partial\theta}. \end{aligned} \tag{5.2.1.1}$$

Using the notation in the proof of Corollary 5.1.1, write,

$$\begin{aligned} \frac{\partial\Delta_{t,0,z_0}(\theta)}{\partial\theta} &= E_\theta \left[\sum_{i=1}^{t} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t)}, Z_0 = z_0 \right] \\ &- E_\theta \left[\sum_{i=1}^{t-1} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t-1)}, Z_0 = z_0 \right] \end{aligned} \tag{5.2.1.2}$$

where

$$\phi(\theta, Q_{i-1}, Q_i) = \phi(\theta, (Z_{i-1}, Y_{i-1}), (Z_i, Y_i)) \qquad (5.2.1.3)$$
$$= \frac{\partial \log(a_{Z_{i-1}, Z_i} f_\theta(Y_i \mid Y_{i-1}, Z_i))}{\partial \theta}$$

is the conditional score function of (Z_i, Y_i) given (Z_{i-1}, Y_{i-1}). Similarly, for $m \geq 0$,

$$\frac{\partial \Delta_{t,m}(\theta)}{\partial \theta} = E_\theta \left[\sum_{i=1}^{t} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(-m,t)} \right] \qquad (5.2.1.4)$$
$$- E_\theta \left[\sum_{i=1}^{t-1} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(-m,t-1)} \right],$$

consider the filtration $\mathcal{F}_t = \sigma(Y_s, s \leq t)$ for all, $t \in \mathbb{Z}$. By the dominated convergence theorem, we can write,

$$E_{\theta^0} \Bigg[\sum_{i=-\infty}^{t-1} \left(E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t)} \right] \right. \qquad (5.2.1.5)$$
$$\left. - E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t-1)} \right] \right) \mid Y^{(-\infty,t-1)} \Bigg] = 0,$$

where

$$E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t-1)} \right] \qquad (5.2.1.6)$$
$$= E_{\theta^0} \left[E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t-1)}, Z_{t-1} \right] \right) \mid Y^{(-\infty,t-1)} \right] = 0$$

So that $\{\frac{\partial \Delta_{t,\infty}(\theta^0)}{\partial \theta}\}_{t=-\infty}^{\infty}$ is an $\mathcal{F}_t = \sigma(Y_s, s \leq t)-$ adapted, stationary, ergodic and square integrable martingale increment sequence for which the CLT for sums of such sequences (see Durrett (1996)) can be applied to show that,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \Delta_{t,\infty}(\theta^0)}{\partial \theta} \to \mathcal{N}(0, I(\theta^0)), \qquad (5.2.1.7)$$

where

$$I(\theta^0) = E_{\theta^0} [\frac{\partial \Delta_{0,\infty}(\theta^0)}{\partial \theta} \frac{\partial \Delta_{0,\infty}(\theta^0)}{\partial \theta}^T] \qquad (5.2.1.8)$$

is the asymptotic Fisher information matrix defined as the covariance matrix of the asymptotic score function (Douc et al. (2004)).

So that

$$\lim_{n\to\infty} E\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\frac{\partial\Delta_{t,0}(\theta^0)}{\partial\theta}-\frac{\partial\Delta_{t,\infty}(\theta^0)}{\partial\theta}\right)\right\|^2=0 \tag{5.2.1.9}$$

and

$$\lim_{n\to\infty} E\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\frac{\partial\Delta_{t,0,z}(\theta^0)}{\partial\theta}-\frac{\partial\Delta_{t,0}(\theta^0)}{\partial\theta}\right)\right\|^2=0 \tag{5.2.1.10}$$

Hence,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial\Delta_{t,0}(\theta^0)}{\partial\theta} \text{ and } \tfrac{1}{\sqrt{n}}\textstyle\sum_{t=1}^{n}\frac{\partial\Delta_{t,0,z}(\theta^0)}{\partial\theta} \text{ have the same limiting distribution.} \tag{5.2.1.11}$$

Therefore , $\frac{\partial\Delta_{t,0}(\theta^0)}{\partial\theta}$ can be approximated in L^2 by a stationary martingale increment sequence. Thus

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial\Delta_{t,0,z}(\theta^0)}{\partial\theta}\to\mathcal{N}(0,I(\theta^0)). \tag{5.2.1.12}$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial l_n(\theta^0,z_0)}{\partial\theta}\to\mathcal{N}(0,I(\theta^0)). \tag{5.2.1.13}$$

5.2.2 Uniform Law of Large numbers for the observed Fisher information

A locally uniform law of large numbers is explored for the observed Fisher information that is, for all possibly random sequences $\{\theta_n^0\}$ such that $\theta_n^0 \xrightarrow{a.s.} \theta^0$ and

$$-\frac{1}{n}\frac{\partial^2 l_n(\theta_n^0,z_0)}{\partial\theta\partial\theta'} \tag{5.2.2.1}$$

converges a.s. to the Fisher information matrix at θ^0.

First express the observed Fisher information in terms of the hessian of the complete log-likelihood, we do this by leaning on the Louis missing information principle [see Louis (1982),Tanner (1993), Wong and Li (2000)]. The basic idea in the principle leads to,

$$\begin{aligned}\frac{\partial^2 \log p_\theta(Y^{(1,n)} \mid Y_0, Z_0 = z_0)}{\partial\theta\theta'} \\ &= E_\theta\left[\sum_{i=1}^{n} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,n)}, Z_0 = z_0\right] \\ &\quad + var_\theta\left[\sum_{i=1}^{n} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,n)}, Z_0 = z_0\right], \quad (5.2.2.2)\end{aligned}$$

where

$$\begin{aligned}\psi(\theta, Q_{i-1}, Q_i) &= \psi(\theta, (Z_{i-1}, Y_{i-1})(Z_i, Y_i)) \quad (5.2.2.3)\\ &= \frac{\partial^2 \log(a_{Z_{i-1},Z_i} f_\theta(Y_i \mid Y_{i-1}, Z_i))}{\partial\theta\partial\theta'}.\end{aligned}$$

Also,

$$\begin{aligned}E_\theta&\left[\sum_{i=1}^{n} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,n)}, Z_0 = z_0\right] \\ &= \sum_{t=1}^{n}\left(E_\theta\left[\sum_{i=1}^{t} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t)}, Z_0 = z_0\right]\right. \\ &\quad \left. - E_\theta\left[\sum_{i=1}^{t-1} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t-1)}, Z_0 = z_0\right]\right) \quad (5.2.2.4)\end{aligned}$$

and

$$var_\theta\left[\sum_{i=1}^{n}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(0,n},Z_0=z_0\right]$$
$$=\sum_{t=1}^{n}\left(var_\theta\left[\sum_{i=1}^{t}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(0,t)},Z_0=z_0\right]\right.$$
$$\left.-\,var_\theta\left[\sum_{i=1}^{t-1}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(0,t-1)},Z_0=z_0\right]\right). \quad (5.2.2.5)$$

As $t\to\infty$ the initial condition on Y_0 becomes more trival.

Thus for $t\geq 1$ and $m\geq 0$,define,

$$\frac{\partial\Delta_{t,m}(\theta)}{\partial\theta}=E_\theta\left[\sum_{i=-m+1}^{t}\psi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t)}\right]$$
$$-\,E_\theta\left[\sum_{i=-m+1}^{t-1}\psi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t)}\right] \quad (5.2.2.6)$$

and

$$\Gamma_{t,m}(\theta)=var_\theta\left[\sum_{i=-m+1}^{t}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t)}\right]$$
$$-\,var_\theta\left[\sum_{i=-m+1}^{t-1}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t-1)}\right] \quad (5.2.2.7)$$

Now, $\frac{\partial\Delta_{t,m}(\theta)}{\partial\theta}$ and $\Gamma_{t,m}(\theta)$ both converge to $\frac{\partial\Delta_{t,\infty}(\theta)}{\partial\theta}$ and $\Gamma_{t,\infty}(\theta)$ respectively in L^1 as $m\to\infty$. It also follows that $\{\frac{\partial\Delta_{t,m}(\theta)}{\partial\theta}\}_{t=1}^{\infty}$ and $\{\Gamma_{t,\infty}(\theta)\}_{t=1}^{\infty}$ are stationary and ergodic.

Thus, the observed Fisher information will converge to

$$-E_{\theta^0}\left[\frac{\partial\Delta_{t,m}(\theta^0)}{\partial\theta}+\Gamma_{t,\infty}(\theta^0)\right] \text{ (see Douc et al. (2004).)} \quad (5.2.2.8)$$

For all z_0, the Fisher Information identity implies that

$$\frac{1}{n}E_\theta\left[\frac{\partial l_n(\theta,z_0)}{\partial\theta}\frac{\partial l_n(\theta,z_0)^T}{\partial\theta}\mid Y_0,Z_0=z_0\right]$$
$$=-\frac{1}{n}E_\theta\left[\frac{\partial^2 l_n(\theta,z_0)}{\partial\theta\partial\theta'}\mid Y_0,Z_0=z_0\right] \quad (5.2.2.9)$$

Finally, the Louis missing information principle (Louis (1982) and Tanner (1993)) show that the limits in n of the two quantities in Equation (5.2.2.9) both coincide with the Fisher information at θ^0 which completes the proof. $\square$

5.3 Summary

In this chapter, we have considered the asymptotic properties of the MLE of the MAR process. We consider a vector of MAR processes (Y_t) as a markov regime autoregressive process with a compact and finite hidden space. We leverage the results of Douc et al. (2004) whose assumptions/proofs are hinged on the paper by Wald (1949) which says that there exisits a deterministic asymptotic criterion function $l(\theta)$ such that $n^{-1}l_n(\theta, z_0) \to l(\theta)$ a.s. uniformly with respect to $\theta \in \Theta$.

Hence, we consider the conditional form of the log likelihood function that is $l_n(\theta, z_0)$ instead of $l(\theta)$ and show that the Maximum Likelihood Estimate of the MAR model is both consistent and asymptotically normal.

Chapter 6

Risk

A major part of decision making involves taking risk. Thus, most decision makers are faced with the question of how to quantify risk. Investors are also faced with the dilemma of how much they can possibly lose on an investment as well as the overall risk exposure of the organisation.

What then is risk? Risk is simply a measure of how volatile the returns on an asset are.

Jorion (1997) defines risk as the volatility of unexpected outcomes. He classified the risk exposure of corporations into 3 main types viz:

- Business Risk.
- Strategic Risk.
- Financial Risk.

Risk taking is necessary for profit making. Hence, it is impossible to eliminate risk not just from financial markets but also from any other profit oriented venture. The need for risk management cannot therefore be over emphasized.

Managing risk, is the ability to recognize and mitigate against future natural, social, political and economic events that can have unfavorable effects on investments. For financial institutions, unfavorable effects usually involve huge losses on a portfolio of assets such as stocks, bonds, etc. Other risk exposures common to other organisations include operational risk (which is the risk associated with incompetent internal processes), fraud or litigation, and many others.

Recent occurrences in the global economy and the substantial losses that companies and major financial houses have suffered in the past decade have made the concept of managing risk extremely vital to businesses.

The unpredictability of future events necessitates the exploration of tools like probability theory, stochastic processes, statistics and the like as critical input for managing risk. Regulators and supervisory authorities require that financial institutions use quantitative techniques to manage risk.

Acerbi (2002), Artzner et al. (1999),Acerbi et al. (2008),Tsay (1997),Tasche (2002),Acerbi and Tasche (2002),Fotios C. Harmantzis (2006), Jorion (1997) and many others have studied various risk measures, the most popular being *Value-at-Risk* and *Expected Shortfall*. This chapter begins with a description of the various categories of risk measures followed by a detailed discussion on Value at Risk and Expected shortfall and the approaches to estimating them.

6.1 Coherent Measures of Risk

Consider a set of real-valued random variables $\mathcal{G}$ on some probability space $(\Omega, \mathcal{A}, P)$ and a fixed $\alpha \in (0,1)$. . A mapping $\rho : \mathcal{G} \to (-\infty, \infty]$ with $\rho(0) = 0$ is called a *coherent risk measure*(Artzner et al. (1999)) if it satisfies the following axioms:

i. **Translation invariance:** $\rho(X+\alpha) = \rho(X) - \alpha$, for all $X \in \mathcal{G}$ and $\alpha \in (0,1)$.

This axiom implies that for each $X, \rho(X + \rho(X)) = 0$. That is including the estimated risk in an investment will cancel out the risk, hence it is optimal to consider building in associated risk into an investment portfolio.

ii. **Subadditivity:** $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$, for all X_1,$X_2 \in \mathcal{G}$. This axiom simply implies that "a merger does not create extra risk". It also means that diversification should reduce risk.

iii. **Positive homogeneity:** $\rho(\lambda X) = \lambda\rho(X)$, for all, $\lambda \geq 0$ and all $X \in \mathcal{G}$. This axiom implies that if position size were to directly influence risk, then lack of liquidity would be a concern when computing future net worth of a position.

iv. **Monotonicity:** $\rho(Y) \leq \rho(X)$, for all, X,$Y \in \mathcal{G}$ with $X \leq Y$. This axiom rules out any risk measure defined by $\rho(X) = -\mathbf{E}_{\mathbb{P}}[X] + \alpha.\sigma_{\mathbb{P}}(X)$. Where $\sigma_{\mathbb{P}}$ is the standard deviation operator, computed under $\mathbb{P}$

Remark 6.1.1. The **subadditivity** and **positive homogeneity** axioms imply that $\rho(nX) \leq n\rho(X)$ for $n = 1, 2, \ldots$.

Remark 6.1.2. The **translation invariant** and **positive homogeneity** axiom imply that, for each $\alpha, \rho(\alpha.(-r)) = \alpha$ where r is the total return on a portfolio.

6.2 Convex Measures of Risk

Multiplying a position by a large factor (say λ) gives rise to additional liquidity risk, so that subadditivity and positive homogeneity axioms are no longer critical.

Follmer and Schied (2002) suggest relaxing the subadditivity and positive homogeneity axioms to the weaker property of convexity.

Definition 6.2.1 (Convex Risk Measure). A risk measure ρ is said to be convex if it satisfies the following axioms:

i It satisfies axioms i and iv in Section 6.1 that is, it is translative invariant and monotonic.

ii **convexity:** $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$ forall $X, Y \in \mathcal{G}$ and $\lambda \in [0, 1]$.

Convexity implies that diversification does not increase risk that is the risk associated with a diversified position $\lambda X + (1-\lambda)Y$ is less than or equal to the weighted average of individual risks associated with the investments that make up the portfolio.

6.3 Spectral Measures of Risk

Acerbi (2002) and Tasche (2002) discuss the class of *spectral risk measures*. This class of risk measures can be viewed as a subclass of coherent risk measures, as they are defined by adding two extra axioms to the set of axioms that define coherency.

The following are some useful definitions of some terms used in this section.

Definition 6.3.1. Comonotonic– A pair of real valued random variables X and Y are said to be *comonotonic* if there exists a real valued random variable Z and two non-decreasing functions $f, g : \mathbb{R} \to \mathbb{R}$ such that $X = f(Z)$ and $Y = g(Z)$.

Definition 6.3.2. Quantiles[Acerbi and Tasche (2002) and Acerbi et al. (2008)]– Let X be a real valued random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\alpha \in (0, 1)$. The elements of the set

$$\mathcal{Q}_\alpha(X) = \{x \in \mathbb{R} : \mathbb{P}[X < x] \leq \alpha \leq \mathbb{P}[X \leq x]\} \tag{6.3.0.1}$$

are the inverse of the distribution function ($F_X(x)$) of a random variable X and is called the $\alpha-$quantile of X, defined as, $\mathcal{Q}_\alpha(X) = [q_\alpha^-(X), q_\alpha^+(X)] \neq \emptyset$ with:

$$q_\alpha^-(X) = \inf\{x \in \mathbb{R} : \mathbb{P}[X \leq x] \geq \alpha\}, \ \alpha \in (0,1] \text{ (lower-}\alpha\text{-quantile)} \tag{6.3.0.2}$$

$$q_\alpha^+(X) = \sup\{x \in \mathbb{R} : \mathbb{P}[X < x] \leq \alpha\}, \ \alpha \in [0,1) \text{ (upper-}\alpha\text{-quantile)} \tag{6.3.0.3}$$

We now give a formal definition of Spectral risk measures.

Definition 6.3.3 (**Spectral Measures of Risk**). A risk measure ρ is said to be a spectral measure of risk if it satisfied the following:

i **Coherence**: it satisfies axioms i-iv in Section 6.1.

ii **Commonotonic additive**: $\rho(X+Y) = \rho(X) + \rho(Y)$ for any commonotonic pair $X, Y \in \mathcal{G}$.

Note that two commonotonic portfolios provide no diversification when added together. Furthermore, if a risk measure is both subadditive and commonotonic additive, then the upper bound $\rho(X) + \rho(Y)$ placed on $\rho(X+Y)$ by subadditivity can be attained for commonotonic variables (Jouini et al. (2006)).

iii **law-invariant**:$\rho(X) = \rho(Y)$ whenever $X, Y \in \mathcal{G}$ have the same probability law that is $\rho(X)$ depends only on the distribution of X.

This axiom is very important in practise as a risk measure can only be estimated from empirical loss data if it is law-invariant. Jouini et al. (2006) show that law invariant convex risk measures have the Fatou Property.

The following are some spectral risk measures found in literature:

1 **Tail Conditional Expectation (TCE)**[Acerbi and Tasche (2002)] Consider a real valued random variable X on a probability space $(\Omega, \mathcal{A}, P)$, and confidence level α. TCE is the measure of risk defined as follows,

$$TCE_\alpha(X) = E[X \mid X \leq q_\alpha^-(X)] \text{ (lower TCE at CL } \alpha) \tag{6.3.0.4}$$

$$TCE^\alpha(X) = -E[X \mid X \leq q_\alpha^+(X)] \text{ (upper TCE at CL } \alpha) \tag{6.3.0.5}$$

By Artzner et al. (1999) TCE^α can also be written as,

$$TCE_\alpha(X) = -\mathbf{E}_\mathbb{P}[X \mid X \leq -VaR_\alpha(X)]. \tag{6.3.0.6}$$

We will define VaR_α in Section 6.4. Note that $TCE_\alpha \geq TCE^\alpha$. Delbaen et al. (2000) was able to show that TCE^α is in general not a coherent measure of risk as it is not sub-additive.

2 **Worst Conditional Expectation (WCE)**[Acerbi and Tasche (2002) and Artzner et al. (1999)] Consider a real valued random variables X on a probability space $(\Omega, \mathcal{A}, P)$, and confidence level α. WCE is the measure of risk defined as follows,

$$WCE_\alpha(X) = -\inf \mathbf{E}_\mathbb{P}[X \mid A] : A \in \mathcal{A} \mid \mathbb{P} > \alpha. \tag{6.3.0.7}$$

The definition of WCE_α implies that WCE is sub-additive that is for any two random variable X and Y on the same probability space, $WCE_\alpha(X + Y) \leq WCE_\alpha(X) + WCE_\alpha(Y)$.

Remark 6.3.1. (Artzner et al., 1999, Proposition 5.1),suggests that $TCE^\alpha \leq WCE_\alpha$, this implies that WCE_α dominates TCE_α.

However, the infimum is not quite effective hence the introduction of the *Conditional Value-at-Risk*.

3 **Conditional Value-at-Risk**[Acerbi and Tasche (2002)]. Consider a real valued random variable X on a probability space $(\Omega, \mathcal{A}, P)$, and confidence level α. The $\alpha-$level $CVaR$ of $X$is the measure of risk defined as follows,

$$CVaR^{\alpha}(X) = \inf\left(\mathbf{E}_{\mathbb{P}}[(X-s)^{-}]/\alpha\right) - s : s \in \mathbb{R}. \tag{6.3.0.8}$$

4 **Tail Mean**[Acerbi and Tasche (2002)] Consider a real valued random variable X on a probability space $(\Omega, \mathcal{A}, P)$, and confidence level α. The $\alpha-$level TM of X is the measure of risk defined as follows,

$$TM_{\alpha}(X) = \alpha^{-1}(\mathbf{E}[X\mathbf{1}_{(X \le q_{\alpha}^{-})}] + q_{\alpha}^{-}(\alpha - P[x \le q_{\alpha}^{-}])), \tag{6.3.0.9}$$

Where $\mathbf{1}_{(\cdot)}$ is an indicator function.

Furthermore, if we have a real-valued integrable random variable X on a probability space $(\Omega, \mathcal{A}, P)$ and confidence level $\alpha \in (0,1)$ fixed, then:

$$TM_{\alpha}(X) = \alpha^{-1}\int_{0}^{\infty} q_{u}^{-}(X)du \tag{6.3.0.10}$$

Note that TM_{α} depends only on the distribution of X and the confidence level α, but not on a particular definition of the quantile (see (Acerbi and Tasche, 2002, Corollary 4.3)).

6.4 Value at Risk

The concept of Value at Risk (VaR) was first introduced by financial companies in the late '80s. However, the underlying mathematics sprung from attempts by Harry Markowitz and others (Markowitz (1952)), to devise optimal portfolios for equity investors. VaR attempts to provide ways of quantifying risks in financial positions/portfolios.

Butler (1999) classifies the contribution of VaR to the science of risk management in three bullet points as follows:

1. It helps to allocate resources more efficiently,so that the organisation is not over exposed to one source of risk.

2. It makes traders and risk managers more accountable for their actions especially where they introduce avoidable risk or fail to hedge against risk.

3. It helps regulators decide capital adequacy requirements for institutions.

Through the glasses of the financial institutions, VaR is seen as the maximal loss associated with a catastrophic event under normal market conditions while the glasses of a regulatory committee reflects VaR as the minimal loss under extraordinary market circumstances (Tsay (1997)).

Value at Risk is formally defined as follows,

Definition 6.4.1. Value at Risk (VaR) is the amount that a portfolio will lose with a given probability-p, over a specified time horizon-t.

That is with probability $(1 - p)$, the potential loss encountered by the holder of the financial position over the time horizon t is less than or equal to VaR (Tsay (1997)).

The probability-p is selected based on how the user and/or the developer of the risk management system wants to interpret the VaR figure. Typically, values for the probability p range between $1\% - 5\%$ (see J.P.Morgan (1995), Jorion (1997)).

The time horizon also known as the holding period should correspond to the maximum period needed for an organization's portfolio liquidation. The value ranges between $1 - 90$ business days (1,5,10 and 90 business days are commonly used).

6.4.1 Hiccups in Value at Risk

There has been some debate as to the sufficiency of VaR as a dependable measure of risk. Some of the arguments that can be found in literature include:

a) VaR does not satisfy the subadditivity axiom which contradicts the framework of modern portfolio theory that is diversification should reduce risk (Jadhav et al. (2009)).

   Acerbi et al. (2008) describes this non-compliance as two fold viz:

   i. Non-additivity by position– When a new instrument is added to a portfolio it often necessitates that VaR be recomputed for the entire portfolio as total VaR is not given by the sum of the partial VaR's of the instruments that make up the portfolio.

   ii. Non-additivity by risk variable– VaR is not a sum of partial VaRs of the multiple risk variables that make up a portfolio, even when the risks are independent of each other.

b) VaR disregards tail risk as it does not consider tail distribution beyond it's value.

c) VaR permits the construction of proxy portfolios having low VaR as a trade-off of heavy tail loss (Mamon and (Eds.) (2007)).

d) Rational investors who wish to maximize expected utility can be misled by the information given by VaR (see Mamon and (Eds.) (2007) for details).

e) VaR is not a coherent measure of risk.

Based on these shortfalls of VaR, a coherent risk measure referred to as *Expected Shortfall* has been suggested as an alternative and/or sometimes a complement for VaR.

6.5 Expected Shortfall

Expected Shortfall (ES) has been proposed as a viable alternative to VaR as it caters for the hiccups in VaR. The concept of Expected Shortfall is defined as follows,

Definition 6.5.1 (**Expected Shortfall**). Fix the confidence level $\alpha \in (0,1)$. Then, the $ES_\alpha(X)$ called the *Expected Shortfall (ES)* at level α of X is the mean loss in the $100\alpha\%$ worst case of a portfolio X, that is,

$$\begin{aligned} ES_\alpha(X) &= E[-X \mid -X \geq VaR_{(\alpha)}(X)] \\ &= \alpha^{-1}\int_0^\alpha VaR_u(X)du = -\alpha^{-1}\int_0^\alpha q_u^+(X)du. \end{aligned} \tag{6.5.0.1}$$

where q_u^+ is as defined in Definition 6.3.2 and $VaR_\alpha(X) = -q_\alpha^+(X)$.

6.5.1 Properties of Expected Shortfall

1. **Coherence**– ES is coherent, that is it satisfies the following axioms (see Section 6.1 for details):

   i. Monotonous

   ii. Sub-additive

   iii. Positively Homogeneous

   iv. Translative invariant

The coherence property is by far the most important property of ES.

2. Expected Shortfall is the excess loss over value at risk that is, $ES_\alpha(X) \geq VaR_\alpha(X)$.

3. ES_α is *law invariant*, that is For any two real-valued random variables X and Y, with $E[X^-] < \infty$ and $E[Y^-] < \infty$ and confidence level $\alpha \in (0,1)$ fixed, the following representation holds,

$$P[X \leq t] = P[Y \leq t], t \in \mathbb{R} \quad \text{implies that} \quad ES_\alpha(X) = ES_\alpha(Y). \quad (6.5.1.1)$$

This property implies that it is possible to determine ES_α from statistical observations only.

4. Tasche (2002) suggest that ES is such that,

$$\frac{\partial^+ ES_\alpha(X)}{\partial_\alpha} = \frac{-q_\alpha^+ + ES_\alpha(X)}{\alpha}, \; \frac{\partial^- ES_\alpha(X)}{\partial_\alpha} = \frac{-q_\alpha^-(X) + ES_\alpha(X)}{\alpha}. \quad (6.5.1.2)$$

5. For a real-valued random variable X with $E[X^-] < \infty$, $\alpha \to ES_\alpha$ is absolutely continuous on $(0,1)$ and non-decreasing. This implies that ES_α is continuous with respect to α.

6. ES_α is *comonotonic additive*, that is,

$$ES_\alpha(f \circ Z + g \circ Z) = ES_\alpha(f \circ Z) + ES_\alpha(g \circ Z) \quad (6.5.1.3)$$

for all non-decreasing f, g and random variables Z with $E[(f \circ Z)^-] < \infty$, $E[(g \circ Z)^-] < \infty$

This property implies that by commontonic additivity, $ES_\alpha(f \circ Z) + ES_\alpha(g \circ Z)$ is the upper bound for the risk $ES_\alpha(f \circ Z + g \circ Z)$ and occurs if $f \circ Z$

and $g \circ Z$ are comonotonic. Thus upper bounds of risk following from sub-addiitvity are proper worst case bounds (Tasche (2002)). So that ES is the proper worst case possible loss.

6.5.2 Other Risk Measures and how they connect to Expected Shortfall

Eberlein et al. (2007) refer to Spectral risk measures as generalisations of Expected Shortfall. Below are some spectral risk measures and how they relate to expected shortfall.

1. Expected Shortfall (ES) and Tail Mean (TM)

$$ES_\alpha(X) = -TM_\alpha(X) \tag{6.5.2.1}$$

Moreover, for a real-valued integrable random variable X, and any $\alpha \in (0, 1)$ and any $\epsilon > 0$ with $\alpha + \epsilon < 1$, we have the following due to Acerbi and Tasche (2002)

$$TM_{\alpha+\epsilon} \geq TM_\alpha \text{ and } ES_{\alpha+\epsilon} \leq ES_\alpha. \tag{6.5.2.2}$$

2. Expected Shortfall (ES) and Conditional Value at Risk (CVaR)

For a real valued integrable random variable X on some probability space $(\Omega, \mathcal{A}, P)$ and for a fixed $\alpha \in (0, 1)$,

$$ES_\alpha(X) = CVaR^\alpha(X) = -\alpha^{-1}(E[X\mathbf{1}_{\{X\leq s\}}]+s(\alpha - P[X \leq s])),$$
$$s \in [q^-_\alpha(X), q^+_\alpha(X)] \tag{6.5.2.3}$$

3. Expected Shortfall (ES) and Worst Conditional Expectation (WCE)

For any two real-valued integrable random variables X and Y on a probability space $(\Omega, \mathcal{A}, P)$ and confidence level $\alpha \in (0, 1)$. Let Y be such that $Y = f(X)$

where f satisfies $f(x) \leq f(q_\alpha^-)$ for $x < q_\alpha^-$, and $f(x) \geq f(q_\alpha^-)$ for $x > q_\alpha^-$. The following proposition by Acerbi and Tasche (2002) holds,

Propositon 6.5.1. i. If $P[X \leq q_\alpha^-]$ then $ES_\alpha(Y) = -\inf_{A\in\mathcal{A},P[A]\geq\alpha} E[Y \mid A]$.

ii. If the distribution of X is continuous, then, $ES_\alpha(Y) = WCE_\alpha(Y)$.

Moreover, if we consider a finite number of real valued integrable random variables in a vector $(X_i, \ldots, X_d)$ on a probability space $(\Omega, \mathcal{A}, P)$. In addition, fix the confidence level $\alpha \in (0, 1)$. Then there exists another random vector $(X_i^{'}, \ldots, X_d^{'})$ also on some probability space $(\Omega^{'}, \mathcal{A}^{'}, P^{'})$ with the following properties:

i. $(X_i, \ldots, X_d)$ and $(X_i^{'}, \ldots, X_d^{'})$ are equally distributed that is,

$$p[X_i \leq x_i, \ldots, X_d \leq x_d] = p^{'}[X_i^{'} \leq x_i, \ldots, PX_d^{'} \leq x_d^{'}] \quad \text{for all} \quad (X_i, \ldots, X_d) \in \mathbb{R}^d. \tag{6.5.2.4}$$

ii. The corresponding WCE and ES coincide for all $i = 1, \ldots, d$ i.e. $WCE_\alpha^{X_i^{'}} = ES_\alpha^{X_i^{'}},\ i = 1, \ldots d$.

Furthermore, for confidence level $\alpha \in (0, 1)$ and a real valued integrable random variable X on some probability space $(\Omega, \mathcal{A}, P)$,

$$TCE^\alpha(X) \leq TCE_\alpha(X) \leq ES_\alpha(X), \tag{6.5.2.5}$$
$$TCE^\alpha(X) \leq WCE_\alpha(X) \leq ES_\alpha(X)$$

6.6 Methodologies for evaluating Value at Risk and Expected Shortfall

All through this section, let X_t represent a financial time-series for example, the daily closing value of stock indices or daily foreign exchange rates. The focus is on the asset returns $r_t = -\log(X_t/X_{t-1}) \times 100$. We describe here existing methodologies for evaluating VaR and hence ES.

Existing methodologies for measuring VaR and ES can be categorized into 3 broad classes viz:

1. Parametric Methods
2. Semi-parametric Methods
3. Non-parametric Methods

6.6.1 Parametric Methods

The parametric models attempt to fit a parametric distribution to the data. VaR is computed directly from the standard deviation based on the fitted distribution.

We now describe some popular parametric methods in practise.

The [JP morgan's] Risk Metrics Model

This method of estimating VaR was introduced by J.P. Morgan in 1995. It is sometimes referred to as the Variance-Covariance approach (Tsay (1997)). The main highlight of the method is that it adapts a practical approach to measuring risk.

The variances-covariance within and between assets are modelled using exponentially weighted Moving average , which corresponds to the Integrated GARCH (IGARCH) model described in Section 2.4.1.8.

The Riskmetrics method assumes that the continuously compounded daily returns of a portfolio follows a conditional normal distribution, while returns themselves may not be normally distributed, but fat tailed with common large outliers. The assumption is that the standardized return (computed as the return divided by the forecasted standard deviation) is normally distributed. The method focuses on the size of the return relative to the standard deviation that is, a large positive or negative return in a period of high volatility may result in a low standardized return, whereas the same return following a period of low volatility will yield an unusually high standardized return.

Denote the daily log return by r_t and the information set available at time $t-1$ by $\mathcal{F}_{t-1}$. The RiskMetrics method assumes that the returns r_t given information up to and including time $t-1$ is distributed as $N(\mu_t, \sigma_t^2)$, where μ_t is the conditional mean and σ_t^2 is the conditional variance of r_t. In addition, the method assumes that the two quantities evolve over time according to the simple model,

$$\mu_t = 0,\ \sigma_t^2 = \alpha\sigma_{t-1}^2 + (1-\alpha)r_{t-1}^2,\ 1 > \alpha > 0. \tag{6.6.1.1}$$

The RiskMetrics method also assumes that the logarithm of the daily price, $r_t = \ln(R_t)$, of the portfolio satisfies the difference equation $r_t - r_{t-1} = a_t$, where $a_t = \sigma_t \epsilon_t$ is an IGARCH(1,1) process without drift also referred to as the random walk IGARCH. Where a often lies between $(0.9, 1)$ with a typical value of 0.94 (Ruppert (2004)).

A useful property of this model is that the conditional distribution of a multiperiod return is easily available. Specifically, for a k-period horizon, the log return from time $t+1$ to $t+k$ (inclusive) is given as follows,

$$r_t[k] = r_{t+1} + \cdots + r_{t+k-1} + r_{t+k} \tag{6.6.1.2}$$

where $[k]$ denotes the k-horizon return. Under the special IGARCH model, the conditional distribution of $r_t[k] \mid \mathcal{F}_{t-1} \sim N(0, \sigma_t^2[k])$. Under the assumption that ϵ is i.i.d and the IGARCH model above, we have,

$$\sigma_t^2[k] = Var(r_t[k] \mid \mathcal{F}_{t-1}) = \sum_{i=0}^{k} Var(a_t \mid \mathcal{F}_t). \tag{6.6.1.3}$$

where $Var(a_t \mid \mathcal{F}_{t-1}) = E(\sigma_t^2 \mid \mathcal{F}_{t-1})$ and can be obtained recursively by $r_{t-1} = a_{t-1} = \sigma_{t-1}\epsilon_{t-1}$. The volatility equation of the IGARCH model is written as,

$$\sigma_t^2 = \sigma_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2(\epsilon_{t-1}^2 - 1) \text{ for all } t. \tag{6.6.1.4}$$

In particular, we have,

$$\sigma_{t+i}^2 = \sigma_{t+i-1}^2 + (1-\alpha)\sigma_{t+i-1}^2(\epsilon_{t+i-1}^2 - 1) \text{ for } i = 2, \ldots, k. \tag{6.6.1.5}$$

since $E(\epsilon_{t+i-1}^2 \mid \mathcal{F}_t) = 0 \quad \text{for} \quad i \geq 2$, Equation (**??**)sigmaigarch) shows that,

$$E(\sigma_{t+i}^2 \mid \mathcal{F}_t) = E(\sigma_{t+i-1}^2 \mid \mathcal{F}_t) \text{ for } i = 2, \ldots, k. \tag{6.6.1.6}$$

For the 1-step ahead forecast, the IGARCH equation shows that $\sigma_{t+i}^2 = \alpha\sigma_t^2 + (1-\alpha)r_t^2$ so that $Var(r_{t+1} \mid \mathcal{F}_t) = \sigma_{t+1}^2$ for $i \geq 1$ and hence, $\sigma_t^2[k] = k\sigma_{t+1}^2$. The results show that $r_t[k] \mid \mathcal{F}_{t-1} \sim N(0, k\sigma_{t+1}^2)$.

Consequently, under the special IGARCH(1,1) model, the conditional variance of $r_t[k]$ is proportional to the time horizon k. The conditional standard deviation of a k-period horizon log return is then $\sqrt{k}\sigma_{t+1}$. The daily VaR of the portfolio under RiskMetrics is computed as,

$$\text{VaR} = \text{Amount of position} \times z_{1-p}\sigma_{t+1} \tag{6.6.1.7}$$

Where $p\%$ is the confidence level and z_{1-p} is the $100(1-p)$th quantile of the standard normal distribution. For example, for a one-sided 95% CI, the daily VaR of the portfolio under RiskMetrics is:

$$\text{VaR} = \text{Amount of position} \times 1.65\sigma_{t+1} \tag{6.6.1.8}$$

and that the k-day horizon is:

$$\text{VaR}(k) = \text{Amount of position} \times 1.65\sqrt{k}\sigma_{t+1} \tag{6.6.1.9}$$

where the argument (k) of VaR is used to denote the time horizon. Consequently, under RiskMetrics, we have

$$\text{VaR}(k) = \sqrt{k} \times \text{VaR} \tag{6.6.1.10}$$

This is referred to as the *square root of time rule* in VaR calculation under RiskMetrics (see Tsay (1997)). Consider a long position so that loss occurs when there is a big price drop (i.e. a large negative return). If the probability is set to 5%, then RiskMetrics uses $1.65\sigma_{t+1}$ to measure the risk of the portfolio; that is, it uses the one-sided 5% quantile of a normal distribution with mean zero and standard deviation σ_{t+1}. The actual 5% quantile is $-1.65\sigma_{t+1}$, but the negative sign is usually ignored with the understanding that it signifies a loss.

Damodaran (1999) argues that focusing on normalized standardized returns exposes the VaR computation to the risk of more frequent large outliers than would be expected with a normal distribution. Hence a more recent variation of the RiskMetrics system was extended to cover mixture of normal distributions, which allow for the assignment of higher probabilities for outliers.

The GARCH Model

The GARCH models discussed in Section 2.4.1 above assumes that the variance of returns follow predictable process (see Jorion (1997)). The model also assumes that the standardized residuals are independently and identically distributed (i.i.d). In the model specification, the conditional variance not only depends on the most recent innovations but also on the immediate past conditional variance. A typical GARCH (1,1) model is written as,

$$a_t = \sigma_t \eta_t, \qquad \eta_t \sim \ iid(0,1) \ \ \sigma_t^2 = \omega + \alpha_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{6.6.1.11}$$

where $\alpha + \beta \leq 1$.

The average unconditional variance is found by setting $E(a_{t-1}^2) = \sigma_t^2 = \sigma_{t-1}^2 = \sigma$ and then substituting $\sigma_t = \sigma$ into Equation (6.6.1.11) we find

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta}, \tag{6.6.1.12}$$

For the GARCH model to be stationary, the *persistence* which is the sum of the parameters $\alpha_1 + \beta$ must be less or equal to unity. The GARCH model provides a parsimonious model useful for financial time series analysis in markets that exhibit volatility clustering

A very important aspect of implementing the GARCH algorithm is the specification of the distribution of η_t. Once we have imposed this distributional assumption, then we are set to write down a likelihood function and get an estimate of the unknown parameters and hence the time series of estimated variance. So that, for a log return r_t of an asset, write the GARCH (p,q) model as,

$$\begin{aligned} a_t &= \sigma_t \eta_t, \qquad \eta_t \sim \ iid(0,1) \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{u} \alpha_i a_{t-i}^2 + \beta_j \sigma_{t-j}^2 \\ r_t^2 &= \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + a_t - \theta_j a_{t-j} \end{aligned} \tag{6.6.1.13}$$

Equation (6.6.1.13) contains the mean and volatility equations for the returns r_t and can be used to compute the 1-step ahead forecasts of the conditional mean and conditional variance of r_t. The pth quantile of the conditional distribution can be obtained by deducting the the product of the pth quantile of the of the assumed distribution of η_t and the 1-step ahead forecast of conditional variance from the 1-step ahead forecast of the conditional mean. VaR is then computed as the amount of the position multiplied by this quantile, for example, for a student t-distribution, we have

$$VaR_{p\%} = \text{Amount of position} \times (\hat{r}_{t,1} - t_{v,p}\hat{\sigma}_{t,1}) \tag{6.6.1.14}$$

For instance, the say 5% quantile under the assumption of a standard normal distribution is simply computed as -1.645 (the 5% quantile of the standard normal) times the estimated standard deviation. so that $VaR_{5\%}$ is computed as amount of position $\times(\hat{r}_{t,1} - z_p\hat{\sigma}_{t,1})$.

A potential handicap of the GARCH model is it's nonlinearity. The parameters of the model have to be estimated by maximization of the likelihood function which involves numerical optimization.

In general it has been found that the GARCH and RiskMetrics methods tend to underestimate VaR. Furthermore, in practice the assumption that the standardized residuals are normal is not consistent with the actual behavior of financial returns.

However, the main advantage of these methods is that they allow a complete characterization of the distribution of returns, there is however room for fine-tuning their performance by avoiding the normality assumption.

In addition, both the GARCH and RiskMetrics methods are subject to three different sources of misspecification viz:

- The specification of the variance equation.
- The distribution chosen to build the log likelihood may be wrong.
- The standardized residuals may not be i.i.d.

These misspecification issues may or may not be relevant for VaR estimation, but are important to note (see Manganelli and Engle (2001)).

Gaussian Method (Fotios C. Harmantzis (2006))

Consider a sample of returns, say X_i, $i = 1, \ldots, n$, i.i.d., such that $X \sim N(\mu, \sigma^2)$. μ and σ are unknown parameters. (typically it is assumed that $\mu = 0$).

The VaR at α confidence level is simply given by $z_\alpha \sigma$ .z_α is such that $P(Z > z_\alpha) = \alpha$, with $Z \sim N(0, 1)$. σ is estimated thus,

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2, \tag{6.6.1.15}$$

with $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$-the mean of X.

ES is computed as,

$$ES = E(X \mid X > VaR) = E(X \mid X > Z_\alpha \sigma_t) = \sigma_t E(X/\sigma_t \mid X_{\sigma_t} > Z_\alpha) \tag{6.6.1.16}$$

where $Z_\alpha = \Phi^{-1}(\alpha)$ represents the $\alpha-$quantile of the standard Normal distribution, Φ is the cumulative distribution function (cdf) of the standard Normal distribution.

Also, $E(Z \mid Z > u) = \varphi(u)/(1 - \Phi(u)))$, where Φ is the probability distribution function (pdf) of the standard Normal distribution and u is a pre-defined threshold.

Stable Paretian Method (Fotios C. Harmantzis (2006))

Under the Stable law, the sum of i.i.d. random variables follow the distribution of the summands (with different parameters).

The distribution is characterized by four parameters:

- The characteristic exponent (or index of stability) $\alpha \in (0, 2)$;
- The scale (or spread) parameter $\sigma \geq 0$;
- The skewness (or symmetry) parameter $\beta \in [-1, 1]$; and
- The shift (or location) parameter $\mu \in \mathcal{R}$.

A random variable X is said to have a Stable distribution i.e. $X \sim S_\alpha(\sigma, \beta, \mu)$ if and only if its characteristic function has the form,

$$E(\exp iXt) = \begin{cases} \exp\{-\sigma^\alpha |t|^\alpha (1 - i\beta sign(t) \tan \frac{\pi\alpha}{2}) + i\mu t\} \text{ for } \alpha \neq 1 \\ \exp\{-\sigma |t| (1 + \frac{2i\beta}{\pi} sign(t) \ln |t|) + i\mu t\} \text{ for } \alpha = 1 \end{cases} \tag{6.6.1.17}$$

1. The characteristic exponent α determines the rate of decay, that is, the heaviness of the tails of the distribution.
2. The parameter β is an indication of the skewness of the distribution. $\beta = 0$ corresponds to the symmetric case.
3. The parameter μ shifts the distribution to the left or right.
4. The parameter σ expands or contracts the distribution around μ.

5. α and β together determine the shape of the distribution while σ and μ both have no effect on its shape.

6. - Standard stable distribution: $\sigma = 1$, $\mu = 0$ and $X \sim S_\alpha(1, \beta, 0)$.
   - Symmetric stable distribution: $\beta = 0$ and $X \sim S_\alpha S$
   - Totally positively skewed stable distribution: $\beta = 1$ and $x \sim S_\alpha(\sigma, 1, \mu$

When $\alpha = 2$, the alpha-Stable distribution reduces to Gaussian distribution with characteristic function,

$$E \exp i\theta X = \exp\{-\sigma^2\theta^2 + i\mu\sigma\}. \tag{6.6.1.18}$$

to compute ES

i Estimate the parameters: this can be done using a Maximum Likelihood (ML) estimator

ii Generate random stable variates using monte carlo simulations

iii Compute

$$ES = E(X \mid X > VaR) = \frac{(\sum_{i=[n\alpha]}^{n} X_{n(i)})}{(n - [n\alpha])} \tag{6.6.1.19}$$

This can also be represented as:

$$\hat{\mu}_\alpha = \frac{\sum_{t=1}^{n} X_t I(X_t \geq \hat{v}_\alpha)}{\sum_{t=1}^{n} I(X_t \geq \hat{v}_\alpha)} = ([n\alpha] + 1)^{-1} \sum_{t=1}^{n} X_t I(X_t \geq \hat{v}_\alpha) \tag{6.6.1.20}$$

where $\hat{v}_p$ is the sample VaR estimator of v_p and X_r is the $r-$th order statistic of $\{X_t\}_{t=1}^{n}$ (see Chen (2008)).

The Stable Paretian model is a fat-tailed model and hence more suitable for financial data while the Gaussian model is quite traditionally thin-tailed and hence not as representative of financial data.

A major draw back in the Gaussian model is that it underestimates ES while the stable paretian model tend to overestimates ES.

Extreme Value Theory Klüppelberg and Mikosch (1997)

In practice, when we do not critically examine extreme values it can have some pretty devastating consequences on decision making. Many researchers and practitioners in modern statistics have tried to build statistical models describing this extreme events.

Extreme Value Theory (EVT) is a branch of statistics that makes an attempt at incorporating information about extreme deviations of probability distribution in model building. Extreme Value Theory is analogous to the Central Limit Theorem in that they both give information about the limiting distributions, that is, what the distribution of extreme values should look like in the limit (as our sample increases or decreases).

Denote the return of an asset measured in a fixed time interval (could be daily, monthly, quarterly etc.) by r_t. For a collection of n returns, $\{r_1, \ldots, r_n\}$, the smallest order statistic is the minimum return denoted by $r_{(1)}$ while the maximum return is the maximum order statistic $r_{(n)}$. In particular we define, $r_{(1)} = \min_{1 \leq j \leq n}\{r_j\}$ and $r_{(n)} = \max_{1 \leq j \leq n}\{r_j\}$. The minimum return is most relevant to estimating VaR for a long position, while for a short position we would be looking at the maximum return. However, properties of the maximum return can be obtained from the minimum by a simple sign change (see Tsay (1997)).

Gili and Këllezi (2006) describe the following two related ways of identifying extreme values in real data. In general, the two related ways of identifying extremes in real data are the *the block maxima* and *the peak over threshold* methods

Klüppelberg and Mikosch (1997). The two methods and the underling theory on which the methods are hinged are discussed below as well as descriptions of how each method is used to measure VaR and ES.

The Block Maxima Method The block maxima method was originally used to analyze data with seasonality.This method considers the maximum value a variable can take in successive periods (these periods could be months, years, weeks or days). These selected observations make up the extreme events and are also referred to as block or per period maxima (Gili and Këllezi (2006)).

Distribution of maxima Denote the limit law of the block maxima by M_n, where n is the size of the block (or subsample). The following theorem is due to Gnedenko (1943).

Theorem 6.6.1. *Let* (X_n) *be a sequence of i.i.d random variables. If there exists constants* $c_n > 0$, $d_n \in \mathbb{R}$ *and some non-degenerate distribution function* H *such that,*

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H, \tag{6.6.1.21}$$

then H *belongs to one of the three standard extreme value distributions:*

$$\begin{aligned} &\textit{Frèchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp[-x^{-\alpha}], & x > 0 \end{cases} \quad \alpha > 0 \\ &\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp[-(-x)^{\alpha}], & x \leq 0 \\ 1, & x > 0 \end{cases} \quad \alpha > 0 \\ &\textit{Gumbel:} \quad \Lambda(x) = \exp[-\exp(-x)], \quad x \in \mathbb{R} \end{aligned} \tag{6.6.1.22}$$

Gnedenko (1943) gave necessary and sufficient conditions for the CDF $F(x)$ of r_t to be associated with one of the three types of limiting distributions listed in

Equation (6.6.1.22). The tail behavior of the CDF determines the limiting distribution of the minimum return. Notice that for the Frèchet family of distributions, the (left) tail of the CDF declines by the power function and for the Weibull family it is the asymptotic distribution of the finite end points, but decays exponentially for the Gumbel family. The polynomially decaying tails of the Frèchet distribution makes it well suited to heavy tailed distributions which include stable and Student-t distributions. The exponentially decaying tail of the Gumbel distribution is better suited to thin tailed distributions such as normal and log-normal distributions. (see Tsay (1997), Klüppelberg and Mikosch (1997) for details).

Consider returns r_t and assume they are serially independent with common cumulative distribution function (CDF) $F(x)$. Let the CDF of the minimum $r_{(1)}$ be given as $F_{n,1}(x)$. In practice, the CDF of r_t is unknown and by implication the CDF of this minimum is also unknown. However as the number of returns increase to infinity, the CDF of the minimum becomes degenerated. Hence, EVT is concerned with finding two sequences, viz: a location series say $\{\mu_n > 0\}$ and a series of scaling factors $\{\sigma_n\}$ such that the distribution of the minimum $r_{(1*)} = (r_{(1)} - \mu_n)/\sigma_n$converges to a non-degenerate distribution as n goes to infinity.

Generalized Extreme Value (GEV) distribution The three standard distributions listed in Equation (6.6.1.22) have been collapsed into the following one-parameter representation (Klüppelberg and Mikosch (1997)),

$$H_\xi(x) = \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & \text{if } \xi \neq 0 \\ \exp[-\exp[-x]], & \text{if } \xi = 0 \end{cases} \tag{6.6.1.23}$$

This generalization is known as the *Generalized Extreme Value (GEV) distribution* and is obtained by setting

1. $\xi = \alpha^{-1}$ for the Frèchet distribution (Φ_α),

2. $\xi = -\alpha^{-1}$ for the Weibull distribution (Ψ_α) and

3. $\xi = 0$ for the Gumbel distribution (Λ).

The standard GEV defined in Equation (6.6.1.23) is the limiting distribution of normalized extrema.

Gili and Këllezi (2006) suggested the following three parameter specification of the GEV.

$$H_{\xi,\mu,\sigma}(x) = H_\xi\left(\frac{x-\mu}{\sigma}\right)\ x \in \mathcal{D},\ \mathcal{D} = \begin{cases}]-\infty, \mu - \frac{\sigma}{\xi}[& \xi < 0 \\ ]-\infty, \infty[& \xi = 0 \\ ]\mu - \frac{\sigma}{\xi}, \infty[& \xi > 0 \end{cases} \tag{6.6.1.24}$$

They called it the limiting distribution of the unnormalized maxima. Where μ and σ are the location and scale parameters representing the unknown norming constants.

We are interested in the return levels (quantiles) of the estimated GEV denoted as,

$$R^k = H^{-1}_{\xi,\sigma,\mu}(1 - \frac{1}{k}) \tag{6.6.1.25}$$

As seen above, the extreme value distribution contains three main parameters viz: ξ-the *shape parameter*, σ_n*-the scale parameter* and μ_n-the *location parameter*. These parameters can be estimated with either parametric or non-parametric methods. As these parameters cannot be estimated based on extreme observations only as there is only a single minimum and maximum for any given sample, we adapt the following idea from Tsay (1997). He suggested that the sample be divided into sub-samples and then the extreme value theory be applied to the sub-samples. As an illustration, assume that there are T returns $\{r_j\}_{j=1}^{T}$ available. The sample can be divided into g non-overlapping sub-samples each with n observations, assuming for ease of computation that $T = ng$, that is, the data is divided as,

$$\{r_1, \ldots, r_n \mid r_{n+1}, \ldots, r_{2n} \mid r_{2n+1}, \ldots, r_{3n} \mid \cdots \mid r_{(g-1)n+1}, \ldots, r_{ng}\} \qquad (6.6.1.26)$$

The observed returns can thus be written as r_{in+j}, where $1 \leq j \leq n$ and $i = 0, \cdots, g-1$. It is noteworthy that each sub-sample corresponds to a sub-period of the data span. In practice, the choice of n can be guided by the financial activities of the company, e.g. for daily returns, $n = 21$ would approximately be the number of trading days in a month, $n = 63$ would be approximately the number of trading days in a quarter and $n = 252$ would correspond to approximately the number of trading days in a year.

Denote the minimum of the ith sub-sample by $r_{n,i}$ ($r_{n,i}$ is the smallest return of the ith sub-sample), where n is the size of the ith sub-sample. When n is sufficiently large, $x_{n,i} = (\frac{r_{n,i}-\mu_n}{\sigma_n})$ should follow an extreme value distribution. Then $\{r_{n,i} \mid i = 1, \ldots, g\}$ (the collection of sample minima) can then be regarded as a sample of g observations from that extreme value distribution.

$$\{r_{n,i} = \min_{1<j<n}\{r_{(i-1)n+j}\},\ i = 1, \ldots, g \qquad (6.6.1.27)$$

The unknown parameters of the empirical distribution are then estimated using the collection of the sub-sample minima $r_{n,i}$. It is thus apparent that the estimates obtained would most probably depend on the choice of n i.e. the length of the sub-period.

The parameters ξ, σ and μ, can be estimated by:

1. Maximum Likelihood method (Klüppelberg and Mikosch (1997), Tsay (1997)).
2. Probability Weighted moments (Klüppelberg and Mikosch (1997)).
3. Regression Method (S.Tsay (1997)).
4. nonparametric Approach (S.Tsay (1997)).

The following are brief descriptions of each of the methods.

1. **Maximum Likelihood Method**

   Let us assume that the subperiod minima $\{r_{n,i}\}$ follow a generalized extreme value (GEV) distribution such that the pdf of $x_i = \frac{(r_{n,i}-\mu_n}{\sigma_n})$ is given by,

$$f(x) = \begin{cases} (1+\xi x)^{1/\xi - 1} \exp\left[-(1+\xi x)^{1/\xi}\right] & \text{if } \xi \neq 0 \ , \\ \exp\left[x - \exp(x)\right] & \text{if } \ \xi = 0, \end{cases} \tag{6.6.1.28}$$

   Where

$$\begin{cases} -\infty < x < \infty & \text{for } \ \xi = 0 \ , \\ x < -1/\xi & \text{for } \ \xi < 0, \\ x > -1/\xi & \text{for } \ \xi > 0. \end{cases} \tag{6.6.1.29}$$

   The pdf of $r_{n,i}$ can be obtained by a simple transformation as,

$$f(r_{n,i}) = \begin{cases} \frac{1}{\sigma_n}\left(1+\frac{\xi_n(r_{n,i}-\mu_n)}{\sigma_n}\right)^{1/\xi_n - 1} \exp\left[-\left(1+\frac{\xi_n(r_{n,i}-\mu_n)}{\sigma_n}\right)^{1/\xi_n}\right] & \text{if } \xi_n \neq 0 \\ \frac{1}{\sigma_n}\exp\left[\frac{r_{n,i}-\mu_n}{\sigma_n} - \exp\left(\frac{r_{n,i}-\mu_n}{\sigma_n}\right)\right] & \text{if } \ \xi_n = 0, \end{cases} \tag{6.6.1.30}$$

where it is understood that $1+\xi_n\frac{(r_{n,i}-\mu_n)}{\sigma_n} > 0$ if $\xi_n \neq 0$ and n is added to the shape parameter k to signify that its estimate depends on the choice of n. Under the independence assumption, the likelihood function of the subperiod minima is,

$$\ell(r_{n,1},\ldots,r_{n,g|\xi_n,\sigma_n,\mu_n}) = \prod_{i=1}^{g} f(r_{n,i}) \tag{6.6.1.31}$$

Nonlinear estimation procedures can then be used to obtain maximum likelihood estimates of ξ_n, μ_n, σ_n. These estimates are unbiased , asymptotically normal, and of minimum variance under proper assumptions (see Klüppelberg and Mikosch (1997), Coles (2001) for more details).

2. **Regression Method**

The regression method assumes that $\{r_{n,i}\}_i^g$ is a random sample from the GEV distribution and makes use of properties of order statistics (Gumbel (2004)). Denote the order statistics of the subperiod minima $\{r_{n,i}\}_i^g$ as,

$$r_{n(1)} \leq r_{n(1)} \leq \cdots \leq r_{n(g)}. \tag{6.6.1.32}$$

Using properties of order statistics (Cox and Hinkley (1979)) we have:

$$E\{F_*[r_n(i)]\} = \frac{i}{g+1},\ i = 1,\ldots,g. \tag{6.6.1.33}$$

where $F_*[r_n(i)]$ is the limiting distribution of the (normalized) minimum For the two cases, $\xi = 0$ and $\xi \neq 0$ we have:

Case 1-$\xi \neq 0$

$$F_*[r_n(i)] = 1 - \exp\left[-\left(1+\xi_n\frac{r_{n(i)}-\mu_n}{\sigma_n}\right)^{1/\xi_n}\right] \tag{6.6.1.34}$$

and then using Equation (6.6.1.33) and Equation (6.6.1.34)and approximating expectation by an observed value we have (S.Tsay (1997))

$$\frac{1}{g+1} = 1 - \exp\left[-\left(1 + \xi_n \frac{r_{n(i)} - \mu_n}{\sigma_n}\right)^{1/\xi_n}\right] \tag{6.6.1.35}$$

Hence,

$$\exp\left[-\left(1 + \xi_n \frac{r_{n(i)} - \mu_n}{\sigma_n}\right)^{1/\xi_n}\right] = 1 - \frac{1}{g+1} = \frac{g+1-i}{g+1},\ i = 1, \dots, g \tag{6.6.1.36}$$

Taking the natural logarithm twice, the prior equation gives

$$\ln\left[-\ln\left(\frac{g+1-i}{g+1}\right)\right] = \frac{1}{\xi_n} \ln\left(1 + \xi_n \frac{r_{n(i)} - \mu_n}{\sigma_n}\right).\ i = 1, \dots, g \tag{6.6.1.37}$$

However, in practice, letting e_i be the deviation between the previous two quantities and assuming that the series $\{e_i\}$ is not serially correlated, we have a regression set up as follows:

$$\ln\left[-\ln\left(\frac{g+1-i}{g+1}\right)\right] = \frac{1}{\xi_n} \ln\left(1 + \xi_n \frac{r_{n(i)} - \mu_n}{\sigma_n}\right) + e_i,\ i = 1, \dots, g \tag{6.6.1.38}$$

The least squares estimate of ξ_n, μ_n, σ_n can be obtained by minimizing the sum of squares of e_i.

Case 2-$\xi = 0$: When $\xi_n = 0$, the regression setup reduces to

$$\ln\left[-\ln\left(\frac{g+1-i}{g+1}\right)\right] = \frac{1}{\alpha_n} r_{n(i)} - \frac{\mu_n}{\sigma_n} + e_i,\ i = 1, \dots, g \tag{6.6.1.39}$$

Note however that although the least squares estimates are consistent, they are less efficient than maximum likelihood estimates, hence the maximum likelihood estimates are more frequently used in literature.

3. **The Nonparametric Approach**

The shape parameter ξ can be estimated using some nonparametric methods. Two such methods are the Hills estimator (Hill (1975)) and Picklands estimator(Picklands (1975)). These estimators do not require us to consider subsamples as they lend themselves directly to the returns $\{r_t\}_{t=1}^{n}$. Again, denote the order statistics of the sample as $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$. The two estimators of the shape parameter ξ are thus defined as follows,

$$\xi_p(q) = -\frac{1}{\ln(2)} \ln\left(\frac{-r_{(q)} + -r_{(2q)}}{-r_{(2q)} + -r_{(4q)}}\right) \tag{6.6.1.40}$$

$$\xi_h(q) = -\frac{-1}{q} \sum_{i=1}^{q} [\ln(-r_{(i)}) i \in (-r_{(q+1)})] \tag{6.6.1.41}$$

where q is a positive integer. The choice of q differs between the Hills and Picklands estimators. There is however still no general agreement on what the best choice of q is, despite the various investigations that have gone into it (see Tsay (1997) for details).

VaR is computed based on the GEV distribution as follows (Tsay (1997)), Consider a sample period of T observed asset returns, divide the sample period into g non overlapping sub periods of length n such that $T = ng$. If T is not a multiple of n, delete the excess over the maximum multiple, that is, if $T = ng + m$ with $1 \leq m \leq n$, delete the first m observations from the sample to obtain equal sized subperiods. Obtain estimates for the location (μ_n), scale(σ_n) and shape(ξ) parameters for each of the sub period minima $\{r_{n,i}\}$, the estimates are then plugged into Equation (6.6.1.23), with $x = (r - \mu_n)/\sigma_n$. The quantile of the GEV distribution for a given probability is then obtained. Thus for a given probability p of potential loss, let r_n be the pth quantile of the subperiod minimum under the limiting generalised extreme value distribution, we have:

$$p = \begin{cases} 1 - \exp\left[-\left(1 + \frac{\xi_n(r_n - \mu_n)}{\sigma_n}\right)^{1/k_n}\right] & \text{if } \xi_n \neq 0 \ , \\ 1 - \exp\left[-\exp\left(\frac{(r_n - \mu_n)}{\sigma_n}\right)\right] & \text{if } \xi_n = 0 \ , \end{cases}$$

where $1 + k_n(r_n - \mu_n)/\sigma_n > 0$ for $\xi_n \neq 0$. The equation can thus be rewritten as,

$$ln(1-p) = \begin{cases} -\left(1 + \frac{k_n(r_n-\mu_n)}{\sigma_n}\right)^{1/k_n} & \text{if } \xi_n \neq 0 \ , \\ -\exp\left(\frac{(r_n-\mu_n)}{\sigma_n}\right) & \text{if } \xi_n = 0 \ , \end{cases}$$

The quantile can be obtained as,

$$r_n = \begin{cases} \mu_n - \frac{\sigma_n}{\xi_n}\{1 - [-\ln(1-p)]^{\xi_n}\} & \text{if } \xi_n \neq 0 \ , \\ \mu_n + \sigma_n \ln[-\ln(1-p)] & \text{if } \xi_n = 0 \ , \end{cases}$$

The case of $\xi_n = 0$ is more suitable to financial applications.

Since most asset returns are either serially uncorrelated or have weak serial correlations, , we can write

$$p = P(r_{n,i} \leq r_n) = 1 - [1 - P(r_t \leq r_n)]^n \tag{6.6.1.42}$$

$$1 - p = [1 - P(r_t \leq r_n)]^n$$

So that,

$$\begin{cases} \mu_n - \frac{\sigma_n}{\xi_n}\{1 - [-n\ln(1-p)]^{\xi_n}\}\text{if } \xi_n \neq 0 \\ \mu_n + \sigma_n \ln[-n\ln(1-p)]\text{if } \xi_n = 0 \end{cases}$$

n is the length of each sub period.

Putting in the estimates, VaR is computed as,

$$\hat{R^k} = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1 - \left(-\log\left(1 - \frac{1}{k}\right)\right)^{-\hat{\xi}}\right) & \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\log\left(p - \log\left(1 - \frac{1}{k}\right)\right) & \hat{\xi} = 0. \end{cases} \tag{6.6.1.43}$$

For example, a value of $\hat{R}^{10}$ of 7 means that *on average, the maximum loss observed during a period of one year will exceed* 7% *once in ten years.*

The Peak Over Threshold Method This method focuses on the realizations exceeding a pre-specified (preferably high) threshold. This method is appealing to modern research as it provides a more efficient way of maximizing available data. The basic ingredients required to produce the estimators as outlined in Klüppelberg and Mikosch (1997) is as follows:

i. reliable models for the point process of exceedences.

ii. a sufficiently high threshold u.

iii. estimators for the shape parameter $\hat{\xi}$, the scale parameter $\hat{\sigma}$ and the location parameter $\hat{\mu}$.

Distribution of Exceedances The peak over threshold (POT) method considers the distribution of exceedances over a predefined threshold.

Consider an unknown distribution function F of a random variable X. F_u is the *conditional excesses distribution function at threshold* u and is represented as,

$$F_u(y) = P(X - u \leq y \mid X > u),\ 0 \leq y \leq x_F - u, \tag{6.6.1.44}$$

where X is a random variable , u is a predefined threshold and $y = x - u$ are the excesses. $x_F \leq \infty$ is the right endpoint of F, so that, F_u can thus be written as,

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}. \tag{6.6.1.45}$$

EVT comes in handy in the estimation of F_u as it equips us with a powerful result about the conditional excess distribution function. as stated in the theorem below:

Theorem 6.6.2 (Picklands (1975), Balkema and Haan (1974)). *For a large class of underlying distribution functions F the conditional excess distribution function $F_u(y)$, for large threshold, u, is well approximated by $F_u(y) \approx G_{\xi,\sigma}(y)$, $u \to \infty$ where*

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma} y\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp[-y/\sigma] & \text{if } \xi = 0 \end{cases}$$

for $y \in [0,\ (x_F - u)]$ if $\xi \geq 0$ and $y \in [0,\ -\frac{\sigma}{\xi}]$ if $\xi < 0$.

And

$$G_{0,\sigma}(x) = \lim_{\xi \to 0} G_{\xi\sigma}(x).$$

$G_{\xi,\sigma}$ is the Generalized Pareto Distribution (GPD) and ξ is the shape parameter or tail index of the distribution, it gives an indication of the heaviness of the tail, the larger the value of ξ, the heavier the tail of the distribution.

VaR is computed as– Amount of Position $\times \hat{x}_p$.

Manganelli and Engle (2001) noted that this estimate is only valid for very low p as the approximation is only asymptotically valid.

Expected Shortfall is measured as follows, Define, ES_α as a function of GPD parameters as follows, from Equation (6.6.1.45) we make $F(x)$ subject of formular and obtain the following expression: $F(x) = (1 - F(u))F_u(y) + F(u)$. Replacing F_u by the GPD and $F(u)$ by the estimate $\frac{n-N_n}{n}$, where $n =$ the total number of observation and $N_u =$ the number of observations above the threshold u, we have,

$$\hat{F}(x) = 1 - \frac{N_u}{n}\left(1 + \frac{\hat{\xi}}{\hat{\sigma}}(x-u)\right)^{-1/\xi} \qquad (6.6.1.46)$$

ES_α is estimated as,

$$\widehat{ES}_\alpha = \widehat{VaR}_\alpha + E(X - \widehat{VaR}_\alpha \mid X > \widehat{VaR}_\alpha) \qquad (6.6.1.47)$$

where the second term on the RHS of Equation (6.6.1.47) defines the expected value of the exceedances over the threshold VaR_α. Klüppelberg and Mikosch (1997, Theorem 3.4.13(e)) refers to it as the *mean excess function* for the GPD with parameter $\xi < 1$ and is represented thus,

$$e(v) = E(X - v \mid X > v) = \frac{\sigma + \xi v}{1 - \xi}, \ \sigma + \xi v > 0 \tag{6.6.1.48}$$

$e(v)$ gives the average of the excesses of X over different values of a threshold v.

If X follows a GPD, then, for all integers r such that $r < 1/\xi$, the r first moments exist (Gili and Këllezi (2006)).

Therefore, by definition of ES and Equation (6.6.1.48), for $v = VaR_\alpha - u$ and X representing the excess y over the threshold u we obtain the following expression for ES_α,

$$\widehat{ES}_\alpha = \widehat{VaR}_\alpha + \frac{\hat{\sigma} + \hat{\xi}(\widehat{VaR}_\alpha - u)}{1 - \hat{\xi}} = \frac{\widehat{VaR}_\alpha}{1 - \hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi} u}{1 - \hat{\xi}} \tag{6.6.1.49}$$

Where

$$\widehat{VaR}_\alpha = u + \frac{\hat{\sigma}}{\hat{\xi}} \left(\left(\frac{n}{N_u} p \right)^{-\hat{\xi}} - 1 \right). \tag{6.6.1.50}$$

This expression is obtained by inverting Equation (6.6.1.46) for confidence level α.

A major drawback of the POT method is in the selection of an optimal threshold u. An overly high value of u would result in limited exceedances and consequently high variance estimators.On the other hand, an overly small u would make the estimators biased (Klüppelberg and Mikosch (1997)).

Klüppelberg and Mikosch (1997) suggests that it is possible to choose u asymptotically optimal by quantification of a trade-off between bias and variance.

In summary, EVT is quite a general approach to tail estimation. Its strong points being in the use of the GEV distribution to parametrize the tail is a seemingly less restrictive assumption considering that it covers most of the commonly used distributions. Despite the strong characteristic of EVT it still has a few draw backs some of which include:

1. The i.i.d. assumption for observations is not quite consistent with the characteristics of financial data. Although generalizations to dependent observations have been proposed (see, for example, Leadbetter et al. (1983) or Klüppelberg and Mikosch (1997)), they either estimate the marginal unconditional distribution or impose conditions that rule out the volatility clustering behavior typical of financial data.

2. EVT works best for very low probability levels. The floor of this is hard to determine on a priori ground. Engle and Manganelli (1999) suggest a Monte Carlo study might help give more insight into the speed with which the performance of the EVT estimators deteriorates as we move away from the tail.

3. Another important white space is the issue of the selection of the cut-off point (threshold u) that determines the number of order statistics to be used in the estimation procedure. This problem is similar to the choice of the number of k upper order statistics that enter the Hill estimator. If the threshold is too high, there are too few exceptions and the result is a high variance estimator. On the other hand, a threshold too low produces a biased estimator, because the asymptotic approximation might become very poor. However, the choice of the threshold cannot be determined purely on statistical theory as different investors/organisations have different risk tolerances.

4. The choice of k based on the minimization of the mean squared error results in a biased estimator is also an important point for consideration. Since the

quantile is a non-linear function of the tail estimator, it is crucial to quantify the size of this bias. In the application of EVT to Value at Risk an erroneous estimation of might have very significant consequences on the profitability of a company/financial institution.

6.6.2 Non-Parametric Methods

Financial risk management focuses majorly with the characteristics of the tail part of the loss distribution. Unfortunately, there is usually not enough data at the tail, hence making it difficult to propose an adequate parametric loss model for the tail. Non-parametric methods come in quite handy in this situation as they are model-free and are thus free from the bias caused from using a mis-specified returns distribution.

Non-parametric methods for computing VaR and ES do not make any distributional assumptions about the the portfolio of returns. The historical simulation method also known as the empirical muantile method (Manganelli and Engle (2001)) is the most common example of this. It provides a simple non-theoretical way of estimating VaR and ES. Apart from assuming that the same distribution holds throughout the prediction period, it makes little or no other assumptions about the statistical distribution of the underlying portfolio returns. Before we discuss some non-parametric estimation methods available in literature, we give a brief commentary on quantile estimation and quantile distribution models.

Quantile Estimation Quantile estimation has fast gained ample recognition in applied statistics, especially in areas where obtaining a good fit to the extreme tails of a distributional model is quite crucial. Since value at risk is concerned with the tail behavior of the cumulative distribution function that is the changes in the value of the portfolio over a time period, quantile estimation is an essential tool in enabling us obtain well-fitted tail distributions. Quantile estimation proves a non-parametric approach to calculating Value at Risk, as, the only assumption it makes is that the same distribution holds all through the holding period.

Quantile Distribution Models A continuous probability distribution can be defined in two equivalent ways viz:

- The probability distribution function

$$p = F(x) = prob(X \leq x) \tag{6.6.2.1}$$

- The quantile distribution function

$$x = F^{-1}(p) = Q(p) \tag{6.6.2.2}$$

It is obvious that $F(\cdot)$ and $Q(\cdot)$ are inverse functions of each other. In the same vein, it can be shown that for the probability density function $f(x)$ and the quantile density function $q(p)$ using the properties of $f(x)$ that,

$$p = F(Q(p)) \qquad \text{where } x = F^{-1}(p) = Q(p) \tag{6.6.2.3}$$

Differentiating with respect to p gives

$$1 = f(Q(p))q(p) \tag{6.6.2.4}$$

so that

$$f(x) = 1/q(p) \tag{6.6.2.5}$$

Historical Method (Empirical Quantile Method)

The empirical quantile method is based on the concept of "rolling windows" (Manganelli and Engle (2001)). The first step is to choose a window of observations , this would usually range from 6months to 2years. The next step is to sort the portfolio returns within the selected window in ascending order and the p-quantile of interest is given by the portfolio return that leaves $p\%$ of the observations on the left side and $(1-p)\%$ on its right side. If the figure falls between two consecutive returns, some interpolation rule is then applied to select a number. one-step ahead VaR is computed by moving the entire window forward by one observation and repeating the entire procedure. The Empirical Quantile of a log returns series r_t is used to calculate VaR viz: Let $\{r_t \mid t = 1, \ldots, n\}$ be sample log returns of a portfolio, assume that these returns are iid random variables with both conditional distribution and probability density function (pdf) find the order statistics of the returns, such that $r_{(i)} \leq r_{(j)}$ for $i < j$. For the order statistic $r_{(l)}$, where $l = np$ with $0 < p < 1$, the following results culled from Tsay (1997) holds.

Let x_p be the pth quantile of $F(x)$, that is, $x_p = F^{-1}(p)$. Assume that the pdf $f(x)$ is not zero at x_p (i.e., $f(x_p \neq 0)$). Then the order statistic $r_{(l)}$ is asymptotically normal with mean x_p and variance $p(1-p)/[nf^2(x_p)]$. That is,

$$f_{(l)} \sim N\left[x_p, \frac{p(1-p)}{n[f(x_p)]^2}\right], \qquad l = np \tag{6.6.2.6}$$

Estimate the empirical quantile of r_t so that for a given probability p, if $l = np$ is an integer, then the empirical quantile is $r_{(l)}$, if np is not an integer, we find the two neighbouring positive integers such that $l_1 < np < l_2$ and then interpolate so that the quantile x_p is estimated by:

$$\hat{x}_p = \frac{p_2 - p}{p_2 - p_1} r(l_1) + \frac{p - p_1}{p_2 - p_1} r(l_2) \tag{6.6.2.7}$$

VaR is computed as: Amount of Position $\times \hat{x}_p$.

The historical method computes ES as a weighted average of excessive losses beyond VaR. The simplicity of this method makes it quite attractive in practice.

Denote observed losses (returns on stocks) as $r_1, \ldots, r_n$, with empirical distribution F_n defined as,

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(r_i \leq t), \tag{6.6.2.8}$$

Where $I(\cdot)$ is an indicator function. The nth $\alpha-$ quantile $F^{-1}(\alpha)$ is estimated as

$$F_n^{-1}(\alpha) = r_{n(i)}, \qquad \alpha \in \left(\frac{i-1}{n}, \frac{i}{n} \right) \tag{6.6.2.9}$$

where

$$r_{n(1)} \leq \cdots \leq r_{n(n)} \qquad \text{are the order statistics.} \tag{6.6.2.10}$$

ES at confidence level α is computed by,

$$ES_\alpha = E(r \mid r > VaR) = \frac{\left(\sum_{i=[n\alpha]}^{n} r_{n(i)}\right)}{(n - [n\alpha])} \tag{6.6.2.11}$$

This can also be represented as,

$$\hat{\mu}_\alpha = \frac{\sum_{t=1}^{n} r_t I(r_t \geq \hat{v}_\alpha)}{\sum_{t=1}^{n} I(r_t \geq \hat{v}_\alpha)} = ([n\alpha] + 1)^{-1} \sum_{t=1}^{n} r_t I(r_t \geq \hat{v}_\alpha) \tag{6.6.2.12}$$

where $\hat{v}_p$ is the sample VaR estimator of v_p and r_r is the $r-$th order statistic of $\{r_t\}_{t=1}^{n}$ (see Chen (2008)).

Although VaR calculation using the empirical method is quite simple to implement and makes no distributional assumptions, it is obvious here that there is an implicit assumption that the distribution of portfolio returns remain unchanged within the specified window, hence a major shortfall in the method.

In the first instance, if all the returns within the window are assumed to have the same distribution, then the logical consequence must be that all the returns of the time series must have the same distribution. Since VaR is essentially concerned with the tail probability this assumption implies that the predicted loss cannot be greater than the historical loss, however, this is not at all the case in practice.

In addition,the empirical quantile estimator is consistent if and only if the window length is sufficiently large enough. Manganelli and Engle (2001) suggest that the length of the window must satisfy two contradictory properties: it must be large enough, in order to make statistical inference significant, and it must not be too large, to avoid the risk of taking observations outside of the current volatility cluster. They conclude that the there is no easy solution to this problem.

Another drawback of the method is that it assigns equal weights to each day in the time series this could pose a problem in VaR computation if there is a trend in the variability of the portfolio returns. For example, if the market is moving from a period of relatively low volatility to a period of relatively high volatility (or vice versa). In this case, VaR estimates based on the historical simulation methodology will be biased downwards (correspondingly upwards), since it will take some time before the observations from the low volatility period leave the window (Manganelli and Engle (2001)).

Moreover, VaR estimates based on historical simulation may present predictable jumps, due to the discreteness of extreme returns. To see why, assume that we are computing the VaR of a portfolio using a rolling window of 180 days and that today's return is a large negative number. It is easy to predict that the VaR estimate will jump upward, because of today's observation. The same effect (reversed) will reappear after 180 days, when the large observation will drop out of the window. This is a very undesirable characteristic and causing some questions as to the reliability of the historical simulation method. Finally, the method is hinged on history repeating itself, with the selected window providing a comprehensive snapshot of the market in other windows, so the critical question to ask is "So what if history does not repeat itself or what if history is interrupted?".

Tsay (1997) suggest that the VaR obtained by the empirical quantile can serve as a lower bound for the actual VaR.

Kernel based Method The kernel based estimation method was first proposed by Scaillet (2004), who defined expected shortfall as follows:

$$ES_\alpha = E[-a^{'}X \mid -a^{'}X > VaR_\alpha] \tag{6.6.2.13}$$

where X is such that $X_t, t \in \mathbb{Z}$ the vector $X_t = (X_{1,t}, \ldots, X_{n,t})^{'}$ corresponds to n risks (returns on n stocks over a pre-specified period of time) VaR_α is such that; $P[-a^{'} > VaR_\alpha] = \alpha$ and $a^{'} = (a_1, \ldots, a_n)^{'}$ is the portfolio structure or composition and α is the loss probability or confidence level. The kernel estimator is represented as follows,

$$[(X_t, a^{'}X_t); \xi] = (Th)^{-1} \sum_{t=1}^{T} X_t K((\xi - a^{'}X_t)/h), \tag{6.6.2.14}$$

and

$$\hat{I}(\xi) = \int_{-\infty}^{\xi} [(X_t, a^{'}X_t); u] du. \tag{6.6.2.15}$$

where $h > 0$ is the bandwidth and depends on the sample size T and it is assumed that $h \to 0$ as $T \to \infty$.

The kernel $K(u)$ is a real valued function that at least integrates to 1. The estimate $q(a, \alpha) = -VaR_\alpha$ is given by the kernel estimator $\hat{q}(a, \alpha)$ of the quantile of level α of the distribution of $a^{'}X$ through the equality:

$$\int_{-\infty}^{\hat{q}(a,\alpha)} [(1, a^{'}X_t; u)] du = \alpha \tag{6.6.2.16}$$

The ratio $\hat{I}(\hat{q}(a,\alpha))/\alpha$ then provides an estimate of the conditional expectation $E[X \mid a^{'}X < q(a,\alpha)]$, and expected shortfall can be estimated as,

$$\widehat{ES}_{\alpha} = -a^{'}\hat{I}(\hat{q}(a,\alpha))/\alpha \tag{6.6.2.17}$$

An empirical study by Chen (2008) compared the kernel based method to the historical method (that is, sample average of excessive losses larger than VaR). They conclude that the extra kernel smoothing does not produce more accurate estimation of expected shortfall. The paper concluded that expected shortfall is effectively a mean parameter and can be estimated rather accurately by simple averaging.

6.7 An MAR approach to measuring VaR and ES

We apply the MAR models described in Chapter 2 to estimating VaR and ES, based on the model described in Equation (2.2.0.1). The parameters of the model is estimated by the Maximum (conditional) likelihood method using the EM algorithm of Dempster et al. (1977). The standard errors of this parameter estimates can be computed using Louis (1982) (see Wong and Li (2001) for a more detailed description). One step ahead predictive distribution is then computed for the returns series based on the MAR model (see Boshnakov (2009)).

VaR is computed as the $100_{\alpha}\%$ quantile of the predictive distribution and ES is computed as, $E[r_t \mid r_t > VaR_{\alpha}]$.

6.8 Backtesting Value at Risk and Expected Shortfall

In the previous section, we described various methods for estimating VaR and ES, it is however important to also examine and compare the predictive performance of these methods over a historical period. This process is referred to as backtesting.

6.8.1 Backtesting VaR models

To backtest the VaR methods examined, we lean on Christoffersen (1998)'s framework for evaluating the accuracy of out-of-sample interval forecasts.

Define the indicator function $H_t = \mathcal{I}(r_t < VaR_{\alpha,t})$ as the VaR violation indicator such that,

$$H_t = \mathcal{I}(r_t < VaR_{\alpha,t}) = \begin{cases} 1 & r_t < VaR_{\alpha,t} \\ 0 & r_t \geq VaR_{\alpha,t} \end{cases} \tag{6.8.1.1}$$

Christoffersen (1998) says that VaR forecasts are efficient with respect to $\mathcal{F}_t$ if,

$$E[H_t \mid \mathcal{F}_{t-1}] = 1 - \alpha = \lambda, \tag{6.8.1.2}$$

H_t is thus uncorrelated with any function of a variable in the information set available at time $t-1$ (Kuester et al. (2006)). So that if Equation (6.8.1.2) holds, then VaR violations will occur with the correct conditional and unconditional probability. If $\mathcal{F}_{t-1}$ is specified to include at least $\{H_1, \ldots, H_{t-1}\}$, then by Christoffersen (1998, lemma 1), efficiency implies correct conditional coverage denoted by,

$$H_t \mid \mathcal{F}_{t-1} \sim Bernoulli(\lambda), \qquad t = 1, \ldots, T. \tag{6.8.1.3}$$

Test of Unconditional Coverage

Let n_1 be the total number of sample VaR violations, $T = n_0 + n_1$ be the total number of observations and $\hat{\lambda}_{mle}$ the maximum -likelihood estimation of λ. $\hat{\lambda}_{mle}$ is the ratio of the total number of violations to the total number of observations that is, $\hat{\lambda}_{mle} = n_1/T$. The test for correct number of violations is carried out by the following hypothesis.

$$H_0 : E[H_t] = \lambda \qquad vs. \qquad H_1 : E[H_t] \neq \lambda. \tag{6.8.1.4}$$

Under the null hypothesis, Equation (6.8.1.3) implies the likelihood-ratio test statistic,

$$LR_{uc} = 2[\mathcal{L}(\hat{\lambda}_{mle} : H_1, \ldots, H_T) - \mathcal{L}(\lambda : H_1, \ldots, H_T)] \sim \chi^2(1) \tag{6.8.1.5}$$

where $\mathcal{L}(\cdot) = \ln(L)$ and $L = f(\lambda : H_1, \ldots, H_T)$ is the bernoulli likelihood.

Test of Independence

The test of unconditional coverage does not consider the possibility that VaR forecast that do not take temporal volatility dependence into account although will produce violation clusters, might still be correct on the average. A test for independence is defined as a test that there are no violation clusters, that is, all the VaR violations are independent. Under the null hypothesis that a violation today has no influence on the probability of a violation tomorrow against an alternative hypothesis of dependence. Christoffersen (1998) models $\{H_t\}$ as a binary first order Markov chain with transition matrix,

$$\Pi = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}, \qquad \pi_{ij} = Pr(H_t = j \mid H_{t-1} = i), \tag{6.8.1.6}$$

the approximate joint likelihood, conditional on the first observation is given by,

$$L(\Pi : H_2, \ldots, H_T \mid H_1) = (1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi_{11})^{n_{10}} \pi_{11}^{n_{11}}, \tag{6.8.1.7}$$

where n_{ij} is the number of transitions form state i to state j, that is , $n_{ij} = \sum_{t=2}^{T} \mathcal{I}(H_t = i \mid H_{t-1} = j)$,

The maximum likelihood of these transition probabilities under the alternative hypothesis are given by,

$$\hat{\pi}_{01,mle} = \frac{n_{01}}{n_{00} + n_{01}}, \qquad \hat{\pi}_{11,mle} = \frac{n_{11}}{n_{10} + n_{11}} \tag{6.8.1.8}$$

Under the null hypothesis of independence, we have that $\pi_{01} = \pi_{11} \equiv \pi_0$ and

$$\begin{aligned} L(\pi_0 : H_2, \ldots, H_T \mid H_1) &= (1 - \pi_{01})^{(n_{00}+n_{10})} \pi_{01}^{n_{01}+n_{11}} \qquad (6.8.1.9) \\ \hat{\pi}_0 = \hat{\lambda}_{mle} &= n_1/T. \end{aligned}$$

So that the LR test for independence of VaR violations is given by,

$$LR_{ind} = 2[\mathcal{L}(\hat{\Pi} : H_2, \ldots, H_T \mid H_1) - \mathcal{L}(\hat{\pi}_0 : H_2, \ldots, H_T \mid H_1)] \sim \chi^2(1) \tag{6.8.1.10}$$

Reject the null hypothesis $(H_0 : \pi_{01} = \pi_{11} \equiv \pi_0)$ if $LR_{ind} > \chi^2_{\alpha}(1)$

Test of Conditional Coverage

Since $\hat{\pi}_{01,mle}$, is unconstrained, the LR test for independence in Equation (6.8.1.10) does not consider correct coverage. To test this, Christoffersen (1998) suggested a test that combines both the test for unconditional coverage (Equation (6.8.1.5)) and the test for independence (Equation (6.8.1.10)) as follows,

$$\begin{aligned} LR_{cc} &= 2[\mathcal{L}(\hat{\Pi} : H_2, \ldots, H_T \mid H_1) - \mathcal{L}(\lambda : H_2, \ldots, H_T \mid H_1)] \sim \chi^2(2) \qquad (6.8.1.11) \\ &= LR_{cc} = LR_{uc} + LR_{ind}, \end{aligned}$$

So that the instances in which the violation series H_t fails the correct conditional coverage property, that is, Equation (6.8.1.3) , can be checked.

6.8.2 Backtesting ES models

Recall that the expected shortfall $ES_{\alpha,t}$ is the conditional loss distribution. Define the residuals

$$\epsilon_t = \frac{r_t - ES_{\alpha,t-1}}{\sigma_t} = Z_{t+1} - E[Z \mid Z > z_q] \tag{6.8.2.1}$$

where $r_t - ES_{\alpha,t-1}$ is the difference between r_t and $ES_{\alpha,t-1}$ in the event of a quantile violation. The process $S_t = r_t - ES_{\alpha,t-1}$ forms a martingale difference series satisfying $E[(r_t - ES_{\alpha,t-1}) \mid \mathcal{F}_t] = 0$. The residuals form an i.i.d zero mean sequence of innovation variables. So that when risk measures and volatility are estimated in practice, the following violation residuals can be formed (McNeil et al. (2010)),

$$\hat{\epsilon}_t = \hat{S}_t/\hat{\sigma}_t, \qquad \hat{S}_t = r_t - \hat{ES}_{\alpha,t-1} \tag{6.8.2.2}$$

These violation residuals are expected to behave like realisations of iid variables from a distribution with mean zero, variance one and an atom of probability mass of size α at zero (McNeil et al. (2010)). To test for the hypothesis of a zero mean behaviour against the alternative of a mean greater than zero,a bootstrap test which makes no assumption about the underlying distribution of the residuals is performed (see Efron and Tibshirani (1993) for a full description). Note that this test can be applied to either the standardized or unstandardized residuals to achieve similar results (McNeil et al. (2010)).

6.9 Summary

A major part of decision making involves risk taking, thus, most decision makers are faced with the question of how to quantify their risks. Investors are also faced with the dilemma of how much they can possibly lose on an investment as well as the loss threshold (that is the maximum that can be lost on an investment). Value at Risk (VaR) attempts to provide answers to quantify risk as well as the causes of risk and in essence suggest ways of reducing risk. VaR does not satisfy the subadditivity axiom which contradicts the framework of modern portfolio theory, that is, diversification should reduce risk (Jadhav et al. [2009]). Based on these shortfalls of VaR, a coherent risk measure referred to in literature as Expected Shortfall has been suggested as an alternative and/or sometimes a complement to VaR. ES is coherent (that is ,monotonous, sub-additive , positively homogeneous and translative invariant).

Existing models for VaR and ES calculation can be categorized into Parametric Methods (The [JP morgan's] Risk Metrics Model (sometimes referred to as the Variance-Covariance approach) (S.Tsay (1997)) ,GARCH Model (Jorion (1997))), Non-Parametric Methods (Historical Simulation (Manganelli and Engle (2001))) and Semi-Parametric Methods (Extreme Value Theory (Klüppelberg and Mikosch (1997))).

We propose the use of the MAR model as a viable approach for evaluating VaR and hence ES. We also describe backtesting methodology for VaR and ES.

In the next chapter, we compare various approaches for measuring VaR and ES to an approach based on the MAR model.

Chapter 7

Mixture Autoregressive models and Financial Risk

We focus here on two popular methods of predicting financial risk, viz. Value-at-Risk and Expected shortfall. We compare the performance of the existing methods of estimating one-step ahead VaR and ES to that based on MAR models. In particular, we consider the MAR(3;2,2,1) model with both Gaussian and Student-t innovations. The MAR(3;2,2,1) is a three–component MAR model. The components are one AR(2) model and two AR(1) models. The methods considered for comparison include:

- Risk metrics,
- Gaussian GARCH(1,1),
- Student-t GARCH(1,1),
- AR(2)-GARCH(1,1) with Gaussian innovations,
- AR(2)-GARCH(1,1) with Student-t innovations,

- Empirical Quantile,
- Traditional Extreme value theory and
- Point over Threshold Extreme value theory.

The data series investigated include two currency exchange rate (GBP/USD and GBP/EUR), two stock market indices (adjusted closing prices of the Standard and Poor (S&P500) and FTSE100), and individual stock (adjusted closing price of IBM stocks). The data sets cover the period between June 2002 and June 2012. The data sets are restricted to daily figures, as in practice regulators require that risk measures are computed on a daily basis.

7.1 Descriptive Statistics of Daily Returns

This study focuses on asset returns rather than asset prices. Here we take the natural logarithm of daily returns. By the logarithmic law a multiple period return is additive, that is, it is the sum of the one-period returns involved. The returns are computed as,

$$r_t = -\log\left(\frac{X_t}{X_{t-1}}\right) \times 100 = -[\log(X_t) - \log(X_{t-1})] \times 100 \qquad (7.1.0.1)$$

where X_t is the daily closing value at day t of stock prices, stock indices, foreign exchange rates, etc.

Some descriptive statistics of the returns series are presented in Table 7.1 and the time series plots, histogram, normal QQ plots and ACF of each of the data sets are presented in Figures 7.4-7.8.

Kurtosis and skewness are of special interest when modelling extreme events in risk management. Kurtosis describes the tails of a probability distribution. A

Table 7.1: Descriptive statistics for daily logarithmic returns

Symbol	n	Mean	Median	Std	Kur	Skw	Max	Min
IBM	2519	0.0460	0.0368	1.5217	5.8607	0.2555	10.9007	-8.6597
S&P500	2519	0.0118	0.0776	1.3761	8.4558	-0.2067	10.9572	-9.4695
FTSE100	2526	0.0077	0.0430	1.3377	6.3262	-0.1262	9.3842	-9.2646
GBP/USD	3658	0.0012	0.0000	0.4550	6.5744	-0.4825	3.1288	-3.9915
GBP/EUR	3658	-0.0062	0.0000	0.3724	5.8485	-0.1909	2.9188	-2.9020

normal distribution has a kurtosis equal to three regardless of the mean or standard deviation. A distribution is called leptokurtic or ”fat-tailed” if its kurtosis is greater than three. From the statistics in Table 7.1 all of the data sets are fat-tailed with kurtosis greater than three. S&P500 has the highest kurtosis (8.4558).

Skewness is a measure of asymmetry of a distribution. A skewness equal to zero suggests a symmetrical distribution, the Gaussian distribution has skewness equal to zero. Table 7.1 infers that none of the data sets are symmetrical, IBM is positively skewed while all the other data sets are negatively skewed.

Note that we do not give standard errors in Table 7.1, since they would be based on unrealistic assumptions, such as, independence. The results of Table 7.1 are buttressed by a close examination of the plot of the third and fourth moments of the data sets. The moments are calculated on a 250 point rolling window. The plots are presented in Figure 7.2 and 7.3 below, the behaviour clearly shows that the kurtosis of all five data sets are a far cry from that of the Gaussian distribution.

The time series plots (see Figure 7.1) show that all the time series examined are non stationary and non uniform with periods of high volatility. The histograms in Figures 7.4–7.8 also show that all the datasets are far from normal.

The heavy tails of the data sets are further revealed in the normal QQ plots of all the series versus a normal distribution. It can be seen that sample quantiles in

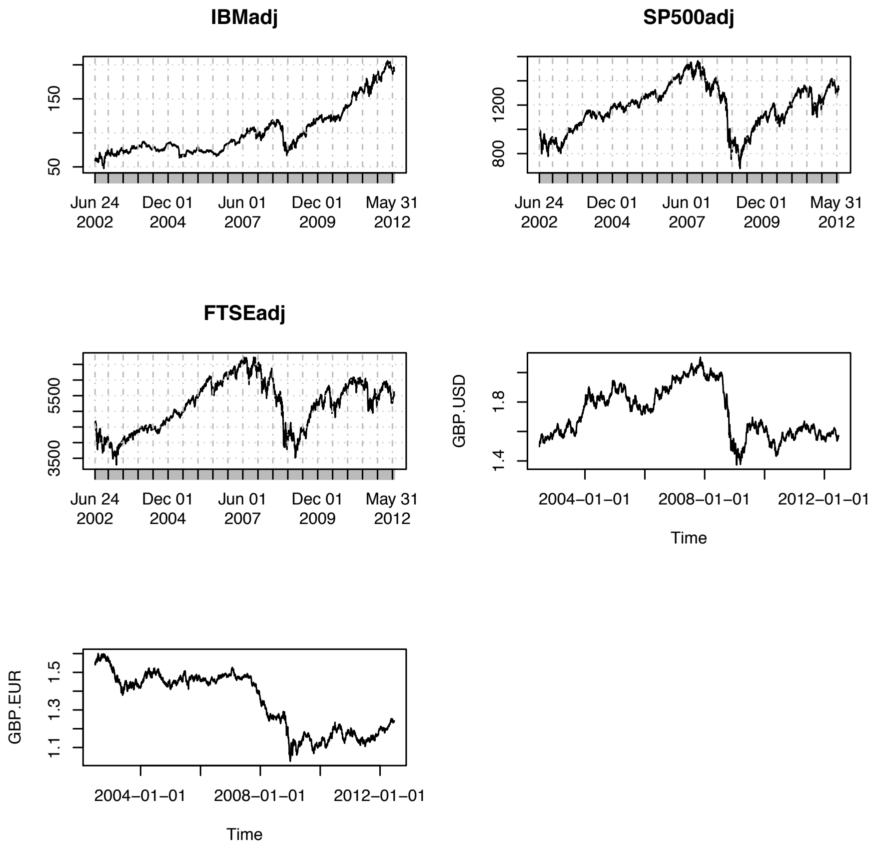

Figure 7.1: Adjusted IBM, FTSE, SP500, GBP/USD and GBP/EUR closing prices from 2002-06-24 to 2012-06-22

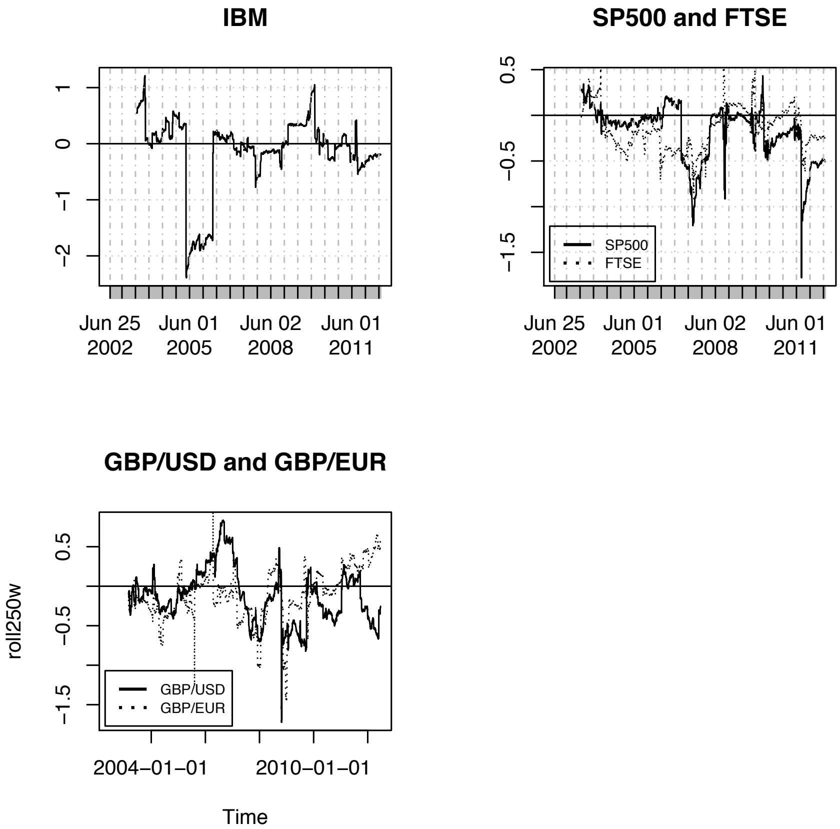

Figure 7.2: Skewness plots of adjusted IBM, FTSE, SP500, GBP/USD and GBP/EUR closing prices from 2002-06-24 to 2012-06-22

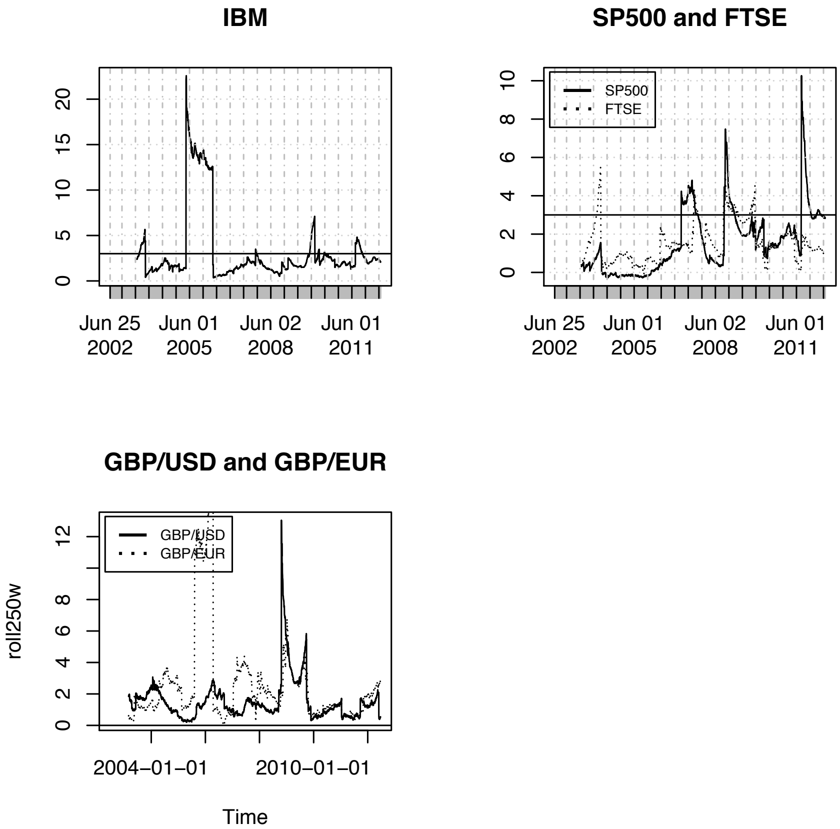

Figure 7.3: Kurtosis plots of adjusted IBM, FTSE, SP500, GBP/USD and GBP/EUR closing price from 2002-06-24 to 2012-06-22

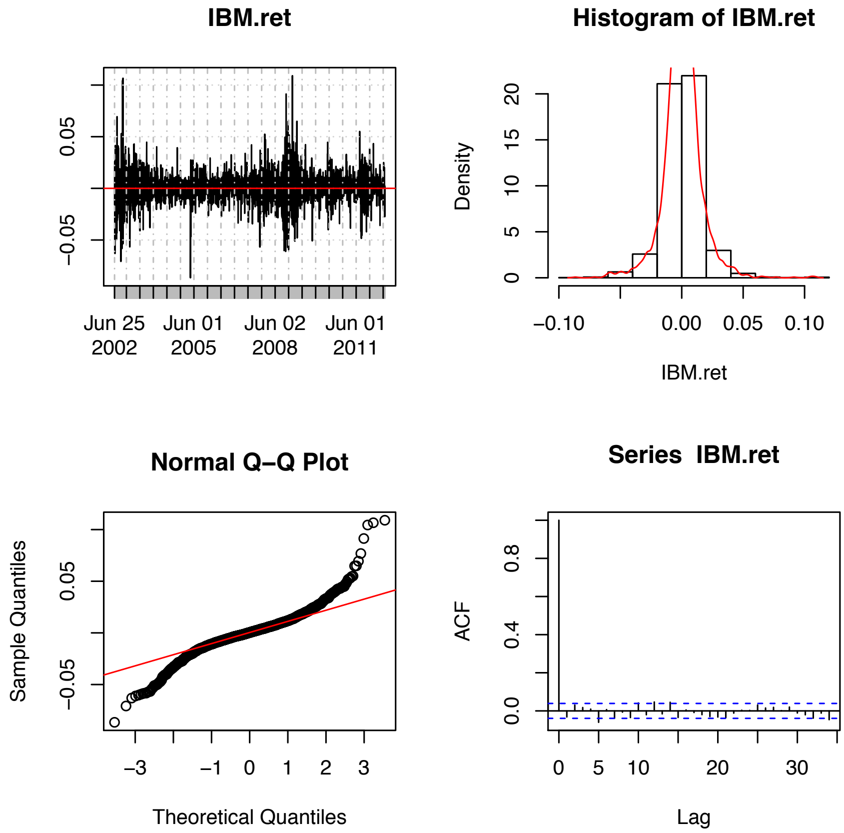

Figure 7.4: Time series plot, histogram, normal QQ plot and ACF of IBM returns from 2002-06-24 to 2012-06-22

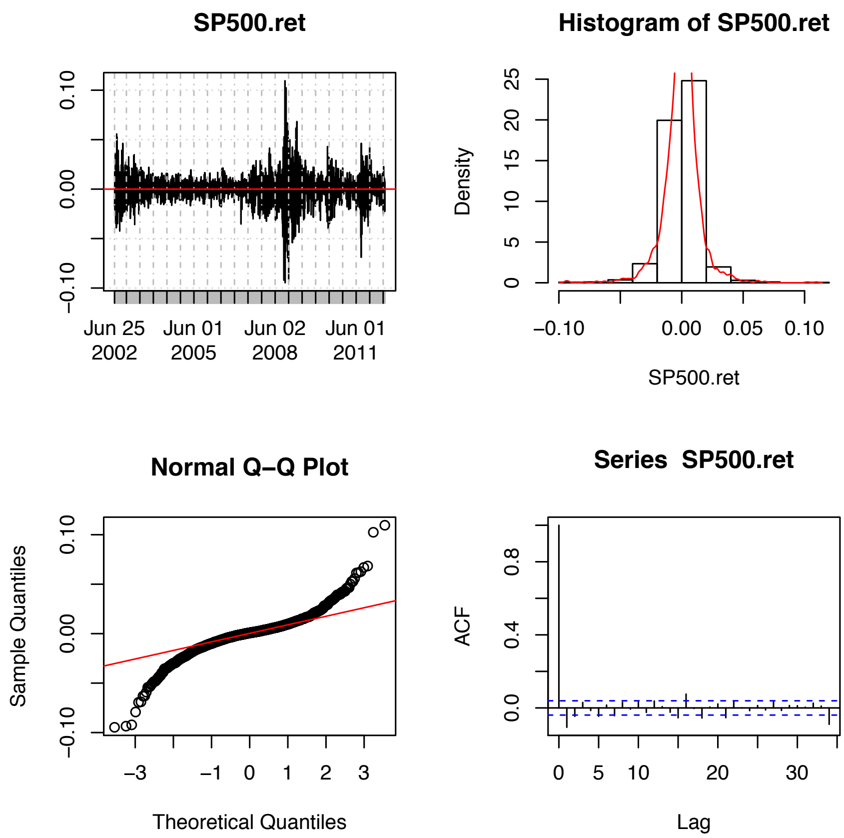

Figure 7.5: Time series plot, histogram, normal QQ plot and ACF of SP500 returns from 2002-06-24 to 2012-06-22

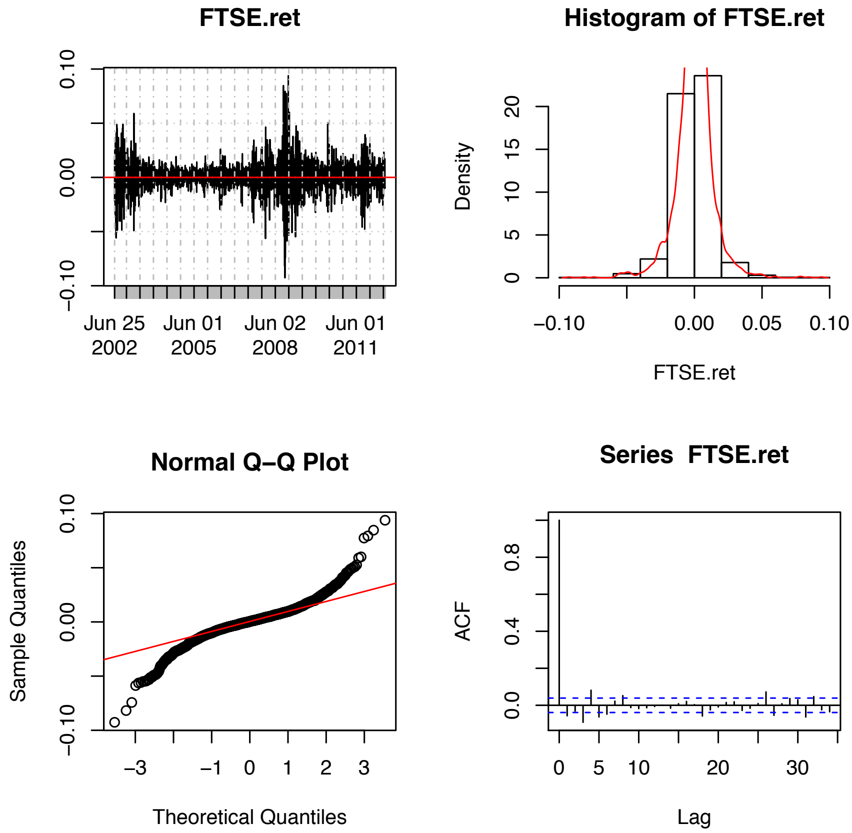

Figure 7.6: Time series plot, histogram, normal QQ plot and ACF of FTSE returns from 2002-06-24 to 2012-06-22

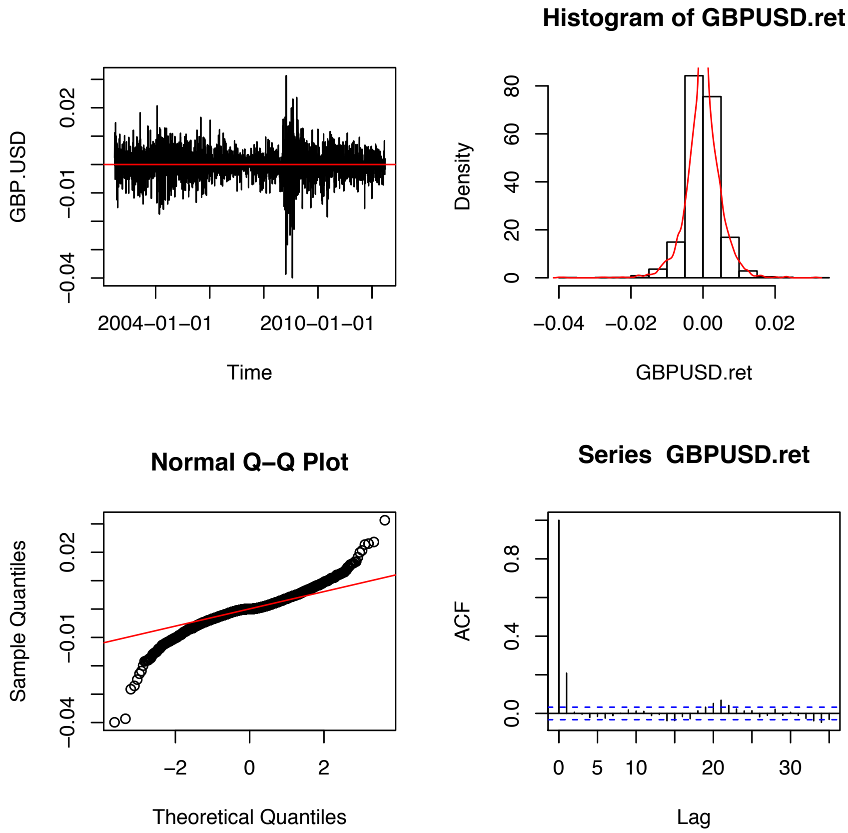

Figure 7.7: Time series plot, histogram, normal QQ plot and ACF of GBP/USD returns from 2002-06-24 to 2012-06-22

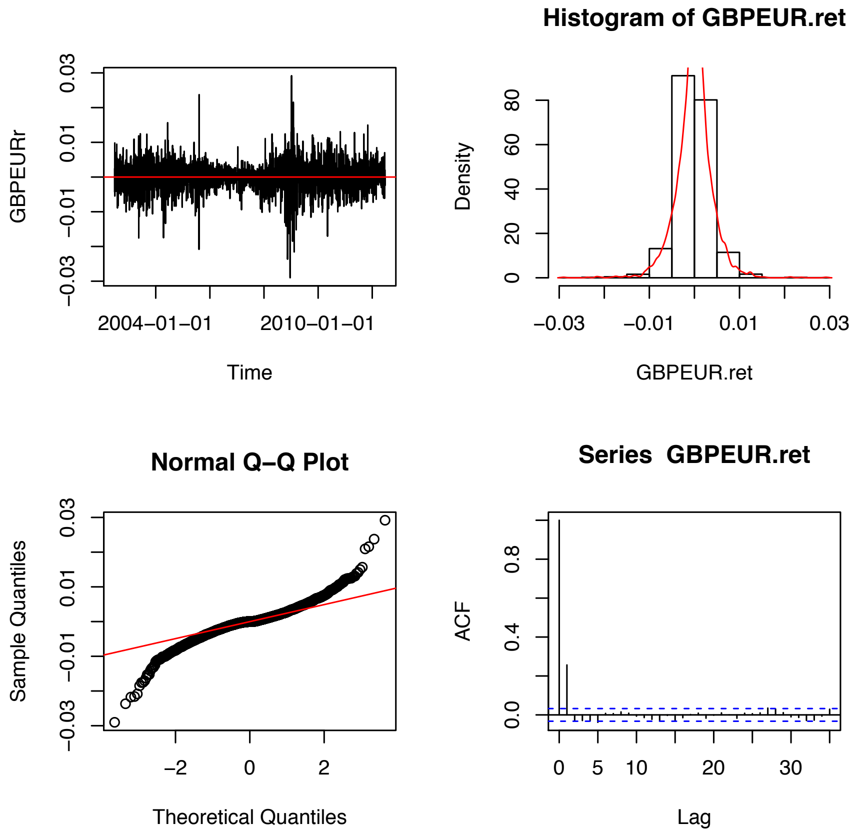

Figure 7.8: Time series plot, histogram, normal QQ plot and ACF of GBP/EUR returns from 2002-06-24 to 2012-06-22

the tails obviously deviate from the corresponding normal quantiles.

7.2 VaR and ES estimation results

In order to adequately compare the performance of the various approaches for estimating VaR and ES described in Section 6.6, we consider the one-step ahead out-of-sample VaR and ES estimates. We calculate VaRs and ESs at $\alpha = 1\%$ and $\alpha = 5\%$ for each of the financial time series described in Section 7.1. All the risk measures are computed on a rolling window of 1000 data points. Tables 7.2–7.6 show the results of the estimated VaR and ES. The results are interpreted thus. If an investor holds a long position worth 100,000GBP in FTSE100 stocks, then the estimated 1-day horizon VaR based on the MAR(3;2,2,1) model with Gaussian innovations at 1% probability is computed as,
100,000X0.0358=3,580GBP and the corresponding 1-day horizonVaR at 5% probability is computed as,
100,000X0.0215=2,150GBP. The corresponding expected shortfalls would then be 4,550GBP and 3,060GBP for probabilities 1% and 5% respectively.

A close examination of the figures reveals distinct differences between the approaches, as well as the value of the tail probability α. We examine here only $\alpha = 1\%$ and $\alpha = 5\%$ because $\alpha = 5\%$ is most commonly used in practise. We have included $\alpha = 1\%$ to give a sounder basis for comparison.

Tsay (1997), suggests that for daily returns, the empirical quantiles of 5% and 1% are decent estimates of the quantiles of the return distribution. We follow this suggestion and hence, treat the results based on empirical quantiles as conservative estimates of the true VaR (i.e., lower bounds). Tsay (1997) noted also that a very small tail probability (say 0.1%), would make the empirical quantile a less reliable estimate of the true quantile, hence the VaR based on empirical quantiles would no longer serve as a lower bound of the true VaR, in which case model based estimates would be better.

Estimating the tail behaviour of a statistical distribution has a lot of uncertainty associated to it, hence different underlying models will inherently give different results. Since there is no benchmark VaR value to pitch the accuracy of each result against, we simply comment on the range of values across all the methods used, and pay close attention to the figures that grossly differ from the others. However, we find that similar classes of models tend to give VaR values within the same range. For example, it can be seen from Tables 7.2–7.6 that computations based on various classes of GARCH models give VaR measures within the range of $0.025 - 0.0265$ at $\alpha = 1\%$ for SP500 and $0.0180 - 0.0187$ at $\alpha = 5\%$.

The tail probability α shows up as very important in VaR and hence ES calculations, as seen in Tables 7.2–7.6 below, VaR and ES at $\alpha = 1\%$ tend to be larger than those for $\alpha = 5\%$. We find that the approaches based on EVT and MAR models give significantly better results as they give values close to the empirical quantiles, while the approaches based on GARCH models tend to underestimate VaR and ES, these results agree with the results in Tsay (1997).

In the next section, we examine the backtest results.

7.3 Backtest Results

We examine here the performance of the various approaches to VaR and ES in Tables 7.2–7.6. We show that the MAR(3;2,2,1) models with both Gaussian and Student-t innovations perform better in more instances than most of the other models. The backtesting procedures adapted here are mostly based on the framework developed by Kupiec (1995) and Christoffersen (1998). They examine whether the

Table 7.2: 1% and 5% VaR/ES computation for daily IBM log returns (returns are in percentages)

	VaR		ES	
	1%	5%	1%	5%
Riskmetrics	0.0268	0.0189	0.0307	0.0237
GARCH(1,1)-norm	0.0305	0.0216	0.0350	0.0271
AR(2)-GARCH(1,1)-norm	0.0303	0.0215	0.0348	0.0269
GARCH(1,1)-t	0.0293	0.0207	0.0350	0.0260
AR-GARCH(1,1)-t	0.0292	0.0207	0.0335	0.0259
Empirical Quantile	0.0458	0.0227	0.0557	0.0359
EVT Threshold (0.02)	0.0427	0.0240	0.0607	0.0365
EVT -GEV	0.0458	0.0227	0.0560	0.0359
MAR(3;2,2,1)-norm	0.0413	0.0252	0.0490	0.0352
MAR(3;2,2,1)-t	0.0401	0.0213	0.0637	0.0350

Table 7.3: 1% and 5% VaR/ES computation for daily S&P500 log returns (returns are in percentages)

	VaR		ES	
	1%	5%	1%	5%
Riskmetrics	0.0254	0.0180	0.0291	0.0226
GARCH(1,1)-norm	0.0264	0.0187	0.0302	0.0234
AR(2)-GARCH(1,1)-norm	0.0261	0.0184	0.0299	0.0231
GARCH(1,1)-t	0.0265	0.0187	0.0302	0.0235
AR-GARCH(1,1)-t	0.0263	0.0186	0.0302	0.0234
Empirical Quantile	0.0409	0.0216	0.0573	0.0341
EVT Threshold (0.019)	0.0391	0.0192	0.0541	0.0319
EVT -GEV	0.0411	0.0217	0.0579	0.0341
MAR(3;2,2,1)-norm	0.0358	0.0215	0.0455	0.0306
MAR(3;2,2,1)-t	0.0349	0.0169	0.0549	0.0294

Table 7.4: 1% and 5% VaR/ES computation for daily FTSE100 log returns (returns are in percentages)

	VaR		ES	
	1%	5%	1%	5%
Riskmetrics	0.0243	0.0172	0.0278	0.0215
GARCH(1,1)-norm	0.0243	0.0172	0.0279	0.0216
AR(2)-GARCH(1,1)-norm	0.0243	0.0172	0.0279	0.0216
GARCH(1,1)-t	0.0245	0.0174	0.0279	0.0218
AR-GARCH(1,1)-t	0.0246	0.0174	0.0281	0.0218
Empirical Quantile	0.0402	0.0218	0.0536	0.0328
EVT Threshold (0.02)	0.0367	0.0195	0.0517	0.0308
EVT -GEV	0.0403	0.0218	0.0540	0.0329
MAR(3;2,2,1)-norm	0.0423	0.0194	0.0503	0.0331
MAR(3;2,2,1)-t	0.0382	0.0162	0.0612	0.0309

Table 7.5: 1% and 5% VaR/ES computation for monthly GBP/USD exchange rates

	VaR		ES	
	1%	5%	1%	5%
Riskmetrics	0.0072	0.0051	0.0082	0.0064
GARCH(1,1)-norm	0.0077	0.0054	0.0088	0.0068
AR(2)-GARCH(1,1)-norm	0.0073	0.0052	0.0083	0.0065
GARCH(1,1)-t	0.0081	0.0057	0.0088	0.0072
AR-GARCH(1,1)-t	0.0076	0.0054	0.0087	0.0067
Empirical Quantile	0.0125	0.0072	0.0179	0.0111
EVT Threshold (0.0072)	0.0118	0.0072	0.0154	0.0102
EVT -GEV	0.0125	0.0072	0.0180	0.0111
MAR(3;2,2,1)-norm	0.0150	0.0081	0.0182	0.0123
MAR(3;2,2,1)-t	0.0137	0.0069	0.0196	0.0113

Table 7.6: 1% and 5% VaR/ES computation for monthly GBP/EUR exchange rates

	VaR		ES	
	1%	5%	1%	5%
Riskmetrics	0.0068	0.0048	0.0077	0.0060
GARCH(1,1)-norm	0.0071	0.0050	0.0081	0.0063
AR(2)-GARCH(1,1)-norm	0.0069	0.0049	0.0080	0.0062
GARCH(1,1)-t	0.0072	0.0051	0.0081	0.0064
AR-GARCH(1,1)-t	0.0069	0.0049	0.0079	0.0061
Empirical Quantile	0.0101	0.0061	0.0142	0.0090
EVT Threshold (0.0057)	0.0099	0.0057	0.0132	0.0084
EVT -GEV	0.0101	0.0061	0.0143	0.0090
MAR(3;2,2,1)-norm	0.0096	0.0054	0.0114	0.0080
MAR(3;2,2,1)-t	0.010	0.0044	0.0161	0.0083

approach adequately forecasts the expected number of violations, generates independent violations and consequently give an ES whose violation residuals have zero mean behaviour. We have given detailed descriptions of these tests in Sections 6.8.1 and 6.8.2.

7.3.1 VaR Backtest Results

Here we implement both the unconditional (Kupiec (1995)) and conditional (Christoffersen (1998)) coverage tests for the correct number of exceedances and independence of these exceedances (see Section 6.8.1 for details). We carry out backtesting procedures for each of the data sets independently for tail probability $\alpha = 1\%$ and $\alpha = 5\%$. For each of the VaR estimation methods considered, we use a rolling window of size 1000 and compute rolling 1-step ahead out-of-sample VaR forecasts.

The results are presented in Tables 7.7–7.16 below. At 99% confidence level, a p-value less than 0.01 is interpreted as evidence against the null hypothesis and similarly at 95% confidence level, a p-value less than 0.05 is interpreted as evidence against the null. The results are generated using the R packages listed in Appendix B.

Test for Correct Exceedances

Here we test the null hypothesis that the model gives the correct number of violations against the alternative that it does not. So that at significance levels 95% and 99%, we would not reject the respective null hypothesis for p-values higher than 0.05 and 0.01 respectively, indicating that the model is a good model. The results of the test for correct exceedances for VaR at tail probabilties $\alpha = 5\%$ and $\alpha = 1\%$ are documented in Tables 7.7–7.11. The first two columns in the tables give results for the expected and actual exceedances, both at $\alpha = 5\%$ and $\alpha = 1\%$. The next two columns give the likelihood ratio statistic and the critical region. The last two columns give the $p-$value and the decision based on the $p-$value. A decision "Fail to Reject H0" is an indication that the model captures the correct number of exceedances.

We find here that across all the data sets examined, the MAR(3;2,2,1) models consistently perform well at both 95% and 99% confidence levels, as we fail to reject the null of correct number of violations (exceedances). In particular, we find based on the $p-$value that for individual stock data set (IBM) and both exchange rate data sets (GBP/USD and GBP/EUR), most of approaches for predicting VaR seem to give correct exceedances. These results are further buttressed by a close examination and comparison of the expected and actual exceedances columns which reveal that for these data sets, the figures in these two columns are not too far from each other at $\alpha = 5\%$ but a bit farther off at $\alpha = 1\%$. These figures are closest for the GARCH based parametric models and the MAR(3;2,2,1) models, than for the empirical quantile (non-parametric) and the EVT (semi-parametric) methods.

However, for the stock indices, viz SP500 and FTSE (see Tables 7.8 and 7.9), only the MAR(3;2,2,1) models give correct exceedances at both significance levels considered. Furthermore, a comparison of the expected and actual exceedances columns reveals that only the MAR(3;2,2,1) models give figures close together. These conclusions can be attributed to the fact that stock indices consist of a portfolio of various individual stocks and hence are most likely to exhibit more volatile behaviour. This strongly supports a claim that the MAR models are better suited to capture dynamics represented in stock indices.

Method	Parameters									
	Exp. Exceed		Act. Exceed		UC.LRstat		UC.LRp		UC.Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	15.00	75.00	29.00	73.00	10.01	0.12	0.00	0.73	Reject H0	Fail to Reject H0
GARCH-norm	15.00	75.00	25.00	68.00	5.36	0.91	0.02	0.34	Reject H0	Fail to Reject H0
AR(2)-GARCH-norm	15.00	75.00	22.00	70.00	2.71	0.50	0.10	0.48	Fail to Reject H0	Fail to Reject H0
GARCH-t	15.00	75.00	74.00	74.00	119.06	0.05	0.00	0.82	Reject H0	Fail to Reject H0
AR(2)-GARCH-t	15.00	75.00	19.00	74.00	0.89	0.05	0.34	0.82	Fail to Reject H0	Fail to Reject H0
Empirical Quantile	15.00	75.00	25.00	102.00	5.36	8.53	0.02	0.00	Reject H0	Reject H0
EVT	15.00	75.00	25.00	98.00	5.36	6.20	0.02	0.01	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	95.00	447.00	0.28	0.00	0.60	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	105.00	455.00	2.40	0.00	0.12	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.7: IBM daily returns: Backtesting VaR Results,Test for Correct exceedances. The $p-$values are obtained by comparison with the $\chi^2(1)$-distribution. The $\chi^2(1)$-distribution has a 5 percent critical value of 3.84 and a 1 percent critical value of 6.23. For a good model, we expect NOT to reject H0

Method	Exp. Exceed 1%	Exp. Exceed 5%	Act. Exceed 1%	Act. Exceed 5%	UC.LRstat 1%	UC.LRstat 5%	UC.LRp 1%	UC.LRp 5%	UC.Decision 1%	UC.Decision 5%
Riskmetrics	15.00	75.00	40.00	99.00	28.25	6.75	0.00	0.01	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	75.00	47.00	102.00	43.23	8.53	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	75.00	46.00	102.00	40.95	8.53	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-t	15.00	75.00	103.00	103.00	223.91	9.17	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	15.00	75.00	44.00	102.00	36.53	8.53	0.00	0.00	Reject H0	Reject H0
EVT	15.00	75.00	41.00	120.00	30.25	23.04	0.00	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	102.00	461.00	1.55	0.00	0.21	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	91.00	460.00	0.01	0.00	0.92	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.8: SP500 daily returns: Backtesting VaR Results,Test for Correct exceedances. The $p-$values are obtained by comparison with the $\chi^2(1)$-distribution. The $\chi^2(1)$-distribution has a 5 percent critical value of 3.84 and a 1 percent critical value of 6.23. For a good model, we expect NOT to reject H0

Method	Exp. Exceed 1%	Exp. Exceed 5%	Act. Exceed 1%	Act. Exceed 5%	UC.LRstat 1%	UC.LRstat 5%	UC.LRp 1%	UC.LRp 5%	UC.Decision 1%	UC.Decision 5%
Riskmetrics	15.00	76.00	33.00	103.00	15.63	8.91	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	76.00	34.00	104.00	17.23	9.55	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	76.00	36.00	105.00	20.60	10.22	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-t	15.00	76.00	105.00	105.00	230.99	10.22	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	15.00	76.00	32.00	106.00	14.10	10.91	0.00	0.00	Reject H0	Reject H0
Empirical Quantile	15.00	76.00	32.00	123.00	14.10	25.59	0.00	0.00	Reject H0	Reject H0
EVT	15.00	76.00	27.00	122.00	7.42	24.58	0.01	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	91.00	462.00	0.01	0.00	0.92	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	92.00	458.00	0.04	0.00	0.83	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.9: FTSE daily returns: Backtesting VaR Results,Test for Correct exceedances. The $p-$values are obtained by comparison with the $\chi^2(1)$-distribution. The $\chi^2(1)$-distribution has a 5 percent critical value of 3.84 and a 1 percent critical value of 6.23. For a good model, we expect NOT to reject H0

Method	Exp. Exceed 1%	Exp. Exceed 5%	Act. Exceed 1%	Act. Exceed 5%	UC.LRstat 1%	UC.LRstat 5%	UC.LRp 1%	UC.LRp 5%	UC.Decision 1%	UC.Decision 5%
Riskmetrics	26.00	132.00	53.00	155.00	20.72	3.80	0.00	0.05	Reject H0	Fail to Reject H0
GARCH(1,1)-norm	26.00	132.00	36.00	132.00	3.09	0.00	0.08	0.96	Fail to Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-norm	26.00	132.00	34.00	127.00	1.96	0.25	0.16	0.62	Fail to Reject H0	Fail to Reject H0
GARCH(1,1)-t	26.00	132.00	138.00	138.00	237.15	0.23	0.00	0.63	Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-t	26.00	132.00	11.00	136.00	11.76	0.09	0.00	0.76	Reject H0	Fail to Reject H0
Empirical Quantile	26.00	132.00	35.00	113.00	2.50	3.19	0.11	0.07	Fail to Reject H0	Fail to Reject H0
EVT	26.00	132.00	34.00	110.00	1.96	4.28	0.16	0.04	Fail to Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	110.00	458.00	4.19	0.00	0.04	1.00	Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	101.00	454.00	1.31	0.00	0.25	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.10: GBP/USD daily returns: Backtesting VaR Results,Test for Correct exceedances. The $p-$values are obtained by comparison with the $\chi^2(1)$-distribution. The $\chi^2(1)$-distribution has a 5 percent critical value of 3.84 and a 1 percent critical value of 6.23. For a good model, we expect NOT to reject H0

Method	Parameters									
	Exp. Exceed		Act. Exceed		UC.LRstat		UC.LRp		UC.Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	26.00	132.00	52.00	156.00	19.34	4.14	0.00	0.04	Reject H0	Reject H0
GARCH(1,1)-norm	26.00	132.00	43.00	130.00	8.72	0.05	0.00	0.82	Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-norm	26.00	132.00	45.00	137.00	10.77	0.16	0.00	0.69	Reject H0	Fail to Reject H0
GARCH(1,1)-t	26.00	132.00	148.00	148.00	271.76	1.83	0.00	0.18	Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-t	26.00	132.00	20.00	146.00	1.76	1.39	0.18	0.24	Fail to Reject H0	Fail to Reject H0
Empirical Quantile	26.00	132.00	34.00	130.00	1.96	0.05	0.16	0.82	Fail to Reject H0	Fail to Reject H0
EVT	26.00	132.00	31.00	127.00	0.73	0.25	0.39	0.62	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	103.00	451.00	1.81	0.00	0.18	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	99.00	449.00	0.88	0.00	0.35	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.11: GBP/EUR daily returns: Backtesting VaR Results,Test for Correct exceedances. The $p-$values are obtained by comparison with the $\chi^2(1)$-distribution. The $\chi^2(1)$-distribution has a 5 percent critical value of 3.84 and a 1 percent critical value of 6.23. For a good model, we expect NOT to reject H0

Test for independence and correct exceedances

We test here the null hypothesis that the model gives the correct number of violations and that there are no violation clusters, that is, all the VaR violations are independent. At significance levels 95% and 99%, we would not reject the respective null hypothesis for p-values higher than 0.05 and 0.01 respectively, indicating that the model is a good model. The results of the test for independence and correct exceedances for tail probabilties $\alpha = 5\%$ and $\alpha = 1\%$ VaR are documented in Tables 7.12–7.16. The first two columns in the tables give results for the expected and actual exceedances both at $\alpha = 5\%$ and $\alpha = 1\%$. The next two columns give the likelihood ratio statistic and the critical region. The last two columns give the $p-$value and the decision based on the $p-$value. A decision "Fail to Reject H0" is an indication that the model captures the correct number of exceedances and that the VaR violations are independent.

We notice here that across all the data sets examined, the MAR(3;2,2,1) model consistently performs well at both tail probabilities $\alpha = 5\%$ and $\alpha = 1\%$, as we not only fail to reject the null hypothesis of correct number of violations (exceedances), but we also agree that the probability that a violation will occur tomorrow does not depend on the violations that have occurred today.

We find that for the individual stock data set (IBM) at $\alpha = 5\%$, all the approaches perform well apart from the EVT and Empirical quantile methods. While at $\alpha = 1\%$, we reject the null hypothesis for the Riskmetrics method and the GARCH-t approaches. However, for the exchange rate data sets (GBP/USD and GBP/EUR) and the the stock indices data sets (SP500 and FTSE100), we find, based on the $p-$values, that the MAR models both give correct and independent exceedances at $\alpha = 5\%$. However, at $\alpha = 1\%$, some of the GARCH based parametric models also give independent and correct exceedances.

These conclusions are consistent with the claim that MAR models are able to capture multiple regimes in financial time series an attribute that is common with both exchange rate and stock indices data.

Method	Parameters									
	Exp. Exceed		Act. Exceed		CC.LRstat		CC.LRp		CC.Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	15.00	75.00	29.00	73.00	296.97	585.52	0.00	0.00	Reject H0	Reject H0
GARCH-norm	15.00	75.00	25.00	68.00	260.22	556.09	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH-norm	15.00	75.00	22.00	70.00	232.66	567.86	0.00	0.00	Reject H0	Reject H0
GARCH-t	15.00	75.00	74.00	74.00	710.41	591.41	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH-t	15.00	75.00	19.00	74.00	205.10	591.41	0.00	0.00	Reject H0	Reject H0
Empirical Quantile	15.00	75.00	25.00	102.00	17.48	12.27	0.00	0.00	Reject H0	Reject H0
EVT	15.00	75.00	25.00	98.00	17.48	10.96	0.00	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	95.00	447.00	2.30	0.00	0.32	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	105.00	455.00	2.44	0.00	0.29	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.12: IBM daily returns: Backtesting VaR Results, Test for Correct exceedances and Independence. The $p-$values are obtained by comparison with the $\chi^2(2)$-distribution. The $\chi^2(2)$-distribution has a 5 percent critical value of 5.99 and a 1 percent critical value of 9.21. For a good model, we expect NOT to reject H0

Method	Exp. Exceed 1%	Exp. Exceed 5%	Act. Exceed 1%	Act. Exceed 5%	CC.LRstat 1%	CC.LRstat 5%	CC.LRp 1%	CC.LRp 5%	CC.Decision 1%	CC.Decision 5%
Riskmetrics	15.00	75.00	40.00	99.00	30.42	11.41	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	75.00	47.00	102.00	43.40	11.60	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	75.00	46.00	102.00	41.08	10.11	0.00	0.01	Reject H0	Reject H0
GARCH(1,1)-t	15.00	75.00	103.00	103.00	227.16	12.41	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	15.00	75.00	44.00	102.00	36.60	10.11	0.00	0.01	Reject H0	Reject H0
EVT	15.00	75.00	41.00	120.00	32.63	25.30	0.00	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	102.00	461.00	1.57	0.00	0.46	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	91.00	460.00	1.87	0.00	0.39	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.13: SP500 daily returns: Backtesting VaR Results, Test for Correct exceedances and Independence. The $p-$values are obtained by comparison with the $\chi^2(2)$-distribution. The $\chi^2(2)$-distribution has a 5 percent critical value of 5.99 and a 1 percent critical value of 9.21. For a good model, we expect NOT to reject H0

Method	Exp. Exceed 1%	Exp. Exceed 5%	Act. Exceed 1%	Act. Exceed 5%	CC.LRstat 1%	CC.LRstat 5%	CC.LRp 1%	CC.LRp 5%	CC.Decision 1%	CC.Decision 5%
Riskmetrics	15.00	76.00	33.00	103.00	333.86	762.82	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	76.00	34.00	104.00	343.05	768.71	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	76.00	36.00	105.00	361.43	774.59	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-t	15.00	76.00	105.00	105.00	995.36	774.59	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	15.00	76.00	32.00	106.00	324.68	780.48	0.00	0.00	Reject H0	Reject H0
Empirical Quantile	15.00	76.00	32.00	123.00	18.78	32.27	0.00	0.00	Reject H0	Reject H0
EVT	15.00	76.00	27.00	122.00	10.28	30.11	0.01	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	91.00	462.00	0.02	0.00	0.99	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	92.00	458.00	0.97	0.00	0.62	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.14: FTSE daily returns: Backtesting VaR Results, Test for Correct exceedances and Independence. The $p-$values are obtained by comparison with the $\chi^2(2)$-distribution. The $\chi^2(2)$-distribution has a 5 percent critical value of 5.99 and a 1 percent critical value of 9.21. For a good model, we expect NOT to reject H0

Method	Exp. Exceed 1%	Exp. Exceed 5%	Act. Exceed 1%	Act. Exceed 5%	CC.LRstat 1%	CC.LRstat 5%	CC.LRp 1%	CC.LRp 5%	CC.Decision 1%	CC.Decision 5%
Riskmetrics	26.00	132.00	53.00	155.00	28.97	53.14	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	26.00	132.00	36.00	132.00	5.83	39.15	0.05	0.00	Fail to Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	26.00	132.00	34.00	127.00	2.85	39.46	0.24	0.00	Fail to Reject H0	Reject H0
GARCH(1,1)-t	26.00	132.00	138.00	138.00	278.87	41.95	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	26.00	132.00	11.00	136.00	11.85	46.85	0.00	0.00	Reject H0	Reject H0
Empirical Quantile	26.00	132.00	35.00	113.00	66.38	57.56	0.00	0.00	Reject H0	Reject H0
EVT	26.00	132.00	34.00	110.00	59.37	57.08	0.00	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	110.00	458.00	4.48	0.00	0.11	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	101.00	454.00	1.32	0.00	0.52	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.15: GBP/USD daily returns: Backtesting VaR Results, Test for Correct exceedances and Independence. The $p-$values are obtained by comparison with the $\chi^2(2)$-distribution. The $\chi^2(2)$-distribution has a 5 percent critical value of 5.99 and a 1 percent critical value of 9.21. For a good model, we expect NOT to reject H0

Method	Parameters									
	Exp. Exceed		Act. Exceed		CC.LRstat		CC.LRp		CC.Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	26.00	132.00	52.00	156.00	22.00	49.35	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	26.00	132.00	43.00	130.00	10.41	27.52	0.01	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	26.00	132.00	45.00	137.00	12.22	23.85	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-t	26.00	132.00	148.00	148.00	303.63	33.69	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	26.00	132.00	20.00	146.00	3.92	44.05	0.14	0.00	Fail to Reject H0	Reject H0
Empirical Quantile	26.00	132.00	34.00	130.00	24.26	72.31	0.00	0.00	Reject H0	Reject H0
EVT	26.00	132.00	31.00	127.00	19.18	71.28	0.00	0.00	Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	103	451	0.55	0.49	0.19	1.00	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	99	449	0.81	0.52	0.34	1.00	Fail to Reject H0	Fail to Reject H0

Table 7.16: GBP/EUR daily returns: Backtesting VaR Results, Test for Correct exceedances and Independence. The $p-$values are obtained by comparison with the $\chi^2(2)$-distribution. The $\chi^2(2)$-distribution has a 5 percent critical value of 5.99 and a 1 percent critical value of 9.21. For a good model, we expect NOT to reject H0

7.3.2 ES Backtest Results

The null hypothesis here is that the excess conditional shortfall (excess of the actual series when VaR is violated), is i.i.d. and has zero mean, against the alternative that the excess shortfall has mean greater than zero and thus that the conditional shortfall is systematically underestimated (see Section 6.8.2). The test is a one sided t-test. The bootstrap method is used to obtain the p-values. This is done to alleviate any bias with respect to assumptions about the underlying distribution of the excess shortfall, (see McNeil et al. (2010)). The results of the test for zero mean behaviour of VaR violations at tail probabilities $\alpha = 5\%$ and $\alpha = 1\%$ for the ES estimation methods described in Section 6.6 are documented in Tables 7.17–7.21. The first two columns in the tables give results for the expected and actual exceedances both at $\alpha = 5\%$ and $\alpha = 1\%$. The last two columns the give the $p-$value and the decision based on the $p-$value. A decision "Fail to Reject H0" is an indication that based on the corresponding model, the excess violations of VaR are i.i.d with zero mean.

The results in Tables 7.17–7.21 reveal that the for the GARCH-t, AR(2)-GARCH-t, EVT, MAR(3;2,2,1) model with Gaussian innovations, MAR(3;2,2,1) with Student-t innovations, and empirical quantile methods, the violation residuals (see Section 6.8.2) do behave like realisations of i.i.d variables from a distribution with zero mean at both $\alpha = 5\%$ and $\alpha = 1\%$. However, the Riskmetrics, GARCH-norm and AR(2)-GARCH-norm methods do not. It is noteworthy that the GARCH-t, AR(2)-GARCH-t, EVT, MAR(3;2,2,1) model with Gaussian innovations, MAR(3;2,2,1) with Student-t innovations, are heavy tailed distributions making them better suited to financial data. The empirical quantile approach does not assume any underlying distribution.

Method	Parameters							
	Exp. Exceed		Act. Exceed		P.Value		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	15.00	75.00	29	72	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	75.00	25	67	0.04	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	75.00	24	72	0.02	0.01	Reject H0	Reject H0
GARCH(1,1)-t	15.00	75.00	74	74	1.00	0.16	Fail to Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-t	15.00	75.00	19	73	0.69	0.15	Fail to Reject H0	Fail to Reject H0
EVT	15.00	75.00	25	98	1.00	0.04	Fail to Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	104	457	0.72	0.43	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	95	463	0.32	0.44	Fail to Reject H0	Fail to Reject H0

Table 7.17: IBM daily returns: Backtesting ES Results, Zero mean test for Excess Violations of VaR. For a good model, we expect NOT to reject H0

Method	Parameters							
	Exp. Exceed		Act. Exceed		P.Value		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	15.00	76.00	33	103	0.01	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	76.00	34	104	0.01	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	76.00	36	105	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-t	15.00	76.00	105	105	1.00	0.01	Fail to Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	15.00	76.00	32	106	0.09	0.01	Fail to Reject H0	Reject H0
Empirical Quantile		76.00		123		0.19		Fail to Reject H0
EVT	15.00	76.00	27	122	1.00	0.18	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	91	462	0.17	0.39	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	92	458	0.22	0.44	Fail to Reject H0	Fail to Reject H0

Table 7.18: FTSE daily returns: Backtesting ES Results, Zero mean test for Excess Violations of VaR. For a good model, we expect NOT to reject H0

Method	Parameters							
	Exp. Exceed		Act. Exceed		P.Value		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	15.00	75.00	40	99	0.01	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	15.00	75.00	47	102	0.03	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	15.00	75.00	46	102	0.02	0.00	Reject H0	Reject H0
GARCH(1,1)-t	15.00	75.00	103	103	1.00	0.00	Fail to Reject H0	Reject H0
AR(2)-GARCH(1,1)-t	15.00	75.00	44	102	0.02	0.00	Reject H0	Reject H0
EVT	15.00	75.00	41	120	1.00	0.01	Fail to Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	102	461	0.75	0.62	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	91	460	0.30	0.41	Fail to Reject H0	Fail to Reject H0

Table 7.19: SP500 daily returns: Backtesting ES Results, Zero mean test for Excess Violations of VaR. For a good model, we expect NOT to reject H0

Method	Parameters							
	Exp. Exceed		Act. Exceed		P.Value		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	26.00	132.00	53	155	0.01	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	26.00	132.00	36	132	0.01	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	26.00	132.00	34	127	0.01	0.00	Reject H0	Reject H0
GARCH(1,1)-t	26.00	132.00	138	138	1.00	1.00	Fail to Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-t	26.00	132.00	11	136	0.99	1.00	Fail to Reject H0	Fail to Reject H0
Empirical Quantile	26.00	132.00	35	113	0.17	0.03	Fail to Reject H0	Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	110	458	0.82	0.39	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	101	454	0.80	0.53	Fail to Reject H0	Fail to Reject H0

Table 7.20: GBP/USD daily returns: Backtesting ES Results, Zero mean test for Excess Violations of VaR. For a good model, we expect NOT to reject H0

Method	Parameters							
	Exp. Exceed		Act. Exceed		P.Value		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	26.00	132.00	52	156	0.00	0.00	Reject H0	Reject H0
GARCH(1,1)-norm	26.00	132.00	43	130	0.00	0.00	Reject H0	Reject H0
AR(2)-GARCH(1,1)-norm	26.00	132.00	45	137	0.01	0.00	Reject H0	Reject H0
GARCH(1,1)-t	26.00	132.00	148	148	1.00	1.00	Fail to Reject H0	Fail to Reject H0
AR(2)-GARCH(1,1)-t	26.00	132.00	20	146	0.97	1.00	Fail to Reject H0	Fail to Reject H0
Empirical Quantile	26.00	132.00	34	130	0.24	0.21	Fail to Reject H0	Fail to Reject H0
EVT	26.00	132.00	31	127	1.00	0.16	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-norm	90.00	450.00	100	455	0.77	0.54	Fail to Reject H0	Fail to Reject H0
MAR(3;2,2,1)-t	90.00	450.00	94	457	0.62	0.63	Fail to Reject H0	Fail to Reject H0

Table 7.21: GBP/EUR daily returns: Backtesting ES Results, Zero mean test for Excess Violations of VaR. For a good model, we expect NOT to reject H0

7.4 Summary

We considered the out of sample VaR and ES measures, at tail probabilities $\alpha = 1\%$ and $\alpha = 5\%$ for each of the financial time series selected. We treated the results based on empirical quantiles as conservative estimates of the true VaR (i.e., lower bounds), and find that the approaches based on EVT and MAR models give significantly better results as they give values close to the empirical quantiles while the approaches based on GARCH models tend to underestimate VaR and ES.

Across all the data sets examined, we find that the MAR(3;2,2,1) models with Gaussian and Student-t innovations consistently perform well at both $\alpha = 5\%$ and $\alpha = 1\%$, as we fail to reject the null of correct number of violations (exceedances), independent VaR violations and excess violations of VaR is i.i.d with zero mean.

We find that for the individual stock data set (IBM) and both exchange rate data sets (GBP/USD and GBP/EUR), most of approaches for predicting VaR seem to give correct excedences based on both the $p-$values and a comparison of the expected and actual exceedances columns.

However, for the stock indices (SP500 and FTSE), only the MAR models consistently give correct exceedances, independent VaR violations and i.i.d excess violations of VaR with zero mean based on both the $p-$values and a comparison of the expected and actual exceedances columns. These conclusions can be attributed to the fact that stock indices consists of a portfolio of various individual stocks and hence most likely to exhibit more volatile behaviour. MAR models are hence better suited to capture multiple regimes and other dynamics associated with this kind of data.

The ES backtest results reveal that the for the GARCH(1,1) and AR(2)-GARCH(1,1) with Student-t innovations, EVT, MAR(3;2,2,1) models with both

Gaussian and Student-t innovations and empirical quantile methods at both $\alpha = 5\%$ and $\alpha = 1\%$, the violation residuals do behave like realisations of i.i.d variables from a distribution with zero mean, while the Riskmetrics, GARCH(1,1) and AR(2)-GARCH(1,1) with Gaussian innovation models do not.

It is noteworthy here that for the EVT based approaches, there are some other improved variations (see Gilli and këllezi (2006), Klüppelberg and Mikosch (1997), and Tsay (1997)) which we have not considered here due to time and computational constraints.

Chapter 8

Density Forecasts and the MAR Model

Forecasts play a very significant role in economics and finance just as they do in any other science. Evaluating accurate/dependable predictions is of primary concern. A large chunk of the existing forecast literature is focused on evaluating point forecasts, a smaller slice on interval forecasts and a much thinner slice on probability forecasts. Point forecasts have been noted to be generally unsuitable for all models as forecasts based on quite a number of financial and economic models are not readily summarized by point forecasts (Berkowitz (2001)). Density forecasts have become more popular as applications in real life scenarios require not only a forecast estimate but also the uncertainty associated to such a forecast.

Applications of density forecasts span across the field of macro economics. A popular example is the 'fan-chart' of inflation and GDP published by the Bank of England and by the Sveriges Riksbank in Sweden in their quarterly inflation reports.

Further applications can be found in finance, the major area being in risk management. Distributional forecasts of a portflio are issued with the purpose of tracking measures of portfolio risk such as Value at Risk (VaR) and Expected

Shortfall (ES) (Diebold et al. (1998)). Another application of density forecast in finance is the extraction of density forecasts from option price data (Diebold et al. (1998)).

We aim here to evaluate the tail forecast density of some financial time series based on the MAR model. We compare these tail density forecasts to those based on some popular GARCH and AR-GARCH models. Evaluating and comparing the tail density forecasts, create an avenue to further assess, the overall quality of the conditional loss distributions of the parametric methods used in estimating risk measures like VaR and ES.

8.1 Density Forecasting

A density forecast is an estimate of the future probability distribution of a random variable, conditional on the information available at the time of the forecast. It gives a complete characterisation of the uncertainty associated with a prediction, as against the point forecast which does not provide information about the uncertainty of the prediction (Diebold et al. (1998)).

8.1.1 Loss Functions and Action Choices

Diebold et al. (1998) show that the problem of density forecast evaluation is intrinsically linked to the forecast user's loss function. They observed that understanding the connection between density forecast, loss functions and action choices will shed more light on what can or cannot be achieved when evaluating density forecasts.

Loss Function, Action Choice and The Decision Environment Consider a stochastic process y_t with corresponding series of realisations $\{y_t\}_{t=1}^{m}$. Let $\{f_t(y_t \mid \Omega_t)\}_{t=1}^{m}$

be the corresponding sequence of conditional densities governing y_t. Also, let $\{p_t(y_t \mid \Omega_t)\}_{t=1}^{m}$ be the sequence of one-step ahead density forecasts of y_t. Where $\Omega_t = \{y_{t-1}, y_{t-1}, \ldots\}$ is the information set up to time $t-1$.

Denote the loss function as $L(a, y)$, where a refers to an action choice. The loss function $L(a, y)$ chooses an action to minimise expected loss computed using the density of the data generating process. If $p(y)$ is the correct density, then the user chooses an action a^* such that:

$$a^*(p(y)) = argmin_{a \in A} \int L(a, y)p(y)dy \tag{8.1.1.1}$$

The action choice defines the loss $L(a^*, y)$ faced for every realization of the process y with density $f(y)$. This loss is in itself a random variable and possesses a probability distribution that depends only on the action choice. The expected loss with respect to the true data generating process is represented as (Diebold et al. (1998)),

$$E[L(a^*, y)] = \int L(a^*, y)f(y)dy \tag{8.1.1.2}$$

Density forecast has quite a glaring effect on the user's expected loss as different density forecasts will in general lead to different action choices and hence different loss distributions. *The better the density forecast, the lower the expected loss computed with respected to the true data generating function.*

The Probrobability integral transform (PIT)

The main idea behind the probability integral transform goes as far back as as Rosenblatt (1952) and was made popular by Diebold et al. (1998). The probability integral transform is simply the cummulative density function corresponding to the sequence of density forecasts evaluated at y_t. The *PITS* of the actual realisation of the variables over the forecast period ($\{y_t\}_{t=1}^n$), $t = 1, \ldots, n$ is calculated with respect to the model's forecast densities. Denote this forecast density by $\{p_t(y_t)\}_{t=1}^n$ so that we evalutate,

$$z_t = \int_{-\infty}^{y_t} p_t(u)du \qquad \text{for } t = 1, \ldots, n. \tag{8.1.1.3}$$

The true predictive density of y_t is given by its data generating process denoted by $f_t(y_t)$. When the model forecast density corresponds to the true predictive density, that is, when $p_t(y_t) = f_t(y_t)$, then the sequence $\{z_t\}_{t=1}^n$ is iid $U(0, 1)$.

Diebold et al. (1998) say that the sequence $\{z_t\}_{t=1}^n$ will consist of i.i.d uniform variables in a time series context when the true densities (conditional on the past process) are used at each time t to transform the realisations of the series y_t's. The density forecast is evaluated by assessing whether there is statistically significant evidence that the realisations do not come form the density, this is as good as testing whether the $\{z_t\}_{t=1}^n$ series depart from the i.i.d uniform assumption, resulting in a joint test of independence and uniformity.

8.2 Testing Density Forecasts—overview of methods

Diebold et al. (1998) use the probability integral transforms (PITS) to evaluate density forecasts. They use graphical tools test whether the resultant series consists of independently and identically distributed uniform random variables $U(0, 1)$. They assess independence by examining the correlogram and plot of the probability density function (PDF) to assess uniformity. They argue that statistical tests do not give insight into the reasons for rejection. Diebold et al. (1998) investigate independence by testing for non-zero autocorrelations in the first three moments of the z_t series. The following statistics can then be used to test the null hypothesis of uniformity $U(0,1)$:

1. Kolmogorov-Smirnov (KS(D)) (see Neave and Worthington (1988)). $D = \max_j\{abs(z_j - A_j)\}$ (z_j is the theoretical CDF under the null, A_j is the empirical CDF and n is the number of observations.

2. Kuiper(V) (see Kuiper (1962)). $D^+ \max[(j/n) - z_j]$ and $D^-[z_j - (j-1)/n], V + D^+ + D^-$.

3. Cramér-von Mises(W^2) (see Cramér (1945)). $W^2 = \sum_{j=1}^{n}[z_j - (2j-1)/2n]^2 + (1/2n)$.

4. Watson(U^2): $U = W^2 - n(\bar{z} - 0.5)^2$($\bar{z}$ is the average cumulative probability under the null).

5. Anderson and Darling(A^2) (see Anderson and Darling (1954)). $A^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)[\log(z_j) + \log(1 - z_{-j})]$.

However, these non-parametric tests are quite data intensive. Research shows the need for at least 1000 observations for a relatively reliable conclusion (Berkowitz (2001)).

Berkowitz (2001) introduces an extension of the Rosenblatt transformation. He advocates for a simple transformation to normality and suggests working with the inverse normal CDF transformation. That is, rather than $\{z_t\}_{t=1}^{n}$, he transforms the observed portfolio returns to create a series, $z_t = \Phi^{-1}(\hat{F}(y_t))$.

Let $\Phi^{-1}(\cdot)$ be the inverse of the standard normal distribution function, for a sequence of forecasts regardless of the underlying distribution of the portfolio returns. The following results are due to Berkowitz (2001).

Propositon 8.2.1 (Berkowitz (2001)). If the series $r_t = \int_{-\infty}^{x_t} f(u)du$ is distributed as an i.i.d $U(0,1)$, then

$$z_t = \Phi^{-1}[r_t] = \int_{-\infty}^{x_t} f(u)du \quad \text{is an i.i.d } N(0,1) \tag{8.2.0.4}$$

Equation (8.2.0.4) is the inverse PIT. This transformation is widely used to generate random numbers in computations. Berkowitz (2001) apply the inverse PIT to time series.

Propositon 8.2.2. Let $h(z_t)$ be the density of the z_t and let $\phi(z_t)$ be standard normal. Then $\log[f(y_t)/\hat{f}(y_t)] = \log[h(z_t)/\phi(z_t)]$.

Propositions (8.2.1) and (8.2.2) above enable the use of Gaussian likelihood tools to test the null hypothesis that the data follows a normal distribution. They also establish that the inaccuracies in the density forecast will be preserved in the transformed data.

Berkowitz (2001) proposes an LR test based on censored likelihood, where the shape of the forecasted tail of the density is compared to the observed tail. He constructs an LR test where he evaluates a restricted likelihood $L(0,1)$. He then compares restricted likelihood to an unrestricted likelihood $L(\hat{\mu}, \hat{\sigma}^2)$. The test statistic is based on the difference between the constrained and the unconstrained values of the likelihood. That is,

$$LR_{tail} = -2(L(0,1) - L(\hat{\mu}, \hat{\sigma}^2)) \tag{8.2.0.5}$$

Under the null hypothesis, $LR_{tail} \sim \chi^2(2)$.

To jointly test the null hypothesis of independence, and mean$= 0$ and variance$=$ 1, he defines the following combined statistic,

$$LR = -2(\mathcal{L}(0,1,0) - \mathcal{L}(\hat{\mu}, \hat{\sigma}^2, \hat{p})) \sim \chi^2(3) \tag{8.2.0.6}$$

He also proposed LR tests based on a censored likelihood where he compares the shape of the forecasted tail of the density to the observed tail. The test statistic is based on the difference between the constrained ($\mathcal{L}(0,1)$) and the unconstrained ($\mathcal{L}(\hat{\mu}, \hat{\sigma}^2)$). He forms an LR tail test that tests the null that the mean and variance of the violations equal those implied by the model as follows,

$$LR_{tail} = -2(\mathcal{L}(0,1) - \mathcal{L}(\hat{\mu}, \hat{\sigma}^2)) \sim \chi^2(2). \tag{8.2.0.7}$$

The test will not only reject if the tails are too large but will also asymptotically reject if the tail has excessively small losses relative to forecast.

In this work, we consider the Berkowitz (2001) tail test approach for testing density forecasts as we are interested in the tails of the forecasts.

8.3 Density Forecast and Mixture Autoregressive (MAR) model

We use two underlying methods to compute the density forecast of financial time series based on the MAR model.

- In Section 8.3, we we obtained the density forecasts by generating a random sample of $y_{t+1}^{(k)}$, $k = 1, \dots, N$. We then take the histogram of $y_{t+1}^{(k)}$ as an estimate of the one-step ahead density forecast, the graphs of which are presented in Figures 8.1-8.5 below.

- In section 8.4, we obtained the one-step ahead density forecast based on the MAR model by applying the following theorem by Boshnakov (2009)

  Theorem 8.3.1. *For each $h \geq 1$ the conditon charsteritc function, $\varphi_{t+h|t}(s) \equiv E(e^{isy_{t+h}} \mid \mathcal{F}_t)$, of the $h-$step predictor at time t of the MAR process is given* *by*

$$\varphi_{t+h|t}(s) \equiv E(e^{isy_{t+h}} \mid \mathcal{F}_t) \tag{8.3.0.8}$$
$$= E(E(e^{isy_{t+h}} \mid \mathcal{F}_t, z_{t+h}, \ldots, z_{t+1}) \mid \mathcal{F}_{t-h})$$
$$= \sum_{k_1,\ldots,k_h=1}^{g} (\pi_{k_1} \ldots \pi_{k_h}) e^{is(\mu_{k_1,\ldots,k_h}(t+h))} \Pi_{i=0}^{h-1} \varphi_{k+h-i}(\theta_{h-i}^{k_1,\ldots,k_h} s),$$

where

$$\mu_{k_1,\ldots,k_h}(t+h)) = \sum_{i=1}^{p} \beta_i^{k_1,\ldots,k_h} y(t+1-i) + \beta_0^{k_1,\ldots,k_h} \tag{8.3.0.9}$$

For the one step ahead density forecast that is $h = 1$, this gives

$$\varphi_{t+1|t}(s) = \sum_{k=1}^{g} \pi_k e^{is\mu_k(t+1))} \varphi_k(\sigma_k, \phi_{k,1} s), \tag{8.3.0.10}$$

The computations were done using the MixAR R-package newly developed (and still work in progress) by my supervisor, Dr. Georgi Boshankov.

Here we generate one-step ahead out-of-sample density forecasts of daily returns on IBM stocks, FTSE and S&P 500 stock indices, GBP/USD and GBP/EUR exchange rate data, based on MAR(3;2,2,1) model with both Gaussian and Student-t innovations.

8.3.1 The MAR(3;2,2,1) model

The MAR(3;2,2,1) model is a mixture autoregressive model with three AR components. The first two AR components are of order two and the third one is of order one, that is, $p_1 = p_2 = 2$, $p_3 = 1$ and $k = 3$.

The MAR(3;2,2,1) is such that,

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1}y_{t-1} + \phi_{1,2}y_{t-2} + \sigma_1\epsilon_1(t) & \text{with probability } \pi_1 \\ \phi_{2,0} + \phi_{2,1}y_{t-1} + \phi_{2,2}y_{t-2} + \sigma_2\epsilon_2(t) & \text{with probability } \pi_2 \\ \phi_{3,0} + \phi_{3,1}y_{t-1} + \sigma_3\epsilon_3(t) & \text{with probability } \pi_3, \end{cases}$$

with conditional distribution

$$F_{t|t-1}(x) = \pi_1 F_1\left(\frac{y_t - \phi_{11}y_{t-1} - \phi_{12}y_{t-2}}{\sigma_1}\right) + \pi_2 F_2\left(\frac{y_t - \phi_{21}y_{t-1} - \phi_{22}y_{t-2}}{\sigma_2}\right) + \pi_3 F_3\left(\frac{y_t - \phi_{31}y_{t-1}}{\sigma_3}\right) \tag{8.3.1.1}$$

We investigate the MAR(3;2,2,1) model for $F_i(\cdot),\ i = 1, 2, 3$ Gaussian and Student-t with 3-degrees of freedom. The parameters estimated for each of the financial returns series based on the two models are given in Tables 8.1 and 8.2.

8.3.2 Density forecasts based on the MAR(3;2,2,1) model

We evaluate one-step ahead density forecast for each of the returns series based on The MAR(3;2,2,1) models with Gaussian and Student-t innovations.The graphs of the density forecasts are presented in Figures 8.1-8.5.

Figures 8.1–8.5 reveal that within the time frame selected, the models with Student t-innovations produce plots that have sharper peaks and fatter tails than those with Gaussian innovations. It is noteworthy here that periods of high volatility will inadvertently result in sharp changes in the shape of the returns series hence we would expect a bimodal or even multimodal distributions. However, if the volatility of the financial time series is low then we would expect moderate changes in the shape of the distribution and hence a unimodal distribution. This seems to be the case in our selected period for most of investigated series. We notice that the shape of the predictive distribution for most of the data sets are all unimodal with

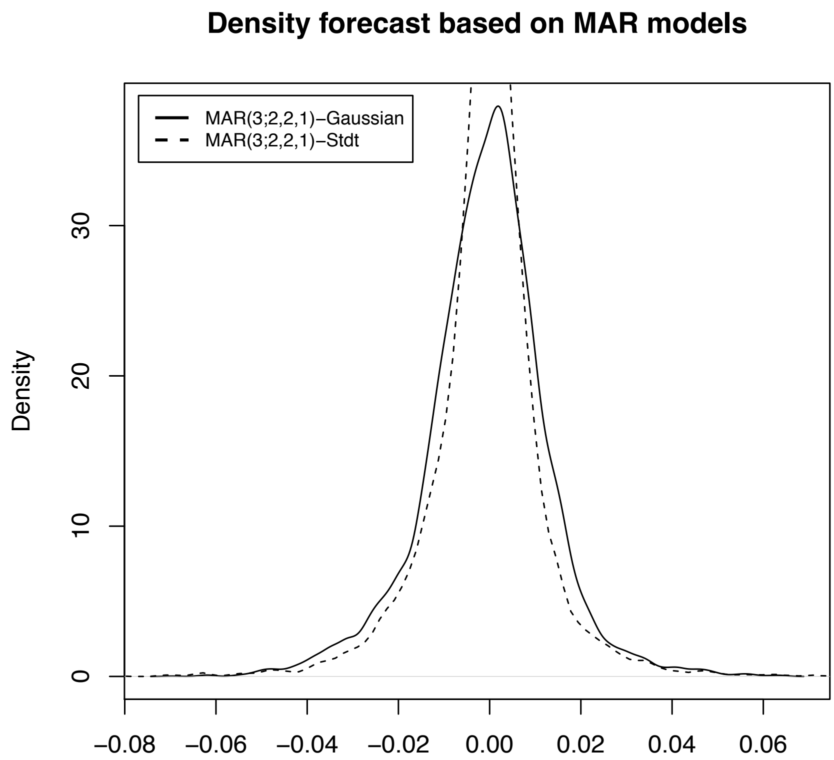

Figure 8.1: One-step ahead density forecast of IBM returns at 2012-06-22 based on an MAR(3;2,2,1) model with Gaussian and Standardised Student-t innovations

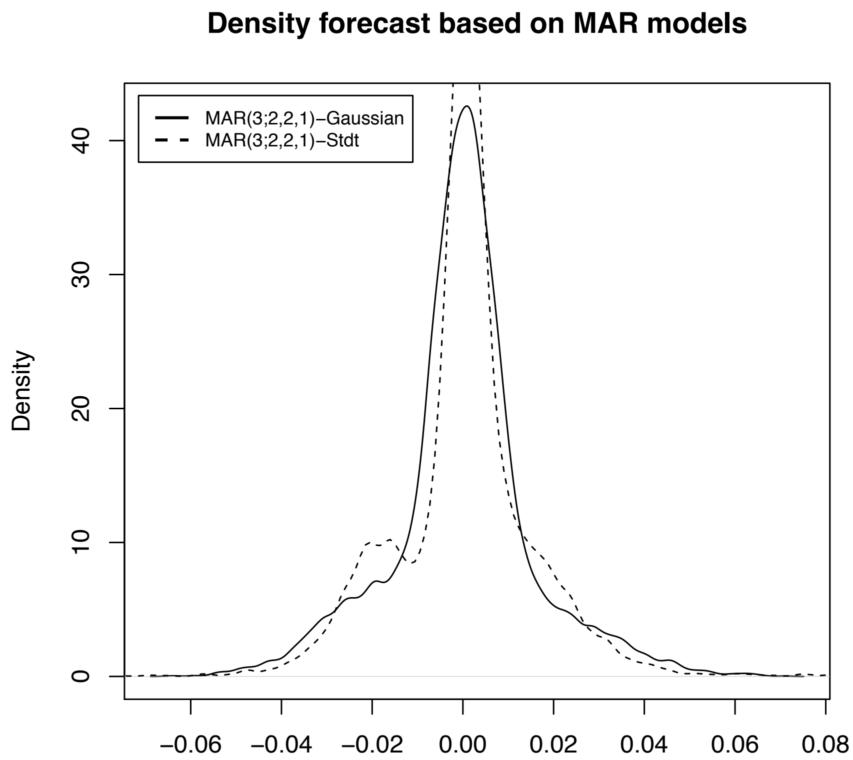

Figure 8.2: One-step ahead density forecast of S&P500 returns at 2012-06-22 based on an MAR(3;2,2,1) model with Gaussian and Standardised Student-t innovations

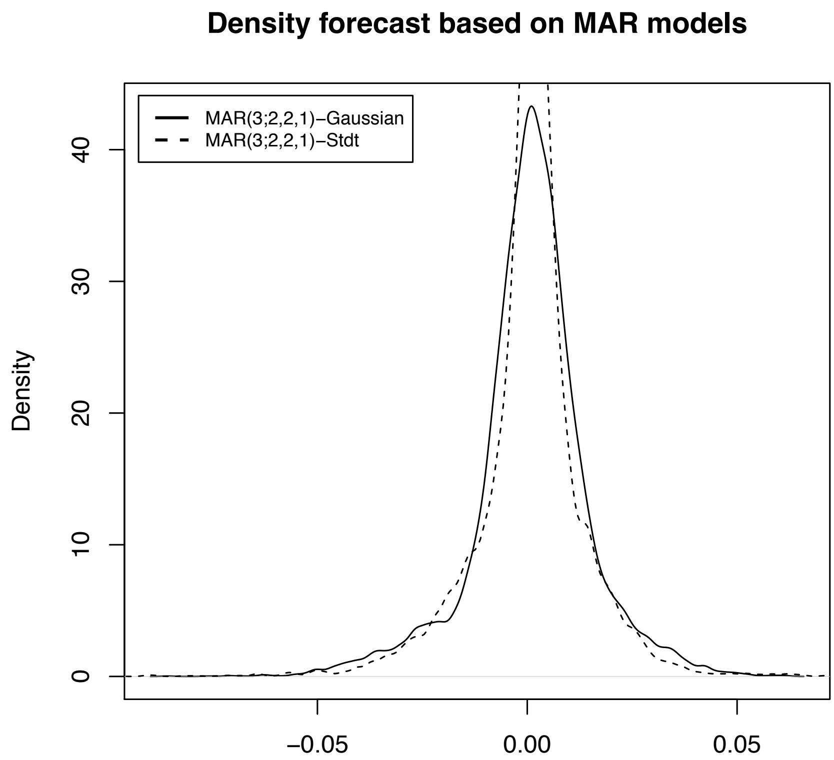

Figure 8.3: One-step ahead density forecast of FTSE returns at 2012-06-22 based on an MAR(3;2,2,1) model with Gaussian and Standardised Student-t innovations

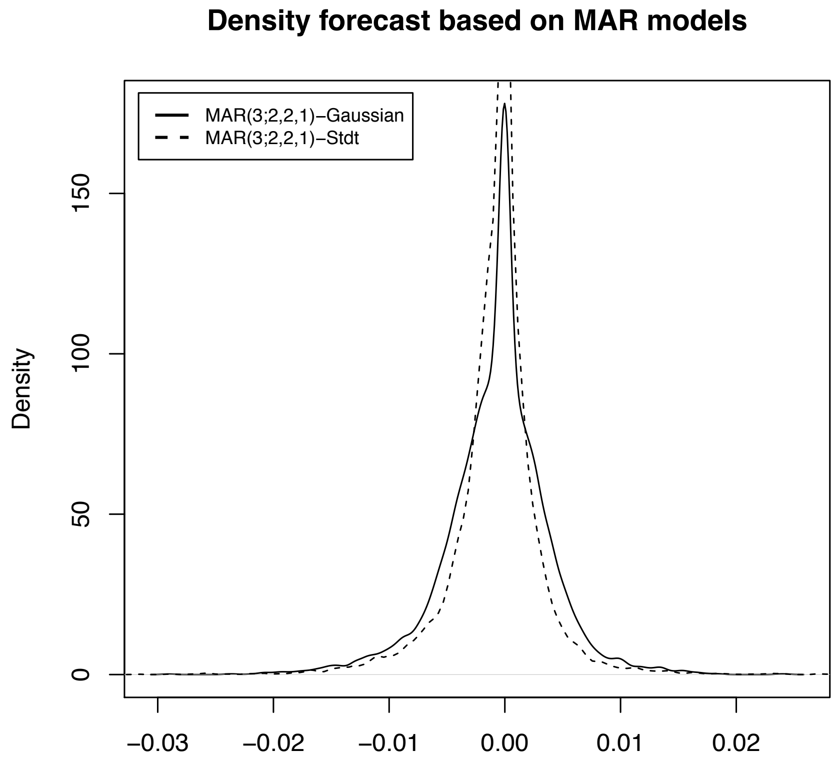

Figure 8.4: One-step ahead density forecast of GBPUSD returns at 2012-06-22 based on an MAR(3;2,2,1) model with Gaussian and Standardised Student-t innovations

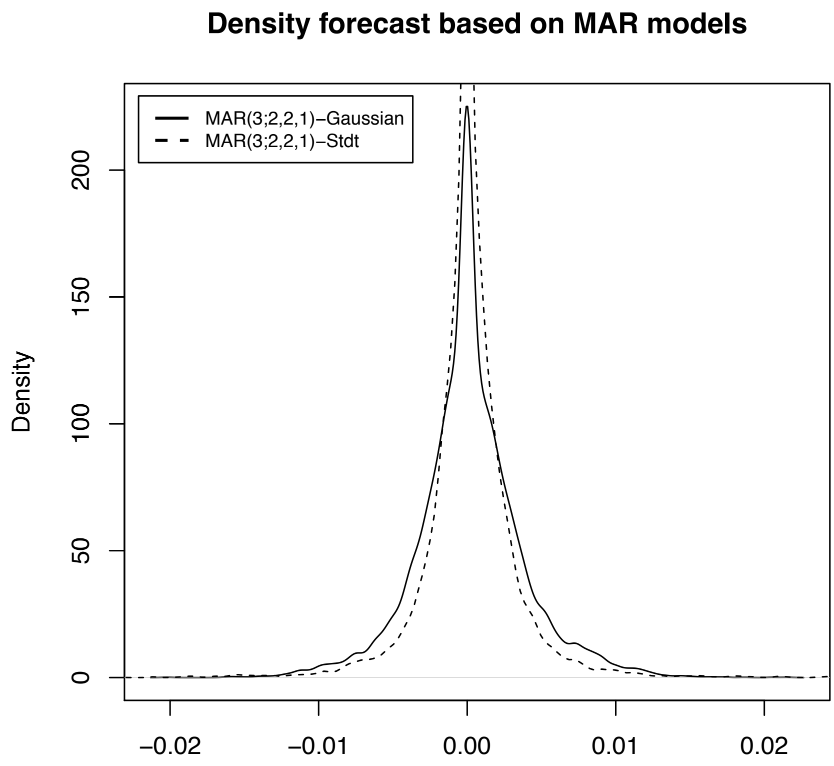

Figure 8.5: One-step ahead density forecast of GBPEUR returns at 2012-06-22 based on an MAR(3;2,2,1) model with Gaussian and Standardised Student-t innovations

Table 8.1: Parameters of the MAR(3;2,2,1) model with Gaussian innovations, for daily logarithmic returns of IBM, FTSE100, S&P500, GBP/USD and GBP/EUR

Parameter	IBM	FTSE100	S&P500	GBP/USD	GBP/EUR
p_1	2	2	2	2	2
p_2	2	2	2	2	2
p_3	1	1	1	1	1
π_1	0.1786	0.185	0.2087	0.2252	0.3339
π_2	0.1786	0.187	0.6638	0.1125	0.5679
π_3	0.6427	0.6281	0.1276	0.6623	0.09824
σ_1	0.0244	0.01918	0.02195	0.007538	0.005313
σ_2	0.01339	0.01287	0.006703	0.00008948	0.002209
σ_3	0.008355	0.006945	0.01145	0.003162	0.00008423
$\phi_{1,0}$	-0.001924	0.00001373	-0.0007612	-0.0003367	-0.0002889
$\phi_{2,0}$	0.002839	-0.003116	0.001173	-0.00001085	0.00008435
$\phi_{2,0}$	0.0003086	0.001006	-0.004454	0.0001412	-0.00001024
$\phi_{1,1}$	0.5127	-0.21	-0.596	0.3826	0.5317
$\phi_{1,2}$	0.4718	-0.7283	-0.1037	-0.1492	-0.2552
$\phi_{2,1}$	-0.6242	0.02879	-0.06614	-0.001028	0.1425
$\phi_{2,2}$	-0.4152	0.7263	0.0238	0.00006171	-0.01684
$\phi_{3,1}$	0.007439	-0.04229	0.8579	0.1552	-0.0004513

slight tail disturbances for FTSE100 data set. The S&P500 data how ever gives clear bimodal shape, this would be expected as the recent global financial crisis has had its toll on quite a number of American stocks.

8.4 Testing the Density Forecasts

We compare the methodology described in Section 8.3 to evaluate density forecasts of some financial time series based on some popular models in the financial industry. The models considered include the IGARCH model, GARCH models with Gaussian and Student-t innovations and AR(2)-GARCH(1,1) models with Gaussian and Student-t innovations. For each of these models, we generate one-step-ahead out-of-sample density forecasts of daily returns on IBM stocks, FTSE100

Table 8.2: Parameters of the MAR(3;2,2,1) model with Student-t innovations, for daily logarithmic returns of IBM, FTSE100, S&P500, GBP/USD and GBP/EUR

Parameter	IBM	FTSE100	S&P500	GBP/USD	GBP/EUR
p_1	2	2	2	2	2
p_2	2	2	2	2	2
p_3	1	1	1	1	1
π_1	0.1866	0.1632	0.2086	0.2227	0.5838
π_2	0.1364	0.5712	0.6638	0.6646	0.3172
π_3	0.677	0.2656	0.1276	0.1127	0.09897
σ_1	0.01762	0.02327	0.02195	0.00755	0.002261
σ_2	0.01994	0.006679	0.006702	0.003174	0.005383
σ_3	0.0085	0.01151	0.01145	0.00008981	0.00008506
$\phi_{1,0}$	0.003268	-0.001259	-0.0007611	-0.0003433	0.0000811
$\phi_{2,0}$	-0.003964	0.001382	0.001173	0.0001421	-0.0003033
$\phi_{3,0}$	0.0004588	-0.001933	-0.004452	-0.00001081	-0.00001032
$\phi_{1,1}$	-0.7556	-0.6319	-0.596	0.3878	0.1455
$\phi_{1,2}$	0.439	0.03735	-0.1037	-0.1645	-0.01967
$\phi_{2,1}$	0.8135	-0.1375	-0.06616	0.1531	0.5448
$\phi_{2,2}$	-0.3147	-0.05243	0.0238	0.01082	-0.2617
$\phi_{3,1}$	0.001504	0.4549	0.8577	-0.001003	-0.0004233

and S&P 500 stock indices, GBP/USD and GBP/EUR exchange rate data. We then pitch the distributional forecasts against the realized returns.

We start the comparision by examining the histogram of the Probability Integral Transform (Diebold et al. (1998)) of each of the financial time series data with respect to the density forecast produced based on each of the models. The histograms are also presented in Figures 8.6–8.10.

Close examination reveals that for the PIT of the IBM returns, the IGARCH(1,1), GARCH(1,1) with Gaussian innovations, GARCH(1,1) with Student-t innovations and AR(2)-GARCH(1,1) with Gaussian innovation models display slightly butterfly shaped histograms while the AR(2)-GARCH(1,1) with Student-t innovations, the MAR(3;2,2,1) with Gaussian innovations and MAR(3;2,2,1) with Student-t innovations, models give histograms that are closer to uniform. For the GBP/USD

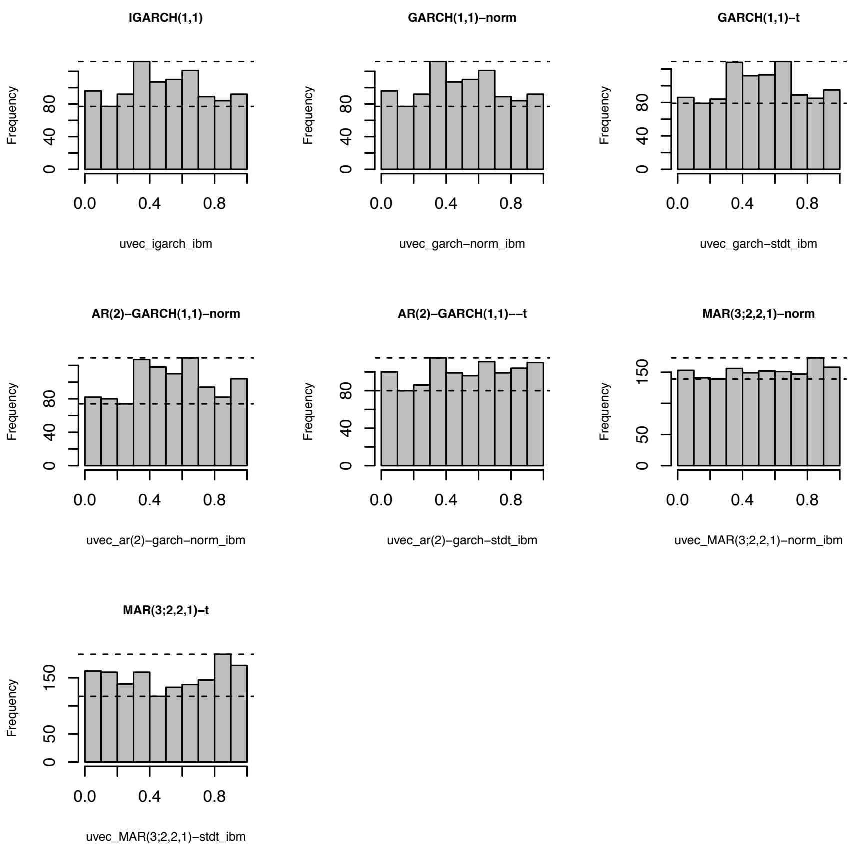

Figure 8.6: Histogram showing the PIT of IBM returns with respect to the density forecast produced under the assumption that IBM returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t

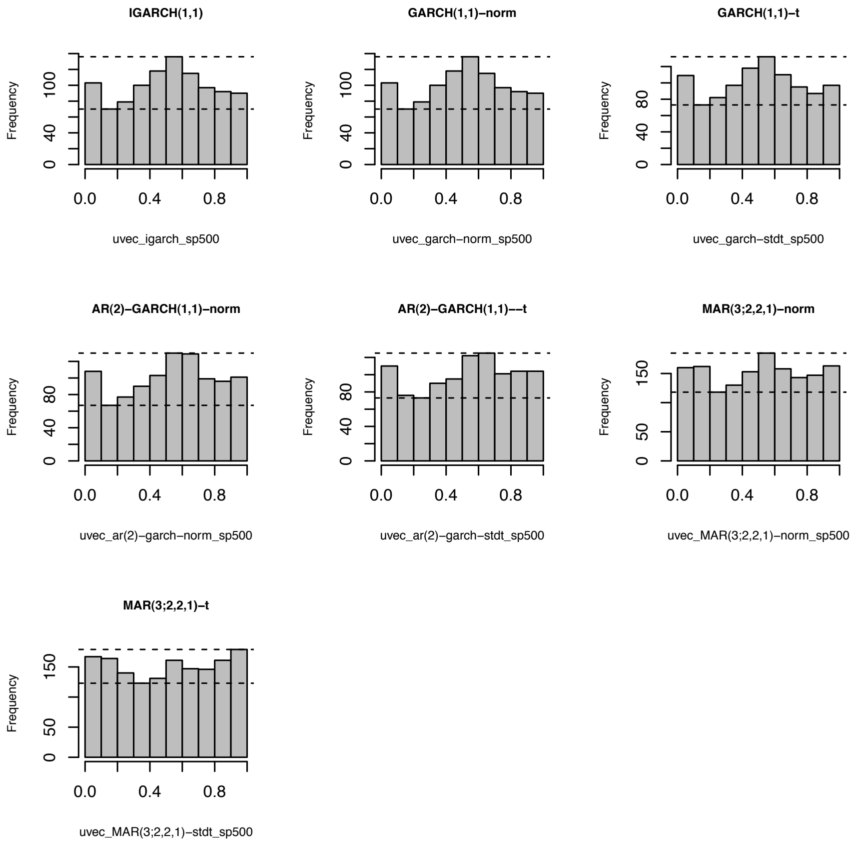

Figure 8.7: Histogram showing the PIT of SP500 returns with respect to the density forecast produced under the assumption that SP500 returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t ,MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t

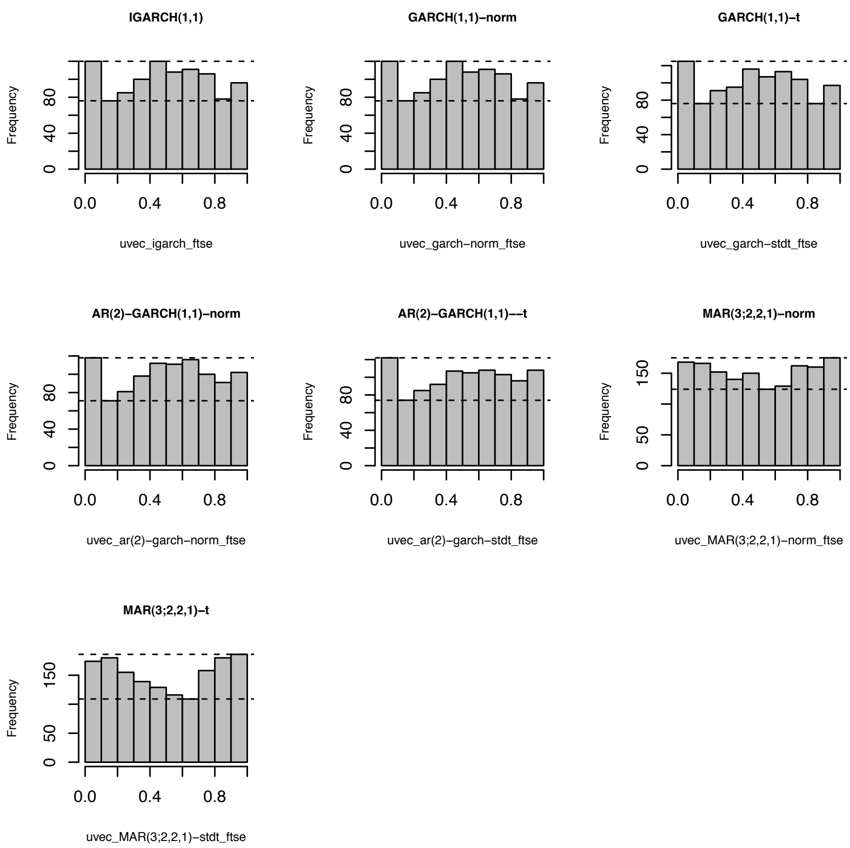

Figure 8.8: Histogram showing the PIT of FTSE returns with respect to the density forecast produced under the assumption that FTSE returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t

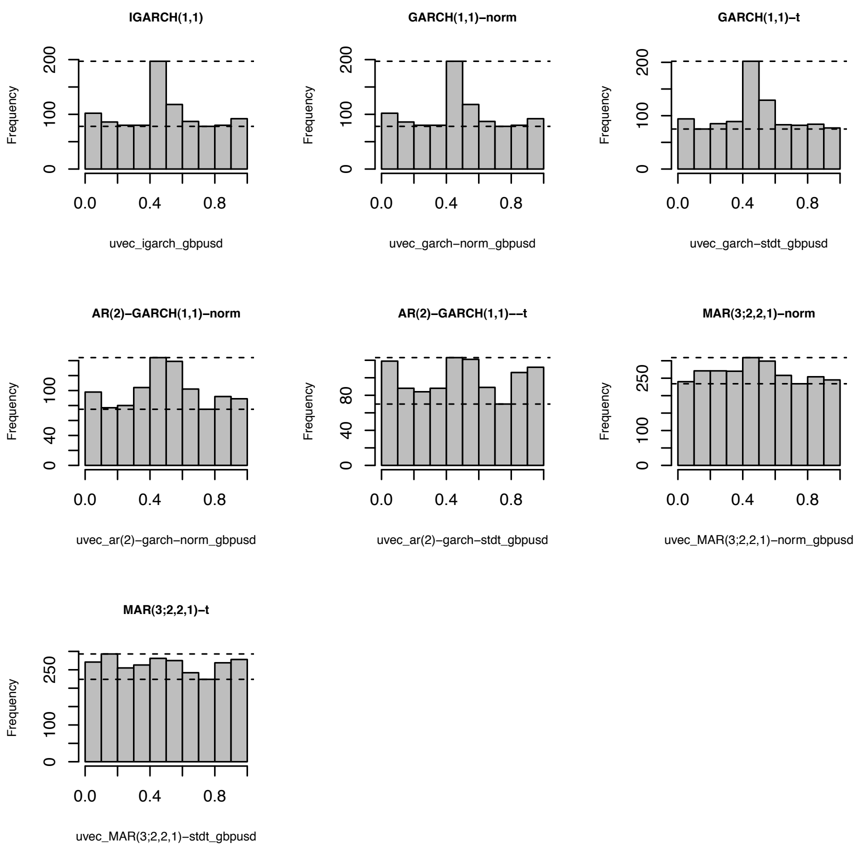

Figure 8.9: Histogram showing the PIT of GBPUSD returns with respect to the density forecast produced under the assumption that GBPUSD returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t,MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t

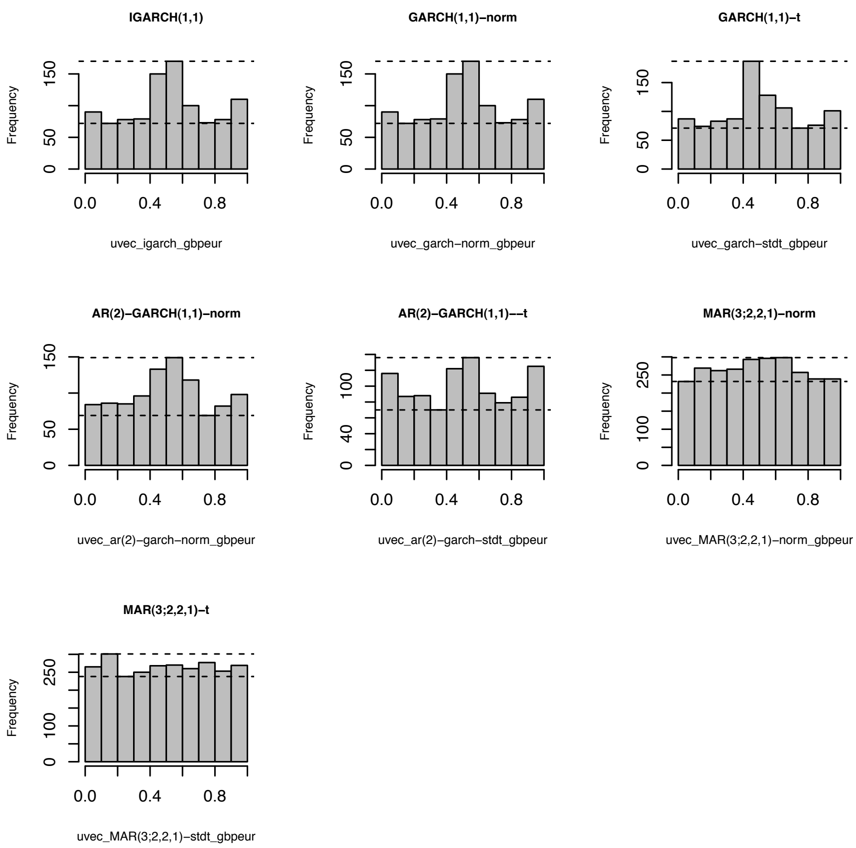

Figure 8.10: Histogram showing the PIT of GBPEUR returns with respect to the density forecast produced under the assumption that GBPEUR returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t"

and the GBP/EUR data sets, only the MAR(3;2,2,1) with Gaussian innovations and MAR(3;2,2,1) with Student-t innovations models give histograms that are close to uniform.

However, for the S&P500 and the FTSE returns, although the MAR(3;2,2,1) with Gaussian innovations and MAR(3;2,2,1) with Student-t innovations models do give somewhat uniform looking PIT histograms, their histograms do not look strikingly different from that generated based on the other models.

We further emphasize the results of the histograms by comparing the PIT of each of the data sets based on the different models using a quantile-quantile (QQ) plot. If the PIT is U(0,1), then the plot should be nearly linear. The plots are presented in Figures 8.11–8.15.

It is obvious here that the MAR models give consistently linear plots hence buttressing our claims based on the histograms that the MAR models better describe the underlying properties of financial time series.

8.4.1 Tail Density Forecast Test

Since we are interested in the forecast density at the tails, we proceed to carry out the Berkowitz tail test (Berkowitz (2001)) for testing the tail density forecast strength of the models. The tables are presented in Tables 8.3–8.12.

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-113.41	-236.07	-125.33	-250.33	23.85	28.52	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-113.41	-236.07	-125.33	-250.33	23.85	28.52	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-75.84	-215.28	-79.72	-220.02	7.76	9.47	0.02	0.01	reject NULL	reject NULL
GARCH(1,1)-t	-99.69	-220.99	-107.98	-233.13	16.58	24.30	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-t	-58.95		-59.03		0.17		0.92		fail to reject NULL	
MAR(3;2,2,1)-norm	-75.90	-316.96	-76.29	-317.30	0.79	0.67	0.67	0.72	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-t	-70.56	-323.55	-71.16	-325.00	1.20	2.90	0.55	0.23	fail to reject NULL	fail to reject NULL

Table 8.3: 95% Berkowitz likelihood ratio tail test for one-step ahead density forecast for IBM daily Returns

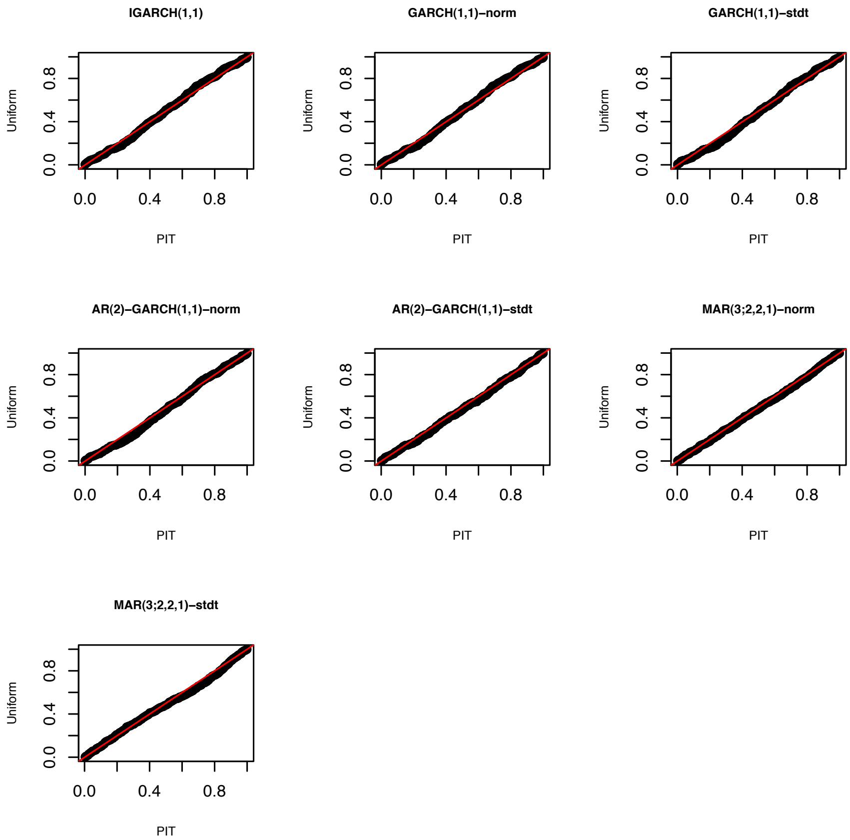

Figure 8.11: QQ plot comparing the PIT of IBM returns with respect to the density forecast produced under the assumption that IBM returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t to the uniform random sample

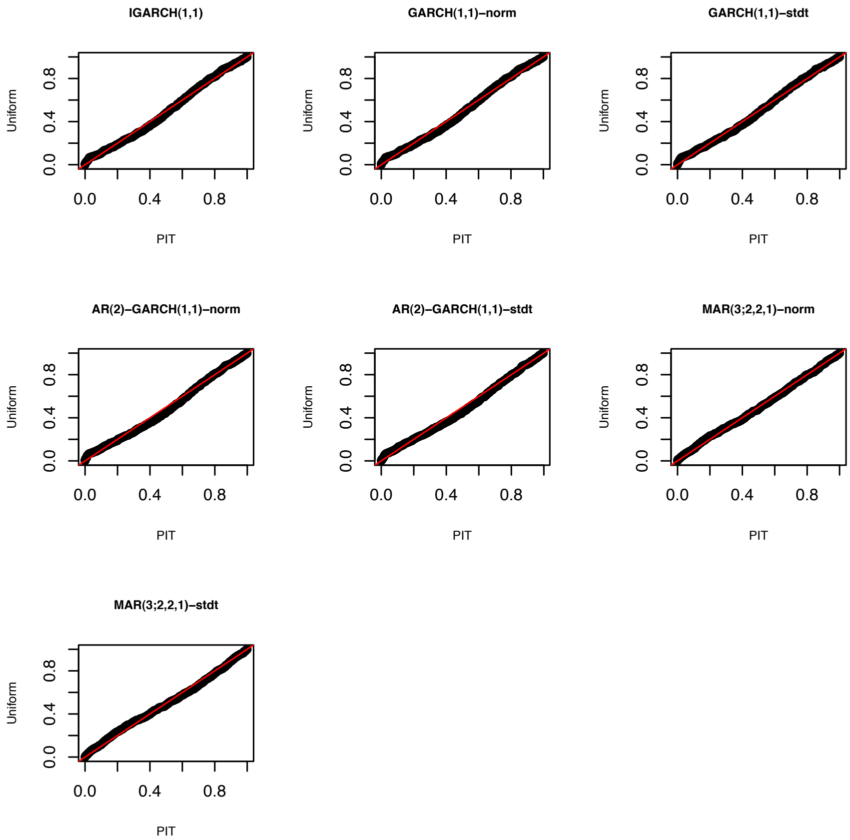

Figure 8.12: QQ plot comparing the PIT of SP500 returns with respect to the density forecast produced under the assumption that SP500 returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t to the uniform random sample

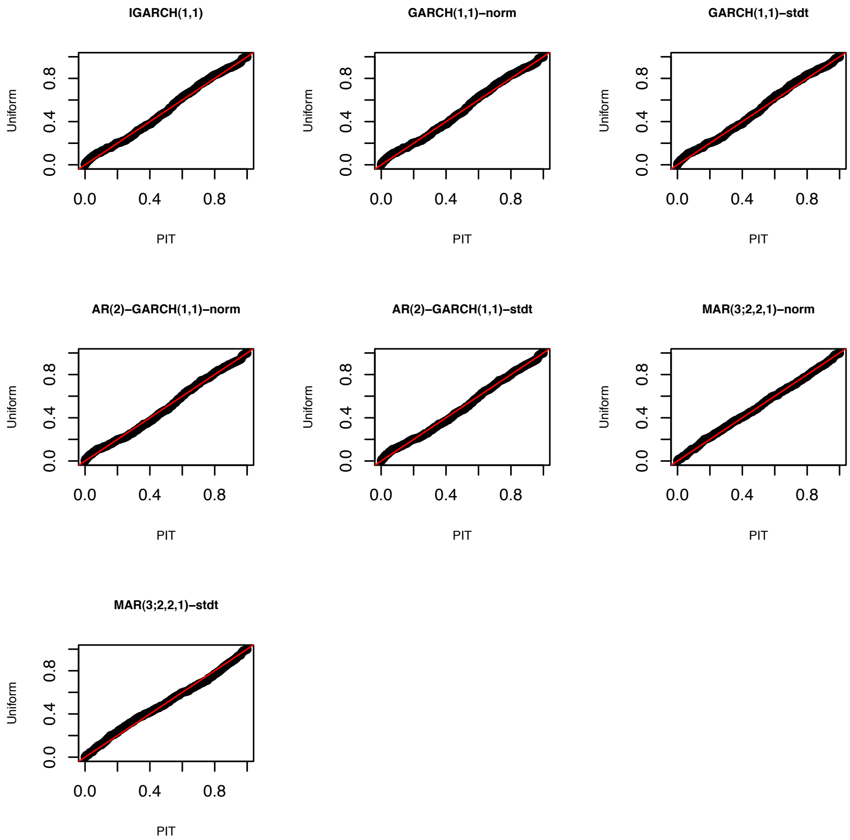

Figure 8.13: QQ plot comparing the PIT of FTSE returns with respect to the density forecast produced under the assumption that FTSE returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t to the uniform random sample

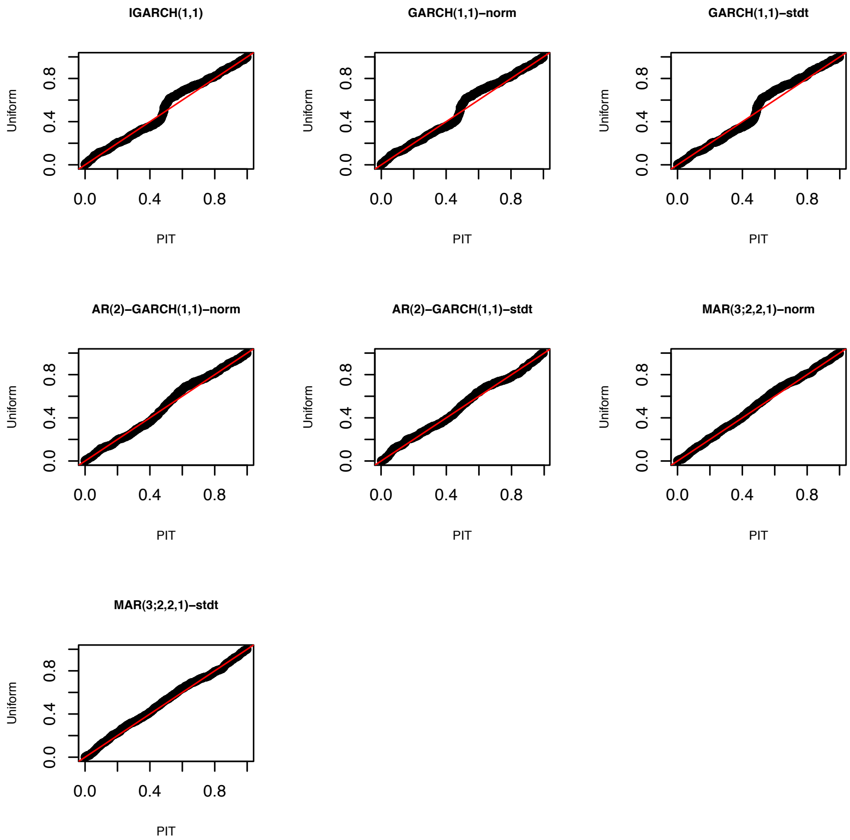

Figure 8.14: QQ plot comparing the PIT of GBPUSD returns with respect to the density forecast produced under the assumption that GBPUSD returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t to the uniform random sample

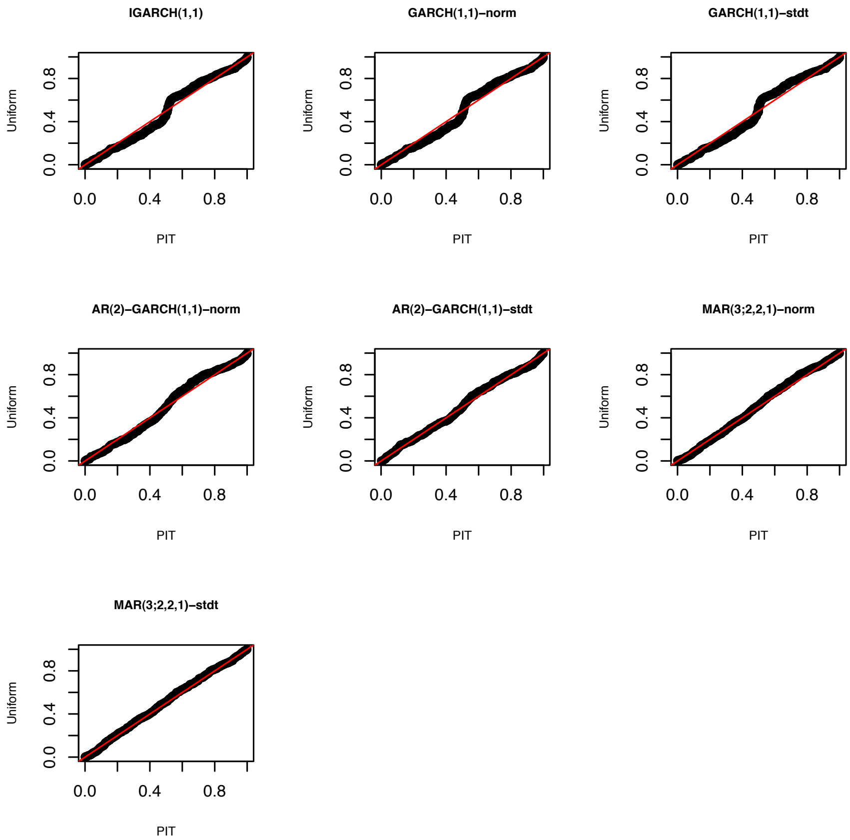

Figure 8.15: QQ plot comparing the PIT of GBPEUR returns with respect to the density forecast produced under the assumption that GBPEUR returns are IGARCH(1,1), GARCH(1,1)-norm, GARCH(1,1)-t, AR(2)-GARCH(1,1)-norm, AR(2)-GARCH(1,1)-t, MAR(3;2,2,1)-norm and MAR(3;2,2,1)-t to the uniform random sample

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-113.41	-236.07	-125.33	-250.33	23.85	28.52	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-113.41	-236.07	-125.33	-250.33	23.85	28.52	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-75.84	-215.28	-79.72	-220.02	7.76	9.47	0.02	0.01	fail to reject NULL	reject NULL
GARCH(1,1)-t	-99.69	-220.99	-107.98	-233.13	16.58	24.30	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-t	-58.95	-206.06	-59.03	-206.75	0.17	1.39	0.92	0.50	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-norm	-75.90	-316.96	-76.29	-317.30	0.79	0.67	0.67	0.72	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-t	-70.56	-323.55	-71.16	-325.00	1.20	2.90	0.55	0.23	fail to reject NULL	fail to reject NULL

Table 8.4: 99% Berkowitz likelihood ratio tail test for one-step ahead density forecast for IBM daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-140.37	-289.40	-161.55	-312.06	42.36	45.32	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-140.37	-289.40	-161.55	-312.06	42.36	45.32	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-148.46	-298.67	-168.83	-321.77	40.75	46.20	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-t	-141.81	-307.49	-160.84	-328.95	38.04	42.92	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-t	-103.39		-108.36		9.95		0.01		reject NULL	
MAR(3;2,2,1)-norm	-129.35	-375.30	-133.60	-378.46	8.50	6.32	0.01	0.04	reject NULL	reject NULL
MAR(3;2,2,1)-t	-121.27	-389.54	-123.45	-394.05	4.37	9.01	0.11	0.01	fail to reject NULL	reject NULL

Table 8.5: 95% Berkowitz likelihood ratio tail test for one-step ahead density forecast for SP500 daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-140.37	-289.40	-161.55	-312.06	42.36	45.32	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-140.37	-289.40	-161.55	-312.06	42.36	45.32	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-148.46	-298.67	-168.83	-321.77	40.75	46.20	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-t	-141.81	-307.49	-160.84	-328.95	38.04	42.92	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-t	-103.39	-278.52	-108.36	-285.81	9.95	14.58	0.01	0.00	reject NULL	reject NULL
MAR(3;2,2,1)-norm	-129.35	-375.30	-133.60	-378.46	8.50	6.32	0.01	0.04	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-t	-121.27	-389.54	-123.45	-394.05	4.37	9.01	0.11	0.01	fail to reject NULL	fail to reject NULL

Table 8.6: 99% Berkowitz likelihood ratio tail test for one-step ahead density forecast for SP500 daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-102.30	-288.94	-107.42	-296.85	10.24	15.82	0.01	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-102.30	-288.94	-107.42	-296.85	10.24	15.82	0.01	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-101.93	-283.86	-106.90	-291.12	9.93	14.51	0.01	0.00	reject NULL	reject NULL
GARCH(1,1)-t	-109.24	-294.98	-115.55	-303.94	12.63	17.92	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-t	-83.90		-86.03		4.26		0.12		fail to reject NULL	
MAR(3;2,2,1)-norm	-94.98	-357.24	-97.59	-358.61	5.23	2.75	0.07	0.25	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-t	-92.77	-361.51	-92.89	-363.94	0.24	4.86	0.89	0.09	fail to reject NULL	fail to reject NULL

Table 8.7: 95% Berkowitz likelihood ratio tail test for one-step ahead density forecast for FTSE daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-102.30	-288.94	-107.42	-296.85	10.24	15.82	0.01	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-102.30	-288.94	-107.42	-296.85	10.24	15.82	0.01	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-101.93	-283.86	-106.90	-291.12	9.93	14.51	0.01	0.00	reject NULL	reject NULL
GARCH(1,1)-t	-109.24	-294.98	-115.55	-303.94	12.63	17.92	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-t	-83.90	-283.03	-86.03	-289.78	4.26	13.50	0.12	0.00	fail to reject NULL	reject NULL
MAR(3;2,2,1)-norm	-94.98	-357.24	-97.59	-358.61	5.23	2.75	0.07	0.25	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-t	-92.77	-361.51	-92.89	-363.94	0.24	4.86	0.89	0.09	fail to reject NULL	fail to reject NULL

Table 8.8: 99% Berkowitz likelihood ratio tail test for one-step ahead density forecast for FTSE daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-92.19	-267.50	-101.34	-275.85	18.30	16.71	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-92.19	-267.50	-101.34	-275.85	18.30	16.71	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-73.12	-221.02	-79.92	-225.98	13.59	9.92	0.00	0.01	reject NULL	reject NULL
GARCH(1,1)-t	-57.77	-215.35	-63.06	-217.48	10.58	4.26	0.01	0.12	reject NULL	fail to reject NULL
AR(2)-GARCH(1,1)-t	-35.01		-36.08		2.14		0.34		fail to reject NULL	
MAR(3;2,2,1)-norm	-139.50	-501.20	-148.63	-504.84	18.26	7.28	0.00	0.03	reject NULL	reject NULL
MAR(3;2,2,1)-t	-124.11	-507.35	-125.12	-507.85	2.02	1.01	0.36	0.60	fail to reject NULL	fail to reject NULL

Table 8.9: 95% Berkowitz likelihood ratio tail test for one-step ahead density forecast for GBP/USD daily Returns

Tables 8.3, 8.5, 8.7, 8.9 and 8.11 present the results of the Berkowitz (2001) test at 95% significance each for tail levels $\alpha = 5\%$ and $\alpha = 1\%$. The AR(2)-GARCH(1,1) with Student-t innovations and MAR(3;2,2,1) with Student-t innovations models tend to perform better than all the other models, as we fail to reject the null hypothesis in most instances for these models except for the S&P500 data set where none of the models seem to have adequately captured the tail density forecast at the 95% confidence level.

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-92.19	-267.50	-101.34	-275.85	18.30	16.71	0.00	0.00	reject NULL	reject NULL
GARCH(1,1)-norm	-92.19	-267.50	-101.34	-275.85	18.30	16.71	0.00	0.00	reject NULL	reject NULL
AR(2)-GARCH(1,1)-norm	-73.12	-221.02	-79.92	-225.98	13.59	9.92	0.00	0.01	reject NULL	reject NULL
GARCH(1,1)-t	-57.77	-215.35	-63.06	-217.48	10.58	4.26	0.01	0.12	reject NULL	fail to reject NULL
AR(2)-GARCH(1,1)-t	-35.01	-212.41	-36.08	-215.59	2.14	6.36	0.34	0.04	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-norm	-125.31	-464.68	-135.93	-471.57	21.23	13.78	0.00	0.00	reject NULL	reject NULL
MAR(3;2,2,1)-t	-114.30	-480.32	-115.80	-481.84	3.01	3.04	0.22	0.22	fail to reject NULL	fail to reject NULL

Table 8.10: 99% Berkowitz likelihood ratio tail test for one-step ahead density forecast for GBP/USD daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-40.42	-204.92	-43.73	-205.16	6.63	0.48	0.04	0.79	reject NULL	fail to reject NULL
GARCH(1,1)-norm	-40.42	-204.92	-43.73	-205.16	6.63	0.48	0.04	0.79	reject NULL	fail to reject NULL
AR(2)-GARCH(1,1)-norm	-70.25	-217.57	-70.96	-217.93	1.42	0.72	0.49	0.70	fail to reject NULL	fail to reject NULL
GARCH(1,1)-t	-39.62	-201.01	-42.16	-201.03	5.08	0.04	0.08	0.98	fail to reject NULL	fail to reject NULL
AR(2)-GARCH(1,1)-t	-20.09		-23.49		6.81		0.03		reject NULL	
MAR(3;2,2,1)-norm	-125.31	-464.68	-135.93	-471.57	21.23	13.78	0.00	0.00	reject NULL	reject NULL
MAR(3;2,2,1)-t	-114.30	-480.32	-115.80	-481.84	3.01	3.04	0.22	0.22	fail to reject NULL	fail to reject NULL

Table 8.11: 95% Berkowitz likelihood ratio tail test for one-step ahead density forecast for GBP/EUR daily Returns

Method	Parameters									
	ull		rll		LR		LRp		Decision	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Riskmetrics	-40.42	-204.92	-43.73	-205.16	6.63	0.48	0.04	0.79	fail to reject NULL	fail to reject NULL
GARCH(1,1)-norm	-40.42	-204.92	-43.73	-205.16	6.63	0.48	0.04	0.79	fail to reject NULL	fail to reject NULL
AR(2)-GARCH(1,1)-norm	-70.25	-217.57	-70.96	-217.93	1.42	0.72	0.49	0.70	fail to reject NULL	fail to reject NULL
GARCH(1,1)-t	-39.62	-201.01	-42.16	-201.03	5.08	0.04	0.08	0.98	fail to reject NULL	fail to reject NULL
AR(2)-GARCH(1,1)-t	-20.09	-213.66	-23.49	-217.66	6.81	7.99	0.03	0.02	fail to reject NULL	fail to reject NULL
MAR(3;2,2,1)-norm	-133.43	-476.66	-140.43	-480.94	14.00	8.57	0.00	0.01	reject NULL	fail to reject NULL
MAR(3;2,2,1)-t	-118.65	-502.40	-119.73	-503.25	2.16	1.70	0.34	0.43	fail to reject NULL	fail to reject NULL

Table 8.12: 99% Berkowitz likelihood ratio tail test for one-step ahead density forecast for GBP/EUR daily Returns

Tables 8.4, 8.6, 8.8, 8.10 and 8.12 also present the results of the test at 99% significance each for tail levels $\alpha = 5\%$ and $\alpha = 1\%$. In this case the MAR(3;2,2,1) with Student-t innovations model does perform better for all the data sets at both $\alpha = 5\%$ and $\alpha = 1\%$ tail levels even for the S&P500 data set.

The results of the 95% and 99% Berkowitz test indicates that the models with Student-t innovations do give better out of sample tail density forecasts.

We observed from the tables that, in general, the MAR models tend to outperform the other models in most instances. The MAR model with Student-t innovations however performs better in all cases even for the SP500 data sets. These findings are consistent with the findings of De Raaij and Raunig (2005) and Diebold et al. (1998) that fat-tailed conditional distributions generally give more satisfactory density forecasts for financial time series.

8.5 Summary

Obtaining good density forecast relies heavily on the ability to make proper distributional assumptions and adequate modelling of the dynamics of the relevant conditional moments of financial returns. From the set of models we investigated, the MAR(3;2,2,1) model with fat-tailed Student-t innovations deliver the best out-of-sample density forecasts. The SP500 data out-of-sample evaluation results indicate that skewed fat-tailed conditional distributions may be needed to obtain entirely satisfactory density forecasts, this peculiarity is adequately captured by the MAR(3;2,2,1) model with Student-t innovations.

Chapter 9

Conclusions/Recommendations for further research

9.1 Conclusion

This thesis extensively studied the class of finite mixture models introduced by Wong and Li (2001), the mixture autoregressive (MAR) model.

In chapter 2, we defined the traditional residuals of the MAR model as the difference between the observed values and their conditional means.These traditional residuals are quite important as they give information on how close the observed values are to the means of the corresponding predictive distribution. We have shown that these residuals form a martingale difference sequence and that the unconditional variance of the traditional residuals is bounded by the expectation of its conditional variance. These are useful properties for establishing some asymptotic properties of the parameter estimates.

We mentioned some extensions of the MAR model and their applications to Financial modelling found in literature. We then compared the MAR Model to the class of GARCH models. We observed that both the GARCH type models and MAR models can be cast into the framework of Random Coefficient Autoregressive model as well as Generalised Hidden Markov models. We also noticed that persistence in the MAR model makes it suitable for capturing changes in volatility persistence in most financial data.

In chapter 3 we extended the work done by Klimko and Nelson (1978) on an estimation procedure for stochastic processes based on the minimization of a sum of squared deviations about conditional expectations to the Mixture Autoregressive model. We gave an MAR(2;1,1) example and showed that for the model, the variance-covariance matrix is positive definite and identical for both the conditional least square and maximum likelihood penalty functions.

In chapters 4 and 5 we established the geometric ergodicity of the MAR model and by implication showed that it satisfies the absolutely regular and strong mixing conditions (that is $\beta - mixing$ and $\alpha - mixing$ conditions). In addition, we showed that the markov chain Y_t (and by implication the MAR process y_t) has stationary distribution with finite second moments. We used the geometric ergodicity property of the MAR model along with the results in Douc et al. (2004) and showed that the Maximum Likelihood Estimator of the MAR model is both consistent and asymptotically normal.

In chapters 6 and 7, we did an extensive study on classes of financial risk measures, we focused on the most popular of these, Value at Risk (VaR) and Expected Shortfall(ES). We discussed various existing approaches to evaluating VaR and ES. We then proposed the use of the MAR model as a viable approach for evaluating VaR and hence ES. We computed out of sample VaR and ES, at $\alpha = 1\%$ and $\alpha = 5\%$ for individual stock (IBM), stock indices (S&P500 and FTSE) and exchange rate (GBP/USD and GBP/EUR). Based on Tsay (1997)'s suggestion to treat the results based on empirical quantiles as conservative estimates of the true VaR (that is, lower bounds), we found that the approaches based on EVT and MAR models give values close to the empirical quantiles while the approaches based on GARCH models tend to underestimate VaR and ES. These agrees with the results in Tsay (1997).

Thereafter, we proceeded to backtest the VaR and ES results using the framework developed by Kupiec (1995) and Christoffersen (1998) which examines whether the each approach adequately forecasts the expected number of violations, generate independent violations and consequently gives ES with violation residuals that exhibit zero mean behaviour. We carried out backtesting procedures for each of the data sets independently at $\alpha = 1\%$ and $\alpha = 5\%$.

Across all the data sets examined, the MAR(3;2,2,1) models consistently performs well at both $\alpha = 5\%$ and $\alpha = 1\%$ p values, as we fail to reject the null of correct number of violations (exceedances), we also agree that the probability that a violation will occur tomorrow does not depend on the violations that have occurred today. In particular, we found that for individual stock (IBM data set) and both forex data sets (GBPUSD and GBPEUR data sets), most of the approaches for evaluating VaR seem to give correct exceedances based on the p-values, this conclusion is further buttressed by a close examination and comparison of the expected and actual exceedances columns which reveal that for these data sets, the figures in these two columns are not too far from each other at $\alpha = 5\%$ but a bit farther off at $\alpha = 1\%$. However, for the stock indices (SP500 and FTSE), based on the $p-$values, only the MAR models give correct exceedances.

Furthermore, our comparison of the expected and actual exceedances columns revealed that the MAR models figures were closer together in the two columns. These conclusion can be attributed to the fact that stock indices consists of a portfolio of various individual stocks and hence most likely to exhibit more volatile behaviour. The MAR models show up to be better suited to capture these kind of data dynamics.

The ES backtest results revealed that the MAR model does perform better as at both $\alpha = 5\%$ and $\alpha = 1\%$, the violation residuals do behave like realisations of i.i.d variables from a distribution with zero mean.

In chapter 9, we used the Berkowitz01 test to compare the predictive density of individual stock (IBM), stock indices (S&P500 and FTSE) and exchange rate (GBP/USD and GBP/EUR) based on the Gaussian GARCH(1,1), Student-t GARCH(1,1), the AR(2)-GARCH(1,1) with Gaussian innovations, the AR(2)-GARCH(1,1) with Student-t innovations against the MAR model(3;2,2,1) with both Gaussian and Student-t innovations. We find that the MAR models consistently perform better than the other models examined.

9.2 Further Research

In this thesis we have based our VaR and ES calculations as well as the prediction density on one-step ahead predictions. The next steps would be to consider multi-step VaR and ES predictions and as well as multi-step prediction densities and see if the MAR model does perform better.

It will also be worth exploring applications of seasonal mixture autoregressive models to financial modelling.

Appendix A

Useful definitions, terms and theorems

1. **Cauchy Schwartz Inequality**

   For any two random varables X and Y, we have that

   $$E(XY)^2 \leq E(X^2)E(Y^2)$$

   with equality attained when $Y = cX$, c being a constant. It is worth mentioning that the Cauchy-Schwartz inequality is a special case of the Hölder's inequality i.e. for $p = 2$.see Davidson (1997).

2. **Martingale Central Limit Theorems**

   Theorem A.0.1. *Let $\{X_{nt}, \mathcal{F}_{nt}\}$ be a martingale difference array with finite unconditional variances $\{\sigma_{nt}^2\}$, and $\sum_{t=1}^n \sigma_{nt}^2 = 1$. If*

   i $\sum_{t=1}^n X_{nt}^2 \xrightarrow{pr} 1$, *and*

   ii $\max_{1 \leq t \leq n} |X_{nt}| \xrightarrow{pr} 0$,

   then $S_n = \sum_{t=1}^n X_{nt} \xrightarrow{D} N(0, 1)$.

Theorem A.0.2. *If $\{X_{nt}, \mathcal{F}_{nt}\}$ is a square integrable martingale difference sequence and $E(X_t^2 \mid \mathcal{F}_{t-1}) = \sigma_t^2$ a.s then, there exists a sequence of positive constants $\{c_t\}$ such that $\{X_t^2/c_t^2\}$ is uniformly integrable and $\sup_n nM_n^2/s_n^2 < \infty$ where $M = \max_{1 \leq t \leq n} C_t^2$, and conditions i and ii in A.0.1 hold for $X_{nt} = X_t/S_n$*

see Hall.P and Heyde.C.C (1980) and Davidson (1997) for details and proof.

3. **Billinglsey's Central Limit Theorem for Martingales**

This is better referred to as "The Lindeberg-Lèvy theorem for martingales"

Theorem A.0.3. *Billingsley (961b) Theorem. Let $u_1, u_2, \cdots$ be a stationary, ergodic stochastic process such that $E\{u_1^2\}$ is finite and $E\{u_n \| u_1, \cdots, u_{n-i}\} = 0$ with probability one. Then the distribution of $n^{1/2} \sum_{k=1}^{n} u_k$ approaches the normal distribution with mean 0 and variance $E\{u_1^2\}$.*

It is noteworthy that the condition imposed is to ensure that the the partial sums $n^{1/2} \sum_{k=1}^{n} u_k$ form a martingale

4. **Strong Laws for Martingales**

Theorem A.0.4. *Let $\{X_t, \mathcal{F}_t\}_0^\infty$ be a martingale difference sequence with variance sequence $\{\sigma_t^2\}$, and $\{a_t\}$ a positive constant sequence with $a_t \uparrow \infty$. $S_n/a_n \xrightarrow{as} 0$ if $\sum_{t=1}^{\infty} \sigma_t^2/a_t^2 < \infty$.*

Theorem A.0.5. *If $\{X_t \mathcal{F}_t\}_1^\infty$ is a martingale difference sequence satisfying the following $\sum_{t=1}^{\infty} E|X_t|^p/a_t^p$ then for $1 \leq p \leq 2$, $S_n/a_n \xrightarrow{as} 0$.*

For details, arguments and proofs, see Hall.P and Heyde.C.C (1980) and Davidson (1997).

5. **Ergodic Theorem**

Theorem A.0.6. *Let* $\{X_t(\omega)\}_1^\infty$ *be a stationary, ergodic, integrable sequence. Then* $\lim_{n\to\infty} S_n(\omega)/n = E(X_1)$ *,a.s.*

see Davidson (1997) for proof.

6. **Fatou Property**Jouini et al. (2006) Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a standard probability space. For a function $U : \mathbb{L}^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ the following are equivalent:

   i U is a law invariant monetary utility function.

   ii There is a law invariant, lower semi-continuous, convex function $V\mathbb{L}^1(\Omega, \mathcal{F}, \mathbb{P}) \to [0, \infty]$ such that $dom(V) \subset \mathcal{P}(\Omega, \mathcal{F}, \mathbb{P})$ and

$$U(X) = \inf_{Y \in \mathbb{L}^1} \{E[XY] + V(Y)\} \text{ for } X \in \mathbb{L}^\infty \tag{A.0.0.1}$$

   iii There is a convex function $v : \mathcal{P}([0,1]) \to [0, \infty]$ such that $\mathbb{L}^1$

$$U(X) = \inf_{m \in \mathcal{P}([0,1])} \{\int_0^1 U_\alpha(X) dm(\alpha) + v(m)\} \text{ for } X \in \mathbb{L}^\infty \tag{A.0.0.2}$$

   If any of these conditions is satisfied, then U satisfies the Fatou property.

7. **Mean Value Theorem** If $f(x)$ is defined and continuous on the interval $[a, b]$ and differentiable on (a, b), then there is at least one number c in the interval (a, b) (that is $a < c < b$) such that

$$f^{'}(c) = \frac{f(b) - f(a)}{b - a} \tag{A.0.0.3}$$

8. **Standardaized t-distribution (Zubrzycki (1972))** The standardized Student-t distribution is a special case of the Student's t distribution. A

random variable has a standardized Student's t distribution with n degrees of freedom if it can be written as a ratio between a standard normal random variable and the square root of a Gamma random variable that is, $X = \frac{Z}{\sqrt{(\Gamma)}}$, where Z is a standard normal random variable and Γ is a Gamma random variable with parameters n and $h = 1$. Z and Γ are independent.

The standardized Student-t distribution can also be written as, $x = \frac{Z}{\sqrt{(\chi_n^2/n)}}$ where χ_n^2 is a Chi-square random variable with n degrees of freedom.

Appendix B

R Packages

All computations were done with R (R Core Development Team (2012)) and the following contributed R packages.

- PerformanceAnalytics (Carl et al. (2012))
- quantmod (Ryan (2011))
- FinTS (Graves (2009))
- car (Fox and Weisberg (2011))
- fExtremes (Wuertz et al. (2012b))
- fBasics (Wuertz et al. (2012a))
- rugarch (Ghalanos (2012))
- mixAR (Boshnakov (2012))
- mixARsim (Boshnakov (2011a))
- tiger (Reusser (2011))

- fImport (Wuertz and many others (2012))
- xtable (Dahl (2012))

Bibliography

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance*, 26(7):1505 – 1518.

Acerbi, C., Nordio, C., and Sirtori, C. (2008). Expected Shortfall as a Tool for Financial Risk Management. *ArXiv Condensed Matter e-prints*.

Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance*, 26(7):1487–1503.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):283–301.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Balkema, A. and Haan, L. D. (1974). Residual lifetime at great age. *Annals of Probability*, 2:792–804.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, 19(4):465–474.

Bhattacharya, R. and Lee, C. (1999). On geometric ergodicity of nonlinear autoregressive models. *Statistics and Probability Letters*, 41(4):439 – 440.

Billingsley, P. (1961b). The lindeberg-levy theorem for martingales. *Proceedings of the American Mathematical Society*, 12(15):788–792.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1986). Arch models. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 49, pages 2959–3038. Elsevier, 1 edition.

Bollerslev, T. and Wooldridge, J. M. (1988). *Quasi-maximum likelihood estimation of dynamic models with time varying covariances*. Massachusetts Institute of Technology in Cambridge, Mass, Cambridge.

Boshnakov, G. (2011a). *mixARsim: Additional utilities for simulations with mixture autoregressive models*. R package version 0.0-1.

Boshnakov, G. (2012). *mixAR: Mixture autoregressive models*. R package version 0.19-12.

Boshnakov, G. N. (2009). Analytic expressions for predictive distributions in mixture autoregressive models. *Statistics and Probability Letters*, 79(15):1704–1709.

Boshnakov, G. N. (2011b). On first and second order stationarity of random coefficient models. *Linear Algebra and its Applications*, 434(2):415 – 423.

Box G. E. P., J. G. M. and C., R. G. (1994). *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliff, 3rd edition edition.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2(2):107–144.

Butler, C. (1999). *Mastering Value at Risk: A Step by Step Guide to Understanding and Applying VaR*. Market Editions, Chicago.

Carl, P., Peterson, B. G., Boudt, K., and Zivot, E. (2012). *PerformanceAnalytics: Econometric tools for performance and risk analysis.* R package version 1.0.4.4.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39.

Chen, S. X. (2008). Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, 6(1):87–107.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics. Springer, London.

Cox, D. and Hinkley, D. (1979). *Theoretical Statistics*. Chapman & Hall, Florida.

Cramér, H. (1945). *Mathematical Methods of Statistics*. Princeton Landmarks in Mathematics and Physics Series. Princeton University Press, United States.

Dahl, D. B. (2012). *xtable: Export tables to LaTeX or HTML.* R package version 1.7-0.

Damodaran, A. (1999). Estimating risk parameters. Finance Working Papers.

Davidson, J. (1997). *Stochastic Limit Theory*. Oxford University Press, New York, 3rd edition. Advanced texts in Econometrics.

De Raaij, G. and Raunig, B. (2005). Evaluating density forecasts from models of stock market returns. *The European Journal of Finance*, 11(2):151–166.

Dedecker, J., Doukhan, P., Lang, G., Leon, J., Louhichi, S., and Prieur, C. (2007). *Weak Dependence: With Examples and Applications*. Springer, New York.

Delbaen, F., Hochschule, E. O. T., and Urich, Z. (2000). Coherent risk measures on general probability spaces.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883.

Douc, R., Moulines, E., and Ryden, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics*, 32(5):2254–2304.

Duan, J.-C. (1997). Augmented garch (p,q) process and its diffusion limit. *Journal of Econometrics*, 79(1):97 – 127.

Durrett, R. (1996). *Stochastic calculus: a practical introduction*. CRC Press, Boca Raton, London.

Eberlein, E., Frey, R., Kalkbrener, M., and Overbeck, L. (2007). Mathematics in financial risk management.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics & Applied Probability. Chapman & Hall, Florida.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. and Manganelli, S. (1999). *CAViaR: Conditional Value at Risk by Quantile Regression*. National Bureau of Economic Research, Cambridge. Working paper 7341.

Feigin, P. D. and Tweedie, R. L. (1985). Random coefficient autoregressive processes:a markov chain analysis of stationarity and finiteness of moments. *Journal of Time Series Analysis*, 6(1):1–14.

Follmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, 6:429–447.

Fotios C. Harmantzis, Linyan Miao, Y. C. (2006). Empirical study of value-at-risk and expected shortfall models with heavy tails. *The Journal of Risk Finance*, 7(2):117–135.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition.

Francq, C. and Roussignol, M. (1998). Ergodicity of autoregressive processes with markov-switching and consistency of the maximum-likelihood estimator. *Statistics*, 32(2):151–173.

Francq, C. and Zakoian, J.-M. (2010). *Garch Models: Structure, Statistical Inference and Financial Applications*. Wiley, United Kingdom.

Fruhirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York, 4th edition edition.

Ghalanos, A. (2012). *rugarch: Univariate GARCH models.* R package version 1.0-12.

Gili, M. and Këllezi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, 24:207–228.

Gilli, M. and këllezi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, 27(2):207–228.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453.

Graves, S. (2009). *FinTS: Companion to Tsay (2005) Analysis of Financial Time Series*. R package version 0.4-4.

Gumbel, E. (2004). *Statistics of Extremes*. Dover Books on Mathematics Series. Dover, New York.

Hall.P and Heyde.C.C (1980). *Martingale limit theory and its application.* Academic Press,INC, New York ,London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hjellvik, V. and Tjostheim, D. (1999). Modelling panels of intercorrelated autoregressive time series. *Biometrika*, 86(3):573–590.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge University Press, "Cambridge".

Hull, J. (2000). *Options, Futures and Other Derivatives*. Prentice Hall, Upper Saddle River, NJ, 4th edition edition.

Jadhav, D., Ramanathan, T., and Naik-Nimbalkar, U. (2009). Modified Estimators of the Expected Shortfall. *Journal of Emerging Market Finance*, 8(2):87–107.

Jin, S. and Li, W. (2006). Modeling panel time series with mixture autoregressive model. *Journal of Data Science*, 4(4):425–446.

Jorion, P. (1997). *Value at Risk: The New Benchmark for Controlling Market Risk*. McGraw-Hill, New York.

Jouini, E., Schachermayer, W., and Touzi, N. (2006). Law invariant risk measures have the fatou property. In *Advances in Mathematical Economics*, volume 9 of *Advances in Mathematical Economics*, pages 49–71. Springer Japan.

J.P.Morgan (1995). *RiskMetrics^{TM} -Technical Document*. Morgan Guaranty Trust Company, New York, fourth edition.

Klimko, L. A. and Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *The Annals of Statistics*, 6(3):629–642.

Klüppelberg, P. E. C. and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer, Berlin.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1):53–89.

Kuiper, N. H. (1962). Tests concerning random points on a circle. *Nederl. Akad. Wetensch. Proc. Ser. A*, 63:38–47.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. Finance and Economics Discussion Series 95-24, Board of Governors of the Federal Reserve System (U.S.).

Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the garch model. *Journal of Business and Economic Statistics*, 8(2):225–234.

Lampinen, J. and Oja, E. (1989). Self-organizing maps for spatial and temporal ar models. Proceedings of the 6th Scandanavian Conference on Image Analysis, pages 120–127, Helsinki Finland.

Lanne, M. and Saikkonen, P. (2003). Modeling the u.s. short-term interest rate by mixture autoregressive processes. *Journal of Financial Econometrics*, 1(1):96–125.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436):1504–1515.

Leadbetter, R., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer series in statistics. Springer-Verlag, New York,London.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models. *Stochastic Processes and their Applications*, 40(1):127 – 143.

Loeve, M. (1977). *Probabilty Theory I*. Springer-Verlag, New York ,London, 4th edition edition. Graduate Texts in Mathematics.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Mamon, R. S. and (Eds.), R. J. E. (2007). *Hidden Markov Models in Finance*. Springer, New York. Expected Shortfall Under a Model With Market and Credit Risks, ch. 6 Kin Bong Siu and Hailiang Yang.

Manganelli, S. and Engle, R. F. (2001). *Value at Risk Models in Finance*. European Central Bank, New York. Working paper series.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1):77–91.

Martinetz, T., Berkovich, S., and Schulten, K. (1993). 'neural-gas' network for vector quantization and its application to time-series prediction. *Neural Networks, IEEE Transactions on*, 4(4):558 –569.

Masanobu Taniguchi, Y. K. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer, New York.

McNeil, A., Frey, R., and Embrechts, P. (2010). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press, New Jersey.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer, Verlag, London.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *J. Math. Psychol.*, 47(1):90–100.

Neave, H. and Worthington, P. (1988). *Distribution-Free Tests*. Unwin Hyman, United Kingdom.

Ni, H. and Yin, H. (2008). A self-organising mixture autoregressive network for non-stationary time series modelling. *International Journal of Neural Systems*, 18(6):469 – 480.

Ni, H. and Yin, H. (2009). A self-organising mixture autoregressive network for fx time series modelling and prediction. *Neurocomputing*, 72(16):3529 – 3537.

Picklands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131.

Poterba, J. M. and Summers, L. H. (1986). The persistence of volatility and stock market fluctuations. *The American Economic Review*, 76(5):1142–1151.

Quinn, B. G. and Nicholls, D. F. (1981). The stability of random coefficient autoregressive models. *International Economic Review*, 22(3):741–44.

R Core Development Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reusser, D. (2011). *tiger: TIme series of Grouped ERrors.* R package version 0.2.2.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Ruppert, D. (2004). *Statistics and finance: an introduction.* Springer texts in statistics, New York.

Ryan, J. A. (2011). *quantmod: Quantitative Financial Modelling Framework.* R package version 0.3-17.

Saikkonen, P. (2007). Stability of mixtures of vector autoregressions with autoregressive conditional heteroskedasticity. *Statistica Sinica*, 17(1):221–239.

Scaillet, O. (2004). Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance*, 14(1):115–129.

Shao, Q. (2006). Mixture periodic autoregressive time series models. *Statistics and Probability Letters*, 76(6):609 – 618.

Stockis, J.-P., Franke, J., and Kamgaing, J. T. (2010). On geometric ergodicity of charme models. *Journal of Time Series Analysis*, 31(3):141–152.

Strickert, M. and Hammer, B. (2005). Merge som for temporal data. *Neurocomputing*, 64(0):39 – 71.

S.Tsay, R. (1997). Extreme values and their applications in finance. Booth School of Business, The University of Chicago and Risk Management Institute, National University of Singapore.

Tanner, M. A. (1993). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions.* Springer-Verlag, New York ,London, 2nd edition edition. Springer series in statistics.

Tasche, D. (2002). Expected shortfall and beyond. *Journal of Banking and Finance*, 26:1519–1533.

Tjostheim, D. (1986). Estimation in nonlinear time series models. *Stochastic Processes and their Applications*, 21(2):251 – 273.

Tjostheim, D. (1990). Non-linear time series and markov chains. *Advances in Applied Probability*, 22(3):587–611.

Tong, H. (1990). *Non-linear Time Series.* Oxford University Press, New York.

Tsay, R. S. (1997). *Analysis of Financial Time Series.* John Wiley & Sons, Inc., New Jersey, second edition. Wiley Series in Probability and Statistics.

Tweedie, R. L. (1988). Invariant measures for markov chains with no irreducibility assumptions. *Journal of Applied Probability*, 25:pp. 275–285.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wong, A. C. S. and Chan, W.-S. Mixture gaussian time series modelling of long-term market returns. Technical report.

Wong, C. and Li, W. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 62(1):95–115.

Wong, C. S., Chan, W. S., and Kam, P. L. (2009). A student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika*, 96(3):751–760.

Wong, C. S. and Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association*, 96(455):982–995.

Wuertz, D., core team members, R., uses code builtin from the following R contributed packages: gmm from Pierre Chauss, gld from Robert King, gss from Chong Gu, nortest from Juergen Gross, from David Scott, H., sandwich from Thomas Lumley, Zeileis, A., fortran/C code from Kersti Aas, and akima from Albrecht Gebhardt (2012a). *fBasics: Rmetrics - Markets and Basic Statistics.* R package version 2160.85.

Wuertz, D. and many others (2012). *fImport: Rmetrics - Economic and Financial Data Import.* R package version 2160.81.

Wuertz, D., many others, and see the SOURCE file (2012b). *fExtremes: Rmetrics - Extreme Financial Market Data.* R package version 2160.78.

Yu.A.Davydov (1973). Mixing conditions for markov chains. *Theory Probability and its Applications*, 18(2):321–338.

Zubrzycki, S. (1972). *Lectures in probability theory and mathematical statistics.* Modern analytic and computational methods in science and mathematics. American Elsevier Pub. Co.