System Identification for Complex Financial System

Liang Zhao

A thesis submitted to the University of Sheffield for the degree of
Doctor of Philosophy

Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield, UK
November 2010

Abstract

The main purpose of this thesis focuses on the investigation of major financial volatility models including the relevant mean model used in the context of volatility estimation, and the development of a systematic nonlinear identification methodology for these problems. Financial volatility is one of the key aspects in financial economics and volatility modelling involves both the mean process modelling, and the volatility process modelling. Although many volatility models have been derived to approximate the volatility process, linear mean models are almost always used and to the best of our knowledge there is no application of fitting the mean process using a nonlinear model with selected structure.

Based on the fact that nonlinearity has been observed in many financial market return data sets, the Nonlinear AutoRegression Moving Average with eXogenous input (NARMAX) modelling methodology with the term selection algorithm Orthogonal Forward Regression (OFR) is proposed to approximate the nonlinear mean process during volatility modelling. However, the assumption of a constant variance is usually violated in financial market return data. A new Weighted OFR algorithm is therefore proposed to correct for the impact of heteroskedastic noise on the term selection of the nonlinear mean model based on the assumption that the variance process is modelled by a Generalized AutoRegressive Conditional Heteroskedastic (GARCH) model. Because the weights to use are unknown, an iterative refined procedure is developed to learn the weights and to simultaneously improve the parameter estimates of both the mean and the volatility models.

New validation methods are proposed to validate the nonlinear selected mean model and the volatility model. During the validation, the assumptions associated with the mean model are tested using a correlation method and the assumptions of the volatility model are tested using a Brock-Dechert-Scheinkman (BDS) independent and identically distributed (i.i.d.) testing method. The prediction performance of the mean and volatility models is evaluated using a hold out Cross Validation (CV) method. A departure in the prediction of the volatility for the linear mean model, when using nonlinear simulated data, is successfully identified by the new validation methods and the nonlinear selected mean model passes the test.

Another application of the NARAMX model, in the very new field of modelling mortality rate, is introduced. A quadratic polynomial mortality rate model selected by the OFR algorithm is developed based on the LifeMetrics male deaths and exposures data for England & Wales from the Office of National Statistics. Comparing the long term prediction of the new model with the Cairns-Blake-Dowd (CBD) statistical mortality rate model indicates the better prediction performance of the quadratic polynomial models. A back-testing method is applied to indicate the robustness of the selected NARMAX type mortality rate models.

The term selection, parameter estimation, validation methods and new identification procedures proposed in this thesis open a new gateway to apply the NARMAX modelling technique in the financial area, and for mortality rate modelling to provide a new empirical practice of the NARMAX modelling method.

Acknowledgements

I would like to express my most sincere gratitude to Professor Stephen A. Billings, my supervisor, for his excellent supervision, guidance, and encouragement in my research and writing. He has provided numerous supports during my studies in the period of this project. This thesis would not have been possible without his financial support. I also want to express great appreciation to Professor Kevin Dowd, my second supervisor, for his guidance in finance. His support inspires me during my research and I have learned a lot from the work in mortality modelling. His thoughts were brainstorms to me and I enjoyed talks with him. I want to thank Dr. Daniel Coca, my third supervisor for his great guidance and valuable suggestions on my research.

I would like to thank the Department of ACSE for the scholarship and equipment in the laboratory. It is a pleasure to thank the staff of the University of Sheffield for their support.

I must also acknowledge Dr. Xiuping Hua for her great help in my internship. Very special thanks to my friends Yuzhu Guo, Dazhi Jiang, Qi Jiang, Jingjing Luo, Shu Wang, and all other friends in Sheffield.

Finally, I would like to dedicate this thesis to my parents and my girlfriend Liu Ye. Without emotional and financial support, this research would have been impossible to carry out.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background and Motivation

The concept of volatility was firstly introduced by Louis Bachelier a century ago in 1900 when he created the term 'coefficient of nervousness' to the price. The term 'financial volatility' is then usually used to refer to the standard deviation or, alternatively, but in practice, equivalently, to the variance of the underlying return data associated with a time series; loosely speaking, it refers to the intensity of the fluctuation which affects the return prices. For example, the volatility of the stock market will obviously increase during the periods of financial turmoil such as the market crash in Oct, 1987, the Asian Financial Crisis starting from July, 1997 and the terrorist attack on 11th, Sep, 2001.

Financial volatility is one of the key aspects in financial economics especially for the pricing of derivative securities for example in the Black-Scholes model (Black and Scholes, 1973), where the volatility of underlying asset is used to price the option. Derivatives with clearly specified measurements of volatility in the contracts are often traded nowadays and investors trend to maximize the expect return subject to a risk constraint of the portfolio. Therefore, the forecast of the volatility of the underlying assets are essential over the defined period and any improvement in the volatility prediction by even one percent can be significant for the investment decision. Volatility is also commonly used to calculate the Value-at-Risk (VaR) estimation for the purpose of risk management. The VaR is one of the commonly used modern risk measure techniques and it measures the probability of the worst expected loss under normal market conditions over a specific time interval at a given confidence level.

Based on the observed features of volatility, 'large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes' (Mandelbrot, 1963) and 'volatility response to a large negative return is often far greater than it is to a large positive return of the same magnitude' (Black, 1976), many different volatility models have been developed. The Exponential Weighted Moving Average (EWMA) model is one of the commonly used volatility models and is used in the famous application of the JP Morgan RiskMetrics®. More sophisticated groups of volatility models have been developed since the AutoRegressive Conditional Heteroskedastic (ARCH) model (Engle, 1982) was introduced. The Nobel Prize for Economics has been awarded to Robert Engle in 2003 for the innovation of analyzing economic time series with time-varying volatility using the ARCH model. Bollerslev (1986) then extended the ARCH model to the more general case as known as the Generalized ARCH (GARCH) model. The feature of volatility described by Black (1976) is now commonly referred to as the leverage effect which means that the equity returns are strongly asymmetric. However, both ARCH and GARCH models are not capable of charaterising asymmetry. The extension of GARCH has been developed in econometric ways to describe the asymmetry. Models such as the Exponential GARCH (EGARCH) introduced by Nelson (1990), Quadratic GARCH (QGARCH) of Engle (1990), GJR-GARCH of Gloasten, Jaganathan and Runkle (1993) and Threshold GARCH of Zakoian (1994) are popular asymmetric GARCH class models. Based on the fact that the effect of shocks in the foreign exchange market may endure for a long period, the Integrated GARCH (IGARCH) of Engle and Bollerslev (1986) was innovated with assumptions of unit roots for the GARCH parameters. The volatility of the next period is usually calculated by squaring the shock of the current period in a standard GARCH model. However, for large shocks, the square operation will produce dramatic increases in the variance. Taylor (1986) and Schwert (1989) argued about this problem and suggested employing absolute residuals which can provide a less drastic approach. The nonlinear ARCH (NARCH) of Higgins and Bera(1992) adopted the suggestion of absolute residuals and parameterized the conditional standard deviation with unknown power as a function of lagged conditional standard deviation and lagged absolute residuals with the same power. Baillie, Bollerslev and Mikkelsen (1996) extended the AutoRegressive Fractional Integrated Moving Average (ARFIMA) model type to a GARCH representation as the Fractional Integrated GARCH (FIGARCH) model which is designed to capture the long-run dynamic dependencies in the volatilities. Other modelling techniques have also been developed such as the Stochastic Volatility modelling framework and they will not be introduced here as the GARCH class models are still among the most popular volatility models nowadays.

Since volatility is usually unknown in realized data, the only way to validate the volatility model is to test the model assumptions. There are usually two assumptions in GARCH modelling. One is the distribution assumption (i.e., the assumption that noise is Gaussian) and the other is the assumption that noise is i.i.d. As a matter of fact, the distribution assumption is usually rejected in empirical practice. The Quasi Maximum Likelihood Estimation (QMLE) method has been commonly investigated as in Bollerslev and Wooldridge (1988) to compensate the impact on MLE estimators when the assumption of normality is violated. However, the distribution assumption is still a major issue for parameter estimation of the GARCH class of models.

From the papers on the GARCH class of models, volatility is usually derived from the residuals of the mean model which is commonly referred to as the model fitted from the underlying time series data. Although the nonlinear modelling techniques have been widely developed and applied, in ARCH literatures most of the mean models are still selected as linear. For example, in the pioneering work on ARCH by Engle (1982), the UK consumer price index data was fitted by an AutoRegressive with eXogenous input (ARX) type model. In a later application of the ARCH model, Engle (1983) used an AR type model with multiple inputs and a time trend to model the US deflation indicator. When Bollerslev (1986) proposed the GARCH model, an AR model with 4 lagged terms was used to approach the US implicit GNP deflator. Nelson (1991) introduced the EGARCH model and applied it to estimate the volatility of value weighted CRSP market index which was fitted by an AutoRegressive Moving Average (ARMA) model. Engle (1989) commented on the article of Schwert (1989) and proposed a QGARCH model to estimate the volatility of the stock market round the crash of 1987 with the mean fitted by a variance in mean Moving Average (MA) model. Higgins and Bera (1992) applied a NARCH model to the US/FF weekly exchange rate and the mean model was fitted using an AR (1) model. Gray (1992) used an AR (1) model to model the short-term interests and proposed the RS-GARCH model to calculate the volatility. Ding et al. (1993) investigated the long memory properties of the volatility that existed in the S&P 500 stock market index data and employed a MA (1) model during the application. Tse (1991) fitted models to the Tokyo Stock Exchange using an AR (1) mean model while Akgiray (1989) also used an AR (1) model to approach the mean process of New York Stock Exchange index data. Hamilton and Susmel(1994) proposed a class of Markov-switching ARCH models and modelled the New York Stock Exchange weekly index data using an AR (1) model during simulation. The UK market volatility properties were then studied by McMillan et al. (2000) and the FTA all share index and FTSE 100 index were fitted using an AR type mean model. Linear mean models were also commonly adopted in the application of Multivariate GARCH as in Bollerslev et al. (1988), Engle and Kroner (1995). Some literatures even treated the mean process as a constant model as in Baillie et al. (1996) and Kawakatsu (2006).

Many nonlinear dynamics have been observed in real time market returns as in Abhyankar et al. (1995) and therefore the use of a nonlinear model can improve the mean model forecast ability and provide more accurate residual estimation for the volatility process. Some types of nonlinear models have already been used to fit the mean process in GARCH literature such as in Bollerslev et al. (1993), an exponential AR mean model was used to fit the US stock market volatility and Cao and Tsay (1992) used a threshold AR model. However, the application of a nonlinear mean model is limited and the structure of the nonlinear model is usually specified during estimation. Therefore, this motives the use of nonlinear models with structure determination methods to fit the mean process in order to improve the accuracy of the residuals for the volatility estimation.

Among the nonlinear modelling techniques, the Nonlinear Auto Regressive Moving Average with eXogenous inputs (NARMAX) model proposed by Leontaritis and Billings (1985) can provide a unified formation for a wide class of nonlinear system processes and comparied with the series expansion approaches such as Volterra and Wiener nonlinear models, NARMAX can approach the underlying process with a more concise representation. NARMAX can also provide a more transparent model format than the Radial Basis Function (RBF) neural network and wavelet network approaches. The NARMAX model has successfully modelled many real world nonlinear systems including chaotic electronic circuits, water management systems, turbocharged diesel engines, etc (Billings and Coca, 2001). The pitfall of linear models and the advantage of NARMAX above inspire the application of NARMAX methodology in financial mean process modelling.

The models of many real world systems are usually unknown and determining the structure of the model is the most difficult part during identification. Especially in the nonlinear case, the number of terms may increase dramatically when the redundant variables are falsely selected. Based on the NARMAX model specification, the Orthogonal Forward Regression (OFR) algorithm and Error Reduction Ration (ERR) definition were introduced by Billings et al.(1988, 1989), Korenberg et al.(1988), Chen et al.(1989), Billings and Zhu (1994) to provide an efficient way to determine the most significant terms among the candidate model term set. The structure of the model can be formatted by selecting the terms with ERR above a chosen cutoff value.

The OFR term selection algorithm assumes that the variance of the noise is homoskedastic. However, during financial volatility estimation, the noise of the mean process is usually heteroskedastic. The breach of the assumption may induce bias in the ERR values and therefore, impact the term selections. The falsely selected model terms will cause inaccurate estimation of the modelling noise and the parameters of the volatility model will be affected. The inaccurate parameters of the volatility model will then produce more forecast errors during prediction.

Accordingly, it is essential to find a method to eliminate the impact of heteroskedastic noise on ERR and term selection.

Weighted Least Squares (WLS) can be used to eliminate the impact of heteroskedastic noise. This motives the application of WLS in the OFR algorithm when determining the unknown structure of the financial mean process. However, WLS is applied based on the known weights and according to our knowledge currently there is no solution available to estimate the weights when the unknown system is nonlinear. Although the GARCH model can produce estimation of volatilities and the square roots of the volatility can be treated as weights, the accuracy of the weights will be highly dependent on the mean model structures and the GARCH model is estimated after the mean model. It is impossible to get an accurate parameter estimation of GARCH model before a mean model has been selected. Once the structure of the mean model has been determined, an iterative reweight calculation can be used to give a numerical refinement of the parameters of both mean the model and volatility model.

In system identification, model validation is one of the most important steps. Because the models are driven by assumptions and finite data inputs, it is essential that the assumption and the fitness of the model are tested. In statistics, the Cross Validation (CV) method is commonly used to analyse the prediction performance of a fitted model and during CV the model can be tested using independent data sets. There are several CV methods available in practice (Devijver and Kittler, 1982) including the holdout method, K-fold CV method and Leave-one-out CV method. The holdout method splits the data into two data sets and one set is used to fit the model while the other set is used to test the prediction performance. Due to the simplicity and the serial dependence of financial time series, the holdout method will be applied in our case.

During CV, the assumptions of the models need to be tested after the mean model and volatility model have been fitted. As the LS estimator can only provide unbiased estimation when the noise is white, the mean model assumptions can be tested by taking autocorrelation of the modelling residuals. For the GARCH class of volatility model, the distribution assumption can be tested using the Jarque and Bera (JB) test (Jarque and Bera, 1980), QQ plot etc. and the i.i.d. assumption can be tested using the Brock, Dechert and Scheinkman (BDS) test (Brock et al., 1987). The BDS test uses a nonparametric technique to test against a wide class of data departing from the i.i.d. requirement and it has been proved to be successful in detecting nonlinearity in economics as in Brock et al. (1991). The first example of using the BDS to test against the GARCH assumption was in Brock et al (1991) where the distribution of the BDS test from the standard residuals was obtained by Monte Carlo simulations. Bollerslev et al. (1993) concluded that the BDS test has the power to test the i.i.d. assumption for ARCH when the volatility model or mean model is miss-specified. The BDS test has been applied in most recent empirical practice as Caporale et al. (2004) tested the adequacy of GARCH specifications using the BDS test and Mangani (2009) used the BDS test to verify the significance of the GARCH model when fitting market data from South Africa. Therefore, either an inaccurate mean model or inaccurate volatility model will lead to failure of the BDS test on standard mean model residuals. This raises the motivation to validate both the mean and the volatility models simultaneously using a CV method. If both models are accurate, the BDS test on standard one-step-ahead prediction errors calculated using the second data set during CV should not be rejected.

Longevity risk now plays a key role for the institutes that provide pensions. The mortality rate which is measured as the death rate in a population is the prime element in longevity risk. If the mortality rate in pricing annuities is overestimated, the profit margin of pension providers will shrink significantly. Many techniques have been developed to model the mortality rate such as the nonparametric Lee-Carter model (Lee and Cater, 1992), Age-Period-Cohort (APC) model of Tabeau et al. (2001), and the Cairns-Blake-Dowd (CBD) model of Cairns et al. (2006). However, currently there is no existing model which is entirely satisfactory. The NARMAX modelling method can be used to fit the mortality rate surface and give a reasonable prediction. In mortality rate literature, the back testing method is employed to test the forecast performance of existed mortality rate models as in Dowd et al. (2008) and it is necessary to compare our new models with popular mortality models using this method.

1.2 Objectives

The main objectives of this thesis are to investigate financial volatility models and to develop a systematic nonlinear mean model identification method using financial return data. This includes developing mean model term selection algorithms under heteroskedastic noise conditions, validating simultaneously the mean and volatility models and comparing the volatility prediction performance of the nonlinear mean model with commonly used linear mean model in the GARCH literature.

The GARCH class of volatility models has developed very fast since the innovation of the ARCH model by Engle (1982) and there are a hundred or more GARCH class of volatility models which exist currently (Bollerslev, 2008). However, since the volatility model is developed to mimic the observed volatility features, the fundamental of the volatility model concept are very similar. The GARCH class of models basically are an extension of the ARCH and GARCH models. Therefore, the objective is to summarize a review of the major GARCH class of models.

However, most ARCH literature treats the mean process as linear and the MLE method is used during model parameter estimation. As a matter of fact, nonlinearity has been observed in most financial return processes and this suggests that a nonlinear model is more appropriate for forecasting and accurate descriptions of the financial returns and volatility. The MLE method is highly dependent on the assumption of the distribution and numerical search methods are usually non-trivial. There are several nonlinear modelling methods available and the NARMAX polynomial model can be a very good candidate as the NARMAX model can approximate a very wide class of nonlinearities. The NARMAX term selection and parameter estimation algorithm which is known as OFR algorithm is independent of the distribution assumption. However, the OFR algorithm is based on the assumption of constant variance and the question addressed in this thesis is to investigate the impact on term selection when the noise is heteroskedastic and the objective is to derive a new method to compensate this impact.

As far as we know, the GARCH literature barely investigates the accuracy of the mean model and validates both the mean model and volatility model simultaneously. As the GARCH class of models are proposed based on the assumption of i.i.d. distributed standard mean model residuals and the accuracy of the mean model residuals are directly impacted by the accuracy of the mean model, the i.i.d. assumption will not be rejected only if both mean and variance models are accurate enough to approximate the process. It is essential to develop such validation procedures to ensure the prediction performance of the selected nonlinear mean model. This is another key achievement in this thesis.

The morality rate is a key factor in hedging longevity risk among the pension issuers. Without considering external impacts, it has been observed that the mortality rate is mainly related with the age and the birth year of the underlying population. Therefore, the mortality rate surface can be treated as a projection of the age and birth years. Since there is no existing model which is entirely satisfactory, in this thesis a new NARMAX modelling method is developed to fit the mortality rate and to predict the future mortality rate. The fitness of the selected model is then checked using back-testing methods. In order to demonstrate the prediction performance of the new models, comparisons with existing mortality predictions are given.

1.3 Layout of this thesis

This thesis is organized into seven chapters. Chapter 2 reviews the major models in GARCH class of volatility models. Chapter 3 briefly reviews the mean models of the major ARCH literature and the NARMAX modelling method. Chapter 4 investigates the impact of heteroskedastic noise on the OFR algorithm. A new algorithm is derived as a solution to correct for the impact and to refine the parameter estimation of both the mean model and the volatility model. Chapter 5 deals with the validation of both mean model and the variance model validations. Chapter 6 is a new development of the NARMAX modelling method to the mortality rate. Chapter 7 gives the conclusions of this thesis.

Chapter 2 begins with a fundamental introduction to the volatility concept. Different volatility forecast models are investigated and several major GARCH class of models are reviewed. An alternative GARCH class model is proposed to give smooth parameter estimation during MLE. The details of parameter estimation methods for the GARCH class of models are also given. The commonly used forecast evaluation methods in the GARCH literature are also reviewed.

Chapter 3 investigates the mean models used in the major GARCH class of models and gives an introduction to the NARMAX modelling methodology. Examples are given to demonstrate the volatility forecast performance comparison between a linear mean model and selected nonlinear mean models when the mean process is nonlinear.

In Chapter 4, firstly the OFR algorithm is introduced based on the NARMAX polynomial model. Next, the impact of heteroskedastic noise on term selection using the OFR algorithm is investigated and a new weighted OFR algorithm is proposed to correct for this impact. An iterative reweighted procedure is then introduced to refine the parameter estimation of both the mean and volatility models. Examples are given to demonstrate the new term selection problem of the OFR algorithm under heteroskedastic noise and to illustrate the application of the new algorithm.

In Chapter 5, the CV method is introduced and commonly used distribution assumption testing methods are given. Next, the impact of the mean model term selection on the ML estimation of the volatility model is analyzed theoretically. A new method to validate simultaneously the mean and volatility models is proposed and examples are given to illustrate the application of the new validation methods.

In Chapter 6, firstly the definition of the mortality rate is given and commonly used mortality rate models are reviewed. The NARMAX modelling technique is then applied to derive a polynomial mortality rate model using realized death and exposures data of England & Wales. A long term forecast comparison is given between the derived mortality model and the CBD mortality rate model. A back-testing analysis is then carried to assess the models' ex post forecasting performance.

The main contributions of this thesis and some suggestions for further research are given in Chapter 8.

Chapter 2: Introduction to financial volatility modelling

2.1 Introduction

In financial systems, volatility is a measure of the dispersion in a probability density function and often refers to the variance or standard deviation of a return series. Volatility is one of the most important variables for evaluating the financial uncertainty and it is often a key input to many investment decisions and the creations of portfolio. Significant features of volatility have been found in financial time series including persistence 'large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes' (Mandelbrot, 1963), leverage effects 'volatility response to a large negative return is often far greater than it is to a large positive return of the same magnitude' (Black, 1976) and reversion to mean 'prices and returns eventually move back towards the mean or average'. Apart from all the features above, volatility cannot be observed directly from the data. Hence, volatility models were proposed to capture these features and models based on historical volatilities were commonly investigated. Among historical volatility models, the Exponential Weighted Moving Average (EWMA) model was commonly used to give volatility predictions. More sophisticated groups of volatility models are the AutoRegressive Conditional Heteroscedasticity (ARCH) family models. The first example of the GARCH class of models was the AutoRegressive Conditional Heteroscedasticity (ARCH) model introduced by Engle (1982) where conditional variance was a function of q past squared residuals. The ARCH process was extended to the more general case of the GARCH process with past conditional volatilities by Bollerslev (1986). Motivated by experimental results in foreign exchange markets that the sum of GARCH parameters were close to one, GARCH was extended to the Intergrated GARCH (IGARCH) (Engle and Bollerslev, 1986). As a matter of fact both ARCH and GARCH models were symmetric in the form of the squared residuals and leverage effects could not be included in these models. Nelson (1991) argued for an asymmetric form according to the finding of Black (1976) and modified the conditional volatility to the Exponential GARCH

(EGARCH) model. The conditional volatility which was specified in logarithmic form in the EGARCH model guaranteed that there was no need to impose estimation constraints which ensured the non-negativity of the conditional variance. Following Engle (1982) it was suggested that conditional variance model could be written in the form of absolute residuals, Taylor (1986) and Schwert (1989) employed the absolute residuals in the conditional standard deviation model. Based on the discussion of Schwert (1990), Engle (1989) suggested using square absolute residuals and derived the Quadratic GARCH (QGARCH) model in order to capture the leverage effect. The Nonlinear ARCH (NARCH) proposed by Higgins and Bera (1992) nested the ARCH model into a nonlinear form as setting the order of every ARCH model term to be a fraction. The GJR-GARCH model proposed by Glosten, Jaganathan and Runkle (1993) added an indicator variable to the GARCH model in order to capture the leverage effect. Based on the fact that squared and absolute returns of financial assets usually have serial correlations that are slow to decay, the Fractionally Integrated GARCH (FIGARCH) model was proposed by Baillie, Bollerslev and Mikkelsen (1996) to reduce the impact of a shock on future volatility over an infinite horizon. More models were proposed recently and many of them have flexible specifications which can include several other models as special cases and hence will not be introduced here.

Since the pioneering work of Engle (1982), the assumption of conditional normality has been commonly used in theoretical and empirical research. Based on this assumption, the Maximum Likelihood Estimation (MLE) method was the standard method used to estimate parameters. Weiss (1986) gave the first study of the asymptotic properties of the ARCH MLE and indicated that MLE is consistent and asymptotically normal with the condition of the finite fourth order moments of the unnormalized data. However, evidence of heavy tails-leptokurtosis suggests that the common assumption of conditional normality is often rejected empirically. The Quasi Maximum Likelihood Estimation (QMLE) method which was then commonly investigated and Bollerslev and Wooldridge (1988) showed that QMLE can still give consistent estimation under assumptions of asymptotic normality of score matrix and uniform weak convergence of likelihood and its second derivative. However, Engle and Gonzalez-Rivera (1991) investigated the loss of efficiency of QMLE when the distribution is falsely assumed to be normal and proposed a nonparametric method to estimate the conditional distribution. Hence, the distribution assumption is still a major issue for parameter estimation of the GARCH class of models.

Beside parameter estimation, the forecasting power of the GARCH class of models has also been studied and forecasting performance comparison of the competing models becomes to one of the major direction of any forecasting research. The squared return is usually used as the proxy to the volatility forecast evaluation and popular evaluation measures include Mean Error (ME), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percent Error (MAPE). However, Andersen and Bollerslev (1997) discussed that squared returns can be a noisy estimator of the actual variance dynamics and Lopze (2001) proved this by the theoretical evidence. Andersen, Bollerslev, Diebold and Labys (2001) proposed the alternative proxy which is the realized volatility calculated by high frequency data. Another issue arises as the distribution of the return is usually skewed because the negative news commonly causes bigger impact than positive news. Accordingly, heteroskedasticity-adjusted MSE (HMSE) was proposed to penalize the asymmetrical volatility forecasts by Bollerslev and Ghysels (1996). With these measurement methods, the GARCH class of models forecast contests exploded in the past twenty years. Most recent developments of volatility forecasting trends to use long memory volatility models which are included in the literatures of Andersen et al. (2002), Vilasuso (2002) and Zumbach (2002) and forecast evaluation based on those models all used intra-day high frequency data to calculate the proxy.

The purpose of this chapter is to review the definition of financial volatility, volatility models, parameter estimation methods and forecast evaluation of these models. Commonly used models and evaluation methods will be given and the remaining sections of this chapter are organized as follows. Section 2.2 provides some definitions. Section 2.3 gives the historical reviews of the volatility models. Section 2.4 focuses on the commonly used GARCH class volatility models and Section 2.5 gives the parameter estimation methods for the GARCH class models. Section 2.6 lists most of the forecast evaluation methods and describes different volatility proxies. Section 2.7 gives the conclusion.

2.2 Definition of financial volatility

Financial volatility is interpreted as uncertainty and usually refers to the degree of unpredictable change over time of a certain variable. Volatility is not observable and it can be measured as standard deviation σ or variance σ^2 of the continuously returns of a financial market with specific time horizon. The return data series are usually converted from a price series using

$$R(t)=\log\frac{P(t+1)}{P(t)}=\log P(t+1)-\log P(t) \qquad (2.1)$$

where the price observation made at sample time t and t+1 are denoted as $P(t)$ and $P(t+1)$ and $R(t)$ is the return.

There are several kinds of financial volatilities which exist in the current literature: (1) unconditional volatility, (2) implied volatility, (3) realized volatility and (4) conditional volatility. Unconditional volatility refers to the constant finite variance σ throughout the whole data generation process and is defined by the assumption of stationary stochastic process. The statistic equation of the unconditional volatility is

$$\sigma=\sqrt{\sum_{t=1}^{T}\left(R(t)-\bar{R}\right)^2/(T-1)} \qquad (2.2)$$

where T is the sample return length and $\bar{R}$ is the sample average return as $\bar{R}=\sum R_t/T$.

The implied volatility is a value calculated from an option price and it is usually associated with the Black and Scholes option model. The realized volatility which is also called historical volatility is the standard deviation of a set of previous returns.

The statistical equation of the realized volatility is similar as (2.2) the length of the set can be varied as 15 days, 30 days etc. The conditional volatility is the standard deviation of a future return that is conditional on known information set such as the history of previous returns. Unlike realized volatility, the conditional volatility is calculated by a proper selected and estimated time series model using appropriate data and it is usually time-varying. According to the fact that most market volatility changes have the characters as mentioned in the introduction, the variance of the return process is usually time-varying. Hence, the volatility mentioned in this thesis is referred to as conditional volatility. If the return is rewritten as $y(t)$, the conditional variance will be

$$\sigma^2(t) = Var_{t-1}(y(t)) \tag{2.3}$$

where Var_{t-1} denotes the variance conditioned on past observations. Since volatility is time varying and not observable, many discrete-time models have been proposed to model the volatility by inferring volatility from either absolute or squared returns.

2.3 Time series volatility forecasting models

Volatility modelling has attracted much attention in recent years, largely motivated by its importance in financial markets. Reliable volatility estimates and forecasts are essential due to the increasing needs in hedging against risk and portfolio management. Different types of volatility models have been developed as moving average models and GARCH class of models to account for different market facts.

2.3.1 Moving average model

Moving average models has been commonly used in financial data modelling because the models are the easiest to manipulate and construct. These kind of models are usually directly built up on historical volatilities and shown good forecasting performance in some literature like Figlewski (1997), Andersen, Bollerslev, Diebold and Labys (2003).

2.3.1.1 Historical average model

Among moving average models, the simplest is the historical volatility model or equally weighted moving average model. The historical average model is calculated on a fixed size data window which is rolling through time samples and oldest return will be substituted by new return at every new sample point. The equally weighted average of squared daily returns are usually used by the historical model and the n-day historical volatility is calculated by

$$\hat{\sigma}^2(t)=\sum_{i=1}^{n} y^2(t-i)/n \tag{2.4}$$

where n is a user chosen length and $\hat{\sigma}^2(t)$ is the volatility estimation for the time t. The reason for using a squared return rather than the square mean deviation as $(y(t)-\bar{y})^2$ is because during empirical research on the accuracy of volatility forecasts, the use of the squared return has demonstrated little disadvantage (Alexander, 2001) and also the mean of the return is usually assumed to be zero. The n-day historical volatility is commonly used to measure the portfolio risk in practice. However, the major problem of the historical average model is that the model only includes the extreme events as important to current estimation without considering when they occurred. Hence, even just one unusual return will cause affection to the n-day historical volatility the same as the extent of n following days after that event. Short-term historical model are supposed to solve the problem above and capture more 'clustering' volatilities, but equal weighting cannot account the dynamic of return properly. Accordingly, the exponentially smoothing average model was proposed.

2.3.1.2 Exponentially smoothing average volatility model

In contrast to equal weighting, the exponentially weighting method is another popular approach to volatility forecasting. It is more robust and accurate in forecasting volatility in the short term (Gardner, 1985) and can pass the shock from an extreme event as an exponential decay to the current volatility.

The exponentially smoothing model usually puts more weights on the most recent observations and hence it is also called Exponential Weighted Moving Average (EWMA) model. The EWMA can capture some dynamic ordering of returns and the EWMA volatility estimation of market events reacts over time with a strength that is determined by smoothing constant λ which is a number between 0 and 1. The larger the value of λ, the more weight is put on past observations and the volatility series becomes smoother. The general EWMA volatility formula can be written as

$$\hat{\sigma}^2(t) = (1-\lambda)\sum_{i=1}^{\infty}\lambda^{i-1}y^2(t-i) \tag{2.5}$$

and the formula can also be written as a recursion format as

$$\hat{\sigma}^2(t) = (1-\lambda)r^2(t-1) + \lambda\hat{\sigma}^2(t-1) \tag{2.6}$$

The term of $(1-\lambda)r^2(t-1)$ determines the degree of reaction of volatility to market events and the smaller the value of λ is, the more yesterday's return will react to market information in volatilities. The other term of $\lambda\hat{\sigma}^2(t-1)$ determines the persistence in volatility. Since λ is between 0 and 1, the effect of a single event will be reduced after some time horizon. A higher λ will give more persistence in volatility to actual market events and a lower λ gives higher reaction on volatility but which will fade away quickly. The main restriction of EWMA model is that the summation of persistence parameter and the reaction parameter are one which means the model should either have volatility persistence or have high reactions. Based on this limitation, EWMA model is usually used in the foreign exchange market. (Alexander, 2001)

2.3.2 ARCH model

In the moving average model, the returns process has been assumed to be independent and identically distributed as the returns are used directly to calculate the volatility and there is no time-varying volatility assumption. Meanwhile, only the current volatility is taken as the prediction. All those features of moving average greatly limit the volatility prediction and the assumptions are hard to satisfy in real market data. The returns in many financial markets are usually not well modelled by an independent and identically distribution process and they may show autocorrelation in some high frequency data. Especially, the squared returns often show autocorrelation which is an indication of volatility clustering (Mandelbrot, 1963). Engle (1982) first described a framework to model the time varying volatility and introduced the ARCH model. The ARCH process is a zero mean, serially uncorrelated process with time varying volatilities conditional on the past and it assumes that conditional volatility of today is a weighted average of past squared unexpected innovations. The ARCH model assumes that the innovation $\varepsilon(t)$ can be formatted by a multiplication of an independent and identically distributed random variable $z(t)$ and a time varying standard deviation as

$$\varepsilon(t)=z(t)\sigma(t) \tag{2.7}$$

The variable $z(t)$ has zero mean and identical variance of one. Engle also assumed that the innovation process follows a normal distribution as

$$\varepsilon(t)\big|\psi_{t-1} \sim N\left(0,\sigma^2(t)\right) \tag{2.8}$$

where ψ_{t-1} denotes all the variable information of past returns $\left(y(t-1),...,y(t-n)\right)$ up through time t-1. If the conditional variance $\sigma^2(t)$ is rewritten as $h(t)$, the ARCH model can be expressed as

$$h(t)=a_0+a_1\varepsilon^2(t-1)+a_2\varepsilon^2(t-2)+\cdots+a_p\varepsilon^2(t-p) \tag{2.9}$$

where $a_0,a_1,a_2,...,a_p$ are unknown parameters which satisfy the conditions as $a_0>0,a_1,...,a_p\geq 0$ and $\sum_{i=1}^{p}a_i<1$, and $\varepsilon^2(t-1),\varepsilon^2(t-2),...,\varepsilon^2(t-p)$ are past innovations derived from the return data. If an innovation variable $v(t)$ is defined as

$$v(t)=\varepsilon^2(t)-h(t) \tag{2.10}$$

, then the ARCH model in equation (2.9) can be rewritten as

$$\varepsilon^2(t)=a_0+\sum_{1}^{p}a_i\varepsilon^2(t-i)+v(t) \tag{2.11}$$

As $E_{t-1}(v(t)) = E_{t-1}(\varepsilon^2(t) - h(t)) = E_{t-1}(\varepsilon^2(t)) - E_{t-1}(h(t))$ and $E_{t-1}(\varepsilon^2(t)) = h(t)$, $E_{t-1}(v(t)) = 0$. Hence, the ARCH model corresponds directly to an AR (q) model with squared innovations. From the stationary condition requirement of the AR model, the sum of the parameters should be less than one and this coincides with the ARCH parameter conditions above.

Before using the ARCH model, it is needed to test the ARCH effect first. The Lagrange multiplier test for ARCH was originally proposed by Engle (1982) and it is simply a regression on the innovation $\varepsilon(t)$ by

$$L = Tf^{0\prime}z(z'z)^{-1}z'f^0 / f^{0\prime}f^0 \tag{2.12}$$

where L is the Lagrange multiplier statistic, T is the sample length, f^0 is a column vector of $\left(\frac{\varepsilon^2(t)}{h^0} - 1\right)$, h^0 is the hypothesis of $h^0 = a_0$, $z(t) = (1, \varepsilon^2(t-1), \ldots, \varepsilon^2(t-p))$ and $z = [z(1), \ldots, z(T)]$. Under the null hypothesis, the statistic will be asymptotically distributed as chi square with p degree of freedom. The intuition behind the test is that if the data are homoskedastic, the variance cannot be predicted and variations in $\varepsilon^2(t)$ will be purely random. Alternatively, if ARCH exists, large values of$\varepsilon^2(t)$will be predicted by large values of the past squared innovations.

In empirical applications of the ARCH model, a long lag length p and a large number of parameters was often used. Accordingly, it becomes more difficult to estimate the parameters because the likelihood function often becomes very flat and non-negative conditionals are usually violated. For example, Lilien and Robins (1987) used a linearly declining structure on the parameters to prevent some of them from being negative. Consequently, a more general case of ARCH model which is called GARCH has been proposed and a bibliography of research papers was published to introduce new models based on GARCH.

2.4 The ARCH Class of volatility models

2.4.1 GARCH model

The GARCH model introduced by Bollerslev (1986) is a more general case of the ARCH model. In order to take account of the typical long memory effect of the volatility shock, the parameter estimations of the linear declining lag structure of ARCH usually violate the conditions. The GARCH model uses a more parsimonious representation to allow the more flexible lag structure than that of ARCH and the GARCH specification provides that the best volatility prediction in the next period is a weighted average of the long-run volatility, the volatility prediction of this period and the newest information in this period which is captured by the most recent squared residuals. The general GARCH (p, q) model is written as

$$h(t) = a_0 + \sum_{i=1}^{q} a_i \varepsilon^2 (t-i) + \sum_{i=1}^{p} \beta_i h(t-i) \qquad (2.13)$$

$$p \geq 0, \ q > 0$$

$$a_0 > 0, a_i \geq 0, \ i = 1,\ldots,q$$

$$\beta_i \geq 0, \ i = 1,\ldots,p$$

$$\sum_{i=1}^{q} a_i + \sum_{i=1}^{q} \beta_i \leq 1$$

where $h(t)$ is conditional variance at sample t, $\varepsilon(t)$ is the innovation from return process at sample t, $a_i, i = 0,\ldots,q$ and $\beta_i, i = 0,\ldots,p$ are unknown parameters. Therefore, the short-run dynamics of the volatility process are determined by the sizes of the parameters a and β. Large β shows that the shocks to conditional volatility take a long time to die out and large a indicates that the volatility reacts intensely to market movements.

The extension from ARCH to GARCH is similar to the extension of a time series AutoRegressive (AR) process to the AutoRegressive Moving Average (ARMA) process. It can be shown that an ARCH (∞) model can be represented as a GARCH (1, 1), viz.

$$\begin{aligned} h(t) &= a_0 + a_1\varepsilon^2(t-1) + \beta_1 h(t-1) \\ &= a_0 + a_1\varepsilon^2(t-1) + \beta_1\left(a_0 + a_1\varepsilon^2(t-2) + \beta_1 h(t-2)\right) \\ &= a_0 + \beta_1 a_0 + a_1\varepsilon^2(t-1) + a_1\beta_1\varepsilon^2(t-2) + \beta_1^2 h(t-2) \\ &= a_0 + \beta_1 a_0 + \cdots + \beta_1^{n-1}a_0 + a_1\varepsilon^2(t-1) + a_1\beta_1\varepsilon^2(t-2) + \cdots + a_1\beta_1^{n-1}\varepsilon_{t-n}^2 \\ &= A + B_1\varepsilon^2(t-1) + B_2\varepsilon^2(t-2) + \cdots + B_n\varepsilon^2(t-n) \end{aligned} \tag{2.14}$$

where A denotes $a_0 + \beta_1 a_0 + \cdots + \beta_1^{n-1}a_0$ and B_n denotes $a_1\beta_1^{n-1}$ for $n = 1,2,\ldots,\infty$. If an innovation variable is assumed to be $v(t) = \varepsilon^2(t) - h(t)$, the GARCH (p, q) model can become to an ARMA model consisting only with $\varepsilon^2(t)$ and $v(t)$ as

$$\begin{aligned} h(t) &= a_0 + \sum_{i=1}^{q} a_i\varepsilon^2(t-i) + \sum_{i=1}^{p}\beta_i h(t-i) \\ \varepsilon^2(t) - v(t) &= a_0 + \sum_{i=1}^{q} a_i\varepsilon^2(t-i) + \sum_{i=1}^{p}\beta_i\left[\varepsilon^2(t-i) - v(t-i)\right] \\ \varepsilon^2(t) &= a_0 + \sum_{i=1}^{q} a_i\varepsilon^2(t-i) + \sum_{i=1}^{p}\beta_i\varepsilon^2(t-i) + v(t) - \sum_{i=1}^{p}\beta_i v(t-i) \end{aligned} \tag{2.15}$$

Although the GARCH model is directly set up for one-step-ahead forecast, the long term prediction of GARCH (1, 1) can be also constructed according to the assumption in equation (2.7) by

$$\begin{aligned} h(t) &= a_0 + a_1\varepsilon^2(t-1) + \beta_1 h(t-1) \\ &= a_0 + a_1 z^2(t-1)h(t-1) + \beta_1 h(t-1) \\ &= a_0 + \left(a_1 z^2(t-1) + \beta_1\right)\left(a_0 + a_1 z^2(t-2)h(t-2) + \beta_1 h(t-2)\right) \\ &= a_0 + \left(a_1 z^2(t-1) + \beta_1\right)\left[a_0 + \left(a_1 z^2(t-2) + \beta_1\right)h(t-2)\right] \\ &\;\vdots \\ &= a_0 + \cdots + a_0\left(a_1 z^2(t-1) + \beta_1\right)\cdots\left(a_1 z^2(t-n) + \beta_1\right) + \left(a_1 z^2(t-1) + \beta_1\right)\cdots\left(a_1 z^2(t-n) + \beta_1\right)h_0 \end{aligned} \tag{2.16}$$

Because $E\left(z^2(t)\right) = 1$, after taking expectation on both side equation (2.15) becomes

$$E\left(h(t)\right) = a_0 + a_0\left(a_1 + \beta_1\right) + \cdots + a_1\left(a_1 + \beta_1\right)^{n-1} + \left(a_1 + \beta_1\right)^{n-1}E\left(h_0\right) \tag{2.17}$$

According to the condition of $a_1 + \beta_1 < 1$, equation (2.16) can be rewritten as

$$E\left(h(t)\right) = \frac{a_0}{\left(1 - a_1 - \beta_1\right)} \tag{2.18}$$

With a similar extension method, the long term prediction of GARCH (p, q) can be written as

$$\frac{a_0}{1-\sum_{i=1}^{q} a_i - \sum_{i=1}^{p} \beta_i} \tag{2.19}$$

Most financial markets have GARCH volatility forecasts that 'mean-revert' as the volatility forecast converges to the long term prediction as in equation (2.19) and the forecast of the GARCH model is stationary. However, in currencies and commodities market, the shock to the volatility trends to have an infinite persistence during forecasting (Engle and Bollerslev, 1986). Hence, the stationary GARCH model can not apply in this case.

2.4.2 IGARCH model

Engle and Bollerslev (1986) introduced a model which captures non-mean-revert effect and the volatility is integrated by the definition of $\sum_{i=1}^{q} a_i + \sum_{i=1}^{p} \beta_i = 1$ in the GARCH model. Relative to the simple GARCH model, the new model is called the Integrated GARCH model and the simple IGARCH (1, 1) can be written as

$$h(t) = a_0 + a_1 \varepsilon^2 (t-1) + (1-a_1) h(t-1) \tag{2.20}$$

As the long term prediction of the GARCH model is listed in equation (2.19), when $\sum_{i=1}^{q} a_i + \sum_{i=1}^{p} \beta_i = 1$ is applied the long term prediction of IGARCH model is infinity. Therefore, the unconditional volatility does not exist. For illustration, based on model (2.20) the expectation of the one-step-ahead unconditional variance is

$$\begin{aligned} E(h(t)) &= a_0 + E\left(a_1 \varepsilon^2 (t-1)\right) + E\left((1-a_1) h(t-1)\right) \\ &= a_0 + E\left(a_1 z^2 (t-1) h(t-1)\right) + (1-a_1) E\left(h(t-1)\right) \\ &= a_0 + a_1 E\left(h(t-1)\right) + (1-a_1) E\left(h(t-1)\right) \\ &= a_0 + E\left(h(t-1)\right) \end{aligned} \tag{2.21}$$

Iteratively substituting the $E(h(t))$ by its previous estimation for n steps, equation (2.21) becomes to

$$E(h(t)) = na_0 + E(h(t-n)) \tag{2.22}$$

It is obvious that the unconditional volatility is integrated in equation (2.22) and when n trends to infinity, the unconditional volatility trends to be infinity. The IGARCH model can become to EWMA model when the constant parameter a_0 is zero. Apart from the volatility clustering, the leverage effect is also found to exist in most of the market data. However, the GARCH and IGARCH model are symmetric models which mean that the impact to the volatility of positive and negative returns is the same.

2.4.3 EGARCH model

Although, GARCH models have been applied with much success to modelling of financial returns, the simple structure imposes important limitations. The symmetric assumption has been questioned empirically and therefore, Nelson (1991) argues for a model in which the conditional variance responds asymmetrically to positive and negative innovations. Black (1976) found evidence that volatility trends to rise in response to bad news and to fall in response to good news, the conditional distribution of the innovations is therefore usually left skewed. Nelson also argued that the nonnegative constraints of parameters can create difficulties in estimating GARCH models. Accordingly, Nelson adopt a similar process for ensuring the conditional volatility remains nonnegative by making $\ln(h(t))$ linear in some function of time and lagged $z(t)$ in some suitable function g as

$$\ln(h(t)) = a(t) + \sum_{k=1}^{\infty} \beta_k g(z(t-k)) \quad \beta_1 = 1 \tag{2.23}$$

where $\{a_t\}_{t=-\infty,\infty}$ and $\{\beta_k\}_{k=1,\infty}$ are real, non-stochastic, scalar sequences and $g(z(t)) = \theta z(t) + \gamma\left[|z(t)| - E|z(t)|\right]$. Because $E(z(t)) = 0$ and $E(|z(t)| - E|z(t)|) = E|z(t)| - E|z(t)| = 0$, $g(z(t))$ is a zero mean and i.i.d. random sequence. Over the range $0 < z(t) < \infty$, $g(z(t))$ is linear in $z(t)$ with slope $\theta + \gamma$ and over the range $-\infty < z(t) \le 0$, $g(z(t))$ is linear with slope $\theta - \gamma$. Therefore, the structure of $g\left(z(t)\right)$ allows the conditional variance process to respond asymmetrically to rises and falls in the stock price. Nelson then introduced an ARMA process to approach the infinite parameters of β_k as

$$\ln\left(h(t)\right)=a_t+\frac{1+\psi_1 L+\cdots+\psi_q L^q}{1-\Delta_1 L-\cdots-\Delta_p L^p}g\left(z(t-1)\right) \tag{2.24}$$

where L is lag operator, ψ,Δ are parameters and the terms $\left[1-\sum_{i=1}^{p}\Delta_i L^i\right]$ and $\left[1+\sum_{i=1}^{q}\psi_i L^i\right]$ are assumed to have no common roots. Therefore, the general EGARCH model is written as

$$\ln\left(h(t)\right)=a_0+\sum_{i=1}^{p}\beta_i\ln\left(h_{t-i}\right)+\frac{1+\psi_1 L+\cdots+\psi_q L^q}{1-\Delta_1 L-\cdots-\Delta_p L^p}\left(\theta z(t-1)+\gamma\left[\left|z(t-1)\right|-E\left|z(t-1)\right|\right]\right) \tag{2.25}$$

where $a_0,\beta_i,\psi,\Delta,\theta,\gamma$ are unknown parameters. In practical the terms $\frac{1+\psi_1 L+\cdots+\psi_q L^q}{1-\Delta_1 L-\cdots-\Delta_p L^p}$ are usually cut to finite terms and the practical EGARCH model can be written as

$$\ln\left(h(t)\right)=a_0+\sum_{i=1}^{p}\beta_i\ln\left(h_{t-i}\right)+\sum_{j=1}^{q}a_j\left[\frac{\left|\varepsilon(t-j)\right|}{\sqrt{h(t-1)}}-E\left\{\frac{\left|\varepsilon(t-j)\right|}{\sqrt{h(t-1)}}\right\}\right]+\sum_{j=1}^{q}c_j\left(\frac{\left|\varepsilon(t-j)\right|}{\sqrt{h(t-1)}}\right) \tag{2.26}$$

Beside the ability of modelling the leverage effect, the other advantage of EGARCH from an implementation perspective is that the estimation of the parameters does not require that the parameters satisfy any inequality constraints. The log operator ensures that the conditional volatility is positive all the time. However, lack of analytic form for the volatility term structure limits the application in forecasting volatility of EGARCH model.

2.4.4 QGARCH model

Schwert (1990) studied the stock market crash of October 19, 1987 and tried to use 22 lagged terms to model the mean and the absolute standard deviation. Based on Schwert's discussion, Engle (1989) introduced the historical absolute innovation term with unknown power and a negative innovation term to the GARCH model as

$$h(t) = a_0 + a_1 \left|\varepsilon(t-1)\right|^b - \lambda\varepsilon(t-1) + \beta_1 h(t-1) \tag{2.27}$$

where $a_0, a_1, \lambda, \beta_1$ are unknown parameters and b is unknown power. According to the simulation results, Engle found that the parameter b is close to 2. By extending the Binomial theorem (Poul, 1955) to equation (2.27), Engle finally proposed a Quadratic GARCH (1, 1) with leverage ratio to compensate the impact of the negative returns as

$$h(t) = a_0 + a_1\left(\varepsilon(t-1) - \gamma\right)^2 + \beta_1 h(t-1) \tag{2.28}$$

where γ is the leverage ratio and a_0, a_1, β_1 are parameters. The squared term in equation (2.28) ensures the positivity of the conditional variance and the positive γ ensures the QGARCH model matches the leverage effect of negative returns. Sentana (1995) discussed that the QGARCH is actually the Taylor series expansion of the conditional volatility and gave a general QGARCH (p, q) model as

$$h(t) = a_0 + \psi' X_{t-1,q} + X'_{t-1,q} A X_{t-1,q} + \sum_{i=1}^{p} \beta_i h(t-i) \tag{2.29}$$

where a_0 is constant parameter, ψ' is a vector of parameters of linear lagged innovations, A is a matrix of parameters of quadratic terms, $X_{t-1,q}$ is a column vector with lagged innovations from $\varepsilon(t-1)$ to $\varepsilon(t-q)$. QGARCH is proposed to contain the leverage effect in modelling the volatilities; however, there was no theoretical explanation to verify the use of a quadratic Taylor expansion as an approach.

2.4.5 NARCH model

Engle (1982) suggested two alternative volatility models-the exponential value model and absolute value model:

$$h(t)=\exp\left(a_0+a_1\varepsilon^2(t-1)+\cdots+a_p\varepsilon^2(t-p)\right) \tag{2.30}$$

$$h(t)=a_0+a_1\left|\varepsilon(t-1)\right|+\cdots+a_p\left|\varepsilon(t-p)\right| \tag{2.31}$$

Then in an empirical application, Engle and Bollerslev (1986) reported one of the best models in modelling U.S.dollar/Swiss franc exchange rate among all competing models was

$$h(t)=a_0+a_1\left|\varepsilon(t-1)\right|^u+\cdots+a_p\left|\varepsilon(t-p)\right|^u \tag{2.32}$$

Geweke (1986) and Pantula (1986) suggested a logarithm ARCH model to avoid the non-negativity restrictions as

$$\log\left(h(t-1)\right)=a_0+a_1\log\left(\varepsilon^2(t-1)\right)+\cdots+a_p\log\left(\varepsilon^2(t-p)\right) \tag{2.33}$$

Since each of the above models has individual limitations and depends upon the particular empirical application, Higgins and Bera (1992) proposed the NARCH model which is the first GARCH class model encompassing some other models as

$$h(t)=\left[a_0\left(\sigma^2\right)^\delta+a_1\left(\varepsilon^2(t-1)\right)^\delta+\cdots+a_p\left(\varepsilon^2(t-p)\right)^\delta\right]^{1/\delta} \tag{2.34}$$

where σ^2 is the unconditional variance of the innovation $\varepsilon(t)$, $a_0,a_1,\cdots,a_p$ are unknown parameters and δ is the unknown power parameter. When $\delta=1$, NARCH becomes the standard ARCH model. Otherwise, the equation can be rewritten as

$$h^\delta(t)=a_0\left(\sigma^2\right)^\delta+a_1\left(\varepsilon^2(t-1)\right)^\delta+\cdots+a_p\left(\varepsilon^2(t-p)\right)^\delta \tag{2.35}$$

According to Box-Cox (1964) power transformation

$$y^\lambda=\begin{cases}\dfrac{y^\lambda-1}{\lambda}, & if \lambda\neq 0\\ \log(y), & if \lambda=0\end{cases} \tag{2.36}$$

where y is a dependent variable, when $\lambda\to 0$ the equation (2.35) can be transformed to

$$\log\left(h(t)\right)=\phi_0\log\left(\sigma^2\right)+\phi_1\log\left(\varepsilon^2(t-1)\right)+\cdots+\phi_p\log\left(\varepsilon^2(t-p)\right) \tag{2.37}$$

which is equivalent to equation (32) with $a_0=\phi_0\log\left(\sigma^2\right)$ and $a_i=\phi_i$ for i=1,..., p.

2.4.6 GJR GARCH and Threshold GARCH model

Based on the empirical finding of Engle and Ng (1993) that negative shocks of similar magnitude lead to larger revision on conditional volatility, Glosten et al. (1993) proposed another popular GARCH class model – the GJR GARCH model. GJR model nested a dummy variable in to the GARCH model and the dummy variable is an indicator of the sign of the innovation. In order to compensate the leverage effect, if the innovation is negative, the indicator variable is set to be 1 and otherwise 0. The GJR model can be written as

$$h(t)=a_0+a_1\varepsilon^2(t-1)+\cdots a_q\varepsilon^2(t-q)+\beta_1h(t-1)+\cdots\beta_ph(t-p)+\gamma\varepsilon^2(t-1)I_{t-1} \tag{2.38}$$

where I_t is the indicator, $a_0,a_1,\cdots a_q,\beta_1,\cdots\beta_p$ are unknown parameters and γ is the leverage parameter. The GJR model introduces an alternative asymmetric component to the EGARCH model since a negative residual $\varepsilon^-(t-1)$ contributes $(a_1+\gamma)\varepsilon^2(t-1)$ to the conditional variance while a positive residual $\varepsilon^+(t-1)$ only contributes $\varepsilon^2(t-1)$ to it. As the indicator term in GJR GARCH is only related with the most recent innovation, Threshold GARCH is the general case of the GJR GARCH and the indicator term has been extended to compensate the negative impact of all past innovations as

$$h(t)=a_0+\sum_{i=1}^{q}a_i\varepsilon^2(t-i)+\sum_{i=1}^{q}\gamma_iI_{t-i}\varepsilon^2(t-i)+\sum_{i=1}^{p}\beta_ih(t-i) \tag{2.39}$$

2.4.7 Logistic GARCH

In GJR GARCH and Threshold GARCH, the indicator function is introduced to model the leverage effect. However, during the parameter estimation of those two models, the non differentiability of the indicator function may cause problems. Here we propose to use a logistic STAR function to approach the indicator function. The logistic STAR function is usually written as

$$L=\frac{-1}{1+\exp\left(-\lambda\left(y_t-c\right)\right)} \tag{2.40}$$

where λ is the smooth parameter and c is the threshold parameter. As in GJR and Threshold GARCH the indicator is switched at 0, the parameter c then should be set to 0. Different logistic functions under different smooth parameters are listed in Figure 1. It can be seen that when λ trends to be large, the logistic function trends to approach the indicator function I_t. Accordingly, the logistic GARCH can be written as

$$h(t) = a_0 + a_1\varepsilon^2(t-1) + \beta_1 h(t-1) + \frac{-b}{1+\exp(-\lambda\varepsilon(t-1))}\varepsilon^2(t-1) \tag{2.41}$$

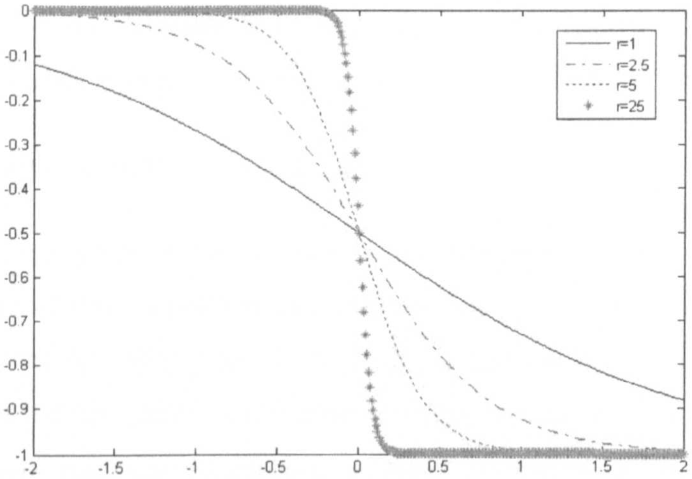


Figure 2.1 Logistic function simulation under different smooth parameters

2.4.8 FIGARCH

The shock in volatility series has been found to be capable to impact the future volatility over very long horizon (Taylor, 1986). According to this stylized fact and based on the time series long memory fractionally integrated process, Baillie et al. (1986) proposed the FIGARCH to approach the long memory effect to the volatility process with a more flexible model structure. Because the fractionally differencing operator term can be expanded in terms of the hypergeometric function as

$$\begin{aligned}(1-L)^d &= \sum_{k=0}^{\infty}\Gamma(k-d)\Gamma(k+1)^{-1}\Gamma(-d)^{-1}L^k \\ &= \sum_{k=0}^{\infty}\pi_k L^k\end{aligned} \tag{2.42}$$

where $\Gamma(x)$ is Gamma function, π_k is the parameter of every lagged term and L is lag operator. The FIGARCH is then given by

$$\left[1-\sum_{i=1}^{p}\beta_i L^i\right]h(t)=a_0+\left[1-\sum_{i=1}^{p}\beta_i L^i-\sum_{i=1}^{q}a_i L^i(1-L)^d\right]\varepsilon^2(t) \qquad (2.43)$$

where a, β are parameters and d is the fractional differencing parameter which satisfies $0 < d \leq 1$. One advantage of the FIGARCH model is that the impact of lagged squared innovations on conditional volatilities can have a slow hyperbolic rate of decay rather than an infinite propagation as in IGARCH model. However, Granger (2001) pointed out that the integrated process which has a time trend in volatility level is not observable in practice. Therefore, it is difficult to test against the FIGARCH model in empirical application.

2.4.9 Summary of GARCH class models

In the recent twenty years, different types of GARCH class model have been applied to a wide range of time series analyses and the applications in finance have been particularly successful. Nearly all of the GARCH class of models have one major assumption that the innovation of the return process consists of the multiplication of an i.i.d. variable and the conditional standard deviation as $\varepsilon(t)=z(t)\sigma(t)$. Although there are some other GARCH class models existed in literature, however the models referred above are widely used and have been commonly tested in the empirical applications. Therefore, further discussion on other GARCH class models will not be introduced here.

2.5 Parameter Estimation of GARCH class models

2.5.1 Maximum likelihood Estimation (MLE) method

Since the ARCH model was proposed by Engle (1982), the parameter estimation method for GARCH class of models has barely changed. Standard practice is to estimate the parameters using the MLE method. There is a pre-assumption before using MLE which requires a certain form of the joint probability density function.

The most commonly used probability function is the conditional Gaussian distribution function

$$\varepsilon(t)\big|\psi_{t-1} \sim f\left(\varepsilon(t)\big|\psi_{t-1}\right)=\frac{1}{\sqrt{2\pi\sigma^2(t)}}\exp\left\{-\frac{\varepsilon^2(t)}{2\sigma^2(t)}\right\} \tag{2.44}$$

where $f\left(\varepsilon(t)\big|\psi_{t-1}\right)$ denotes the conditional density function for $\varepsilon(t)$. If the average log likelihood is denoted by l and the log likelihood of t^{th} observation is denoted by l_t, then $l=\frac{1}{T}\sum_{t=1}^{T} l_t$ where T is the sample size. Since

$$\begin{aligned} l_t &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2(t)}}\exp\left\{-\frac{\varepsilon^2(t)}{2\sigma^2(t)}\right\}\right) \\ &= \frac{1}{2}\log\left(2\pi\sigma^2(t)\right)+\left(-\frac{\varepsilon^2(t)}{2\sigma^2(t)}\right) \\ &= -\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\left(\sigma^2(t)\right)-\frac{\varepsilon^2(t)}{2\sigma^2(t)} \end{aligned} \tag{2.45}$$

and the term $\log(2\pi)$ can be ignored because it is a constant, the log likelihood function at single observation then becomes to

$$l_t=-\frac{1}{2}\log\left(h(t)\right)-\frac{1}{2}\frac{\varepsilon^2(t)}{h(t)} \tag{2.46}$$

. Then the first derivative of equation (2.46) with respect to the parameter vector θ (the parameters of the GARCH model) is

$$\begin{aligned} \frac{\partial l_t}{\partial\theta} &= -\frac{1}{2}\frac{\partial\left(\log\left(h(t)\right)\right)}{\partial\theta}-\frac{\varepsilon_t^2}{2}\frac{\partial\left(\frac{1}{h(t)}\right)}{\partial\theta} \\ &= -\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\theta}+\frac{1}{2}\frac{\varepsilon^2(t)}{h^2(t)}\frac{\partial h(t)}{\partial\theta} \\ &= \frac{1}{2h(t)}\frac{\partial h(t)}{\partial\theta}\left(\frac{\varepsilon^2(t)}{h(t)}-1\right) \end{aligned} \tag{2.47}$$

Therefore, according to Bolloslev (1986) a Newton-Raphson searching method can be implemented to give the parameter updating procedure as

$$\theta_{k+1} = \theta_k - \lambda_k \left(\sum_{k=1}^{T} \frac{\partial^2 l_t}{\partial \theta_k \partial \theta'_k} \right)^{-1} \sum_{k=1}^{T} \frac{\partial l_t}{\partial \theta_k} \tag{2.48}$$

where θ_k is a vector of parameters estimated at the k^{th} iteration, λ_k is the step length, and ∇ denotes the first derivative of function $f(\theta_k)$. According to Berndt et al. (BHHH, 1974) the term $\left(\sum_{k=1}^{T} \frac{\partial^2 l_t}{\partial \theta_k \partial \theta'_k} \right)^{-1}$ can be approximated by $\left(\sum_{t=1}^{T} \frac{\partial l_t}{\partial \theta_k} \frac{\partial l_t}{\partial \theta'_k} \right)^{-1}$.

Then the parameters can be calculated from

$$\theta_{k+1} = \theta_k - \lambda_k \left(\sum_{t=1}^{T} \frac{\partial l_t}{\partial \theta_k} \frac{\partial l_t}{\partial \theta'_k} \right)^{-1} \sum_{k=1}^{T} \frac{\partial l_t}{\partial \theta_k} \tag{2.49}$$

Since the parameters of GARCH model have constraints, the Lagrangian function can be used during optimization to include the constraints in the object function as

$$L(x, \lambda, \rho) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{n} \rho_j h_j(x) \tag{2.50}$$

If the GARCH (1, 1) model with constraints $a_0 > 0, a_1, \beta_1 \geq 0, a_1 + \beta_1 < 1$ is taken as an example, the Lagrangian function for the estimation procedure is

$$L(\theta, \lambda, \rho) = f(\theta) - \lambda a_1 - \lambda \beta_1 + \frac{1}{\rho} \left(\frac{1}{a_0} + \frac{1}{1 - a_1 - \beta_1} \right) \tag{2.51}$$

For the GJR (1, 1) and QGARCH (1, 1) models there is an additional constraint because the leverage effect coefficient should be greater than zero. In empirical applications, the numerical derivative is usually used to approach the first derivative function as $\frac{\partial l_t}{\partial \theta_k} \approx \frac{\nabla l_t}{\nabla \theta_k}$ where ∇ is taken to be a very small value such as 10e-10.

Although during parameter estimation $h(t)$ is an unknown variable, it can be calculated recursively by the GARCH class model equation with initial settings. During numerical searching, different GARCH class models may give different recursive calculations of $h(t)$ but the procedure of parameter updating is the same as in equation (2.49).

As the normal distribution assumption is usually rejected by practical data, a plausible alternative to the normal is the student's t distribution which allows for heavier tails than the normal distribution. Bollerslev (1987) allowed the conditional distribution of GARCH model to be the t distribution function as

$$f\left(\varepsilon(t)|\psi_{t-1}\right)=\frac{1}{\sqrt{\pi}}\Gamma\left(\frac{v+1}{2}\right)\Gamma\left(\frac{v}{2}\right)^{-1}\left((v-2)h(t)\right)^{-1/2}\left(1+\frac{\varepsilon^{2}(t)}{h(t)(v-2)}\right)^{-(v+1)/2} \tag{2.52}$$

where Γ is Gamma function, v is the number of Degree of Freedom (DoF) and should be bigger than 4 as kurtosis and skewness statistics of the t distribution are defined under the condition of DoF>4. The corresponding log likelihood function is

$$\begin{aligned} l_t &= \log\left(\frac{1}{\sqrt{\pi}}\Gamma\left(\frac{v+1}{2}\right)\Gamma\left(\frac{v}{2}\right)^{-1}\left((v-2)h(t)\right)^{-1/2}\left(1+\frac{\varepsilon^{2}(t)}{h(t)(v-2)}\right)^{-(v+1)/2}\right) \\ &= \frac{1}{2}\log(\pi)+\log\left(\Gamma\left(\frac{v+1}{2}\right)\right)-\log\left(\Gamma\left(\frac{v}{2}\right)\right)-\frac{1}{2}\log(v-2)-\frac{1}{2}\log\left(h(t)\right) \\ &\quad -\frac{v+1}{2}\log\left(1+\frac{\varepsilon^{2}(t)}{h(t)(v-2)}\right) \end{aligned} \tag{2.53}$$

Then during MLE the normal likelihood function (2.46) can be substituted by equation (2.53) and the DoF can also be treated as an unknown parameter.

The implementation of the GARCH MLE procedure can be summarised as follows:

(1) Initialize the parameters. As a rule of thumb, the persistence coefficients β_1 of the GARCH (1, 1) model are usually in excess of 0.8 and reaction coefficients a_1 is usually no more than 0.2 (Alexander, 2001). The initial parameters of GARCH (p, q) can then be set as

$$h_1=\text{var}\left(\varepsilon(t)\right);a_1=a_2=\cdots a_q=\frac{0.05}{q};\beta_1=\beta_2=\cdots\beta_p=\frac{0.85}{p};a_0=0.1h_1$$

For the GJR and QGARCH models, along with the above initial parameters, the initial value of the leverage effect coefficient γ should be set to zero.

For the EGARCH model, the parameters should be set initially as

$$h_1 = \mathrm{var}(\varepsilon(t)); a_1 = a_2 = \cdots a_q = \frac{0.2}{q}; \beta_1 = \beta_2 = \cdots \beta_p = \frac{0.9}{p}; c_1 = c_2 = \cdots = c_q = 0$$

The step length λ_k in equation (2.49) should be set to be less than one and could vary depending on the different models. The penalty parameters λ in equation (2.51) should be set to extremely small values as 10E-30 because the likelihood should not be affected much by penalty terms. ρ in equation (2.51) should be set to extremely big values as 10E30for the same reason.

(2) The conditional variances $h(t)$ (t=1,..,T)are constructed recursively using appropriate GARCH class models under initial parameters and the likelihood l_t^1 of the each sample is calculated by the logarithm likelihood function (2.46). In order to calculate ∇l_t, initial parameters θ are saved and multiplied by $\nabla + 1$ as $(\nabla + 1)\theta$ and the likelihood l_t^2 is calculated by the updated parameters. ∇l_t is then calculated as $\nabla l_t = l_t^2 - l_t^1$. The term $\nabla\theta$ can be calculated as the multiplication of the initial parameters $\theta$and ∇. Therefore, $\frac{\nabla l_t}{\nabla \theta_k}$ can be determined and $\sum_{t=1}^{T} \frac{\partial l_t}{\partial \theta_k} \frac{\partial l_t}{\partial \theta_k'}$ in equation (2.49) can be calculated.

(3) Treat the new updated parameters by equation (2.49) as the initial parameters and repeat procedure (2) until some stop condition is achieved.

2.5.2 The Quasi-Maximum Likelihood Estimation (QMLE) method

During practical application of the MLE on estimating parameters of GARCH class of models, the assumption of the conditional normality is always breached and the tails of the conditional distribution has always been found to be fatter than that of the normal distribution. Bollerslev and Wooldridge (1988) discussed those facts and proposed the QMLE method to give a consistent estimation under weak regularity conditions. In particular, Bollerslev and Wooldridge showed that

$$\left(A_T^{0-1} B_T^0 A_T^{0-1}\right)^{-1/2} \sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \overset{d}{\sim} N(0, I) \qquad (2.54)$$

where $\hat{\theta}_T$ is the estimated parameters under assumption of normal distribution and

$$B_T^0 = \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{\partial l_t}{\partial \theta_0}\frac{\partial l_t}{\partial \theta_0'}\right) \tag{2.55}$$

$$A_T^0 = \frac{1}{T}\sum_{t=1}^{T} E\left(-\frac{\partial^2 l_t}{\partial \theta_0 \partial \theta_0'}\right) \tag{2.56}$$

However, they did not give any efficiency analysis of QMLE when the distribution was falsely assumed to be normal. Addressing this issue Engle and Gonzalez-Rivera (1991) discussed that QMLE will lose efficiency when the conditional distribution is not normal and they defined the notion of Relative Efficiency (RE) of QMLE in order to describe this efficiency loss. By definition, the RE is the ratio of asymptotic variance of the parameters when the true density function is known to its asymptotic variance when normality has been assumed. RE can be written as

$$RE_\theta = \frac{\mathrm{var}\left(\hat{\theta}_{MLE}\right)}{\mathrm{var}\left(\hat{\theta}_{QMLE}\right)} \tag{2.57}$$

where the $\mathrm{var}\left(\hat{\theta}_{MLE}\right)$ is the asymptotic variance of parameters from the MLE method when the conditional distribution is correctly specified and it is calculated as $\sqrt{B_T^0}$. If the true conditional distribution is normal then $A_T^0 = B_T^0$ (Weiss, 1982) and the RE will be equal to one. However if the true conditional distribution is non-normal, then the RE will be less than 1.

In order to show the efficiency losses, the RE is derived theoretically in the cases where the true conditional distribution is a symmetric fat-tailed Student's t distribution. Here, only GARCH (1, 1)$h_t = (1-a-\beta) + a\varepsilon_{t-1}^2 + \beta h_{t-1}$ is used during calculation of RE for simplicity. In QMLE, the first derivative of the likelihood function with respect to the GARCH parameters is calculated as

$$\frac{\partial l_t}{\partial \alpha} = -\frac{1}{2}\frac{1}{h_t}\frac{\partial h_t}{\partial \alpha}\left(1-\frac{\varepsilon^2(t)}{h(t)}\right) \tag{2.58}$$

$$\frac{\partial l_t}{\partial \beta} = -\frac{1}{2}\frac{1}{h_t}\frac{\partial h_t}{\partial \beta}\left(1 - \frac{\varepsilon^2(t)}{h(t)}\right) \tag{2.59}$$

By the law of iterated expectation proposed by Patrick (1995, theorem 34.4), the matrices A_T^0 and B_T^0 are given by

$$A_T^0 = -\frac{1}{T}\sum_{t=1}^{T} E\left(\frac{\partial^2 l_t(\theta_0)}{\partial\theta\partial\theta'}\right) = -\frac{1}{T}\sum_{t=1}^{T} E\left(E\frac{\partial^2 l_t(\theta_0)}{\partial\theta\partial\theta'}\middle|\psi_{t-1}\right) \tag{2.60}$$

$$B_T^0 = \frac{1}{T}\sum_t E\left(\frac{\partial l_t(\theta_0)}{\partial\theta}\frac{\partial l_t(\theta_0)}{\partial\theta'}\right) = \frac{1}{T}\sum_t E\left(E\frac{\partial l_t(\theta_0)}{\partial\theta}\frac{\partial l_t(\theta_0)}{\partial\theta'}\middle|\psi_{t-1}\right) \tag{2.61}$$

Therefore, substituting equation (2.58) and equation (2.59) into B_T^0 yields

$$\begin{aligned}
B_{11} &= \frac{1}{T}\sum_t^T E\left(E\left(-\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial \alpha}\left(1-\frac{\varepsilon^2(t)}{h(t)}\right)\bullet -\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial \alpha}\left(1-\frac{\varepsilon^2(t)}{h(t)}\right)\right)\middle|\psi_{t-1}\right)\\
&= \frac{1}{T}\sum_t^T E\left(E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2\left(1-\frac{\varepsilon^2(t)}{h(t)}\right)^2\right)\middle|\psi_{t-1}\right)\\
&= \frac{1}{T}\sum_t^T E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2 E\left(\left(1-\frac{\varepsilon^2(t)}{h(t)}\right)^2\right)\middle|\psi_{t-1}\right)\\
&= \frac{1}{T}\sum_t^T E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2 E\left(1-2\frac{\varepsilon^2(t)}{h(t)}+\left(\frac{\varepsilon^2(t)}{h(t)}\right)^2\right)\middle|\psi_{t-1}\right)\\
&= \frac{1}{T}\sum_t^T E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2\left(1-2E\left(\frac{\varepsilon^2(t)}{h(t)}\right)+E\left(\frac{\varepsilon^4(t)}{h^2(t)}\right)\right)\middle|\psi_{t-1}\right)\\
&= \frac{1}{T}\sum_t^T E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2(k-1)\middle|\psi_{t-1}\right)\\
&= \frac{1}{T}\sum_t^T \frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2(k-1)
\end{aligned} \tag{2.62}$$

where k is the coefficient of conditional kurtosis as $E\left(\varepsilon_t^4\middle|\psi_{t-1}\right)/h_t^2$. Similarly, the other parts of the matrix B_T^0 can be given by

$$B_{22} = \frac{1}{T}\sum_{t}\frac{1}{4}\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial \beta}\right)^{2}(k-1) \tag{2.63}$$

$$B_{12} = B_{21} = \frac{1}{T}\sum_{t}\frac{1}{4}\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial \beta}\right)\left(\frac{\partial h(t)}{\partial \alpha}\right)(k-1) \tag{2.64}$$

The elements of the matrix A_T^0 can be calculated by substituting equations (2.58) and (2.59) into equation (2.60) using the following equation

$$\begin{aligned}
A_{11} &= -\frac{1}{T}\sum_{t=1}^{T}E\left(\frac{\partial\left(-\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\alpha}\left(1-\frac{\varepsilon^{2}(t)}{h(t)}\right)\right)}{\partial\alpha}\right)\\
&= -\frac{1}{T}\sum_{t=1}^{T}E\left(\frac{\partial\left(-\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\alpha}+\frac{1}{2}\frac{\varepsilon^{2}(t)}{h^{2}(t)}\frac{\partial h(t)}{\partial\alpha}\right)}{\partial\alpha}\right)\\
&= -\frac{1}{T}\sum_{t=1}^{T}E\left(-\frac{1}{2}\left(-\frac{\partial h(t)}{\partial\alpha}\frac{1}{h^{2}(t)}\right)\frac{\partial h_{t}}{\partial\alpha}+\frac{1}{2}\frac{\partial h(t)}{\partial\alpha}\left(-2\frac{\varepsilon_{t}^{2}}{h^{3}(t)}\frac{\partial h(t)}{\partial\alpha}\right)\right)\\
&= -\frac{1}{T}\sum_{t=1}^{T}E\left(\frac{1}{2}\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}-\frac{\varepsilon^{2}(t)}{h^{3}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}\right)\\
&= -\frac{1}{T}\sum_{t=1}^{T}E\left(\frac{1}{2}\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}-\frac{\varepsilon_{t}^{2}}{h^{3}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}\right)\\
&= -\frac{1}{T}\sum_{t=1}^{T}E\left(E\left(\frac{1}{2}\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}-\frac{\varepsilon_{t}^{2}}{h^{3}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}\right)\middle|\psi_{t-1}\right)\\
&= -\frac{1}{T}\sum_{t=1}^{T}E\left(E\left(\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}\left(\frac{1}{2}-\frac{\varepsilon^{2}(t)}{h(t)}\right)\right)\middle|\psi_{t-1}\right)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2 E\left(\frac{1}{2}-\frac{\varepsilon^2(t)}{h(t)}\right) \middle| \psi_{t-1} \right) \\
&= -\frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2 \left(\frac{1}{2}-1\right) \middle| \psi_{t-1} \right) \qquad (2.65) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2
\end{aligned}
$$

The other elements of the matrix A_T^0 are given by

$$
A_{22} = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \beta}\right)^2 \qquad (2.66)
$$

$$
A_{12} = A_{21} = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)\left(\frac{\partial h(t)}{\partial \beta}\right) \qquad (2.67)
$$

It can be seen that $B_{11} = A_{11}\frac{1}{2}(k-1), B_{22} = A_{22}\frac{1}{2}(k-1), B_{12} = A_{12}\frac{1}{2}(k-1)$. Hence the asymptotic variances of the parameters of QMLE- $\text{var}(\theta_{QMLE})$ are given by

$$
\begin{aligned}
A_T^{0-1}B_T^0A_T^{0-1} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} B_{11} & B_{12} \\ B_{12} & B_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}^{-1} \frac{1}{2}(k-1) \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}^{-1} \qquad (2.68) \\
&= \frac{1}{2}(k-1) \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}^{-1} \\
&= \frac{1}{2}(k-1) \begin{bmatrix} A_{22} & -A_{12} \\ -A_{12} & A_{11} \end{bmatrix} / \left(A_{11}A_{22} - A_{12}^2\right)
\end{aligned}
$$

The diagonal elements of the above matrix are the asymptotic variances of the parameters as in the following

$$
\begin{aligned}
\text{var}(\hat{\alpha}) &= \frac{1}{2}(k-1)A_{22} / \left(A_{11}A_{22} - A_{12}^2\right) \qquad (2.69) \\
&= B_{22} / \left(A_{11}A_{22} - A_{12}^2\right)
\end{aligned}
$$

$$\begin{aligned}\operatorname{var}\left(\hat{\beta}\right) &= \frac{1}{2}(k-1)A_{11}/\left(A_{11}A_{22}-A_{12}^2\right)\\ &= B_{11}/\left(A_{11}A_{22}-A_{12}^2\right)\end{aligned} \tag{2.70}$$

If the true conditional distribution of the residuals is a Student t distribution, the likelihood function and the first order derivative are given by

$$f\left(\varepsilon_t|\psi_{t-1}\right)=\frac{1}{\sqrt{\pi}}\Gamma\left(\frac{v+1}{2}\right)\Gamma\left(\frac{v}{2}\right)^{-1}\left((v-2)h(t)\right)^{-1/2}\left(1+\frac{\varepsilon^2(t)}{h(t)(v-2)}\right)^{-(v+1)/2} \tag{2.71}$$

where v is the number of degrees of freedom and should be bigger than 2 and $\Gamma(\ )$ is the Gamma function. The corresponding log likelihood function is

$$\begin{aligned}l_t &= \log\left(\frac{1}{\sqrt{\pi}}\Gamma\left(\frac{v+1}{2}\right)\Gamma\left(\frac{v}{2}\right)^{-1}\left((v-2)h_t\right)^{-1/2}\left(1+\frac{\varepsilon^2(t)}{h(t)(v-2)}\right)^{-(v+1)/2}\right)\\ &= \frac{1}{2}\log(\pi)+\log\left(\Gamma\left(\frac{v+1}{2}\right)\right)-\log\left(\Gamma\left(\frac{v}{2}\right)\right)-\frac{1}{2}\log(v-2)-\frac{1}{2}\log(h(t))\\ &\quad -\frac{v+1}{2}\log\left(1+\frac{\varepsilon^2(t)}{h(t)(v-2)}\right)\end{aligned} \tag{2.72}$$

The first derivative with respect to the GARCH parameter θ is then given by

$$\begin{aligned}\frac{\partial l_t}{\partial\theta} &= -\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\theta}-\frac{v+1}{2}\frac{1}{1+\dfrac{\varepsilon^2(t)}{h(t)(v-2)}}\left(-\frac{\varepsilon^2(t)}{h^2(t)(v-2)}\frac{\partial h(t)}{\partial\theta}\right)\\ &= -\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\theta}+\frac{v+1}{2}\frac{h(t)(v-2)}{h(t)(v-2)+\varepsilon^2(t)}\frac{\varepsilon^2(t)}{h^2(t)(v-2)}\frac{\partial h(t)}{\partial\theta}\\ &= -\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\theta}+\frac{1}{2}\frac{1}{h(t)}\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\frac{\partial h(t)}{\partial\theta}\\ &= -\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial\theta}\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)\end{aligned} \tag{2.73}$$

The elements of matrix B_T^{0MLE} are then given by

$$
\begin{aligned}
B_{11}^{MLE} &= \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{\partial l_t(\theta_0)}{\partial \alpha}\frac{\partial l_t(\theta_0)}{\partial \alpha'}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} E\left(-\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial \alpha}\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)\cdot-\frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial \alpha}\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2
\end{aligned}
\tag{2.74}
$$

$$
B_{22}^{MLE} = \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \beta}\right)^2\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2\right) \tag{2.75}
$$

$$
B_{12}^{MLE} = \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)\left(\frac{\partial h(t)}{\partial \beta}\right)\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2\right) \tag{2.76}
$$

Hence, the asymptotic variances of the parameters estimated by MLE under t distribution are the diagonal elements of the inverse B_T^{0MLE} matrix:

$$
\operatorname{var}(\hat{\alpha}_{MLE}) = \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \beta}\right)^2\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2\right) / \left(B_{11}^{MLE}B_{22}^{MLE}-B_{12}^{2\,MLE}\right) \tag{2.77}
$$

$$
\operatorname{var}(\hat{\beta}_{MLE}) = \frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \alpha}\right)^2\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2\right) / \left(B_{11}^{MLE}B_{22}^{MLE}-B_{12}^{2\,MLE}\right) \tag{2.78}
$$

According to the definition of RE, the RE of parameters a,β can be calculated as

$$
RE(\hat{\alpha}) = \frac{\frac{1}{T}\sum_{t=1}^{T} E\left(\frac{1}{4}\frac{1}{h^2(t)}\left(\frac{\partial h(t)}{\partial \beta}\right)^2\left(1-\frac{(v+1)\varepsilon^2(t)}{h(t)(v-2)+\varepsilon^2(t)}\right)^2\right) / \left(B_{11}^{MLE}B_{22}^{MLE}-B_{12}^{2\,MLE}\right)}{B_{22}/\left(A_{11}A_{22}-A_{12}^2\right)} \tag{2.79}
$$

$$RE\left(\hat{\beta}\right)=\frac{\frac{1}{T}\sum_{t=1}^{T}E\left(\frac{1}{4}\frac{1}{h^{2}(t)}\left(\frac{\partial h(t)}{\partial\alpha}\right)^{2}\left(1-\frac{(v+1)\varepsilon^{2}(t)}{h(t)(v-2)+\varepsilon^{2}(t)}\right)^{2}\right)/\left(B_{11}{}^{MLE}B_{22}{}^{MLE}-B_{12}^{2\,MLE}\right)}{B_{11}/\left(A_{11}A_{22}-A_{12}^{2}\right)} \tag{2.80}$$

If the parameters of GARCH and conditional distribution are known, then the RE can be used to evaluate the efficiency losses of QMLE. Therefore, although MLE and QMLE is widely applied in parameter estimation of the GARCH class of models, but when the conditional distribution is falsely assumed, both MLE and QMLE cannot give consistent and efficienct estimation results. This is one of the major issues among the GARCH class of models.

2.6 Forecast Evaluation of GARCH class models

Since so many different types of GARCH class models have been proposed to model the conditional volatility, it is essential to have an evaluation statistic to compare the forecast performance of those models. Because the volatility is not observable and the only observation of the market data is the returns, one common statistical measure of accuracy for a volatility forecast is the likelihood of the return. However, the effectiveness of this method does rely on the correct specification of the conditional distributions. This means that the distribution assumption needs to be tested first; otherwise the test statistical based on likelihood will be unreliable.

Another popular evaluation measure used in literature is to use the squared innovations from the return process as the proxy of the actual volatility and different error statistics are used as the criterion. Popular evaluation measures include Mean Error (ME), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE). As most investors treat gains and losses differently, the error statistics which treat positive errors differently from negative ones have also been proposed to use during practical analysis. For example, Granger (1999) described a LinEx loss function as

$$LinEx = \frac{1}{N}\sum_{t=1}^{N}\left[\exp\left(-a\left(\hat{h}(t)-\varepsilon^2(t)\right)\right)+a\left(\hat{h}(t)-\varepsilon^2(t)\right)-1\right] \qquad (2.81)$$

where $\hat{h}(t)$ is the predicted conditional variance, a is parameter whose sign determines whether positive errors have more or less weight than negative ones. If $a>0$, over predictions $\hat{h}(t)>h(t)$ will have less weight because the term $\exp\left(-a\left(\hat{h}(t)-h(t)\right)\right)$ is less than 1 and the total sum of $\exp\left(-a\left(\hat{h}(t)-h(t)\right)\right)+a\left(\hat{h}(t)-h(t)\right)-1$ is less than $a\left(\hat{h}(t)-h(t)\right)$. If $a<0$, over predictions will have more weight.

Before high frequency data becomes widely available, most of the researchers use the squared innovation from daily return process which is calculated from the daily closing price as the proxy to the daily volatility. However, Lopez (2001) discussed that although squared innovation $\varepsilon^2(t)$ is an unbiased estimator of $h(t)$, the error statistic is very imprecise due to its asymmetric distribution because $\varepsilon(t)=z(t)\sqrt{h(t)}$ and $E\left(\varepsilon^2(t)|\psi_{t-1}\right)=h(t)E\left(z^2(t)|\psi_{t-1}\right)=h(t)$ where $z(t)\sim i.i.d.(0,1)$ and $z^2(t)\sim\chi^2(1)$. However, the median of $\chi^2(1)$ distribution is less than 0.5 which means that $\varepsilon^2(t)<h(t)$ is more than 50% of the time. Therefore, the high frequency intraday return data is proposed by Andersen et al. (2001) to use as the proxy rather than daily return. Further suggestion by Bollerslve and Ghysels (1996) included a proposal to use a Heteroskedasticity adjusted version of MSE (HMSE) as

$$HMSE = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{h(t)}{\hat{h}(t)}-1\right] \qquad (2.82)$$

Because there are several volatility features which already have been observed in different market data, GARCH class models are introduced to match one or more of those features. Accordingly, the evaluation of forecasting accuracy will depend on what criterion is used and it is unlikely to choose a best forecasting model with all possible statistics and evaluation criterions.

2.7 Conclusion

In this chapter, the commonly used volatility models including EWMA, ARCH, GARCH and GARCH class of models have been introduced and reviewed and these models are proposed to model the different volatility features. Much research has been contributed into the properties, estimation and analysis of these models. However, one obvious omission is the adoption of a linear mean model to fit the return process. This will be reviewed in the next chapter and the motivation of using a nonlinear mean model is therefore raised to encounter the problem.

Chapter 3: Introduction to the financial mean model

3.1 Introduction

Following the introduction of ARCH model, many papers have been published which attempt to model the time varying variance in financial time series. Although nonlinear modelling has been widely applied in time series analysis, there has been little theoretical work to investigate the use of nonlinear models to represent the mean model as part of ARCH and GARCH model estimation. Linear models can provide good first order approximations and because linear statistical theory is now well developed, the linear mean model is commonly used in many publications which study aspect of volatility estimation. In the pioneering work by Engle (1982) who introduced the ARCH model, an example of an ARCH model was given together with an ARX like mean model to model inflation in the U.K. In a later ARCH application paper by Engle (1983), a linear AR type mean model with multiple inputs and a time trend was used to model inflation in the U.S. Bollerslev (1986) used an AR (4) model to model the rate of growth in the implicit GNP deflator in the U.S. together with a GARCH variance model. When EGARCH was introduced by Nelson (1991), a variance in the mean AR (1) model was used to model the return series and Engle (1989) used a variance in the mean MA (1) model to model the return series while the variance was modelled using QGARCH. Higgins and Bera (1992) applied a linear AR model with a NARCH to model the variance of foreign exchange rates of five countries while Baillie et al. (1996) just employed a constant term to represent the mean model term in the FIGARCH model. Gray (1992) introduced the RS-GARCH to model the volatility of short-term interest rates and an AR (1) model was used to model the mean process. Ding et al. (1993) argued for a long memory property of the volatility and used a MA (1) model as the mean model. Lee (1991) investigated the out-of-sample forecast accuracy of a GARCH class of models without using any mean model. Instead, he just used the return data as the mean residuals. Tse (1991) favoured an AR (1) model to model the mean associated with the Tokyo stock return while Akgiray (1989) suggested the use of an AR (1) model to fit the New York market returns. Hamilton and Susmel (1994) introduced a class of Markov-switching ARCH models and modelled the mean using an AR (1) model. McMillan et al. (2000) analysed the forecast performance of a variety of statistical variance models using UK stock market data and the mean process was fitted using an AR model. Clearly therefore the linear mean model is very commonly used in univariate GARCH model analysis In later multivariate GARCH model research, a linear mean model is still massively employed in many publications which include Vec-GARCH introduced by Bollerslev et al. (1988), the Multivariate GARCH in mean model of Engle and Kroner (1995), and the constant mean model by Kawakatsu (2006).

The use of a nonlinear mean model example can be found in Lebaron (1992) where an exponential AR mean model was used to model the relation between the autocorrelation and the conditional variance. Bollerslev et al. (1993) also used an exponential AR mean model in fitting the U.S. stock market volatility. In Cao and Tsay (1992), the threshold AR (2) which is a group of two linear AR (2) models switched according to a threshold value was used to model the return process. Although a linear model is easy to estimate, there has been overwhelming evidence of non-linear structures across many financial data sets (Willey, 1992). Hinich and Patterson (1985) tested the different stock market returns with nonlinearity tests and all of the testing results indicated a nonlinear dependence and non Gaussian processes. Therefore, the use of a nonlinear model instead of a linear model to model the return process should improve the forecast accuracy and provide more accurate residual estimation for the conditional variance process.

There are several nonlinear modelling techniques which are commonly used in system identification including the Volterra series (Volterra, 1930)/the Wiener series (Wiener, 1958) expansion approach, the Radial Basis Function (RBF) neural network method (Moody and Darken, 1989) and wavelet networks (Antoniadis and Oppenheim, 1995). However, the Volterrra and the Wiener series require excessive parameters to give an adequate approximation to the underlying nonlinear system process (Ogunfunmi, 2007) and the RBF and wavelet networks approach are not very transparent models. The Nonlinear Auto Regressive Moving Average with eXogenous inputs (NARMAX) model which is a generalization of the ARMAX model proposed by Leontaritis and Billings (1985) can provide a unified formation for a wide class of nonlinear system processes with a more concise representation than the Volterra and the Wiener nonlinear models. The NARMAX model has been proved to be successful in modelling numerous real world nonlinear systems including chaotic electronic circuits, water management systems, turbocharged diesel engines, etc(Billings and Coca, 2001). Accordingly, this chapter proposes to use NARMAX to model the nonlinear mean model of financial return data.

The purpose of this chapter is to review the mean model used in the literature related to GARCH model research and the commonly used mean models will be investigated. The NARMAX model method will be introduced and examples will be given to illustrate the differences of variance estimation when the nonlinear mean process is falsely fitted using a linear model. Section 3.2 provides a review of the mean model. Section 3.3 gives an introduction to the NARMAX modelling method. Section 3.4 gives an example showing the impact on the variance estimation when the nonlinear mean process is falsely estimated using linear model. Section 3.5 is the conclusions.

3.2 Commonly used return models

3.2.1 The linear mean model

In the financial application area, modelling methods based on time varying variance have developed very quickly since the ARCH model was introduced by Engle (1982). ARCH and the generalised form of the model GARCH can very well represent the observed effect of financial time varying variance. However, most of the financial literature was concentrated on modelling the variance and ignores the importance of the mean model and therefore, due to simplicity a linear mean is very frequently quoted in most GARCH research papers. In Engle (1982), a linear mean model was used during the simulations as defined by

$$\Delta p(t)=\beta_1\Delta p(t-1)+\beta_2\Delta p(t-4)+\beta_3\Delta p(t-5)+\beta_4\left(p(t-1)-w(t-1)\right)+\beta_5 \qquad (3.1)$$

where $\Delta p(t)$ is the first difference of the log of the quarterly consumer price index, w is the log of the quarterly index of manual wage rates and β are parameters. As $\Delta p(t)=p(t)-p(t-1)$ and w is like an external input to the models, the model (3.1) can be expressed as

$$\begin{aligned}\Delta p(t)&=\beta_1\Delta p(t-1)+\beta_2\Delta p(t-4)+\beta_3\Delta p(t-5)+\beta_4\left(p(t-1)-w(t-1)\right)+\beta_5\\ p(t)-p(t-1)&=\beta_1\left(p(t-1)-p(t-2)\right)+\beta_2\left(p(t-4)-p(t-5)\right)+\beta_3\left(p(t-5)-p(t-6)\right)\\ &\quad+\beta_4 p(t-1)-\beta_4 w(t-1)+\beta_5\\ p(t)&=\left(\beta_1+1+\beta_4\right)p(t-1)-\beta_1 p(t-2)+\beta_2 p(t-4)\\ &\quad+\left(\beta_3-\beta_2\right)p(t-5)-\beta_3 p(t-6)+\beta_5-\beta_4 w(t-1)\end{aligned}$$

$$(3.2)$$

It is clear that the model (3.2) is a linear ARX model with one lagged input term and five AR terms. In a later ARCH model application paper by Engle (1983), a linear model with a time trend was used to fit the consumer price index data as

$$\Delta p(t)=\beta_1\Delta p(t-1)+\beta_2\Delta p(t-2)+\beta_3\Delta PM(t-1)+\beta_4\Delta w(t-1)+\beta_5\Delta m(t-1)+\beta_6 t+\beta_0$$

$$(3.3)$$

where $\Delta p(t)$ is the deflator, $\Delta w(t)$ is the rate of change of money supply, $\Delta m(t)$ is the rate of change of wages, $\Delta PM(t)$ is the rate of change of the import deflator and t is the time trend. Therefore, the model (3.3) can be expressed as

$$\begin{aligned}\Delta p&=\beta_1\Delta p_{-1}+\beta_2\Delta p_{-2}+\beta_3\Delta PM_{-1}+\beta_4\Delta w_{-1}+\beta_5\Delta m_{-1}+\beta_6 t+\beta_0\\ p_t-p_{t-1}&=\beta_1\left(p_{t-1}-p_{t-2}\right)+\beta_2\left(p_{t-2}-p_{t-3}\right)+\beta_3\left(PM_{t-1}-PM_{t-2}\right)+\beta_4\left(w_{t-1}-w_{t-2}\right)+\beta_6 t+\beta_0\\ p_t&=\left(\beta_1+1\right)p_{t-1}+\left(\beta_2-\beta_1\right)p_{t-2}-\beta_2 p_{t-3}+\beta_3\left(PM_{t-1}-PM_{t-2}\right)+\beta_4\left(w_{t-1}-w_{t-2}\right)+\beta_6 t+\beta_0\end{aligned}$$

$$(3.4)$$

The model (3.4) is a linear model with multiple inputs and a time trend. Bollerslev (1986) extended the ARCH model to a more general case-GARCH model and he used a mean model in an example as

$$\pi(t)=\beta_1+\beta_2\pi(t-1)+\beta_3\pi(t-2)+\beta_4\pi(t-3)+\beta_5\pi(t-4)+\varepsilon(t) \qquad (3.5)$$

where $\pi(t)=100\times\ln(GD(t)/GD(t-1))$, $GD(t)$ is the implicit price deflator for GNP and β are parameters. The model (3.5) is a typical AR (4) model. Nelson (1991) proposed a new GARCH class model-EGARCH and fittd the return with model as

$$R(t)=\beta_0+\beta_1 R(t-1)+\beta_2\sigma^2(t)+\varepsilon(t) \tag{3.6}$$

where $R(t)$ is the excess return, $\sigma^2(t)$ is the conditional variance and β are parameters. The model (3.6) is an AR (1) with variance in the mean and Lo and Mackinlay (1988) noted that such a simple model cannot adequately explain the short term autocorrelation behaviour of the market indices. Nelson adopted the variance in mean terms because there is evidence of a positive correlation between the return series and the conditional variance series as in French et al. (1987) and Chou (1987). Engle (1989) modelled the daily return of a stock index by using Nonlinear ARCH model and the mean model used was

$$y(t)=\beta_0+\beta_1\sigma^2(t)+\beta_2\varepsilon(t-1)+\varepsilon(t) \tag{3.7}$$

where y_t is stock index return. The model (3.7) is a MA (1) model with variance in the mean. Apart from the mean model listed above, most of the researchers frequently use the simplest mean model which takes the form

$$y(t)=c+\varepsilon(t) \tag{3.8}$$

where c is a constant term. The application of model (3.8) can be found in Baillie et al. (1996) and Kawakatsu (2006).

3.2.2 Nonlinear mean model

Cao and Tsay (1992) proposed the use of a Threshold AR (TAR) model to model the mean process as

$$y(t)=\begin{cases}\beta_0+\beta_1 y(t-1)+\beta_2 y(t-2)+\beta_3 y(t-3)+\varepsilon^{(1)}(t) & if\ \ y(t)<T\\ \alpha_0+\alpha_1 y(t-1)+\alpha_2 y(t-2)+\alpha_3 y(t-3)+\varepsilon^{(2)}(t) & if\ \ y(t)\geq T\end{cases} \tag{3.9}$$

where α,β are parameters, $y(t)$ is return series and T is a threshold value. The TAR model is simply two AR models switch by threshold values. As noted in Chapter 2

Section 2.4.7, the threshold process can be described by the logistic STAR function as

$$L=\frac{-1}{1+\exp\left(-\lambda\left(y(t)-c\right)\right)} \tag{3.10}$$

and the model (3.9) then takes the form

$$y(t)=a_0+a_1y(t-1)+a_2y(t-2)+a_3y(t-3)+\frac{-(\beta_0-a_0)}{1+\exp\left(-\lambda\left(y(t)-T\right)\right)}+\frac{-(\beta_1-a_1)}{1+\exp\left(-\lambda\left(y(t)-T\right)\right)}y(t-1)$$
$$+\frac{-(\beta_2-a_2)}{1+\exp\left(-\lambda\left(y(t)-T\right)\right)}y(t-2)+\frac{-(\beta_3-a_3)}{1+\exp\left(-\lambda\left(y(t)-T\right)\right)}y(t-3)+\varepsilon(t)$$

(3.11)

Lebaron (1992) proposed the use of the exponential AR model to model the mean process as

$$y(t)=\beta_0+\left(\alpha_0+\alpha_1\exp\left(-\frac{\sigma^2(t)}{\alpha_3}\right)\right)y(t-1)+\varepsilon(t) \tag{3.12}$$

where β and α are parameters. The term $\left(\alpha_0+\alpha_1\exp\left(-\frac{\sigma^2(t)}{\alpha_3}\right)\right)$ is time varying as the conditional variance is time varying. Bollerslev et al. (1993) adopted a similar mean model in the simulation of the US stock market index as

$$y(t)=\beta_0+y(t-1)\left(\beta_1+\beta_2\exp\left(-\frac{\sigma^2(t)}{\beta_3}\right)\right)+\beta_4\sigma^2(t)+\varepsilon(t) \tag{3.13}$$

where $y(t)$ is the return, β are parameters and $\sigma^2(t)$ is conditional variance. Compared with model (3.12), model (3.13) has an extra term $\beta_4\sigma^2(t)$. According to the Taylor series expansion, the term $\exp\left(-\frac{\sigma^2(t)}{\beta_3}\right)$ can be expanded to

$$\exp\left(-\frac{\sigma^2(t)}{\beta_3}\right)=1-\frac{\sigma^2(t)}{\beta_3}+\frac{1}{2!}\left(-\frac{\sigma^2(t)}{\beta_3}\right)^2+\frac{1}{3!}\left(-\frac{\sigma^2(t)}{\beta_3}\right)^3+\cdots \tag{3.14}$$

After substituting (3.14) into (3.13), equation (3.13) becomes

$$y(t)=\beta_0+y(t-1)\left(\beta_1+\beta_2\left(1-\frac{\sigma^2(t)}{\beta_3}+\frac{1}{2!}\left(-\frac{\sigma^2(t)}{\beta_3}\right)^2+\frac{1}{3!}\left(-\frac{\sigma^2(t)}{\beta_3}\right)^3\right)\right)+\beta_4\sigma^2(t)+\varepsilon(t) \tag{3.15}$$

According to GARCH, $\sigma^2(t)$ consists of lagged $\varepsilon^2(t)$, and $\varepsilon^2(t)$ can be substituted using $y(t)$and$y(t-1)$, therefore model (3.15) can contain higher order lagged $y(t)$and shows nonlinearity.

Apart from the linear mean model, most publications use a specified nonlinear model format which may not fully reflect the nonlinearity that exists in the process. The use of NARMAX can give a more general choice of nonlinear model term selection and provide a universal approach to the nonlinear mean process.

3.3 NARMAX model and its polynomial representation

Most real life problems involve nonlinear systems. For most of the practical applications the nonlinear model usually has advantage to describe the nonlinear relationships rather than a linear model and nonlinear models are designed to provide a better mathematical instrument to characterize the nonlinearity in real dynamic systems. Nonlinear model representations can be generally classified into three types: (1) System Input-Output representation, (2) State-space representation and (3) Model-free representation (Chow et al., 2001). The discrete time Input-Output representation approach can usually be written as

$$y(t)=f(x)+e(t) \tag{3.16}$$

where x represents the system input, t is the time sample,$e(t)$is noise and $y(t)$is the system output and $f(\;)$denotes a mathematical relationship. When the system is linear, $f(\;)$represents a linear mapping between the input and the output and a linear differential equation is commonly used to approximate the process. The ARMAX model is usually employed to provide a unified input-output representation. When the system is nonlinear, there are several methods which exist to give an approximation to the nonlinear function $f(\ )$. The Volterra/Wiener representation is one technique to model the input-output nonlinearity based on a Volterra series mathematical function. Although Volterra system models can be used to represented a large range of nonlinear systems, in order to give an adequate approximation the number of the parameters usually exceeds many hundreds and the Volterra kernels which are nth-order impulse responses have to be estimated. Therefore, the Volterra nonlinear representation procedure can be computationally complex. The NARMAX model proposed by Leontaritis and Billings (1985) extends the ARMAX model to the nonlinear input/output case and NARMAX usually takes the form of a set of nonlinear equations as

$$y(t)=f\left(y(t-1),\ldots,y\left(t-n_y\right),u(t-d),\ldots,u\left(t-n_u\right),e(t-1),\ldots,e\left(t-n_e\right)\right)+e(t) \tag{3.17}$$

where $u(t)$ is the input vector, n_y and n_u are maximum output and input lag, n_e is the maximum noise lag, $y(t)$ is output vector, and $f(\ )$ is unknown nonlinear mapping. The noise variable $e(t)$ which accommodates the effects of measurement noise, modelling errors and unmeasured disturbances are assumed to be bounded and uncorrelated with the input.

Since $f(\ )$ is unknown, the identification of the NARMAX model involves not only determining the parameters of the models but also the structure a model terms from the input/output data. The polynomial representation of $f(\ )$ is one of the common implementations and it has received great attention because of the good approximation properties and the simple model structure what this choice yields. Therefore, the nonlinear mapping $f(\ )$ here is considered to be approximated by a polynomial representation with a finite degree in all variables and the structure is assumed to be linear-in-parameters. Accordingly, the general form of the polynomial NARMAX representation can be written as

$$y(t)=\sum_{i=1}^{M}\theta_i p_i\left(x(t)\right)+\varepsilon(t) \tag{3.18}$$

where $x(t)$ represents $y(t-1),...,y(t-n_y),u(t-d),...,u(t-n_u),\varepsilon(t-1),...,\varepsilon(t-n_\varepsilon)$, $p_i(\ )$ are model terms which are a linear or a nonlinear combination of the variables , $\varepsilon(t)$ is the modelling error, M is the number of all the distinct terms and θ_i are unknown parameters related. The matrix format of model (4.3) can be written as

$$Y=P\Theta+\Xi \tag{3.19}$$

where $Y=\left[y(1),y(2),...,y(N)\right]^T$, $P=\left[p_1,p_2,...,p_M\right]$, $p_i=\left[p_i(x(1)),p_i(x(2)),...,p_i(x(N))\right]^T$, $\Theta=\left[\theta_1,\theta_2,...,\theta_M\right]^T$ and $\Xi=\left[\varepsilon(1),\varepsilon(2),...,\varepsilon(N)\right]^T$. In this thesis proposes to use NARMAX model could be used to fit the nonlinear finance return process.

3.4 Simulations

Since most of the GARCH publications use a linear model to fit the return series, it is intriguing to illustrate the impact on variance estimation when a nonlinear return series is falsely fitted by a linear mean model. Consider the nonlinear mean model are formatted as

$$y(t)=a_0+a_1y(t-1)^2+a_2y(t-2)+\varepsilon(t) \tag{3.20}$$

where a_0,a_1,a_2 are parameters, $\varepsilon(t)$ is residual and where it is assumed that the time varying variance is generated by a GARCH model as

$$\sigma^2(t)=A_0+A_1\varepsilon^2(t-1)+B_1\sigma^2(t-1) \tag{3.21}$$

where A_0,A_1,B_1 are GARCH parameters. The parameters of the model (3.20) and (3.21) are listed in Table 3.1.

Table 3.1 Parameters of the simulated models

Parameter of mean model (3.20)	Value
a_0	0.001
a_1	12
a_2	-0.1
Parameter of variance model (3.20)	Value
A_0	3e-6
A_1	0.075
B_1	0.920

The residual $\varepsilon(t)$ is assumed as $\varepsilon(t)=z(t)\sigma(t)$ where $z(t)$ is an i.i.d. (0, 1) random variable. Therefore, the simulated variance and mean process are drawn in Figure 3.1. In order to illustrate the impact of the mean model on the variance estimation, a linear mean model is used to fit the simulated mean data and the variance is then estimated from the modelling residuals using a GARCH model. Assume that the linear mean model is chosen as the commonly used AR (1) model as

$$y(t)=\beta_0+\beta_1 y(t-1)+\varepsilon(t) \tag{3.22}$$

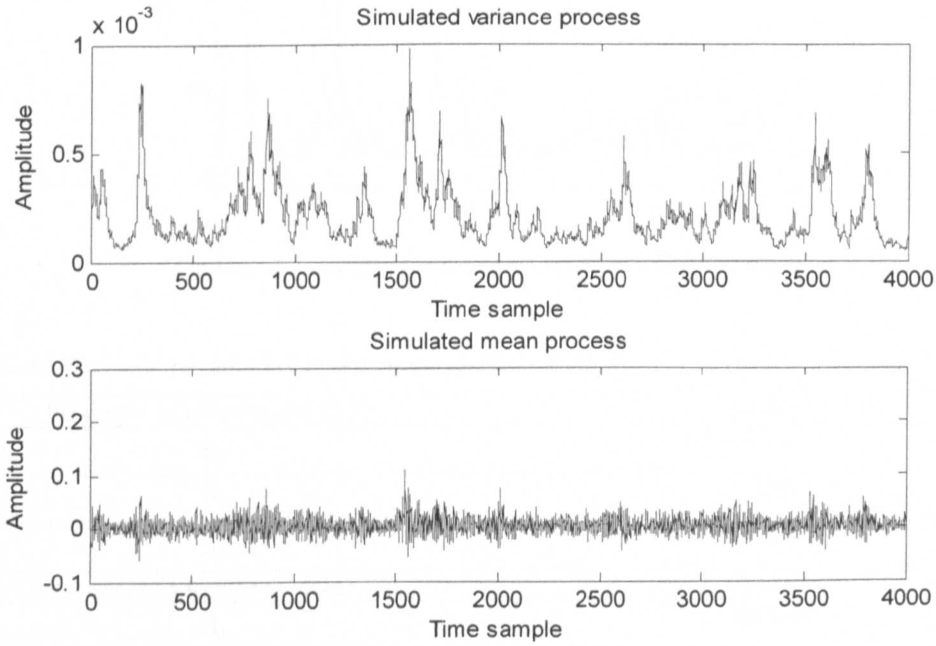


Figure 3.1 Simulated variance and mean process

The estimation of parameters β_0, β_1 are $\beta_0 = 0.0033$ and $\beta_1 = 0.1758$. Then a GARCH (1, 1) model was used to fit the mean model residuals based on equation (3.21) and the estimated results were $A_0 = 4.9134e-6$, $A_1 = 0.1016$, $B_1 = 0.8720$. The estimated variances are then drawn in Figure 3.2. It is obvious that around sample point 1500 the estimated variance shows a significant difference away from the simulated variance as in Figure 3.1. If the mean model is now correctly selected, parameter estimates for the model (3.20) and model (3.21) are listed in Table 3.2. The estimated variance is drawn in Figure 3.3. In order to give a comparison of the estimated variance, the absolute differences between the variance estimated from the linear mean model residuals and the simulated variance is drawn together with the differences between the variance estimated from the nonlinear mean model residuals and the simulated variance in Figure 3.4.

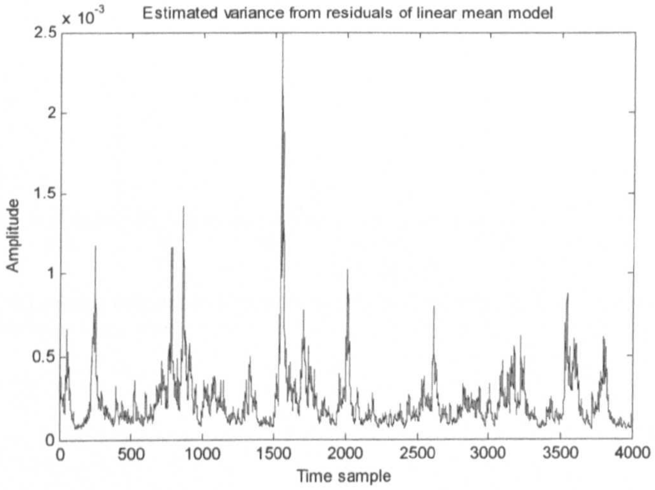


Figure 3.2 Variance from residuals of linear mean model

Table 3.2 Estimate parameters of the model (3.20) and (3.21)

Parameter of mean model (3.20)	Value
a_0	0.0016
a_1	11.4709
a_2	-0.1049
Parameter of variance model (3.20)	Value
A_0	3.3489e-6
A_1	0.0870
B_1	0.8966

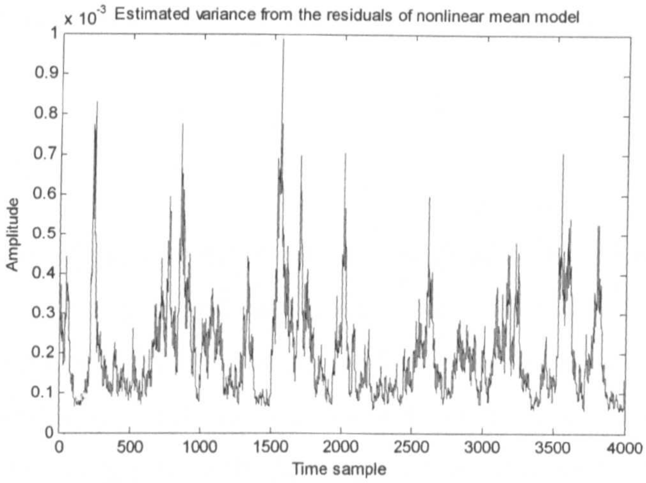


Figure 3.3 Estimated variance from residuals of nonlinear mean model

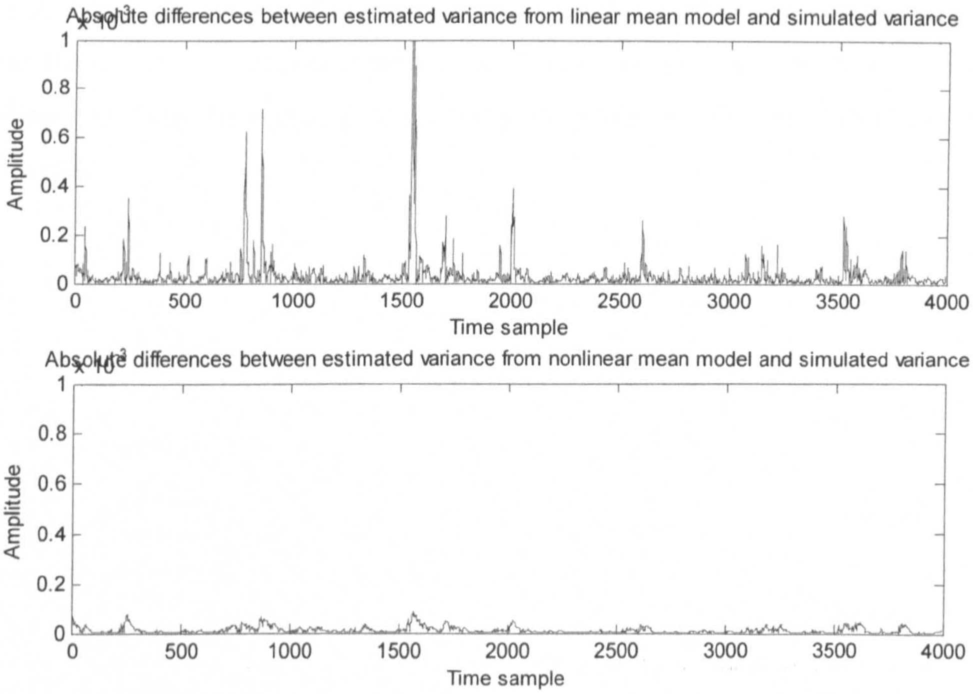


Figure 3.4 Absolute differences between the estimated variance and the simulated variance of linear and nonlinear mean model

It is obvious from Figure 3.4 that the nonlinear model can lead to much more accurate variance estimation especially at the extreme event. The comparison results indicate that if a nonlinear return process has been incorrectly fitted by a linear mean model, the accuracy of the estimation of variance will be affected. Therefore, it is essential to have an accurate mean model before the estimating the variance from the residuals.

3.5 Conclusions

Although fixed terms nonlinear models have already been applied to model the mean process, most of the literature still uses linear models to model the mean process as explained in the introduction. As far as we are aware, there is no existing paper which concentrates on term selection for the non linear mean model. However, it is widely accepted that most real world data is nonlinear. The use of linear models may therefore induce forecast accuracy problems as shown in the example in Section 3.4. However, when the model is nonlinear, higher orders may cause the number of terms in the models to increase significantly. This therefore raises the motivation of modelling the financial return process using nonlinear models but with selected model terms.

Chapter 4: Weighted Orthogonal Forward Regression in the presence of heteroskedastic (GARCH) noise

4.1 Introduction

System identification is commonly used approach to derive mathematical models of unknown dynamical process. Mathematical models are essential for analysis, controller design and forecasting. The identification of linear systems is based on the popular Auto Regressive Moving Average with eXogenous (ARMAX) inputs model (Box and Jenkins, 1970). However, in practice most systems in the real world are nonlinear. The most comprehensive methodology for nonlinear systems identification is based on the Nonlinear Auto Regressive Moving Average with eXogenous inputs (NARMAX) (Billings and Leontaritis, 1981) model. The NARMAX model can describe a wide range of nonlinear systems and includes other popular classes of models such as Volterra, Wiener etc. as special cases.

During NARMAX model estimation, the most difficult part is to decide the structure of the model i.e. which variables and model terms should be included in the model. If redundant variables are falsely selected, the number of terms of the nonlinear model may increase dramatically and the model may turn out to be overestimation of the underlying process and sensitive to the training data set. Model structure selection which is an essential part of the NARMAX system identification methodology ensures that only the relevant model terms are selected in the model. This results in a parsimonious model which describes the underlying dynamical process rather than the estimation data set.

The NARMAX model structure selection is based on the Orthogonal Forward Regression (OFR) algorithm (Billings et al., 1988, 1989, Korenberg et al., 1988, Chen et al., 1989, Billings and Zhu, 1994). The OFR algorithm is also used to estimate the unknown parameter simultaneously with the term selection. The

NARMAX system identification methodology is arguably the most powerful nonlinear modelling methodologies available at the moment.

One of the major assumptions made in the formulation of the NARMAX model and the associated model structure selection and parameter estimation algorithms is that the noise is homoskedastic. However, there are many situations in which the assumption of homoskedastic noise is not valid such as when dealing with the econometrics data, where the variance of the noise is not constant and in many cases can be described by a GARCH process. If the noise is heteroskedastic, this will have a negative impact on the performance of the model term selection and parameter estimation algorithms, which have been derived under the homoskedastic assumption. Specifically, this chapter demonstrates that the ranking of the candidate model terms using the Error Reduction Ratio criteria (Billings et al. 1988, 1989) will be affected leading to an incorrect model structure being selected.

The effects of heteroskedastic noise when performing ordinary least squares (OLS) are well known (Bjorck, 1996) and can be addressed by using weighted least squares (WLS). However, up to now the problem of model structure selection in the presence of heteroskedastic noise has not be investigated or addressed.

The aim of this chapter is to investigate how heteroskedasticity affects the model structure selection and parameter estimation algorithms used to identify NARMAX models and to introduce a new Weighted Orthogonal Forward Regression (WOFR) for NARMAX system identification in the presence of heteroskedastic noise. The main assumption in this work is that the variance of the noise can be modelled by a GARCH process. However, the proposed algorithm can also be used for other types of variance models. The chapter is organised as follows. Section 4.2 introduces the classical OFR algorithm. Section 4.3 investigates analytically the effect of heteroskedastic noise on classical OFR model term selection algorithm and introduces the new WOFR solution to this problem. Section 4.4 describes in detail the iterative implementation of the WOFR algorithm. Section 4.5 presents numerical simulation studies demonstrating the applicability of the proposed algorithm and the conclusion is given in Section 4.6.

4.2 Model structure selection and parameter estimation for NAMRAX models

4.2.1 The Orthogonal Least Squares algorithm

According to Chapter 3, the polynomial NARMAX representation can be written as

$$y(t)=\sum_{i=1}^{M}\theta_i p_i\left(x(t)\right)+\varepsilon(t) \tag{4.1}$$

where $x(t)$ represents $y(t-1),\ldots,y(t-n_y),u(t-d),\ldots,u(t-n_u),\varepsilon(t-1),\ldots,\varepsilon(t-n_\varepsilon)$, $p_i(\ )$ are model terms which are a linear or a nonlinear combination of the variables , $\varepsilon(t)$ is the modelling error, M is the number of all the distinct terms and θ_i are unknown parameters related. And the matrix format of model (4.1) can be written as

$$Y=P\Theta+\Xi \tag{4.2}$$

where $Y=\left[y(1),y(2),\ldots,y(N)\right]^T$, $P=\left[p_1,p_2,\ldots,p_M\right]$, $p_i=\left[p_i(x(1)),p_i(x(2)),\ldots,p_i(x(N))\right]^T$, $\Theta=\left[\theta_1,\theta_2,\ldots,\theta_M\right]^T$ and $\Xi=\left[\varepsilon(1),\varepsilon(2),\ldots,\varepsilon(N)\right]^T$.

Model (4.1) includes all possible polynomial terms for a given polynomial order. In practice, only a small subset of terms is relevant for describing a particular nonlinear dynamical system. Fitting a more complex model than required usually results in overfitting and even instability. It is therefore essential to have in place a method for selecting from the initial set of candidate terms and only the relevant model terms are needed to construct a faithful representation of the underlying dynamical process. One of the first model selection procedures for NARMAX models is based on the Orthogonal Least Squares (OLS) algorithm (Korenberg et al.,1988). Assuming that the matrix P in equation (4.2) is full rank it can be orthogonally decomposed as

$$P=WA \tag{4.3}$$

where A is an $M \times M$ unit upper triangular matrix and W is an $N \times M$ matrix with orthogonal columns $w_1, w_2, ..., w_M$ such as $W^T W = D = diag[d_1, d_2, ..., d_M]$ with $d_i = \langle w_i, w_i \rangle = \sum_{t=1}^{N} w_i(t) w_i(t)$. Equation (4.2) then becomes

$$Y = P(A^{-1}A)\Theta + \Xi = (PA^{-1})(A\Theta) = WG + \Xi \tag{4.4}$$

where $G = [g_1, g_2, ..., g_M]^T$ is an auxiliary parameter vector given by

$$G = D^{-1}W^T Y - D^{-1}W^T \Xi \tag{4.5}$$

. The estimated $\hat{G}$ is therefore given by

$$\hat{G} = D^{-1}W^T Y \tag{4.6}$$

which gives

$$\hat{g}_i = \frac{\langle Y, w_i \rangle}{\langle w_i, w_i \rangle} \tag{4.7}$$

as the original estimates according to Korenberg et al. (1988). The OLS procedure can be summarized as follows:

$$\begin{aligned} & w_0(t) = p_0(t) = 1 \\ & w_i(t) = p_i(t) - \sum_{r=0}^{i-1} \alpha_{ri} w_r(t), \qquad i = 1, ..., M \\ & \alpha_{ri} = \frac{\sum_{t=1}^{N} p_i(t) w_r(t)}{\sum_{t=1}^{N} w_r^2(t)}, \qquad 0 \le r \le i-1 \end{aligned} \tag{4.8}$$

and

$$\hat{g}_0 = \frac{1}{N} \sum_{t=1}^{N} y(t) \tag{4.9}$$

$$\hat{g}_i = \frac{\sum_{t=1}^{N} y(t) w_i(t)}{\sum_{t=1}^{N} w_i^2(t)} \tag{4.10}$$

The estimated parameters $\hat{\theta}_i$, $i = 0, 1, ..., M$ can be calculated as

$$\hat{\theta}_i = \sum_{i=0}^{M} \hat{g}_i c_i \tag{4.11}$$

where

$$\begin{aligned} c_i &= 1 \\ c_m &= -\sum_{r=i}^{m-1} \alpha_{rm} c_r, \quad i < m \le M \end{aligned} \tag{4.12}$$

Multiplying equation (4.4) by itself of both sides the equation becomes

$$Y^T Y = G^T W^T W G + \Xi^T \Xi + G^T W^T \Xi + \Xi^T W G \tag{4.13}$$

After taking expectation of both side, the expectation of two terms $G^T W^T \Xi$ and $\Xi^T W G$ are zero and equation (4.13) becomes

$$\frac{1}{N}\sum_{t=1}^{N} y^2(t) = \frac{1}{N}\sum_{t=1}^{N}\left(\sum_{i=1}^{M} g_i^2 w_i^2(t)\right) + \frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t) \tag{4.14}$$

The contribution to the variance of the output of regressor w_i is given by

$$\frac{1}{N}\sum_{t=1}^{N} g_i^2 w_i^2(t) \tag{4.15}$$

Therefore, the Error Reduction Ratio (ERR) due to the term i can be defined as

$$ERR_i(\%) = \frac{\frac{1}{N}\sum_{t=1}^{N} g_i^2 w_i^2(t)}{\frac{1}{N}\sum_{t=1}^{N} y^2(t)} \times 100, \ i = 1,2,\ldots,M \tag{4.16}$$

The significance of the model terms can then be determined by the value of ERR of each term. The structure of the models can be decided by choosing the terms with ERR bigger than threshold value.

The OLS algorithm has one major drawback which is the algorithm depends on the entire orthogonalization path which means that the ERR for a regressor $p(k)$ depends on its position in the orthononalization sequence. As a result, the ERR does not capture accurately the true significance of a particular model term. Therefore, the Orthogonal Forward Regression (OFR) is proposed to remove the drawback of OLS by Billings et al. (1988).

4.2.2 The Orthogonal Forward Regression algorithm

The OFR is a modified version of the OLS. In the initial stage of OFR, all terms $p_i(t)$ $i=1,2,...,M$ are considered as potential candidates for $w_1(t)$. Then at first iteration of the algorithm, $w_1^i(t)$ is assumed to equal to $p_i(t)$ for all $i=1,2,...,M$. The initial $\hat{g}_1^i$ and ERR_1^i are calculated as

$$\hat{g}_1^i = \frac{\sum_{t=1}^{N} y(t) w_1^i(t)}{\sum_{t=1}^{N} \left(w_1^i(t)\right)^2} \text{ , } ERR_1^i(\%) = \frac{\left(\hat{g}_1^i\right)^2 \sum_{t=1}^{N} w_1^i(t)^2}{\sum_{t=1}^{N} y(t)^2} \tag{4.17}$$

The term with maximum ERR is then selected as the most significant term of the model. The term is then removed from the candidate terms and in the second iteration, the ERR of all remaining candidate terms are re-evaluated. Assuming that $p(j)$ was selected in the first iteration, the second iteration involves computing the following quantities

$$w_2^i(t) = p_i(t) - \alpha_{12}^i w_1(t), \hat{g}_2^i = \frac{\sum_{t=1}^{N} w_2^i(t) y(t)}{\sum_{t=1}^{N} w_2^i(t)^2}, ERR_2^i = \frac{\hat{g}_2^i \sum_{t=1}^{N} w_2^i(t)^2}{\sum_{t=1}^{N} y(t)^2} \tag{4.18}$$

where $i=1,...,M$, $i \neq j$ and

$$\alpha_{12}^i = \frac{\sum_{t=1}^{N} w_1(t) p_i(t)}{\sum_{t=1}^{N} w_1(t)^2} \tag{4.19}$$

The second most significant term will be chosen as the term with the largest ERR from the remaining candidate terms. Subsequently this term will be removed from the candidate terms and the selection process will be continued in a similar manner until the unexplainable variance of the system $1-\sum_{i=1}^{m} ERR_i$ is less than pre-set desired tolerance. In practice, usually $m<M$. The selected orthogonalized model is given by

$$y(t)=\sum_{i=1}^{m} w_i(t)\hat{g}_i+\varepsilon(t) \tag{4.20}$$

which is equivalent to

$$y(t)=\sum_{i=1}^{m} \hat{\theta}_i p_i(x(t))+\varepsilon(t) \tag{4.21}$$

The parameters $\hat{\theta}_i$ can be calculated by equation $\Theta = A\hat{G}$ with $\hat{G}=[\hat{g}_1,\hat{g}_2,...,\hat{g}_m]^T$ and

$$A=\begin{bmatrix} 1 & a_{12} & \cdots & a_{1m} \\ 0 & 1 & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & a_{m-1,m} \\ 0 & \cdots & \cdots & 1 \end{bmatrix} \tag{4.22}$$

The structure of the model can be obtained once the selected model has passed the validation tests such as higher order correlation test introduced by Billings and Voon (1986).

4.3 Weighted Orthogonal Forward Regression

The NARMAX model together with OFR algorithm has been used in practical applications proving to be one of most effective nonlinear system identification methodology. However, there are many situations, particularly when dealing with econometrics data, when the constant noise variance assumption is violated.

4.3.1 OFR performance in the presence of heteroskedastic noise

Let's assume that the variance of the noise is time varying. Specifically the noise $\varepsilon(t)$ is assumed to be given by

$$\varepsilon(t)=z(t)\sigma(t) \tag{4.23}$$

where $z(t)$ is i.i.d. random variables with zero mean and unit variance and $\sigma(t)$ is the time varying standard deviation. Since $z(t)$ is not correlated with $\sigma(t)$, the mean of residuals is

$$\frac{1}{N}\sum_{t=1}^{N}\left(z(t)\sigma(t)\right)=0 \tag{4.24}$$

Let's assume the following regression model terms is

$$y(t)=\sum_{i=1}^{m}\theta_i p_i(t)+\varepsilon(t) \tag{4.25}$$

where $p_i(t), i=1,...,m$ are known model terms and $\theta_1,...,\theta_m$ is unknown parameter vectors. The model (4.25) can be orthogonalized as

$$y(t)=\sum_{i=1}^{m}g_i w_i(t)+\varepsilon(t) \tag{4.26}$$

where g_i and w_i are defined as in section 4.2. The heteroskedastic noise is assumed to be described by a GARCH (1, 1) model (Bollerslev, 1986) as

$$\sigma^2(t)=A_0+A_1\varepsilon(t-1)^2+B_1\sigma^2(t-1) \tag{4.27}$$

where A_0, A_1, B_1 are unknown parameters and the time varying standard deviation can be derived from model (4.27) as

$$\sigma(t)=\sqrt{A_0+A_1\varepsilon(t-1)^2+B_1\sigma^2(t-1)} \tag{4.28}$$

By substituting (4.28) into (4.26), it follows that

$$y(t)=\sum_{i=1}^{m}g_i w_i(t)+z(t)\sqrt{A_0+A_1\varepsilon(t-1)^2+B_1\sigma^2(t-1)} \tag{4.29}$$

The unexplained variance ratio according to the definition of ERR is given by

$$
\begin{aligned}
\frac{\langle \Xi,\Xi\rangle}{\langle Y,Y\rangle} &= \frac{\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t)^2}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \\
&= \frac{\frac{1}{N}\sum_{t=1}^{N}\left(\left(z(t)\sqrt{A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2}\right)^2\right)}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \\
&= \frac{\frac{1}{N}\sum_{t=1}^{N}\left(z(t)^2\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)\right)}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \\
&= \frac{\frac{1}{N}\sum_{t=1}^{N}\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \\
&= \frac{\frac{A_0}{N}+\frac{A_1}{N}\sum_{t-1}^{N}\left(y(t-1)-\sum_{i=1}^{m}g_iw_i(t-1)\right)^2+\frac{B_1}{N}\sum_{t=1}^{N}\sigma(t-1)^2}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \\
&= \frac{\frac{A_0}{N}+\frac{A_1}{N}\sum_{t=1}^{N}\left(y(t-1)^2+\left(\sum_{i=1}^{m}g_iw_i(t-1)\right)^2-2y(t-1)\sum_{i=1}^{m}g_iw_i(t-1)\right)+\frac{B_1}{N}\sum_{t=1}^{N}\sigma(t-1)^2}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \\
&= \frac{\frac{A_0}{N}+\frac{A_1}{N}\sum_{t=1}^{N}y(t-1)^2+\sum_{t=1}^{N}\left(\sum_{i=1}^{m}g_iw_i(t-1)\right)^2-2\sum_{t=1}^{N}y(t-1)\sum_{i=1}^{m}g_iw_i(t-1)+\frac{B_1}{N}\sum_{t=1}^{N}\sigma(t-1)^2}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)}
\end{aligned}
\tag{4.30}
$$

where non-regressor-related terms are

$$
\frac{\frac{A_0}{N}+\frac{A_1}{N}\sum_{t=1}^{N}y(t-1)^2+\frac{B_1}{N}\sum_{t=1}^{N}\sigma(t-1)^2}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \tag{4.31}
$$

The term $\sum_{t=1}^{N}\sigma(t-1)^2$ is assumed to be known because we assumed that the model terms are known. Therefore, according to the definition, the ERR of each regressor when the noise is heteroskedastic is given by

$$\widehat{ERR}_i = \frac{\sum_{t=1}^{N}\left(g_i w_i(t)\right)^2 + \sum_{t=1}^{N}\left(g_i w_i(t-1)\right)^2 - 2\sum_{t=1}^{N} y(t-1) g_i w_i(t-1)}{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)^2\right)} \tag{4.32}$$

Equation (4.32) is derived based on the assumption of known model terms. If the model term is unknown, a simple example is used to demonstrate the impact on ERR of heteroskedastic noise. Considering a simple model with two regressors as

$$y(t) = a_1 p_1(t) + a_2 p_2(t) + \varepsilon(t) \tag{4.33}$$

where $p_1(t), p_2(t)$ are regressors related with lagged $y(t)$ and $\varepsilon(t)$ is heteroskedastic noise and the time varying variance is assumed to be formulated by GARCH (1, 1) model. According to the OFR algorithm, in the first step the ERR for first term is calculated as

$$\begin{aligned}
ERR_1^{1st} &= \frac{\langle Y, p_1\rangle^2}{\langle Y, Y\rangle\langle p_1, p_1\rangle} \\
&= \frac{\left(\sum_{t=1}^{N} y(t) p_1(t)\right)^2}{\sum_{t=1}^{N} y(t)^2 \sum_{t=1}^{N} p_1(t)^2} \\
&= \frac{\left(\sum_{t=1}^{N}\left(a_1 p_1(t) + a_2 p_2(t) + \varepsilon(t)\right) p_1(t)\right)^2}{\sum_{t=1}^{N}\left(a_1 p_1(t) + a_2 p_2(t) + \varepsilon(t)\right)^2 \sum_{t=1}^{N} p_1(t)^2} \\
&= \frac{\left(\sum_{t=1}^{N}\left(a_1 p_1(t) p_1(t) + a_2 p_2(t) p_1(t) + \varepsilon(t) p_1(t)\right)\right)^2}{\sum_{t=1}^{N}\left(\left(a_1 p_1(t)\right)^2 + \left(a_2 p_2(t)\right)^2 + \left(\varepsilon(t)\right)^2 + 2a_1 a_2 p_1(t) p_2(t) + 2a_1 p_1(t)\varepsilon(t) + 2a_2 p_2(t)\varepsilon(t)\right)\sum_{t=1}^{N} p_1(t)^2} \\
&= \frac{\left(\sum_{t=1}^{N}\left(a_1 p_1(t) p_1(t) + a_2 p_2(t) p_1(t)\right)\right)^2}{\sum_{t=1}^{N}\left(\left(a_1 p_1(t)\right)^2 + \left(a_2 p_2(t)\right)^2 + \left(\varepsilon(t)\right)^2 + 2a_1 a_2 p_1(t) p_2(t)\right)\sum_{t=1}^{N} p_1(t)^2}
\end{aligned} \tag{4.34}$$

After substituting $\left(\varepsilon(t)\right)^2$ by $z(t)^2\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)$, the (4.34) becomes

$$ERR_1^{1st}=\frac{\left(\sum_{t=1}^{N}a_1p_1(t)p_1(t)+a_2p_2(t)p_1(t)\right)^2}{\sum_{t=1}^{N}\left(\left(a_1p_1(t)\right)^2+\left(a_2p_2(t)\right)^2+z(t)^2\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)+2a_1a_2p_1(t)p_2(t)\right)\sum_{t=1}^{N}p_1(t)^2}$$

$$=\frac{\left(\sum_{t=1}^{N}a_1p_1(t)p_1(t)+a_2p_2(t)p_1(t)\right)^2}{\left(\sum_{t=1}^{N}\left(a_1p_1(t)\right)^2+\sum_{t=1}^{N}\left(a_2p_2(t)\right)^2+\sum_{t=1}^{N}z(t)^2\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)+2a_1a_2\sum_{t=1}^{N}p_1(t)p_2(t)\right)\sum_{t=1}^{N}p_1(t)^2}$$

$$=\frac{\left(\sum_{t=1}^{N}a_1p_1(t)p_1(t)+a_2p_2(t)p_1(t)\right)^2}{\left(\sum_{t=1}^{N}\left(a_1p_1(t)\right)^2+\sum_{t=1}^{N}\left(a_2p_2(t)\right)^2+\sum_{t=1}^{N}\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)+2a_1a_2\sum_{t=1}^{N}p_1(t)p_2(t)\right)\sum_{t=1}^{N}p_1(t)^2}$$

(4.35)

In the case that the variance of the noise is constant, the amplitude of the noise is different with that of the heteroskedastic noise. Therefore, the value of the regressor $\overline{p}_1(t),\overline{p}_2(t)$ under constant variance noise is different with the value of $p_1(t),p_2(t)$ as the lagged $y(t)$ are assumed to be contained in these regressors. The ERR of the first term in the first step of OFR under constant variance noise is given by

$$
\begin{aligned}
\overline{ERR}_1^{1st} &= \frac{\langle \bar{Y},\bar{p}_1\rangle^2}{\langle \bar{Y},\bar{Y}\rangle\langle \bar{p}_1,\bar{p}_1\rangle} \\
&= \frac{\left(\sum_{t=1}^{N}\bar{y}(t)\bar{p}_1(t)\right)^2}{\sum_{t=1}^{N}\bar{y}(t)^2\sum_{t=1}^{N}\bar{p}_1(t)^2} \\
&= \frac{\left(\sum_{t=1}^{N}\left(a_1\bar{p}_1(t)+a_2\bar{p}_2(t)+\varepsilon(t)\right)\bar{p}_1(t)\right)^2}{\sum_{t=1}^{N}\left(a_1\bar{p}_1(t)+a_2\bar{p}_2(t)+\bar{\varepsilon}(t)\right)^2\sum_{t=1}^{N}\bar{p}_1(t)^2} \\
&= \frac{\left(\sum_{t=1}^{N}\left(a_1\bar{p}_1(t)\bar{p}_1(t)+a_2\bar{p}_2(t)\bar{p}_1(t)\right)\right)^2}{\sum_{t=1}^{N}\left(\left(a_1\bar{p}_1(t)\right)^2+\left(a_2\bar{p}_2(t)\right)^2+\left(\bar{\varepsilon}(t)\right)^2+2a_1a_2\bar{p}_1(t)\bar{p}_2(t)\right)\sum_{t=1}^{N}\bar{p}_1(t)^2}
\end{aligned}
\qquad (4.36)
$$

Comparing with (4.35), both the numerator and denominator of the ERR are different and the term $\sum_{t=1}^{N}\left(A_0+A_1\varepsilon(t-1)^2+B_1\sigma(t-1)^2\right)$ in equation (4.35) is not equal to $\left(\bar{\varepsilon}(t)\right)^2$ in equation (4.36). This indicates that the fact that the noise variance is not constant and will affect the value of the ERR associated with the term $p_1(t)$ in the first step. The ERR of term $p_2(t)$will also be affected in a similar way. The changes of the ERR values may change the order in which the candidate regressors are ranked according to their ERR value and will impact on the term selection procedure. Ultimately, this will lead to an incorrect model structure being identified. This will be demonstrated later in this chapter using numerical simulations.

In order to obtain the correct ERR values in the presence of heteroskedastic noise, it is essential to implement a Weighted Least Squares (Bjorck, 1996) approach where the weighting sequence is selected as the inverse of the time varying standard deviation of the noise. Multiplying $\frac{1}{\sigma(t)}$ on both side of equation (4.26) gives

$$\frac{y(t)}{\sigma(t)}=\frac{\sum_{i=1}^{m}\theta_i p_i(t)}{\sigma(t)}+z(t) \tag{4.37}$$

Denoting $\frac{y(t)}{\sigma(t)}$ by $y'(t)$ and $\frac{p_i(t)}{\sigma(t)}$ as p_i', equation (4.3) can be written as

$$y'(t)=\sum_{i=1}^{m}\theta_i p_i'(t)+z(t) \tag{4.38}$$

The relative unexplainable variance of the system is given by $\frac{E\left(z(t)^2\right)}{E\left(y'(t)^2\right)}$. Of course, in practice the time-varying standard deviation is unknown. To deal with this problem, a weighted orthogonal forward regression algorithm is introduced in the next section. Under the assumption that the structure of the variance model is known, the WOLS algorithm allows the identification of the correct NARMAX mean model structure and the estimation of the weighting sequence and of the parameters of both the variance and mean models.

4.3.2 Weighted Orthogonal Forward Regression

The model (4.37) can be rewritten in matrix form as follows

$$Q^{-1}Y=Q^{-1}P\Theta+Z \tag{4.39}$$

where Q^{-1} is a diagonal matrix whose elements are the inverse of the standard deviation $\sigma(t)$ at each sample point and Z is the $z(t)$ vector in (4.37). The part $Q^{-1}P\Theta$ can be rewritten as $Q^{-1}P\Theta=Q^{-1}P\left(\bar{A}^{-1}\bar{A}\right)\Theta$ and where $Q^{-1}P\bar{A}^{-1}$ can be represented by $\bar{W}$ and $\bar{A}\Theta$ by $\bar{G}$ and (4.39) can be rewritten as

$$Q^{-1}Y=\bar{W}\bar{G}+Z \tag{4.40}$$

The auxiliary parameter vector $\bar{G}$ can be approximated as

$$\hat{\bar{G}}=\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^TQ^{-1}Y \tag{4.41}$$

The estimation of $\bar{G}$ is unbiased and sufficient because

$$\begin{aligned}
E\left(\hat{\bar{G}}\right) &= E\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T Q^{-1}Y\right) \\
&= E\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T\left(\bar{W}\bar{G}+Z\right)\right) \\
&= E\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T\bar{W}\bar{G}+\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T Z\right) \\
&= E\left(\bar{G}\right)+E\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T Z\right) \\
&= \bar{G}
\end{aligned} \tag{4.42}$$

and

$$\begin{aligned}
\operatorname{var}\left(\hat{\bar{G}}\right) &= E\left(\left(\hat{\bar{G}}-\bar{G}\right)^2\right) \\
&= E\left(\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T Q^{-1}Y-\bar{G}\right)^2\right) \\
&= E\left(\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T\left(\bar{W}\bar{G}+Z\right)-\bar{G}\right)^2\right) \\
&= E\left(\left(\bar{G}+\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T Z-\bar{G}\right)^2\right) \\
&= E\left(\left(\bar{W}^T\bar{W}\right)^{-1}\bar{W}^T Z^T Z\bar{W}\left(\bar{W}^T\bar{W}\right)^{-1}\right) \\
&= \left(\bar{W}^T\bar{W}\right)^{-1}
\end{aligned} \tag{4.43}$$

where $E(\ )$ operation is the expectation and $E\left(Z^T Z\right)$ is equal to 1. According to the procedure of OFR, the first most important term is selected by letting $\bar{g}_i = Q^{-1}p_i$ and the $\bar{g}_i$ for each term in the first iteration is

$$\bar{g}_i^{1st} = \frac{\left\langle Q^{-1}Y, Q^{-1}p_i\right\rangle}{\left\langle Q^{-1}p_i, Q^{-1}p_i\right\rangle} \tag{4.44}$$

. Therefore, the weighted ERR for each term in the first iteration is

$$WERR_i^1 = \frac{\left\langle Q^{-1}Y, Q^{-1}p_i\right\rangle^2}{\left\langle Q^{-1}Y, Q^{-1}Y\right\rangle\left\langle Q^{-1}p_i, Q^{-1}p_i\right\rangle}\times 100\% \tag{4.45}$$

Then, in later nth iteration the OFR will be applied by using

$$\bar{g}_i^{nth} = Q^{-1}p_i - \sum_{k=1}^{n-1}\bar{a}_{ki}\bar{g}_k \tag{4.46}$$

where $\bar{a}_{ki} = \dfrac{\left\langle \bar{g}_k, Q^{-1} p_i \right\rangle}{\left\langle \bar{g}_k, \bar{g}_k \right\rangle}$ and $\bar{g}_k$ is the selected term from step 1 to n-1. The weighted ERR in step n becomes

$$WERR_i^n = \frac{\left\langle Q^{-1}Y, \bar{g}_i^n \right\rangle}{\left\langle Q^{-1}Y, Q^{-1}Y \right\rangle \left\langle \bar{g}_i^n, \bar{g}_i^n \right\rangle} \times 100\% \tag{4.47}$$

Assuming the $\sigma_{\min}$ is the minimal value among all the diagonal elements of the matrix Q and $Q_{\min} = \sigma_{\min} I$ where I is an identity matrix with the same rank as Q, there is

$$\left\| Q^{-1} \right\| \le \left\| Q_{\min}^{-1} \right\| \tag{4.48}$$

where $\left\| \ \right\|$ denotes the norm of the matrix. Similarly, there is

$$\left\| Q^{-1} \right\| \ge \left\| Q_{\max}^{-1} \right\| \tag{4.49}$$

Therefore, the term $\left\langle Q^{-1}Y, Q^{-1}p_i \right\rangle$ in equation (4.47) has the property of

$$\left\langle Q_{\max}^{-1}Y, Q_{\max}^{-1}p_i \right\rangle \le \left\langle Q^{-1}Y, Q^{-1}p_i \right\rangle \le \left\langle Q_{\min}^{-1}Y, Q_{\min}^{-1}p_i \right\rangle \tag{4.50}$$

As $Q_{\max}^{-1} = \dfrac{1}{\sigma_{\max}} I$ and $Q_{\min}^{-1} = \dfrac{1}{\sigma_{\min}} I$, the equation (4.50) can be written as

$$\frac{1}{\sigma_{\max}^2} \left\langle Y, p_i \right\rangle \le \left\langle Q^{-1}Y, Q^{-1}p_i \right\rangle \le \frac{1}{\sigma_{\min}^2} \left\langle Y, p_i \right\rangle \tag{4.51}$$

Accordingly,

$$\begin{aligned} \frac{1}{\sigma_{\max}^2} \left\langle Y, Y \right\rangle \le \left\langle Q^{-1}Y, Q^{-1}Y \right\rangle \le \frac{1}{\sigma_{\min}^2} \left\langle Y, Y \right\rangle \\ \frac{1}{\sigma_{\max}^2} \left\langle p_i, p_i \right\rangle \le \left\langle Q^{-1}p_i, Q^{-1}p_i \right\rangle \le \frac{1}{\sigma_{\min}^2} \left\langle p_i, p_i \right\rangle \end{aligned} \tag{4.52}$$

Therefore, in the first step of the weighted OFR the weighted ERR can have a bound as

$$\frac{\dfrac{1}{\sigma_{\max}^2} \left\langle Y, p_i \right\rangle \dfrac{1}{\sigma_{\max}^2} \left\langle Y, p_i \right\rangle}{\dfrac{1}{\sigma_{\min}^2} \left\langle Y, Y \right\rangle \dfrac{1}{\sigma_{\min}^2} \left\langle p_i, p_i \right\rangle} \le WERR_i^{1st} \le \frac{\dfrac{1}{\sigma_{\min}^2} \left\langle Y, p_i \right\rangle \dfrac{1}{\sigma_{\min}^2} \left\langle Y, p_i \right\rangle}{\dfrac{1}{\sigma_{\max}^2} \left\langle Y, Y \right\rangle \dfrac{1}{\sigma_{\max}^2} \left\langle p_i, p_i \right\rangle} \tag{4.53}$$

As $ERR_i^1 = \dfrac{\langle Y, p_i \rangle^2}{\langle Y, Y \rangle \langle p_i, p_i \rangle}$, the weighted ERR in the first step can change within the bounds as

$$\frac{\sigma_{\min}^4}{\sigma_{\max}^4} ERR_i^{1st} \le WERR_i^{1st} \le \frac{\sigma_{\max}^4}{\sigma_{\min}^4} ERR_i^{1st} \tag{4.54}$$

Therefore, the ERR_i^{1st} can be written as $ERR_i^{1st} = WERR_i^{1st} + e_i^{1st}$ where e_i is the fluctuation caused by the heteroskedastic noise. Considering the extreme case in previous example model (4.33), if the $ERR_1^{1st} > ERR_2^{1st}$ however $\dfrac{\sigma_{\min}^4}{\sigma_{\max}^4}$ $WERR_1^{1st} = \dfrac{\sigma_{\min}^4}{\sigma_{\max}^4} ERR_1^{1st}$ and $WERR_2^{1st} = \dfrac{\sigma_{\max}^4}{\sigma_{\min}^4} ERR_2^{1st}$, there is very likely that the rank of the significance will change because $\dfrac{\sigma_{\max}^4}{\sigma_{\min}^4}$ may be far larger than $\dfrac{\sigma_{\min}^4}{\sigma_{\max}^4}$. Even the shift e_i^{1st} may not cause term selection changes in the first step we think it may shift the ERR value in later steps the term selection in later step may be impacted. Simulations will be given to indicate this situation. Therefore, the introducing of weights allows the OFR algorithm evaluate the real contribution of each candidate terms in the presence of heteroskedsatic noise and produce efficient and unbiased parameter estimation.

However, the time varying variance is usually unknown. If the structure of the variance model is know, (a GARCH model in our case) the variance can be estimated iteratively. The iterative WOFR procedure can be summarized as follows:

(1) Estimate parameters for a candidate model using LS method and derived the modelling residuals.

(2) Estimate the time varying variances from the modelling error of step (1).

(3) Apply WOFR to the reference mean model and select the most important terms.

(4) Re-estimate the time varying variance and re-calculate the parameter of the selected mean model with WOFR.

(5) Repeat step (4) until the parameters of the mean model converge.

4.4 Simulations

In this section, several examples are provided to illustrate the efficiency of the new WOFR algorithm for model term detection under heteroskedastic noise. In all examples, only one step ahead prediction is considered.

4.4.1 Example 1: linear AR model

The first example considers a simple linear mean model which is given by

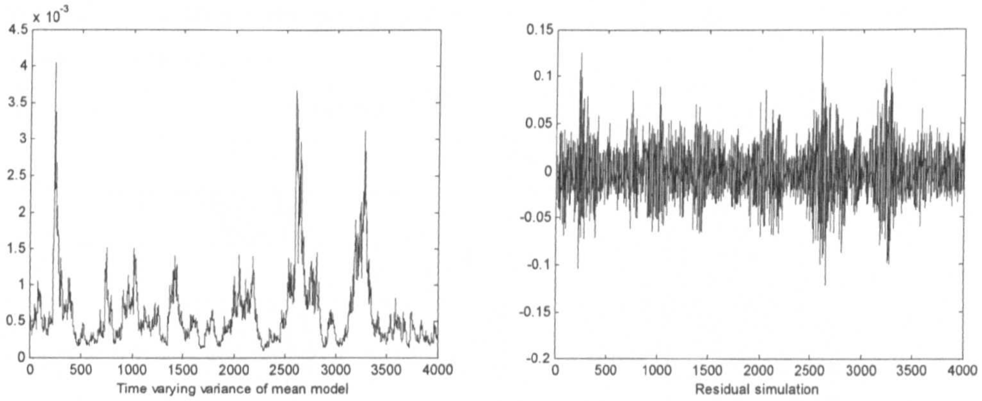
$$y(t)=a_0+a_1y(t-1)+a_2y(t-2)+a_3y(t-3)+\varepsilon(t) \tag{4.55}$$

It is assumed that the time-varying variance of the noise is described by a GARCH(1,1) model. The variances and the residuals are simulated iteratively as $\varepsilon(t)=z(t)\sigma(t)$and can be given by the variance $\sigma(1)$and a normally distributed variable $z(t)$ with unit variance. Parameters of linear mean model and GARCH model are listed in the Table 4.1.

Table 4.1 Parameters of simulated mean model (4.55) and GARCH(1,1) model

Parameters of linear mean model (4.55)			
a_0	a_1	a_2	a_3
1e-2	0.6	-0.55	0.25
Parameters of GARCH(1,1) model			
A_0	A_1	B_1	
3e-6	0.075	0.92	

The number of data points that were generated was 4000. The generated residuals and the time varying variances are shown in Figure 4.1.



(a) (b)

Figure 4.1 (a) Time varying variances simulation figure and (b) residuals simulation figure

Firstly, the simulated data was used to identify the model using the WOFR algorithm. The set of the candidate model terms was generated based on a second order nonlinear NAR(5) model. The candidate model term set are given by

$$P = [Constant\ ,\ y(t-1), y(t-1)^2, y(t-2), y(t-2)^2, y(t-3), y(t-3)^2, y(t-4), y(t-4)^2, y(t-5),\ y(t-5)^2, y(t-1)y(t-2), y(t-1)y(t-3), y(t-1)y(t-4), y(t-1)y(t-5)\ ,\ y(t-2)y(t-3)\ ,\ y(t-2)y(t-4)\ ,\ y(t-2)y(t-5)\ ,\ y(t-3)y(t-4), y(t-3)y(t-5), y(t-4)y(t-4)] \tag{4.56}$$

Table 4.2 shows ranking of the terms according to the standard and weighted OFR algorithm.

Table 4.2 Term ranking generated by OFR and WOFR based on the candidate terms equation (4.56) using data in Figure 4.1

Rank	Standard OFR		Weighted OFR	
	Term	ERR (%)	Term	ERR (%)
1	$y(t-1)$	20.665	Constant	25.761
2	$y(t-4)$	9.4014	$y(t-1)$	6.9721
3	$y(t-2)$	5.2491	$y(t-2)$	10.440
4	Constant	6.4203	$y(t-3)$	3.2534
5	$y(t-3)$	2.3972	$y(t-4)$	0.10805
6	$y(t-1)y(t-4)$	0.20181	$y(t-1)y(t-4)$	0.42568e-1
7	$y(t-4)y(t-5)$	0.10697	$y(t-5)$	0.32366e-1
8	$y(t-1)y(t-1)$	0.94316e-1	$y(t-1)y(t-1)$	0.14393e-1
9	$y(t-1)y(t-5)$	0.87810e-1	$y(t-1)y(t-5)$	0.27351e-1
10	$y(t-4)y(t-4)$	0.46674e-1	$y(t-1)y(t-2)$	0.26762e-1
11	$y(t-1)y(t-2)$	0.41264e-1	$y(t-1)y(t-3)$	0.22775e-1
12	$y(t-1)y(t-3)$	0.13911	$y(t-2)y(t-3)$	0.29313e-1
13	$y(t-2)y(t-2)$	0.14985e-1	$y(t-2)y(t-2)$	0.20029e-1
14	$y(t-5)$	0.1649e-1	$y(t-2)y(t-5)$	0.10955e-1
15	$y(t-5)y(t-5)$	0.68806e-2	$y(t-5)y(t-5)$	0.72217e-2
16	$y(t-3)y(t-5)$	0.14162e-1	$y(t-4)y(t-5)$	0.11567e-1
17	$y(t-3)y(t-4)$	0.41362e-2	$y(t-4)y(t-4)$	0.20384e-2
18	$y(t-2)y(t-5)$	0.37594e-2	$y(t-3)y(t-5)$	0.16895e-2
19	$y(t-2)y(t-4)$	0.53579e-2	$y(t-3)y(t-4)$	0.10955e-1
20	$y(t-2)y(t-3)$	0.49603e-2	$y(t-3)y(t-3)$	0.72218e-2
21	$y(t-3)y(t-3)$	0.46408e-2	$y(t-2)y(t-4)$	0.24510e-2

It is very clear that OFR algorithm ranks incorrectly the term $y(t-4)$ which is not part of the model (4.55). When WOFR is used, the value of WERR associated with the term $y(t-4)$ is very small and the term is ranked after all four correct terms in the model given in (4.55). Based on the WOFR initial ranking, the first 4 terms are obviously more significant than the rest of the terms in term of ERR values. Therefore, the cutoff value can be selected as 1 (Wei and Billings, 2004) and the first four terms are treated as selected as mean model terms. The parameters are then re-estimated iteratively until these converge. The estimated parameters for mean model and variance model after 10 iterations are listed in Table 4.4.

Table 4.3 Estimates of the parameters of the mean and variance models after 10 iterations of the WOFR algorithm

Mean Model			
Term	Parameter estimates	Standard Deviation	Real parameter
Constant	0.010245	2.1783e-7	1e-2
$y(t-1)$	0.55698	2.6569e-4	0.6
$y(t-2)$	-0.52717	2.7283e-4	-0.55
$y(t-3)$	0.25355	2.6253e-4	0.25
GARCH (1,1) model			
Parameter	Parameter estimates	Standard Deviation	Real parameter
A_0	6.05929e-6	2.7744e-12	3e-6
A_1	0.078558	7.8478e-5	0.075
B_1	0.91229	8.9073e-5	0.92

The parameters of mean model and variance model at each iteration are plotted in Figure 4.2 and Figure 4.3. It can be seen clearly from Figure 4.2 and 4.3 that the parameters estimation converged very quickly for both the mean model and GARCH model. The autocorrelation of squared residuals $\varepsilon(t)$ and squared standard error $z(t)^2 = \frac{\varepsilon(t)^2}{\hat{\sigma}(t)^2}$ where $\hat{\sigma}(t)$ is the estimated standard deviation from the last iteration of WORF are plotted in Figure 4.4 with blue line are the 95% confidence interval.

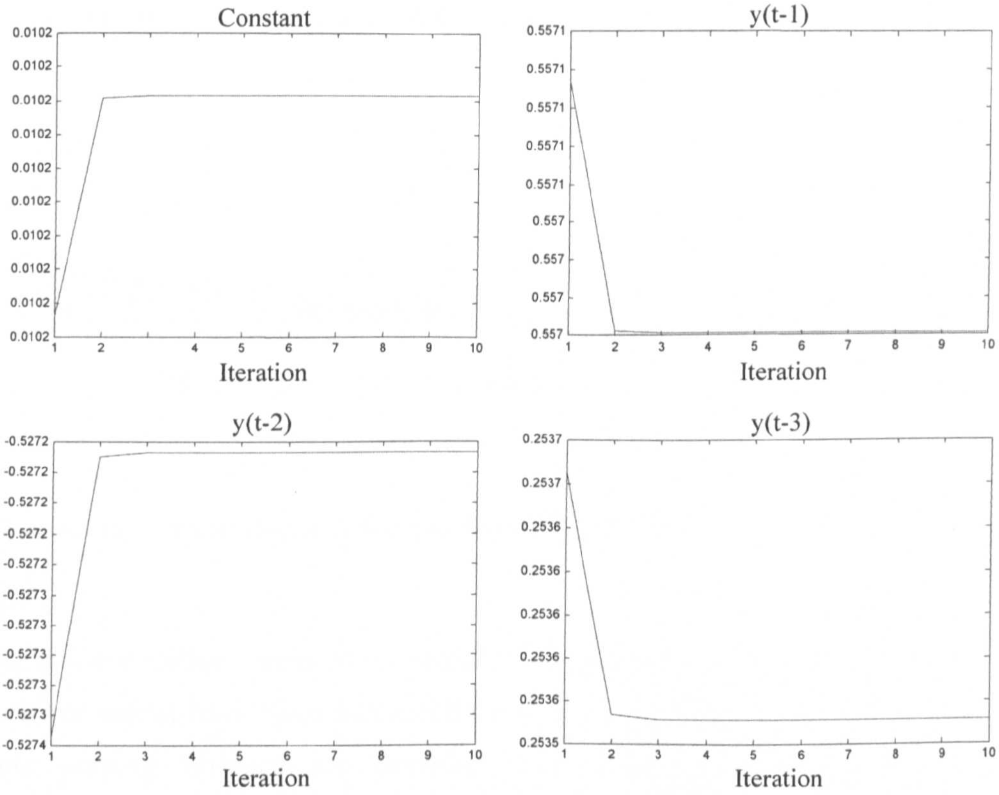


Figure 4.2 Evolution of the parameter estimates at each iteration for the selected linear mean model

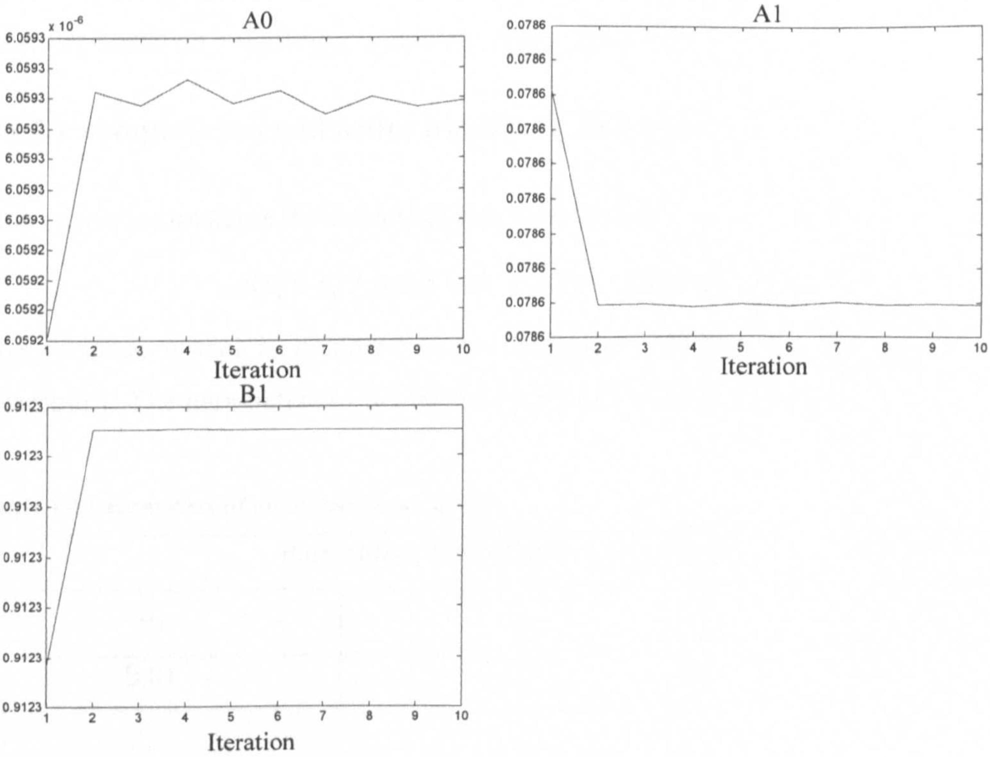
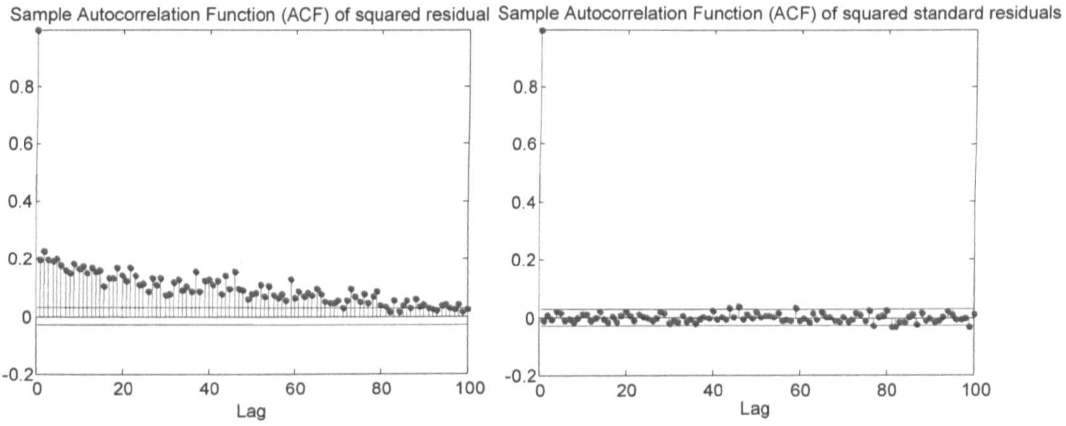


Figure 4.3 Evolution of the parameter estimates at each iteration for GARCH (1, 1) model



(a) (b)

Figure 4.4 (a) Autocorrelation of the squared residuals $\varepsilon(t)^2$ and (b) squared standard residuals $z(t)^2$

The autocorrelation results of the standard residuals indicate that both the mean and variance model have been sufficiently modelled and the information related with time varying variance are removed after modelling. This example clearly demonstrates the need to use the WOFR algorithm in the presence of heteroskedastic noise and the performance of the proposed algorithm in identifying a linear model structure based on a nonlinear candidate model structure.

4.4.2 Example 2: second order nonlinear AR model

This example considers the following nonlinear model

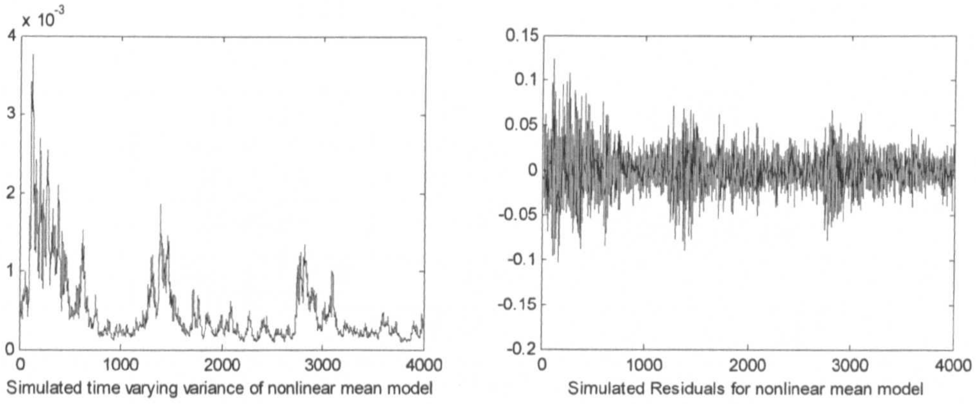
$$y(t) = a_0 + a_1 y(t-1)^2 + a_2 y(t-2) + \varepsilon(t) \tag{4.57}$$

The variance model is assumed to be the same GARCH (1, 1) model used in Example 1. The parameters of the model (4.57) are listed in Table 4.4

Table 4.4 Parameters of nonlinear mean model

Parameters of nonlinear mean model		
a_0	a_1	a_2
0.01	2.5	-0.5

The sample length is 4000 and the simulated time varying variance and residuals are drawn in Figure 4.5.



(a) (b)

Figure 4.5 (a) Simulated time varying variance and (b) simulated residuals for nonlinear mean model

The candidate model terms generated based on a NAR (2, 5) model are the same as the ones used in the previous example in equation (4.56). The results of applying the OFR and WOFR term selection algorithms are listed in Table 4.5 with all the terms ranked in order of significance as measured by the ERR and WERR respectively.

Table 4.5 Term ranking generated by the OFR and WOFR term selection algorithms based on the candidates in equation (4.56)

Rank	Standard OFR		Weighted OFR	
	Term	ERR (%)	Term	ERR (%)
1	$y(t-4)$	10.618	Constant	12.987
2	$y(t-2)$	5.1401	$y(t-2)$	19.560
3	Constant	12.925	$y(t-1)y(t-1)$	1.3224
4	$y(t-1)y(t-1)$	3.4736	$y(t-1)$	0.86001e-1
5	$y(t-1)$	0.18118	$y(t-3)y(t-5)$	0.47232e-1
6	$y(t-3)y(t-5)$	0.11344	$y(t-5)$	0.31229e-1
7	$y(t-2)y(t-2)$	0.45682e-1	$y(t-2)y(t-2)$	0.29616e-1
8	$y(t-4)y(t-5)$	0.29381e-1	$y(t-1)y(t-4)$	0.26163e-1
9	$y(t-1)y(t-5)$	0.18653e-1	$y(t-2)y(t-4)$	0.16261e-1
10	$y(t-1)y(t-3)$	0.26716e-1	$y(t-3)y(t-4)$	0.10894e-1
11	$y(t-3)$	0.25510e-1	$y(t-4)$	0.20663e-1
12	$y(t-3)y(t-3)$	0.37283e-1	$y(t-5)y(t-5)$	0.39993e-2
13	$y(t-5)y(t-5)$	0.38987e-1	$y(t-1)y(t-3)$	0.40746e-2
14	$y(t-3)y(t-4)$	0.27419e-1	$y(t-2)y(t-3)$	0.38241e-2
15	$y(t-2)y(t-3)$	0.19808e-1	$y(t-1)y(t-5)$	0.42474e-2
16	$y(t-2)y(t-5)$	0.27419e-1	$y(t-4)y(t-5)$	0.26509e-2
17	$y(t-4)y(t-4)$	0.17887e-2	$y(t-2)y(t-5)$	0.33273e-2
18	$y(t-1)y(t-2)$	0.11393e-2	$y(t-3)$	0.16864e-2
19	$y(t-5)$	0.85652e-3	$y(t-1)y(t-2)$	0.15553e-3
20	$y(t-2)y(t-4)$	0.35365e-3	$y(t-3)y(t-3)$	0.23747e-4
21	$y(t-1)y(t-4)$	0.29914e-4	$y(t-4)y(t-4)$	0.19142e-5

It can be seen from Table 4.5 that the heteroskedastic noise can impact the OFR term selection. The term $y(t-4)$ is wrongly selected as the most important term by the OFR algorithm. The weighted OFR algorithm however, correctly identified the correct set of terms and the terms having grey background in Table 4.5 are selected as the mean model terms. Subsequently, the iterative reweighted procedure is used to refine the parameter estimates for both the mean model and the variance model. The parameters of selected mean model and variance model at each iteration are drawn in Figure 4.6 and Figure 4.7. It can be seen from Figure 6 that the estimated parameters converge very quickly as in the linear mean model example. The parameters in the last iteration are listed in Table 4.6 and the autocorrelation test of squared residuals and squared standard errors are drawn in Figure 4.7. The autocorrelation tests of the standard squared residuals indicate that both the mean and variance model have been sufficiently modelled and the information related with time varying variance are removed after modelling. The simulation results of nonlinear AR models can verify the efficiency of the iterative WOFR algorithm in model term selection when the noise is heteroskedastic.

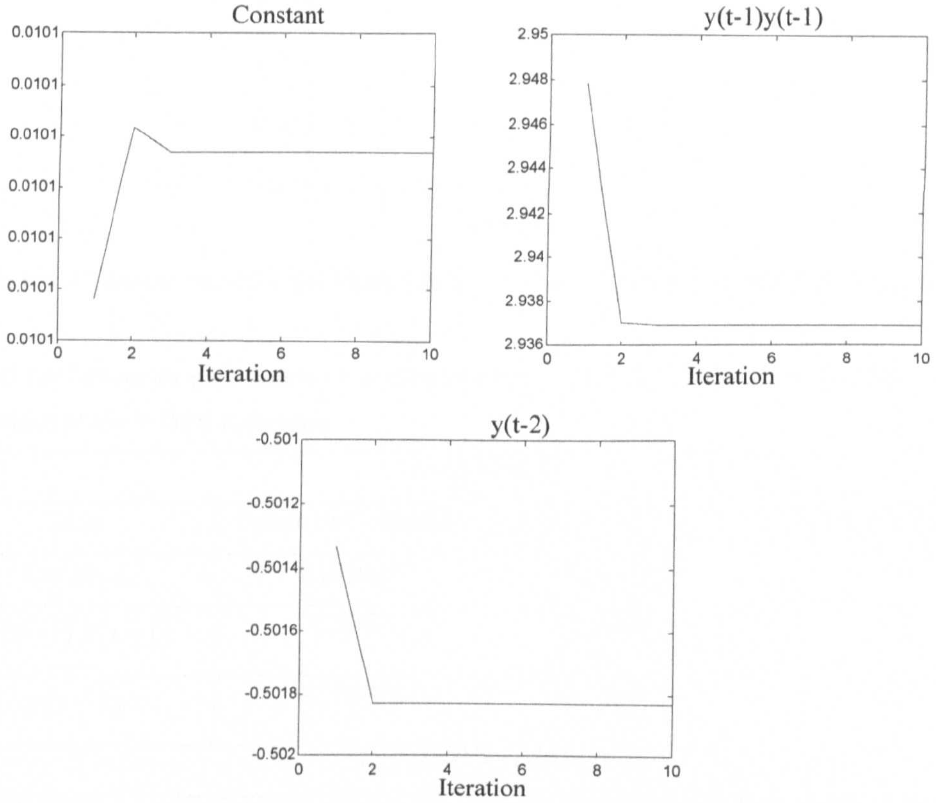


Figure 4.6 Evolution of the parameter estimates at each iteration for the selected nonlinear mean model as in Table 4.5

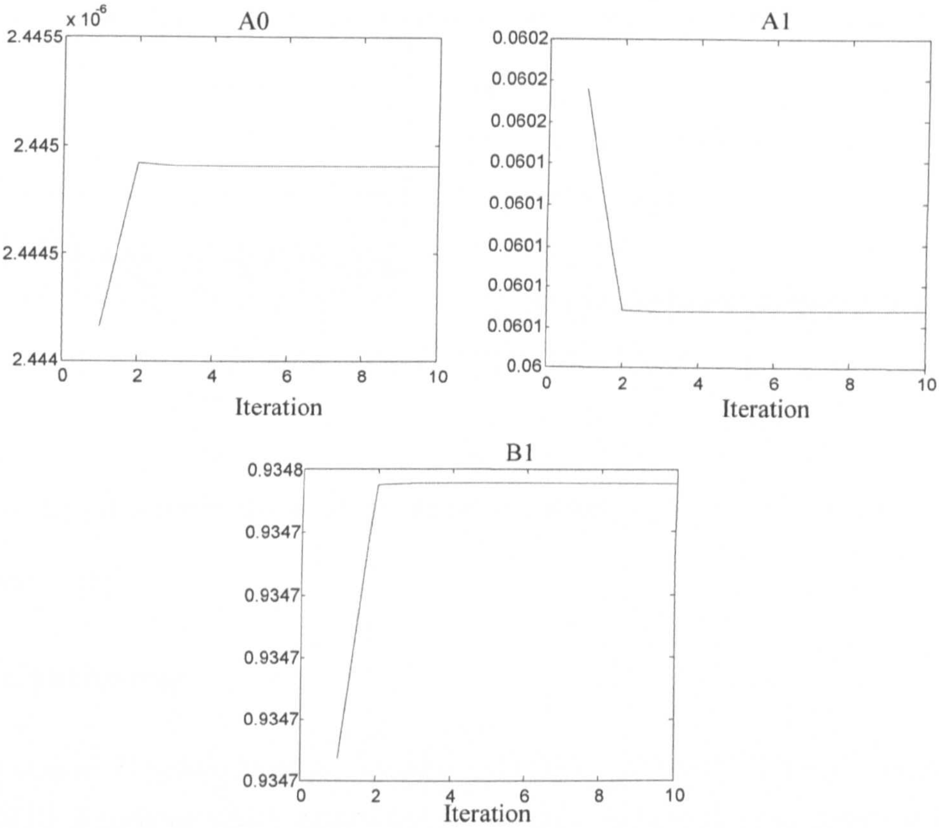
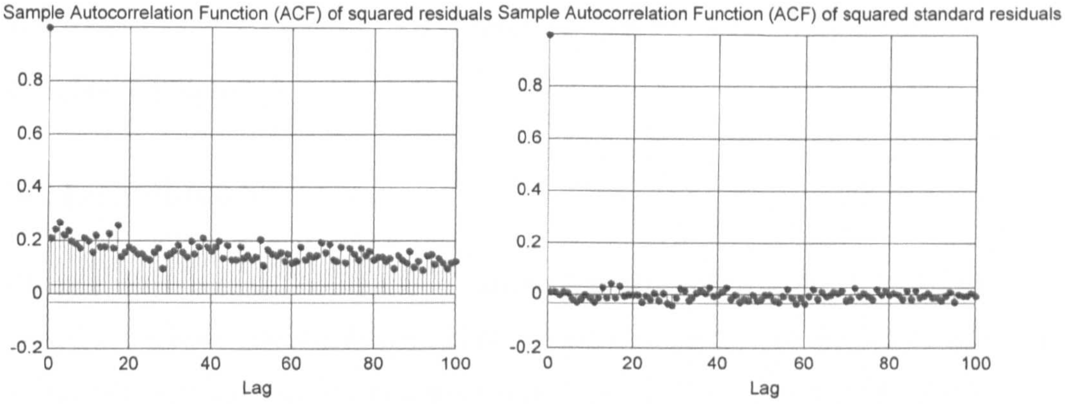


Figure 4.7 Evolution of the parameter estimates at each iteration for GARCH (1, 1) model

Table 4.6 Estimated parameters of nonlinear mean model and GARCH (1, 1) model after 10 iterations of the WOFR algorithm

Mean Model			
Term	Parameter estimates	Standard Deviation	Real parameter
Constant	0.010127	1.0144e-7	0.01
$y(t-1)y(t-1)$	2.9369	2.6569e-4	2.5
$y(t-2)$	-0.50182	2.7283e-4	-0.5
GARCH (1,1) model			
Parameter	Parameter estimates	Standard Deviation	Real parameter
A_0	2.44491e-6	6.1360e-13	3e-6
A_1	0.060068	4.5975e-5	0.075
B_1	0.93476	5.0408e-5	0.92



(a) (b)

Figure 4.8 (a) Autocorrelation of the squared residuals $\varepsilon(t)^2$ and (b) squared standard residuals $z(t)^2$

4.5 Conclusions

This chapter has introduced an iterative Weighted Orthogonal Forward Regression (WOFR) algorithm which addresses the problem of model term selection in the presence of heteroskedastic noise. The main assumption here is that the variance model structure is known. Specifically, we have investigated the case in which the variance is described by a GARCH (1, 1) mode. This is often the case when dealing with econometric data sets. Theoretical results demonstrating the effects of heteroskedasticity on conventional OFR which assumed that the noise has constant variance has been derived. Once the model terms have been selected, the parameters and variance are re-estimated iteratively to achieve convergence. Correlation tests are used to indicate the sufficiency of both the variance model and mean model. Two numerical simulation examples were used to illustrate the negative impact of the heteroskedastic noise on conventional OFR and to demonstrate the effectiveness of the WOFR algorithm for model structure selection and parameter estimation in the presence of such noise.

Chapter 5: Cross validation between NARMAX and GARCH model

5.1 Introduction

Chapter 4 has shown the impact of heteroskedastic noise on the mean model term selection. When a significant term in the mean model is not selected or a linear mean model is falsely used to fit a nonlinear mean process, the resulting residuals will also cause inaccurate estimation using Maximum Likelihood (ML). Therefore, the variance model estimated using ML will be affected simultaneously. Since there appears to be no relevant publication to explain this problem, this chapter derives theoretical results to show the impact on the ML estimator when the mean model is not well selected and where some information is still contained in the residuals. In a practical application, it is natural to try and develop a statistical method to verify this impact and to simultaneously validate the mean model and the variance model. Cross Validation (CV) (Devijver and Kittler, 1982) is a statistical technique which is used to analyse the prediction performance of a fitted model and is suitable for both large data and small data sets. CV can give an indication of how accurately a model can forecast over independent test data and there are several CV methods which are commonly used in practice. Those methods include the holdout method, K-fold CV method and Leave-one-out CV method. The holdout method is the simplest and the data set is split into two parts. One part is used to fit the model while the other part is used to test the forecast performance. The K-fold CV method is an improved method based on the holdout method as the data set is divided into k subsets and the holdout method is repeated k times. The leave-one-out CV is an extreme case of the K-fold CV method with K equal to the data length. Since the purpose of this chapter is to validate time series models, leave-one-out CV is not appropriate here as autocorrelation will always exist in the time series data and removing one time sample in the middle of the data may lead to discontinuity of the data. Therefore, the holdout CV method will be used in this chapter.

Once the mean model and the variance model are fitted, the i.i.d. assumption of the standard residuals will be tested to validate the fitted models. The Brock, Dechert and Scheinkman (BDS) test which was firstly introduced by Brock, Dechert, and Scheinkman (1987) uses a nonparametric technique and has good testing power against a wide class of data departing from i.i.d. as nonstationarity, nonlinearity, and deterministic chaos. Many researchers have analyzed this method for example Abhyankar et al. (1995), Barnett et al. (1993), Chavas and Holt (1991), etc. and the BDS test has also been proved to have the ability to detect nonlinearity in econometric models (Brock et al., 1991). Brock, Hsieh and LeBaron (1991) used Monte Carlo simulations to obtain the distribution of the BDS test from the standardised residuals of a specified GARCH model and Bollerslev et al. (1993) concluded that the BDS test has the power to test the ARCH effect and the i.i.d. assumption of standardised residuals when the variance model or mean model is miss-specified. Hsieh (1993) applied the BDS test to the logarithm of currency prices and concluded that none of the currency prices exhibited i.i.d. Barnett et al. (1997) showed that BDS test has power against a wide range of nonlinearity and i.i.d. Brock et al. (1996) applied the BDS test to the standardized residuals of GARCH models and Brooks and Heravi (1999) suggested using the BDS test jointly with other tests to detect the mis-specifications of the model. Ahlstedt (1998) tested standardized residuals of a GARCH (1, 1) model to currency data and Caporale et al. (2004) applied the BDS test to test the adequacy of the GARCH specifications by using Monte Carlo analysis. Mangani (2009) used the BDS test to verify the significant of the GARCH (1, 1) model when fitting to data from the JSE Securities Exchange of South Africa. Therefore, the BDS test has been commonly adopted by many publications to test the i.i.d. assumption of the GARCH class of models. However, when the test fails it is usually difficult to distinguish whether the mean model is not sufficient or the GARCH model is not sufficient. If the most significant terms of the mean model are accurately selected, the autocorrelation of the residuals should be below the 95% significance line and the one step ahead prediction errors should be close to random. When the variance sequence is well approximated by a GARCH model, the i.i.d. assumption of the standard residuals and standard prediction errors should not be rejected. When the mean model is not correctly selected, the autocorrelation of the prediction error will have outliers and the i.i.d. assumption will be rejected when the variance is not well predicted. Accordingly, this chapter proposes to combine cross validation with the BDS test to validate both the mean model and the GARCH model. It was shown in Chapter 4 that the WOFR algorithm can improve the term selection of the mean model when heteroskedastic noise exists and therefore, during the simulations in this study, the WOFR algorithm will be employed.

The purpose of this chapter is to determine the impact of the term selection of the mean model on the ML estimator, extend the application of the WOFR algorithm, and propose a new method to validate both the mean model and the variance model. Simulations will be given to indicate the effectiveness of this new approach. Section 5.2 introduces the normal distribution testing methods. Section 5.3 derives theoretically the impact of the mean model term selection on the ML estimator. Section 5.4 provides a description of the BDS test and Section 5.5 gives the general procedure of CV to validate both the fitted mean model and the variance model. Section 5.6 illustrates the effectiveness of the above methods by simulation examples and Section 5.7 is the conclusions.

5.2 Testing the distribution assumption

Since the pioneering work of Engle (1982), a time varying variance is commonly fitted using a GARCH class of models and the ML estimation method. The ML method is a popular statistical method but the distribution function of the residuals has to be known apriori to formulate the likelihood function. Violation of the distribution assumption may lead to an inaccurate statistical inference. Therefore, before applying the ML method, the distribution assumption needs to be tested. The normal distribution is one of the most commonly used distributions and the distribution function is

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.1}$$

where σ is the standard deviation and μ is the mean. The density function is symmetric with the shape of a bell curve which is drawn in Figure 5.1 with $\mu=0,\sigma=1$. However, in the real world, the normal distribution assumption will not be satisfied. Chapter 2 has introduced the QMLE method which can compensate the estimation errors for a fat tail distribution. Therefore, if the fat tail has been detected, the QMLE method will be adopted to compensate the estimation error.

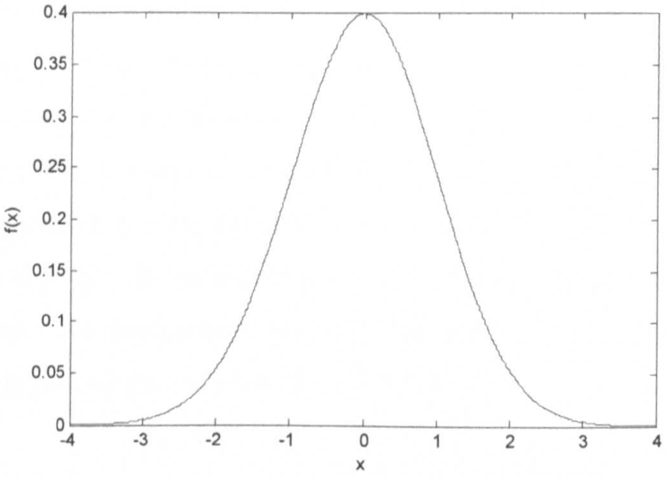


Figure 5.1 Normal distribution density function plot

5.2.1 The JB test

There are many methods which have been developed to test for normality. The JB test which was first proposed by Jarque and Bera (1980) uses Lagrange multiplier procedure to derive the test statistic. The JB test statistic can be expressed as

$$JB=\frac{n}{6}\left(S^2+\frac{1}{4}K^2\right) \tag{5.2}$$

where S is the skewness statistic and K is kurtosis statistic. S and K can be calculated from

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{\frac{3}{2}}} \qquad (5.3)$$

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{2}} - 3 \qquad (5.4)$$

where x_i is the sample value at time i, $\bar{x}$ is the mean of the data and N is data length. The JB statistic has a Chi-squares distribution with two degrees of freedom and the null hypothesis is that the data is normally distributed. If the data is normal, the sample skewness statistic is 0 and the kurtosis statistic is 0 and any departure of those two statistics will increase the JB test statistic. However, when the sample length is small, the chi-square distribution becomes right skewed and the test statistic is over sensitive. Therefore, the JB test is usually used to test large data sets. The Table 5.1 lists the JB testing results on matlab simulated normally distributed data and Student t distributed data. Table 5.1 shown that the JB statistics rejected the Student t distribution as the value of JB test is bigger than 3.

Table 5.1 Tests on simulated normal data and Student t distributed data

Data distribution	normal	Student t
Data length	4000	4000
Skewness	0.0226	0.0449
Kurtosis	2.8911	4.5008
JB statistic	2.3173	376.7372

5.2.2 The QQ plot

The QQ plot is a graphical method to compare the probability distributions of two groups of data based on quantiles. The quantile means the percentage of points below a given value. A reference line which is the line $y = x$ is drawn before the test and if the distributions of the two data sets are similar, the quantiles should approximately overlap the reference line. Otherwise, the greater the departure is, the more different the two testing distributions will have. There are several advantages of the QQ plot. As the QQ plot is non parametric, the tested data lengths are not necessarily equal. The QQ plot can test many distributions and it can also test the distribution aspect simultaneously as the symmetry, and the presence of the outliers. Figure 5.2 demonstrates the QQ plot of a normal distributed data with the normal reference line and Figure 5.3 shows that the normality has been rejected by a group of data generated from a Student t distribution. Since the probability density of the Student t distribution has fatter tails, the quantiles of the sample data departs from the reference line on both sides. Therefore, the QQ plot can indicate directly about the tail behaviour of the tested distribution.

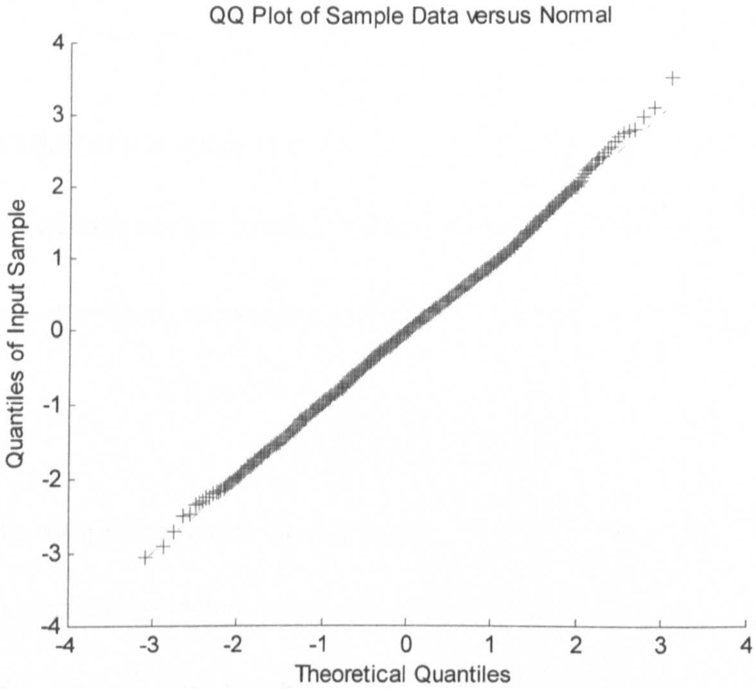


Figure 5.2 QQ plot of data generated from normal distribution versus normal

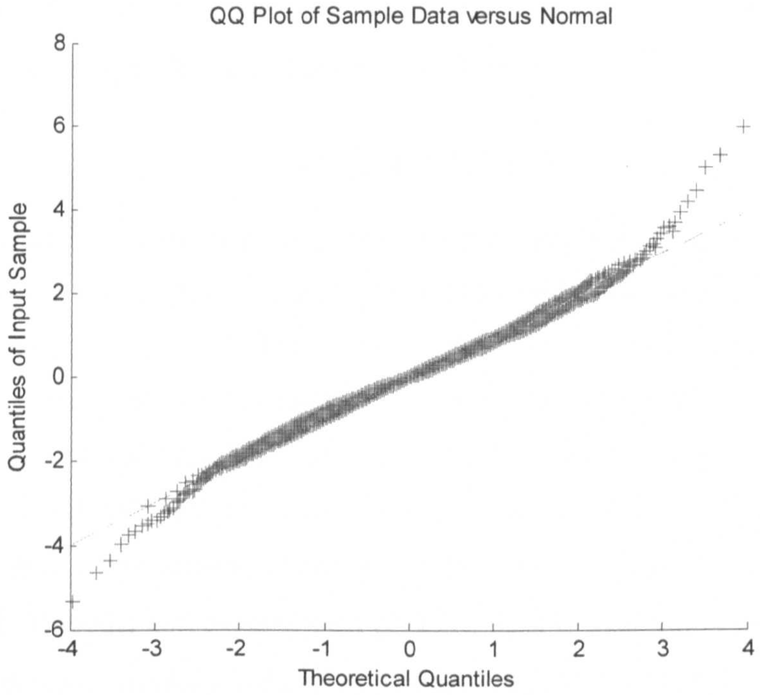


Figure 5.3 QQ plot of data generated from Student t distribution versus normal

5.3 Impact of the mean model term selection on the ML estimator

Consider an orthogonalized true process model defined in Chapter 4 equation (4.13) as

$$y(t)=\sum_{i=1}^{M} g_i w_i(t)+\varepsilon(t) \tag{5.5}$$

If part of the significant terms have not been selected and contained in the terms $\sum_{i=m+1}^{Mp} g_i w_i(t)$ and the selected terms are assumed to be $\sum_{i=1}^{m} g_i w_i(t)$ where the m represents the selected term number, Mp represents the number of unselected terms and $M=m+Mp$, then equation (5.5) can be written as

$$y(t)=\sum_{i=1}^{m} g_i w_i(t)+\sum_{i=m+1}^{Mp} g_i w_i(t)+\varepsilon(t) \tag{5.6}$$

If the terms used to model the mean do not contain all the significant terms, especially as a linear model is commonly used to model nonlinear process, therefore information of related to unselected significant terms may be contained in the residuals. Accordingly, the modelling error $e(t)$ becomes

$$e(t)=\sum_{i=m+1}^{Mp} g_i w_i(t)+\varepsilon(t) \tag{5.7}$$

Then this modelling error will be used to estimate the parameters of the GARCH model using the ML method. According to the ML estimation routine, the likelihood of the logarithm of the probability density function will be maximized and the parameter updating algorithm-Berndt, Hall, Hall and Hausman (1974) (BHHH) algorithm is commonly used. The first step of when using the ML estimator is to specify the initial parameters to calculate the likelihood values at the first iteration. In order to give a comparison, consider that the variance process is generated by a GARCH (1, 1) model and assume that the initial specified parameters are a_0^1,a_1^1,β_1^1. Then the calculated variance in the first iteration of BHHH algorithm is

$$\hat{h}^1(t)=a_0^1+a_1^1 e(t-1)^2+\beta_1^1\hat{h}^1(t-1) \tag{5.8}$$

where $\hat{h}^1$ represents the calculated time varying variance in the first iteration with the modelling error $e(t)$. The true time varying variance calculated in the first iteration is

$$h^1(t)=a_0^1+a_1^1\varepsilon(t-1)^2+\beta_1^1 h^1(t-1) \tag{5.9}$$

where h^1 represents the volatility calculated from the correct model residuals $\varepsilon(t)$. According to the BHHH algorithm, the initial specified variance is $\frac{a_0^1}{1-a_1^1-\beta_1^1}$.

Therefore, the initial variances $\hat{h}^1(1)$ of model (5.8) and $h^1(1)$ of model (5.9) are equal. As the variance is iteratively generated by the GARCH model during the estimation, the estimated variance $\hat{h}^1(2)$ from the modelling error $e(t)$ at the second sample point is

$$\hat{h}^1(2)=a_0^1+a_1^1 e(1)^2+\beta_1^1\hat{h}^1(1) \tag{5.10}$$

Assume that the sample length is N , then after $N-1$ iterations the estimated variance at sample point N is

$$\begin{aligned}\hat{h}^1(N) &= a_0^1 + \beta_1^1 a_0^1 + \cdots + \left(\beta_1^1\right)^{N-2} a_0^1 + a_1^1 e(N-1)^2 + \beta_1^1 a_1^1 e(N-2)^2 \\ &\quad + \cdots + \left(\beta_1^1\right)^{N-2} a_1^1 e(1)^2 + \left(\beta_1^1\right)^{N-1} \hat{h}^1(1)\end{aligned} \tag{5.11}$$

The true variance estimated in the first iteration of the BHHH algorithm at sample point N is

$$\begin{aligned}h^1(N) &= a_0^1 + \beta_1^1 a_0^1 + \cdots + \left(\beta_1^1\right)^{N-2} a_0^1 + a_1^1 \varepsilon(N-1)^2 + \beta_1^1 a_1^1 \varepsilon(N-2)^2 \\ &\quad + \cdots + \left(\beta_1^1\right)^{N-2} a_1^1 \varepsilon(1)^2 + \left(\beta_1^1\right)^{N-1} h^1(1)\end{aligned} \tag{5.12}$$

Therefore, the differences between $\hat{h}^1(N)$and$h^1(N)$is

$$\begin{aligned}\hat{h}^1(N) - h^1(N) &= a_1^1\left(e(N-1)^2 - \varepsilon(N-1)^2\right) + \beta_1^1 a_1^1\left(e(N-2)^2 - \varepsilon(N-2)^2\right) \\ &\quad + \cdots + \left(\beta_1^1\right)^{N-2} a_1^1\left(e(1)^2 - \varepsilon(1)^2\right)\end{aligned} \tag{5.13}$$

According to equation (5.7),

$$\begin{aligned}(e(t))^2 &= \left(\varepsilon(t) + \sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2 \\ &= \varepsilon^2(t) + 2\varepsilon(t)\sum_{i=m+1}^{Mp} g_i w_i(t) + \left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2\end{aligned} \tag{5.14}$$

Therefore,

$$e(t)^2 - \varepsilon(t)^2 = 2\varepsilon(t)\sum_{i=m+1}^{Mp} g_i w_i(t) + \left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2 \tag{5.15}$$

After substituting equation (5.15) in, equation (5.13) becomes

$$\begin{aligned}\hat{h}^1(N) - h^1(N) &= a_1^1\left(2\varepsilon(N-1)\sum_{i=m+1}^{Mp} g_i w_i(N-1) + \left(\sum_{i=m+1}^{Mp} g_i w_i(N-1)\right)^2\right) \\ &\quad + \beta_1^1 a_1^1\left(2\varepsilon(N-2)\sum_{i=m+1}^{Mp} g_i w_i(N-2) + \left(\sum_{i=m+1}^{Mp} g_i w_i(N-2)\right)^2\right) \\ &\quad + \cdots + \left(\beta_1^1\right)^{N-2} a_1^1\left(2\varepsilon(1)\sum_{i=m+1}^{Mp} g_i w_i(1) + \left(\sum_{i=m+1}^{Mp} g_i w_i(1)\right)^2\right) \\ &= \Delta^1(N)\end{aligned} \tag{5.16}$$

where $\Delta^1(N)$ represents the difference between the real estimated variance and the inaccurate variance estimation in the first iteration of the BHHH algorithm. The parameters of the GARCH model will then be updated by the equation

$$\theta^{(i+1)} = \theta^{(i)} + \lambda_i \left(\sum_{t=1}^{N} \frac{\partial l_t}{\partial \theta} \frac{\partial l_t}{\partial \theta'} \right)^{-1} \sum_{t=1}^{N} \frac{\partial l_t}{\partial \theta} \tag{5.17}$$

where $\theta^{(i)}$ represents the parameter estimates of the GARCH model parameters a_0^i, a_1^i, β_1^i at iteration i of the BHHH routine and $\frac{\partial l_t}{\partial \theta}$ is the first order differentiation of the log likelihood function. The logarithm likelihood function is given by $l_t(\theta) = -\frac{1}{2}\log h(t) - \frac{1}{2}\frac{\varepsilon(t)^2}{h(t)}$ and the first partial differential equation of likelihood subject to the parameter is $\frac{\partial l_t}{\partial \theta} = \frac{1}{2}\frac{1}{h(t)}\frac{\partial h(t)}{\partial \theta}\left(\frac{\varepsilon(t)^2}{h(t)} - 1\right)$. In the case of an inaccurate mean model, the first differentiation at the first iteration of the BHHH routine becomes

$$\begin{aligned}
\frac{\partial l_t}{\partial \theta} &= \frac{1}{2}\frac{1}{\hat{h}^1(t)}\frac{\partial \hat{h}^1(t)}{\partial \theta}\left(\frac{e(t)^2}{\hat{h}^1(t)} - 1\right) \\
&= \frac{1}{2}\frac{1}{h^1(t)+\Delta^1(t)}\frac{\partial \hat{h}^1(t)}{\partial \theta}\left(\frac{\left(\sum_{i=m+1}^{Mp} g_i w_i(t) + \varepsilon(t)\right)^2}{h^1(t)+\Delta^1(t)} - 1\right) \\
&= \frac{1}{2}\frac{1}{h^1(t)+\Delta^1(t)}\left(\frac{\left(\sum_{i=m+1}^{Mp} g_i w_i(t) + \varepsilon(t)\right)^2}{h^1(t)+\Delta^1(t)} - 1\right)\frac{\partial \hat{h}^1(t)}{\partial \theta}
\end{aligned} \tag{5.18}$$

where

$$\begin{aligned}\frac{\partial \hat{h}^1(t)}{\partial \theta} &= \Theta^1(t)+\beta_1^1\frac{\partial \hat{h}^1(t-1)}{\partial \theta}\\ &= \Theta^1(t)+\beta_1^1\left(\Theta^1(t-1)+\beta_1^1\frac{\partial \hat{h}^1(t-2)}{\partial \theta}\right)\\ &= \Theta^1(t)+\beta_1^1\Theta^1(t-1)+\ldots+\left(\beta_1^1\right)^{N-2}\Theta^1(t-N+2)+\left(\beta_1^1\right)^{N-1}\frac{\partial \hat{h}(t-N+1)}{\partial \theta}\end{aligned} \tag{5.19}$$

, vector $\Theta^1(t)=\left[1,e(t)^2,\hat{h}^1(t-1)\right]^T$ and $\frac{\partial \hat{h}^1(t-N+1)}{\partial \theta}$ is the initial differentiation.

As the initial differentiation is usually a set value, the value can be the same as in the true mean model case. Therefore, the differences of the first differential of the variance between the inaccurate mean model and the real mean model is

$$\begin{aligned}\frac{\partial \hat{h}^1(t)}{\partial \theta}-\frac{\partial h^1(t)}{\partial \theta} = &\left[0,\left(\sum_{i=m+1}^{Mp} g_i w_i(t-1)\right)^2+2\sum_{i=m+1}^{Mp} g_i w_i(t-1)\varepsilon(t-1),\Delta^1(t)\right]^T+\\ &\beta_1^1\left[0,\left(\sum_{i=m+1}^{Mp} g_i w_i(t-1)\right)^2+2\sum_{i=m+1}^{Mp} g_i w_i(t-2)\varepsilon(t-2),\Delta^1(t-1)\right]^T\\ &+\ldots+\left(\beta_1^1\right)^{N-2}\begin{bmatrix}0,\left(\sum_{i=m+1}^{Mp} g_i w_i(t-1)\right)^2\\ +2\sum_{i=m+1}^{Mp} g_i w_i(t-N+1)\varepsilon(t-N+1),\Delta^1(t-N+2)\end{bmatrix}^T\end{aligned} \tag{5.20}$$

Equation (5.18) and (5.20) indicate that the slope of the GARCH parameter convergence is different at the beginning of the BHHH algorithm between the inaccurate mean model and the real mean model. The calculation of the likelihood of the inaccurate mean model therefore becomes

$$\hat{l}_t^1(\theta)=-\frac{1}{2}\log\left(h^1(t)+\Delta^1(t)\right)-\frac{1}{2}\frac{\left(\sum_{i=m+1}^{Mp} g_i w_i(t)+\varepsilon(t)\right)^2}{h^1(t)+\Delta^1(t)} \tag{5.21}$$

Comparing equation (5.21) with $l_t^1(\theta)=-\frac{1}{2}\log h^1(t)-\frac{1}{2}\frac{\varepsilon(t)^2}{h^1(t)}$, the likelihood calculation in equation (5.21) at the first iteration is different. Therefore, the slope of the parameter updating of the GARCH model will be different from the first iteration between the real mean model and the inaccurate mean model. According to the sensitivity of the ML algorithm, the convergence of the parameters will be affected by the trend and initial conditions. The final estimation of the GARCH parameters will be different accordingly. For the correct mean model, the time varying variance $h(t)$ is the conditional expectation of the squared residuals $E\left(\varepsilon(t)^2\right)$ because $h(t)=E\left(\left(\varepsilon(t)-E\left(\varepsilon(t)\right)\right)^2\right)$ and $E\left(\varepsilon(t)\right)=0$. However, when the mean model is inaccurate the information of any missing significant terms will be contained in the modelling errors, and the conditional expectation of the modelling error will no longer be zero as

$$E\left(e(t)\right)=E\left(\varepsilon(t)+\sum_{i=m+1}^{Mp} g_i w_i(t)\right)=E\left(\varepsilon(t)\right)+E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\neq 0 \qquad (5.22)$$

Therefore, the conditional variance becomes

$$\begin{aligned}
\hat{h}(t) &= E\left(\left(e(t)-E\left(e(t)\right)\right)^2\right)\\
&= E\left(e(t)^2-2e(t)E\left(e(t)\right)+\left(E\left(e(t)\right)\right)^2\right)\\
&= E\left(\begin{array}{l}\varepsilon(t)^2+2\varepsilon(t)\sum\limits_{i=m+1}^{Mp} g_i w_i(t)+\left(\sum\limits_{i=m+1}^{Mp} g_i w_i(t)\right)^2-\\ 2\left(\varepsilon(t)+\sum\limits_{i=m+1}^{Mp} g_i w_i(t)\right)E\left(\varepsilon(t)+\sum\limits_{i=m+1}^{Mp} g_i w_i(t)\right)+\left(E\left(\varepsilon(t)+\sum\limits_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2\end{array}\right)\\
&= E\left(\varepsilon(t)^2\right)+0+E\left(\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2\right)-2E\left(\left(\varepsilon(t)+\sum_{i=m+1}^{Mp} g_i w_i(t)\right)E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)\\
&\quad +E\left(E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2\\
&= E\left(\varepsilon(t)^2\right)+E\left(\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2\right)-2E\left(\varepsilon(t)E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)\\
&\quad -2E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)+\left(E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2
\end{aligned}$$

$$= E\left(\varepsilon(t)^2\right) + E\left(\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2\right) - \left(E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2$$

$$= h(t) + E\left(\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2\right) - \left(E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2 \tag{5.23}$$

According to Jensen's inequality,

$$\left(E\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2 \le E\left(\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)^2\right) \tag{5.24}$$

Therefore, $\hat{h}(t) \ge h(t)$. If there is no significant term contained in $\sum_{i=m+1}^{Mp} g_i w_i(t)$, $E\left(\left(\sum_{i=m+1}^{Mp} g_i w_i(t)\right)\right)^2$ will be insignificant compared with the variance of $y(t)$ and $\hat{h}(t) \approx h(t)$, otherwise the estimated variance will be inaccurate.

It is clearly shown by the above theoretical derivation that the variance estimation and parameters of the GARCH model will be seriously affected when the mean model is not accurate because some significant terms are not selected. Accordingly, it is essential to develop a statistic method to test for this scenario.

5.4 BDS test for i.i.d. assumption

The BDS test which was initially proposed by Brock, Dechert and Scheinkman (1987) is one of the commonly used methods to test the i.i.d. assumption of the underlying data series. It is also well known that the BDS test can be used to test the modelling residuals to check the goodness of fit (Brock et al., 1991). Since one assumption of the GARCH model is that the standardized residuals are distributed as i.i.d., the BDS test appears to be the right tool to test this assumption. The general BDS procedure is

(1) Assume that the standardized residual $z(t)$ is calculated from $z(t)=\frac{\varepsilon(t)}{\sqrt{h(t)}}$, where $h(t)$represents the estimated variance from the mean model residual $\varepsilon(t)$, the data length is N and n is the embedding dimension. Then the residual series are embedded into n-dimensional vectors by taking each n successive points in the series. The residual series is then converted into a series of vectors as

$$\begin{aligned} z_1^n &= \left[z_1, z_2, \ldots, z_n\right] \\ z_2^n &= \left[z_2, z_3, \ldots, z_{n+1}\right] \\ &\vdots \\ z_{N-n}^n &= \left[z_{N-n}, z_{N-n+1}, \ldots, z_N\right] \end{aligned} \tag{5.25}$$

(2) The correlation integral which measures the spatial correlation is then calculated by adding a number of points $(i,j)$where $1 \le i \le N$and $1 \le j \le N$ in the n-dimensional space within a radius $\gamma$of each other as

$$C_{\gamma,n}=\frac{1}{N(N-1)}\sum_{i\neq j} I_{i,j;\gamma} \tag{5.26}$$

where

$$\begin{aligned} I_{i,j;\gamma} &= 1 \qquad if \left\| z_i^n - z_j^n \right\| \le \gamma \\ &= 0 \qquad otherwise \end{aligned} \tag{5.27}$$

(3) If $C_{\gamma,n} \approx \left[C_{\gamma,1}\right]^n$, then the underlying data series is distributed as i.i.d. As pointed out by Lin (1997), if the ratio of $\frac{N}{n}$ is greater than 200, the value of $\frac{\gamma}{\sigma}$ where σ is the standard deviation ranges from 0.5 to 2 and the value of dimension n is between 2 and 5, the quantity $\left[C_{\gamma,n}-\left(C_{\gamma,1}\right)^n\right]$ will be distributed as asymptotic normal with zero mean and variance $V_{\gamma,n}$ defined as

$$V_{\gamma,n}=4\left[K^{n}+2\sum_{j=1}^{n-1}K^{n-j}C_{\gamma}^{2j}+(n-1)^{2}C_{\gamma}^{2n}-n^{2}KC_{\gamma}^{2n-2}\right] \tag{5.28}$$

where

$$\begin{aligned} K=K_{\gamma}&=\frac{6}{N(N-1)(N-2)}\sum_{i<j<N}h_{i,j,N;\gamma} \\ h_{i,j,N;\gamma}&=\frac{\left[I_{i,j;\gamma}I_{j,N;\gamma}+I_{i,N;\gamma}I_{N,j;\gamma}+I_{j,i;\gamma}I_{i,N;\gamma}\right]}{3} \end{aligned} \tag{5.29}$$

(4) The BDS test statistic will be defined as

$$BDS_{\gamma,n}=\frac{\sqrt{N}\left[C_{\gamma,n}-\left(C_{\gamma,1}\right)^{n}\right]}{\sqrt{V_{\gamma,n}}} \tag{5.30}$$

The hypothesis of i.i.d. will be rejected when BDS test statistic is greater than 1.96 or less than -1.96.

Therefore, if the BDS statistic of the standardized residuals is greater than 1.96 or less than -1.96, the i.i.d. assumption will be rejected and this indicates that the GARCH model may not be accurately fitted.

5.5 Cross validation between the mean model and the variance model

CV is a statistical method to estimate how well the model is fitted to the underlying data. Commonly used CV methods include the holdout method, K-fold CV method and Leave-one-out CV method. Holdout method is the simplest kind of CV and in the holdout method the data is usually split into two parts. One part of the data is used to fit the model and the other part is used to test the prediction ability of the fitted model. Since the standardized residuals are derived by the mean model residuals and the standard deviation, either an inaccurate mean model or variance model will induce rejection of the i.i.d. testing statistic. Therefore, there are three situations which will result in the rejection of the i.i.d. assumption as

(1) Biased mean model approach but unbiased GARCH model fitting.

(2) Unbiased mean model approach but biased GARCH model fitting.

(3) Biased mean model approach and biased GARCH model fitting.

Note that the biased mean model means that the significant terms of the mean model are not fully selected and the information related by unselected significant terms is contained in the modelling errors. The biased GARCH model means that the variance estimation from GARCH is inaccurate. Inaccurate variance estimation may be caused by selecting the wrong structure of the GARCH model or using the wrong distribution density function during ML estimation.

In this chapter, the data will be split into two and the first half data will be used to estimate the mean model and the variance model. Then the estimated mean model and variance model parameters from the first half data are used to predict and the one-step-ahead prediction errors will be calculated. If the mean model and the variance model is accurate enough, the standardized prediction error will not reject the i.i.d. assumption. Therefore, the general procedure of cross validation between the mean model and the variance model is:

For the first half data

(1) Fit the data using the WOFR algorithm and select the significant terms the mean model.

(2) Test the autocorrelation of the residuals and validate the fitted mean model.

(3) Test the assumption of i.i.d. using standardized residuals.

For the second half data

(1) Use the parameters of the mean model and the variance model which are estimated from the first half data to calculate the one-step-ahead prediction errors.

(2) Estimate the variance from the prediction errors and test the i.i.d. assumption using standardized prediction errors.

(3) Recursively calculate the conditional variance with the prediction errors and the GARCH parameters estimated from the first half and test the i.i.d. of the standardized prediction errors.

When all the tests are satisfied, the mean model and the variance model should be simultaneously valid. It is essential to test the second half data twice as only when the prediction from the variance model is consistent with the variance estimated from the prediction error, that the mean model and the variance model are valid to compute the predictions.

According to Chapter 3, most of the GARCH model related publications ignored the significance of the mean model accuracy. As proved in Chapter 4 and this chapter, an inaccurate mean model will induce inaccurate estimation of the GARCH parameters and variance estimation. Therefore, the CV of the mean model and the variance model has significant meaning for improving the forecast abilities of the variance model.

5.6 Simulations

In order to demonstrate the effectiveness of the method described in Section 5.5, a nonlinear mean model is used to generate the simulation data. The nonlinear model is

$$y(t)=a_0+a_1y(t-1)+a_2y(t-1)^2+z(t)\sqrt{h(t)} \tag{5.31}$$

where a_0,a_1,a_2 are parameters, $z(t)$ is random variable distributed as $N(0,1)$ and $h(t)$ represents the time varying variance. The parameters for model (5.31) are listed in Table 5.2.

Table 5.2 Parameters for model (5.31)

a_0	a_1	a_2
0.007	-0.11	12

A GARCH (1, 1) model is also used to generate the time varying variance and the model is

$$h(t)=A_0+A_1\left(z(t-1)\sqrt{h(t-1)}\right)^2+B_1h(t-1) \tag{5.32}$$

and the parameters of GARCH model are listed in Table 5.3.

Table 5.3 Parameters for GARCH

A_0	A_1	B_1
3e-7	0.075	0.924

A series of data with length 5000 was then generated using the mean model 5.31 with $y(1)$=0 and $h(1)=\frac{A_0}{1-A_1-B_1}$. In order to avoid any initial value effects, the first half data used for CV was taken from sample point 1500 to 3500 and the second half data was taken from sample point 3501 to 5000. In order to give a comparison, a linear model will be used to model the nonlinear mean process to demonstrate that an inaccurate mean model can cause inaccurate prediction of the variance. A linear reference AR (4) model are firstly used to model the mean process and the estimation results from the OFR algorithm are listed in Table 5.4.

Table 5.4 ERR of linear reference mean model

Term	ERR
Constant	0.48960e-1
y(t-1)	0.71236e-2
y(t-4)	0.42205e-2
y(t-3)	0.11110e-4
y(t-2)	0.47821e-4

According to Table 5.4, the terms, Constant, $y(t-1)$, $y(t-4)$ should be selected as the significant terms of the linear mean model and the linear mean model is given by

$$y(t)=a_0+a_1y(t-1)+a_2y(t-4)+\varepsilon(t) \tag{5.33}$$

where $\varepsilon(t)$ are the residuals. As described in Section 5.5, the autocorrelation of the residuals needs to be checked initially and therefore the autocorrelation of linear mean model residuals are drawn in Figure 5.4.

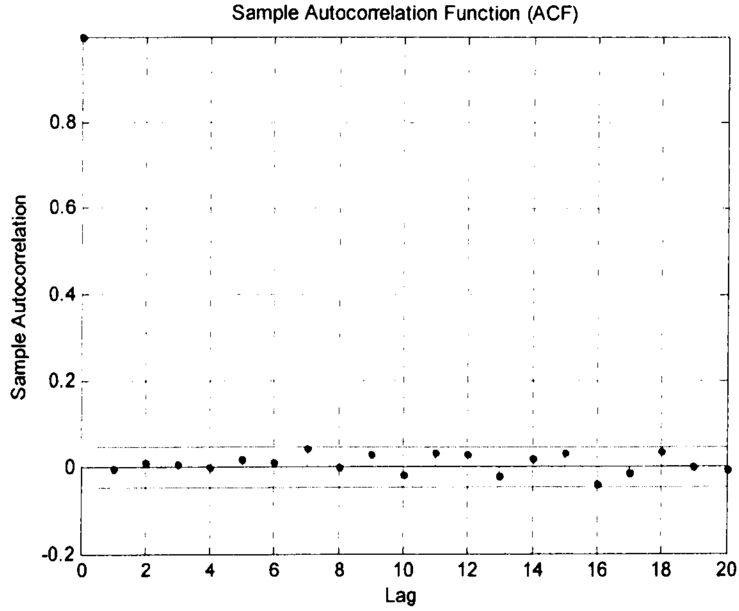


Figure 5.4 Autocorrelation of the linear mean model residuals

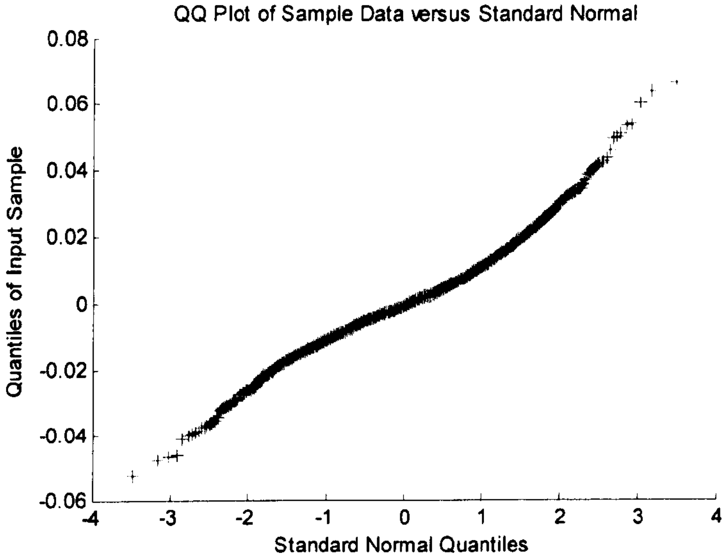


Figure 5.5 QQ plot of the linear mean model residuals

Table 5.5 Statistics of residuals estimated from linear mean model

Skewness	Kurtosis	JB statistic
0.4079	5.1683	446.5998

There is no outlier outside the 95% significant line and the parameter estimation of the mean model is therefore not biased. The QQ plot of the linear mean residuals is drawn in Figure 5.5 and the statistics of the residuals are listed in Table 5.5. The JB test and QQ plot reject the normality of the residuals. However, as the Quasi ML method is employed in estimating the GARCH model, a normal distribution density function may still be used. Then a GARCH (1, 1) model was estimated from the linear mean residuals and the estimated parameters are listed in Table 5.6.

Table 5.6 GARCH model parameters estimated from linear mean model residuals

Parameter	Estimation	Standard Error
A_0	5.2302e-7	2.0441e-7
A_1	0.0835	0.00105
B_1	0.9167	0.0097

Then a nonlinear reference model was used to give the mean model term selection and the results of the WOFR algorithm are listed in Table 5.7. The estimation results using the normal OFR algorithm based on the same nonlinear reference model are also listed in Table 5.7 to give a comparison.

Table 5.7 WOFR and ordinary OFR estimation results of the nonlinear reference mean model

Term	WERR	Term	ERR
y(t-1)y(t-1)	0.68190e-1	y(t-1)y(t-1)	0. 13374
Constant	0.78918e-2	Constant	0. 52270e-2
y(t-1)	0.20116e-2	y(t-4)	0.25221e-2
y(t-4)	0.89675e-3	y(t-1)	0.15988e-2
y(t-1) y(t-4)	0.53990e-3	y(t-2) y(t-3)	0.12209e-2
y(t-3) y(t-4)	0.36171e-3	y(t-3) y(t-4)	0.13312e-2
y(t-2) y(t-3)	0.45212e-3	y(t-1) y(t-3)	0.44957e-3
y(t-3) y(t-3)	0.42843e-3	y(t-1) y(t-4)	0.47574e-3
y(t-2) y(t-2)	0. 87281e-4	y(t-2) y(t-2)	0.32065e-3
y(t-3)	0. 56122e-4	y(t-4) y(t-4)	0.40903e-3
y(t-4) y(t-4)	0.53844e-4	y(t-3) y(t-3)	0.22787e-3
y(t-1) y(t-2)	0.44948e-4	y(t-2) y(t-4)	0.14792e-3
y(t-1) y(t-3)	0.47715e-4	y(t-3)	0.15701e-3
y(t-2)	0.54501e-5	y(t-1) y(t-2)	0.30006e-5
y(t-2) y(t-4)	0.37269e-5	y(t-2)	0.49769e-5

The results in Table 5.7 clearly demonstrate that the first three most significant terms from WOFR estimation are the terms of the real mean model as in equation (5.31) and the ordinary OFR estimation cannot give a correct selection. Therefore, according to the ranking of the WERR, the first three terms with grey background in Table 5.7 are selected as the mean model terms. After an iterative reweighted calculation, the converged parameter estimation for the selected mean model terms are listed in Table 5.8. The autocorrelation and the QQ plot of the residuals of nonlinear selected model are drawn in Figure 5.6 and Figure 5.7. The statistics of the residuals and the JB test results are listed in Table 5.9. Comparing these with the values in Table 5.5, the statistics are improved significantly. According to the autocorrelation, the parameter estimation of the nonlinear mean model is unbiased.

Table 5.8 Parameters of selected nonlinear mean model terms after iterative reweighted calculation

Terms	Parameter estimation	Standard Error
y(t-1)y(t-1)	11.0632	1.2643
Constant	8.825e-4	4.4406e-8
y(t-1)	-4.8802e-2	6.0930e-4

Table 5.9 Statistics of residuals estimated from nonlinear mean model

Skewness	Kurtosis	JB statistic
-0.0355	4.5644	204.3555

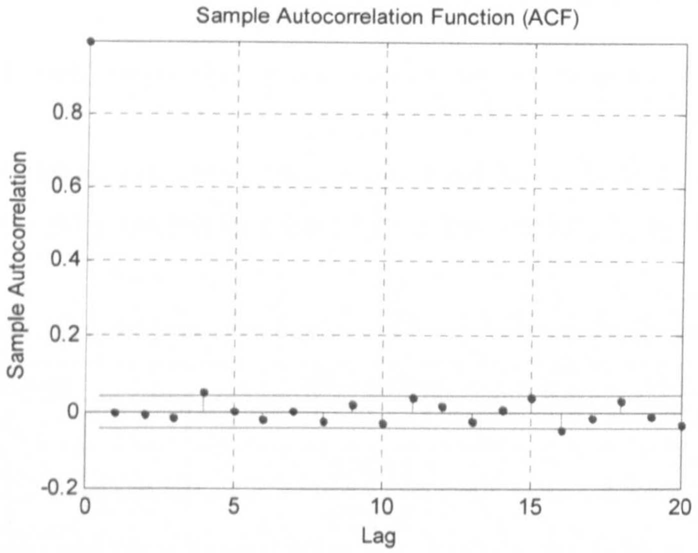


Figure 5.6 Autocorrelation of residuals estimated from nonlinear mean model

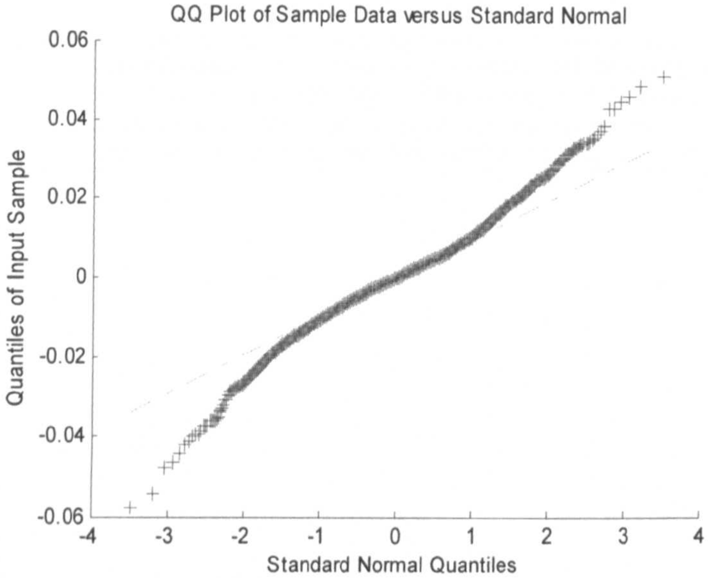


Figure 5.7 QQ plot of residuals estimated from nonlinear mean model

As the GARCH model is simultaneously estimated during WOFR, the parameters of the GARCH model in the last iteration of WOFR are listed in Table 5.10.

Table 5.10 GARCH model parameters estimated from nonlinear mean model residuals

Parameter	Estimation	Standard Error
A_0	3.0750e-7	2.0257e-7
A_1	0.0751	0.0106
A_1	0.9258	0.0097

In order to demonstrate the impact of an inaccurate mean model on the estimation of the variance, the absolute differences between the real variance and the estimated variance, estimated from the residuals of the linear and nonlinear selected mean model, are drawn in Figure 5.8.

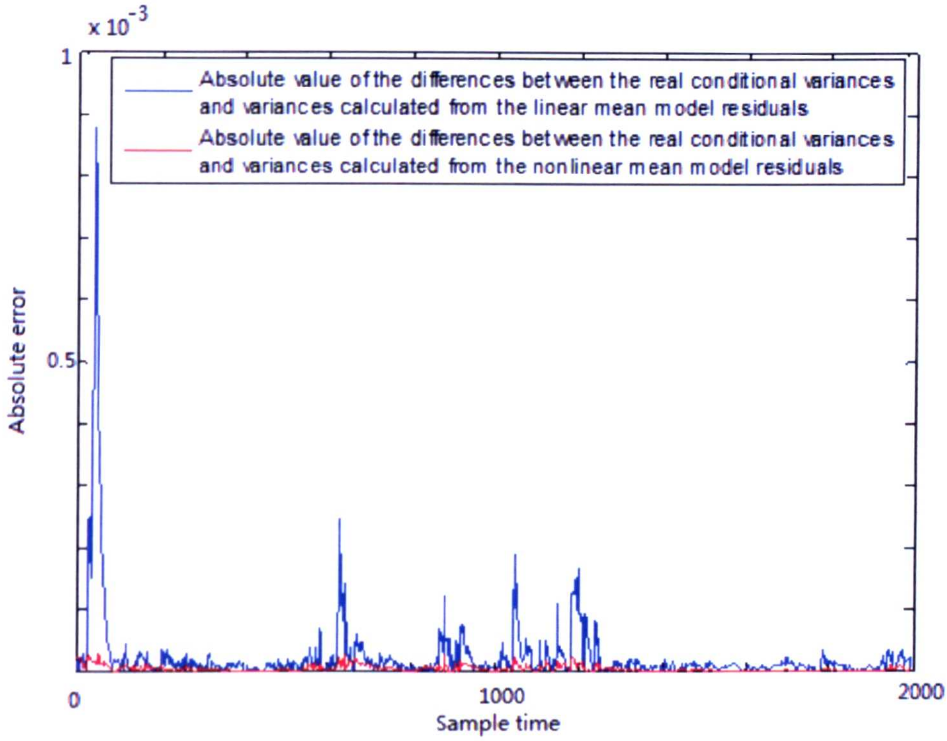


Figure 5.8 Absolute of volatilities differences between the real variances and the variances estimated from the linear (blue) and nonlinear (red) mean models

Figure 5.8 illustrates clearly that the variance calculated from nonlinear mean model is more accurate than that from the linear mean model. Then the BDS tests are applied to the standardized residuals of the linear and nonlinear mean model residuals and the testing results are listed in Table 5.11

Table 5.11 BDS test for standardized linear and nonlinear residuals

γ / σ	Embedding Dimension(m)	Linear residuals	Nonlinear residuals	γ / σ	Embedding Dimension(m)	Linear residuals	Nonlinear residuals
2	2	3.0207	0.1769	1	2	2.1371	0.1594
2	3	2.5012	0.6431	1	3	2.0096	0.2265
2	4	2.0239	0.2486	1	4	1.5337	-0.0518
2	5	1.7139	0.2323	1	5	1.4829	0.0535
1.5	2	2.5739	0.2074	0.5	2	1.4849	-0.1068
1.5	3	2.2328	0.4100	0.5	3	1.6230	-0.0392
1.5	4	1.7455	0.0560	0.5	4	1.0172	-0.5665
1.5	5	1.5104	0.0459	0.5	5	0.9310	-0.2972

The grey background of the test statistics in Table 5.11 indicates that the hypothesis of i.i.d. has been rejected at a specified dimension and γ / σ ratio. Almost half of the test statistics of the residuals estimated from the linear mean model rejects the i.i.d. assumption and there is no rejection for the nonlinear mean model residuals. Obviously the linear mean model did not even pass the test for the first half of data.

Then both the linear and nonlinear mean model parameters are used to calculate the one-step-ahead prediction errors for the second half data which is sampled from 3001 to 5000. The parameters of the GARCH models in Table 5.6 and Table 5.10 are also employed to calculate the one-step-ahead prediction of the variance using the prediction errors. According to the procedure of CV listed in Section 5.5, a GARCH (1,1) model is also used to fit the prediction error. The estimated GARCH parameters, from the prediction errors of the linear and nonlinear mean models, are listed in Table 5.12.

Table 5.12 GARCH model parameters estimated from linear and nonlinear mean prediction error

Parameter(linear)	Estimation	Std. Error	Parameter(nonlinear)	Estimation	Std. Error
A_0	5.7842e-7	2.9860e-7	A_0	5.75060e-7	2.7735e-7
A_1	0.0688	0.0112	A_1	0.0793	0.0131
B_1	0.9281	0.0120	B_1	0.9179	0.0131

Therefore, there will be two groups of variance prediction for each mean model. One group is estimated from the prediction error and the other is recursively calculated using GARCH parameters estimated from the first half data. The absolute differences of the two group variances are then drawn in Figure 5.9 to demonstrate the prediction ability of the GARCH model estimated from the first half data.

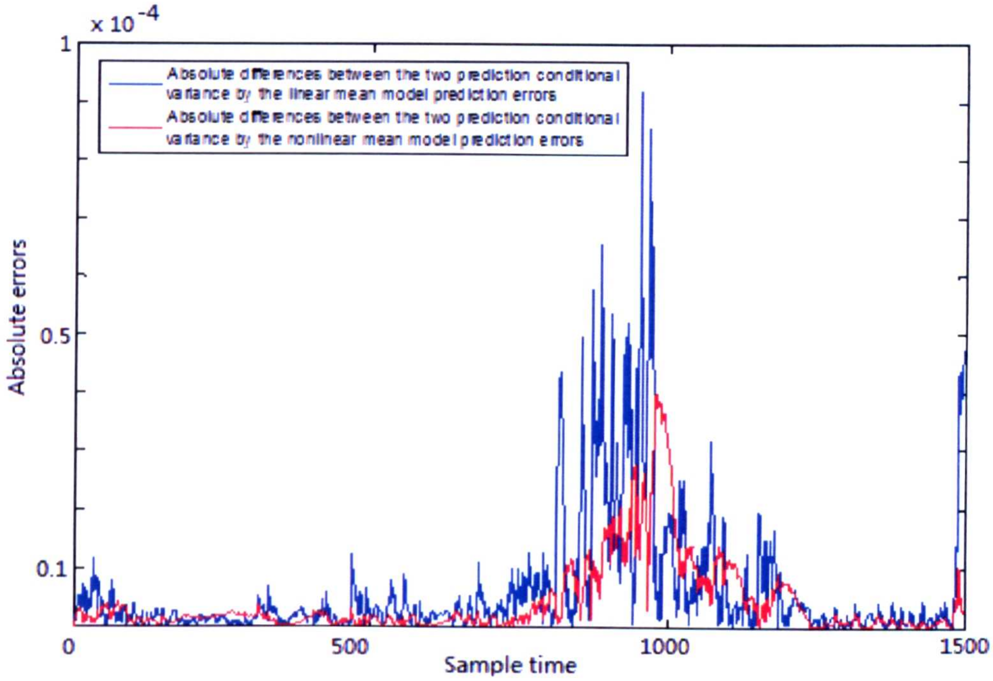


Figure 5.9 Absolute errors between the variances predicted by the GARCH model from the first half data and the variances calculated from the prediction errors. Blue represents that the mean model is linear and red represents that the mean model is nonlinear.

It is clear in Figure 5.9 that the variance prediction of the nonlinear mean model stays closer most of the time than that of linear mean model. Since the real variance is known for the second half data, it is necessary to compare the predicted variance with the real variance. Therefore, the absolute differences between the predicted variance and the real variance are drawn in Figure 5.10. It is clear that the predicted variance estimated from the nonlinear mean model is more accurate than that of linear mean model. Then the BDS tests are applied to the standardized prediction errors and the test results are listed in Table 5.13 and 5.14. The predicted variances of linear and nonlinear mean models are drawn together with real variances in Figure 5.11.

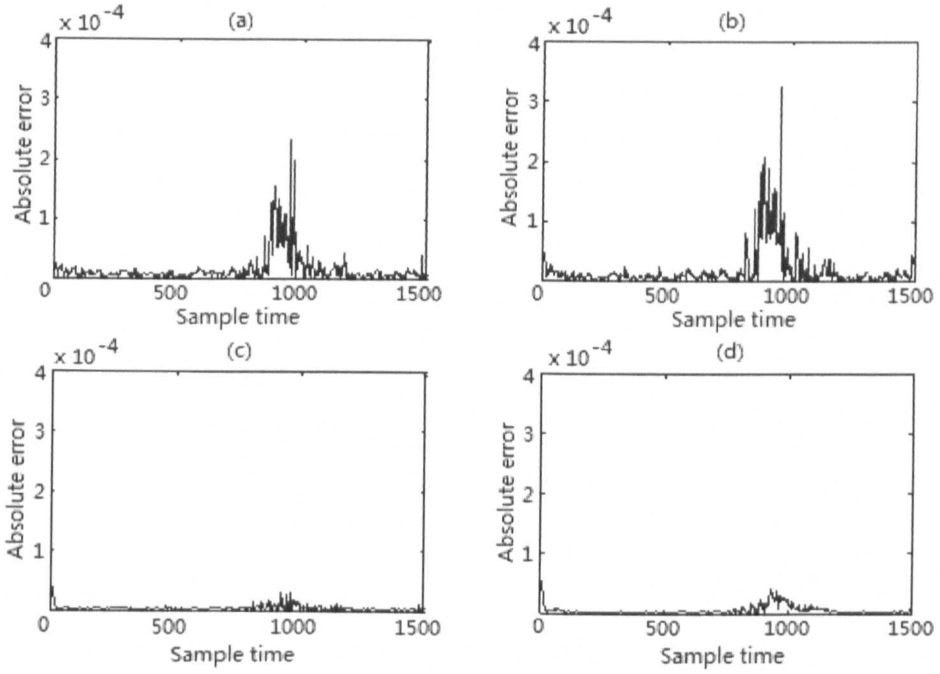


Figure 5.10 Absolute differences between the real variance and the estimated variance

(a) Absolute differences between the real variance and the variance estimated from the prediction errors of linear mean model

(b) Absolute differences between the real variance and the variance calculated using the GARCH parameters estimated from first half data and the prediction errors of the linear mean model

(c) Absolute differences between the real variance and the variance estimated from the prediction errors of nonlinear mean model

(d) Absolute differences between the real variance and the variance calculated using the GARCH parameters estimated from first half data and the prediction errors of the nonlinear mean model

Table 5.13 BDS test statistics of standardized prediction errors using estimated variance

γ / σ	Embedding Dimension(m)	Linear mean model	Nonlinear mean model	γ / σ	Embedding Dimension(m)	Linear mean model	Nonlinear mean model
2	2	3.5936	0.8567	1	2	3.0084	0.7994
2	3	2.8780	0.5191	1	3	2.3364	0.5261
2	4	2.8975	0.5761	1	4	2.5959	0.8768
2	5	3.0066	0.8114	1	5	2.9038	1.3447
1.5	2	3.4602	0.9352	0.5	2	2.7112	1.0152
1.5	3	2.6330	0.5628	0.5	3	2.1486	1.3828
1.5	4	2.6963	0.7231	0.5	4	2.5495	1.6788
1.5	5	2.8755	0.9984	0.5	5	3.2662	1.9357

Table 5.14 BDS test statistics of standardized prediction errors using predicted variance

γ / σ	Embedding Dimension(m)	Linear mean model	Nonlinear mean model	γ / σ	Embedding Dimension(m)	Linear mean model	Nonlinear mean model
2	2	3.1379	0.9816	1	2	2.6473	0.9025
2	3	2.2318	0.6719	1	3	1.8129	0.6390
2	4	2.1615	0.7410	1	4	1.9542	0.9991
2	5	2.2023	0.9946	1	5	2.1444	1.4922
1.5	2	3.0049	1.0348	0.5	2	2.3975	1.0546
1.5	3	2.0155	0.7073	0.5	3	1.7124	1.4781
1.5	4	1.9898	0.8909	0.5	4	1.9607	1.8136
1.5	5	2.1156	1.1884	0.5	5	2.8020	2.2461

According to the test results in Table 5.13 and 5.14, the BDS test statistics of standardized prediction error from the linear mean model indicate that the i.i.d. assumption is rejected as most of the statistics are over 1.96. However, for the nonlinear mean model the i.i.d. assumption has not been rejected as there is only one statistic which is over 1.96.

The simulation indicates that CV between mean model and variance model is very effective and that an inaccurate mean model will affect the variance prediction. Although the distribution assumption is rejected, the QMLE method can compensate fat tails of the distribution and the variance estimation is barely affected. In Figure 5.11, the variance prediction of the nonlinear mean model is improved especially in volatile periods. Since the variance is directly related to risk in the financial area, even 1% improvement may have significant impact on risk evaluation. Therefore, the work on improving variance prediction and CV between the mean model and the variance model indicates that the accuracy of the mean model is essential in variance forecasting.

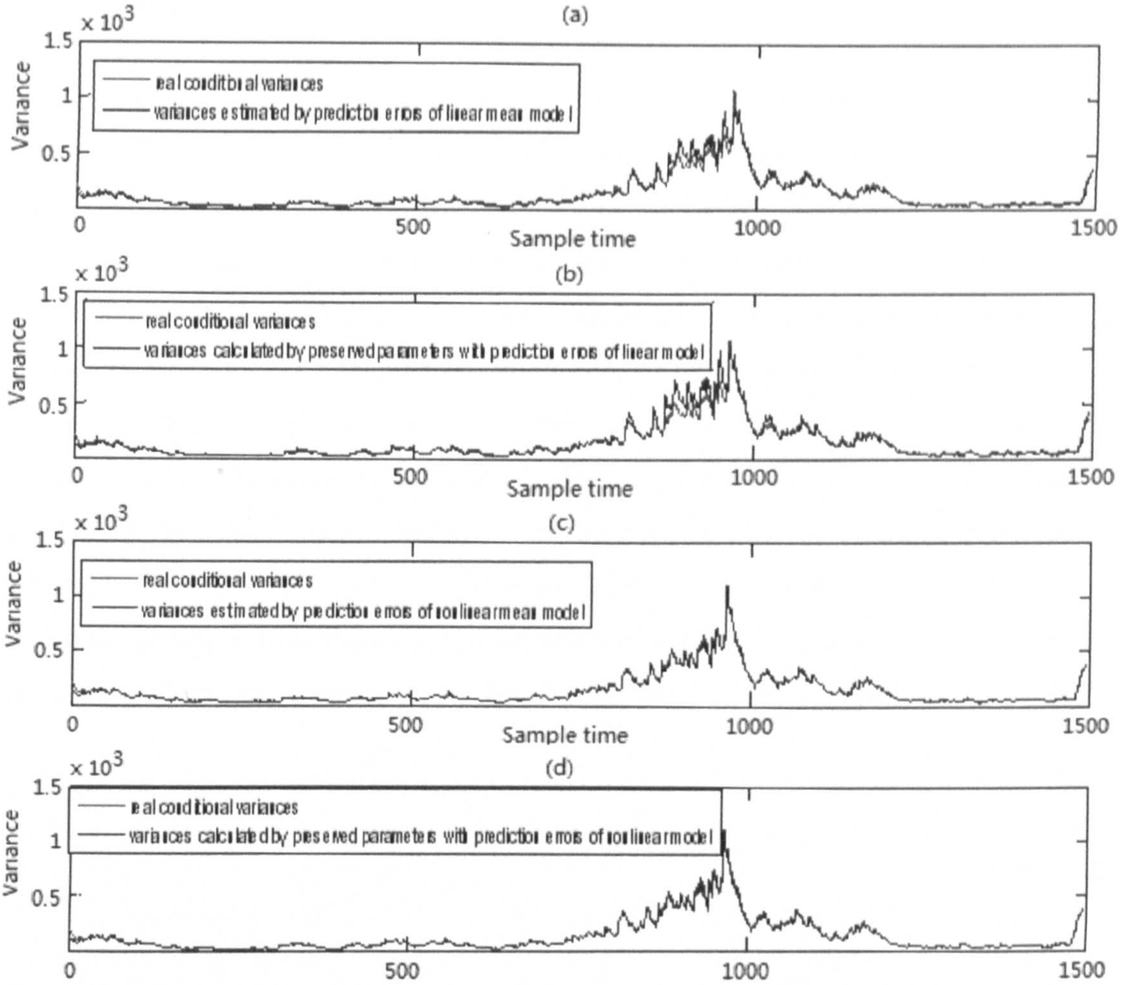


Figure 5.11 The Red line represents the real variances and blue line represents the estimated variances.

(a) The variances are estimated from the prediction errors of linear mean model

(b) The variances are iteratively calculated from the prediction errors of linear mean model using the parameters of GARCH model estimated from the first half data.

(c) The variances are estimated from the prediction errors of nonlinear mean model.

(d) The variances are iteratively calculated from the prediction errors of nonlinear mean model using the parameters of GARCH model estimated from the first half data.

5.7 Conclusions

In this chapter, the impact of the accuracy of the mean model on variance forecasting has been derived theoretically and a CV method between the mean model and variance model has been proposed. The WOFR algorithm proposed in Chapter 4 was used to refine the term selection of the mean model during CV. According to the simulations results, the CV method is very effective in detecting an inaccurate mean model and variance model. As far as we known, there is no work which is related to validate the mean model and variance model simultaneously especially when the mean model is nonlinear. Since variance forecasts are widely applied in the finance area, the work in this chapter has very good application potential. Therefore, the method proposed in this chapter provides a statistical technique to apply a nonlinear mean model with selected terms in predicting the variances when the variance is fitted by the GARCH class of models.

Chapter 6: Mortality rate modelling and forecasting using the NARMAX model

6.1 Introduction

The mortality rate is defined as a measure of the death rate in a population and mortality risk is the key risk factor driving the values of mortality and longevity linked securities. If the mortality in pricing the annuities is overestimated, the profit margin of pension providers will shrink significantly. Insurance products sold by private companies are also influenced by the mortality rate (Brouhns, Denuit and Vermunt, 2002). Therefore, in order to value mortality-related positions and to reduce their exposure to mortality improvements, actuaries employ mortality models to predict the future mortality.

There are many techniques that have been developed to model mortality since Cramer and Wold (1935) firstly modelled mortality rate curves using extrapolation methods. Benjamin and Soliman (1993) fitted the mortality rate using technique based on the projection of parameters while Lee and Cater (1992) proposed a simple model (Lee-Carter model) which can describe mortality changes using both age-dependent and time-dependent terms. Renshaw and Haberman (1996) successfully used Poisson distributed random variables as the additive error term in Lee-Carter model and Brouhns, Denuit and Vermunt (2002) improved the Lee-Cater model using a generalised linear model with Poisson errors. A more complex Age-Period-Cohort (APC) model which adds the cohort factor to the common age structure was then introduced by Tabeau et al. (2001). Some time-series approaches were used by McNown and Rogers (1989), Rogers and Gard (1991). Most recently, there was the CBD model proposed by Cairn, Black and Dowd (2006) and its various generalisations to encompass a cohort effect. However, it is widely accepted in mortality modelling circles that no existing model is entirely satisfactory.

This chapter investigates the use of a NARMAX polynomial representation to model the mortality rate. According to Chapter 3, NARMAX polynomial representation can be used to approach the nonlinear system effectively and precisely with selected terms using the OFR algorithm.

Once the mortality model is estimated, its prediction ability needs to be assessed. Dowd et al. (2008) used the back testing method to evaluate the forecast performance of a number of existing stochastic mortality models. Back testing method can indicate whether the underlying model can give a good out-of-sample prediction. Therefore, in this chapter the back testing method will be employed to check the ex post forecast performance of the fitted mortality models. The models considered are the CBD model and the fitted NARMAX model.

This chapter is organized as follows. Section 6.2 gives the mortality rate related definition and notation. Section 6.3 introduces the most recently mortality models and Section 6.4 applies the NARMAX modelling method to the realised mortality rate data and derives the nonlinear mortality model, where the smoking rate is also involved in the term selection. Section 6.5 evaluates the forecast performance of the NARMAX type models using back testing method and gives prediction comparison with the CBD mortality model. Section 6.6 concludes.

6.2 Definition and notification

In this chapter, the mortality rate is treated as discrete and the calendar year is represented by t and running from t to $t+1$.

Therefore, the death rate $m(t,x)$ for age x is defined as

$$m(t,x)=\frac{deaths\ during\ calendar\ year\ t\ aged\ x(last\ birthday)}{average\ population\ during\ calendar\ year\ t\ aged\ x(last\ birthday)} \quad (6.1)$$

The mortality rate $q(t,x)$ is defined as the probability that an individual aged exactly x at exact year t will die between t and $t+1$.

Another measure of the death rate is the force of mortality rate $\mu(t,x)$ which is defined as the death rate at exact time t for people at exact age x. For small time dt, the probability of death between t and $t+dt$ is $\mu(t,x)\times dt$. If the force of mortality remains the same over each calendar year and over each year of integer age which is for all $0 \leq s,u \leq 1$, $\mu(t+s,x+u)=\mu(t,x)$ and the population remains stationary, the relationships between death rate $m(t,x)$, the force of mortality rate $\mu(t,x)$ and the mortality rate $q(t,x)$ are

$$m(t,x)=\mu(t,x) \tag{6.2}$$

$$q(t,x)=1-\exp(-m(t,x))=1-\exp(-\mu(t,x)) \tag{6.3}$$

Some mortality models also use cohort effect. The cohort effect is usually used to describe some shared life experience among the individuals over some certain times. It is already observed that cohorts born around 1930 have obvious improvement between age 40 and 70 and cohorts born around 1950 have worst mortality in England & Wales (Cairns et al., 2007).

6.3 Introduction to the commonly used mortality models

6.3.1 Lee-Carter mortality model

One of the mostly commonly used mortality model is the Lee-Carter mortality model which was proposed by Lee and Carter (1992). The Lee-Carter model combines a demographic model with statistical time series approach and the model is specified for the logarithmic transformation of the death rate at age x and year t using three parameters: a period related effect, an age-specific parameter that represents the general shape across age of the mortality schedule and a second age specific parameter that is related to the changes in mortality level. The Lee-Carter model can be written as

$$\ln(m(t,x))=\beta_x^{(1)}+\beta_x^{(2)}\kappa_t \tag{6.4}$$

where $\beta_x^{(1)}, \beta_x^{(2)}$ are the age-specific parameters and κ_t reflects the period effect. The constraints for estimating the parameters are

$$\sum_t \kappa_t = 0 \text{ and } \sum_x \beta_x^{(2)} = 1 \tag{6.5}$$

6.3.2 Extend Lee-Carter model

The Lee-Carter model was extended to a generalised case by Renshaw and Haberman (1996) who use Poisson distributed random variation as the error term. In later work, Renshaw and Haberman (2006) added a cohort effect term to the Lee-Carter model:

$$\ln\left(m(t,x)\right) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t + \beta_x^{(3)}\gamma_{t-x} \tag{6.6}$$

where γ_{t-x} is the cohort effect term with the birth year $t-x$ and $\beta_x^{(3)}$ is an additional age related term. The constraints of the parameters are

$$\sum_t \kappa_t = 0,\ \sum_x \beta_x^{(3)} = 1,\ \sum_{x,t} \gamma_{t-x} = 0 \text{and } \sum_x \beta_x^{(3)} = 1 \tag{6.7}$$

Although the additional cohort effect term can provide better modelling accuracy, comparing with Lee-Carter model the parameter estimation converges much more slowly during ML estimation.

6.3.3 Age-Period-Cohort model

The APC model was firstly introduced by Tabeau et al. (2001) and the parameters of APC model described the trajectories of time and cohort effect given the age pattern. A linear APC model can be written as

$$\ln\left(m(t,x)\right) = \beta_x^{(1)} + \kappa_t + \gamma_{t-x} \tag{6.8}$$

and the APC model is a special case of the extended Lee-Carter model with $\beta_x^{(2)} = \frac{1}{N}, \beta_x^{(3)} = \frac{1}{N}$ where N is the number of ages used in the sample on which the model is calibrated. The constraints of the parameters are

$$\sum_t \kappa_t = 0, \sum_{x,t} \gamma_{t-x} = 0 \tag{6.9}$$

However, APC model still has identification problem as the three parameters are linearly dependent. The solution to this problem is to impose identifiability constraints which lead to the following:

$$\ln\left(m(t,x)\right) = \tilde{\beta}_x^{(1)} + \tilde{\kappa}_t + \tilde{\gamma}_{t-x} \tag{6.10}$$

where $\tilde{\kappa}_t = \kappa_t - \delta\left(t - \bar{t}\right), \tilde{\gamma}_{t-x} = \gamma_{t-x} + \delta\left(\left(t-\bar{t}\right) - \left(x - \bar{x}\right)\right)$, $\tilde{\beta}_x^{(1)} = \beta_x^{(1)} + \delta\left(x - \bar{x}\right)$.

6.3.4 CBD model

The CBD model is firstly introduced by Cairns et al. (2006) to fit the mortality rate at higher ages. The CBD model is a two-factor model and can be written as

$$\log it\left(q(t,x)\right) = \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} \tag{6.11}$$

where $\beta_x^{(1)}, \beta_x^{(2)}$ are age related parameters and $\kappa_t^{(1)}, \kappa_t^{(2)}$ are period effects. It is usually assumed that the parameters $\beta_x^{(1)} = 1$ and $\beta_x^{(2)} = \left(x - \bar{x}\right)$. Therefore, CBD model in equation (6.1) can be written in a simpler form as

$$\log it\left(q(t,x)\right) = \kappa_t^{(1)} + \kappa_t^{(2)}\left(x - \bar{x}\right) \tag{6.12}$$

If the cohort effect is considered, an extra term will be added to equation (6.13) to give us one of the generalisations of the model:

$$\log it\left(q(t,x)\right) = \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x} \tag{6.13}$$

where γ_{t-x} is the cohort effect term and $\beta_x^{(3)}$ is usually equal to 1. One advantage of the CBD model is that there is no identification problem because there is no constraint on parameters.

6.3.5 Quadratic regression model

Heathcote and Higgins (2001) used a quadratic regression model to fit the Dutch mortality rates. The model year variable t and age variable x and can be written as

$$\log it\left(q\left(t,x\right)\right)=\beta_0+\beta_1 x+\beta_2 t+\beta_3 x^2+\beta_4 tx+\varepsilon\left(t,x\right) \qquad (6.14)$$

where $\beta_0,\beta_1,\beta_2,\beta_3,\beta_4$ are parameters and $\varepsilon\left(x,t\right)$ is the modelling error which is assumed to be normally distributed.

6.4 Modelling the mortality rate using the NARMAX method

6.4.1 Data

The data set used to model the mortality is the LifeMetrics deaths and exposures data for England & Wales males that originally comes from the Office of National Statistics.The mortality rates for England & Wales males between 1961 and 2006 are drawn in Figure 6.1.

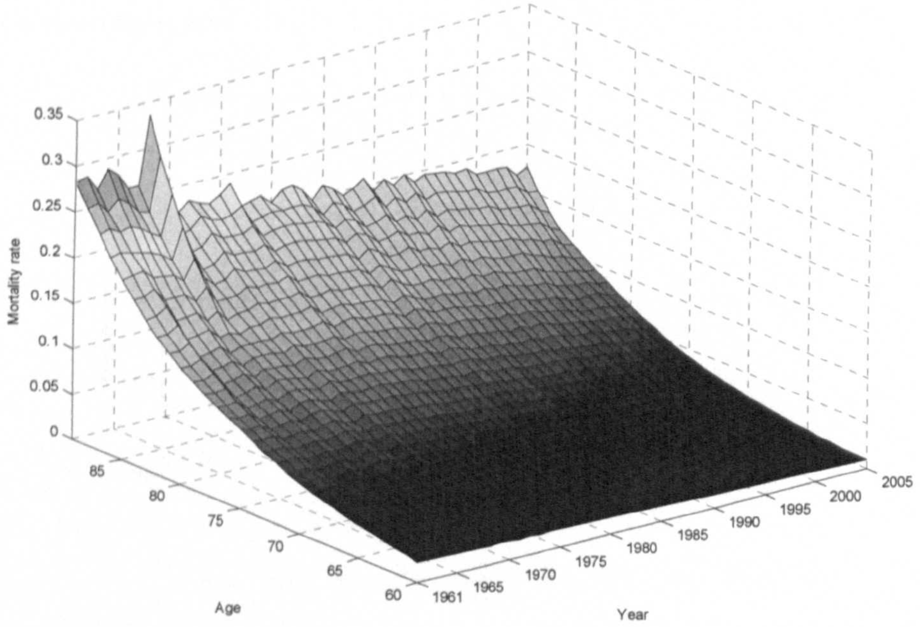


Figure 6.1 Mortality rates for England & Wales Male between age 60 and 89 from year 1961 to 2006

It is clear that the mortality rate in Figure 6.1 rises with age but falls over time. The corresponding q rates for ages 65, 75 and 85 are shown in Figure 6.2.

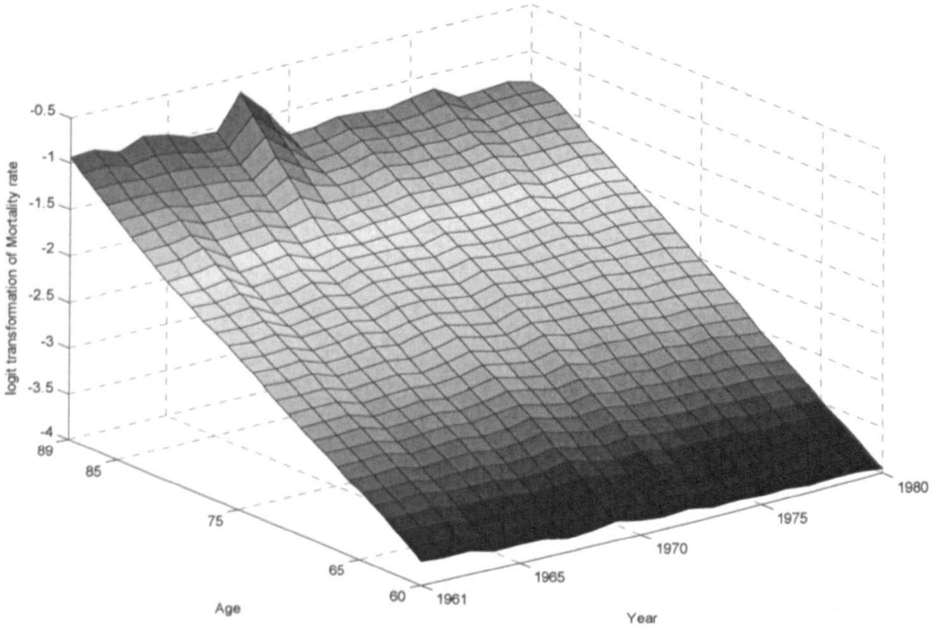


Figure 6.2 Logit transformation of male mortality rate of England & Wales between age 60 and 89 from year 1960 to 2006

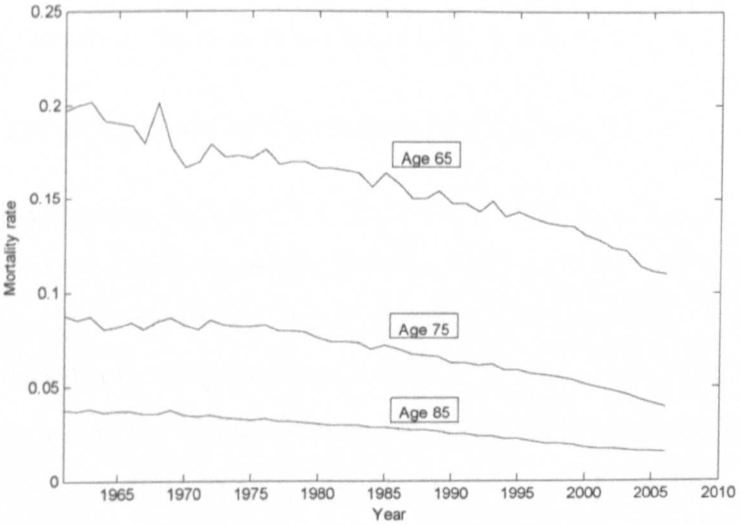


Figure 6.3 logit transformation of mortality rates at Age 65, 75, 85 from year 1961 to 2006

6.4.2 Model specification

Given the linear relationship in both age and year directions observed in Figure 6.2 and Figure 6.3, the following quadratic model is proposed

$$\log it\left(q\left(t,x\right)\right)=a_0+a_1x^2+a_2t^2+a_3xt+a_4x+a_5t+\varepsilon\left(t,x\right) \tag{6.15}$$

where $a_0,a_1,...,a_5$ are parameters, x is the age factor, t is the year factor and $\varepsilon\left(t,x\right)$ is modelling residual. If the cohort effect is considered which is the factor of $\left(t-x\right)$, the model (6.17) can then be transferred to

$$\log it\left(q\left(x,t\right)\right)=a_0+\left(b_1x^2+a_4x\right)+\left(b_2t^2+a_5t\right)-\frac{a_3}{2}\left(x-t\right)^2+\varepsilon\left(x,t\right) \tag{6.16}$$

where $b_1=a_1+\frac{a_3}{2}$, $b_2=a_2+\frac{a_3}{2}$, $\left(b_1x^2+a_4x\right)$ is the age factor related term, $\left(b_2t^2+a_5t\right)$ is year factor related term and $-\frac{a_3}{2}\left(x-t\right)^2$ is the cohort factor related term.

The model (6.15) is a pure quadratic polynomial model with all possible terms included and when the model is used to fit the realised mortality data, some terms may be insignificant and can be removed. Therefore, NARMAX modelling method is adopted to determine the term selection of model (6.15).

6.4.3 Data pre-processing and parameter estimation

Since the mortality data of England & Wales start from year 1961 to 2006, the first 20 years' data is used to fit the model and select the terms. In order to apply the OFR algorithm, the inputs need to be pre-processed to contain both year factor and age factor. Assumed that the length of ages is represented by m and the length of years is represented by n and the inputs of the model (6.15) are formatted by u_1 (age factor) and u_2 (year factor) where u_1 is a column vector with x repeated n times and u_2 (year factor) is a column vector and consisted with $t\left(1\right)$ repeated m times, $t\left(2\right)$ repeated m times until $t\left(n\right)$ repeated m times. Therefore, the inputs u_1 and u_2 share the same length as $m\times n$. The output y can be formatted as a column vector with logit transformation of mortality rate as

$$\left[\log it(q(1,1)),\log it(q(2,1)),...,\log it(q(m,1)),\log it(q(1,2)),...,\log it(q(m,2)),...,\log it(q(1,n)),...,\log it(q(m,n))\right]^{T}$$

Then the model (6.15) can be written as

$$y(T)=X(T)\Phi+\varepsilon(T) \qquad (6.17)$$

where $X=\left[C,u_1^2(T),u_2^2(T),u_1u_2(T),u_1(T),u_2(T)\right]$, C is a column of 1 with length $m\times n$, T is nominal sample time with length $m\times n$ and Φ is the parameter vector $\left[a_0,a_1,a_2,a_3,a_4,a_5\right]^{T}$. The model (6.19) is then treated as a reference model and after applying OFR algorithm, the parameter estimations results and the ERR of each terms are listed in Table 6.1.

Table 6.1 Parameter estimation and ERR results after applying OFR algorithm to model (6.17) with data set England and Wales male age 60 to 89 from year 1961 to 1980

Term	ERR	Estimated parameter value
u_2^2	90.5014	a_2 = - 5.6294 e-5
u_1	9.4537	a_4 = -0. 3307
u_1^2	21636 e-4	a_1 = - 6.1644 -5
Constant	3.5908e-5	a_3 = -2.2656e2
$u_1\times u_2$	49539e-4	a_5 = -1.1635e-4
u_2	40799e-5	a_0 = 2.2099e-1

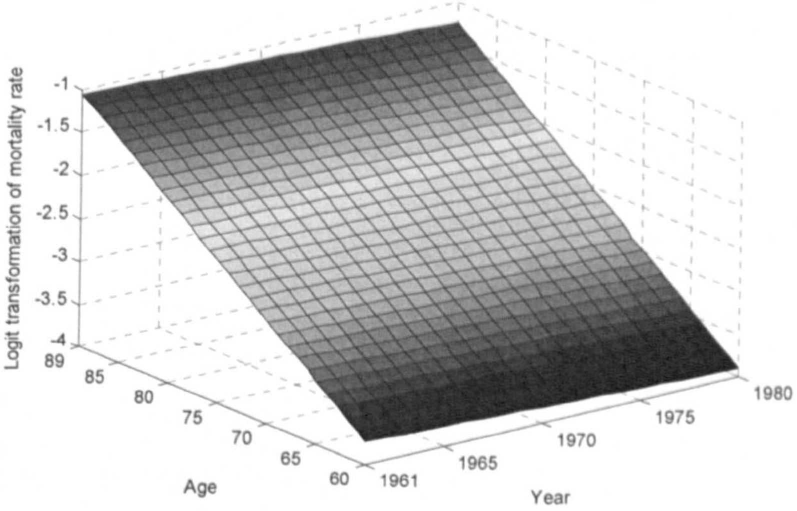
According to Table 6.1, the term u_2^2 and u_1 are the most significant terms and since these two terms represent the t^2 and x in model (6.17), the selected mortality model is

$$\log it(q(x,t))=a_0+a_2t^2+a_4x+\varepsilon(x,t) \qquad (6.18)$$

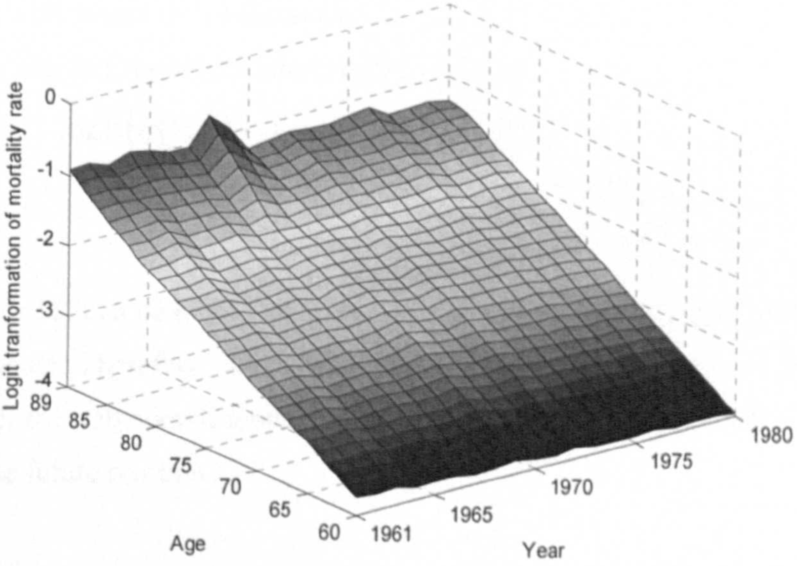
The parameters of model (6.18) are then re-estimated and listed in Table 6.2. Estimated logit transformation of mortality rate surface is drawn in Figure 6.4 and the estimated mortality rate at age 65, 75, 85 are drawn together with realized mortality rate in Figure 6.5.

Table 6.2 Parameters estimation for model (6.18)

Parameter	a_0	a_2	a_4
Estimated Value	4.1938e-2	-2.4183e-6	9.2272e-2



(a)



(b)

Figure 6.4 (a) Mortality rate surface between age 60 and 89 from year 1961 to 1980 calculated from estimated model (6.18). (b) Realized mortality rate between of same age range and year range.

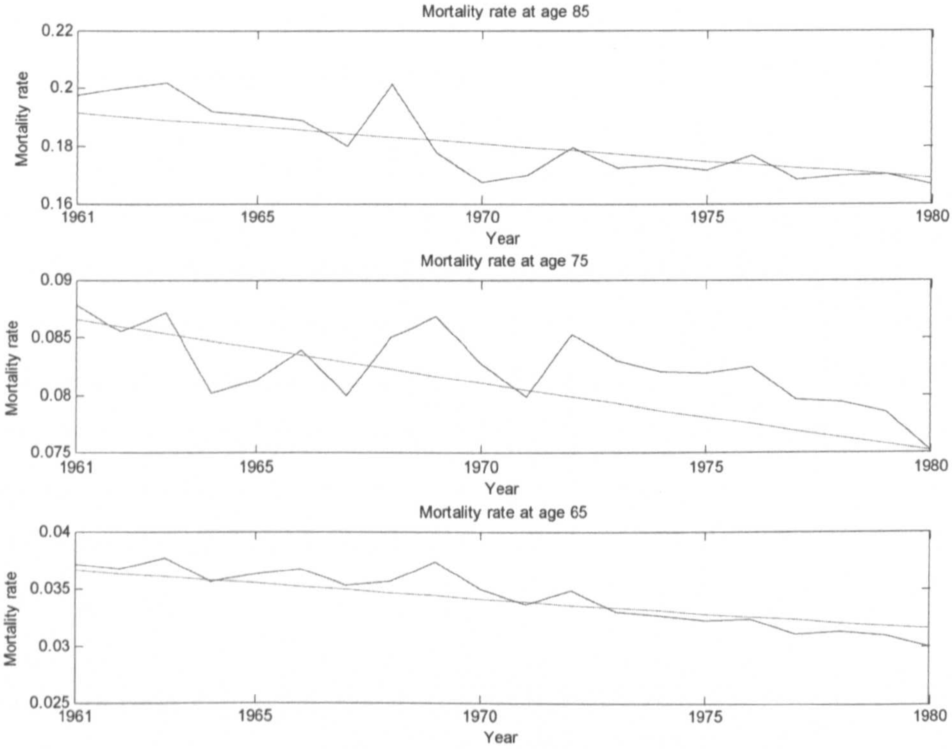


Figure 6.5 Estimated mortality rates in red and realized mortality rate at age 65, 75 and 85

Comparing with model (6.15) the model (6.18) is much simpler and the model (6.18) can also transfer to a model contained cohort effect as

$$\begin{aligned}\log it\left(q(x,t)\right) &= a_0 + a_2 t^2 + a_4 x + \varepsilon(x,t)\\ &= a_0 + a_2 t^2 + a_4 t - a_4 (t-x) + \varepsilon(x,t)\\ &= a_0 + \left(a_2 t^2 + a_4 t\right) - a_4 (t-x) + \varepsilon(x,t)\end{aligned} \tag{6.19}$$

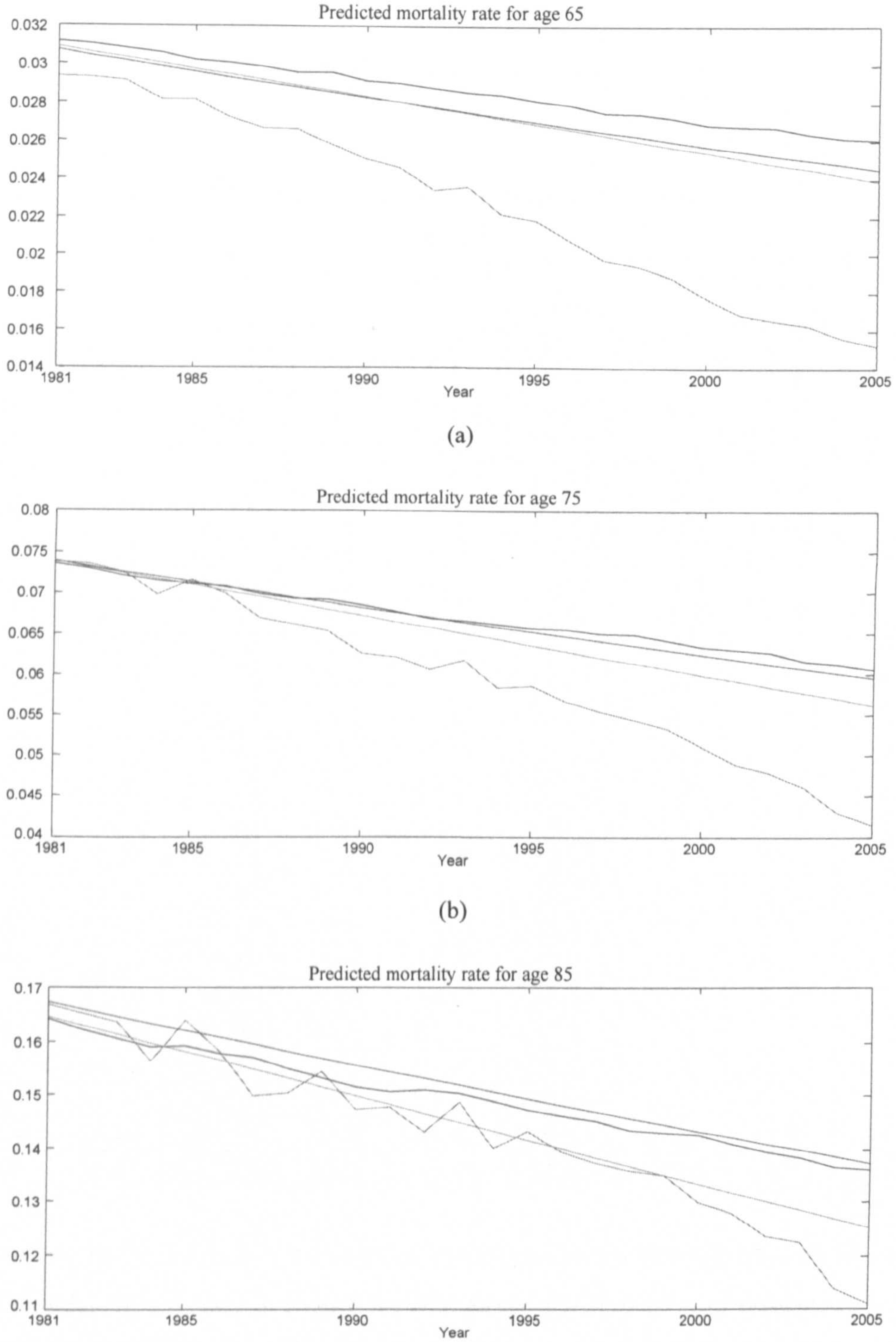
The model (6.18) can be employed to predict the future mortality rate using just age and year inputs. Therefore, comparing with statistical models like Lee-Carter, or CBD model, the polynomial model (6.18) is very convenient to estimate and apply to predict the future mortality.

6.5 Forecasting and back testing

6.5.1 Long term forecast comparison

Since the mortality rate model is fitted for predicting the future mortality rate, it is essential to compare the long term forecast performance of the simplified polynomial model (6.19) and the statistical mortality model CBD model (6.13). The polynomial model (6.15) is also used during prediction to indicate the differences between selected and unselected models. As the data from year 1961 to 1980 is used to fit the model, the data left from years 1981 to 2005 is used to check the long term forecast ability of the fitted models. Accordingly, the 25 year-ahead forecast of mortality rates of model (6.13), model (6.15) and model (6.19) at age 65, 75, 85 are drawn in Figure 6.6.

It can be seen from Figure 6.6 that the prediction from model (6.15) represented by magenta line is outstanding comparing with the prediction from selected model (6.18) and CBD model. The prediction from model (6.18) represented by blue line is better than that from CBD model in younger ages as in Figure 6.6 (a) and (b) but CBD model beats the model (6.18) in predicting older ages as in Figure 6.6 (c). In order to quantify the differences of prediction errors between model (6.18) and CBD model,



(a)

(b)

(c)

Figure 6.6 Predicted mortality rates for age 65 (a), age 75 (b) and age 85 (c) from model (6.18) in blue, CBD model in black and model (6.15) in magenta. The red line represents the realized mortality rate.

the prediction error percentages are calculated using $\frac{-\left(q_{pre} - q_{realized}\right)}{q_{pre}} \times 100\%$ and listed in Table 6.3

Table 6.3 Prediction error percentages of model (6.18) an d CBD model at age 65, 75, 85

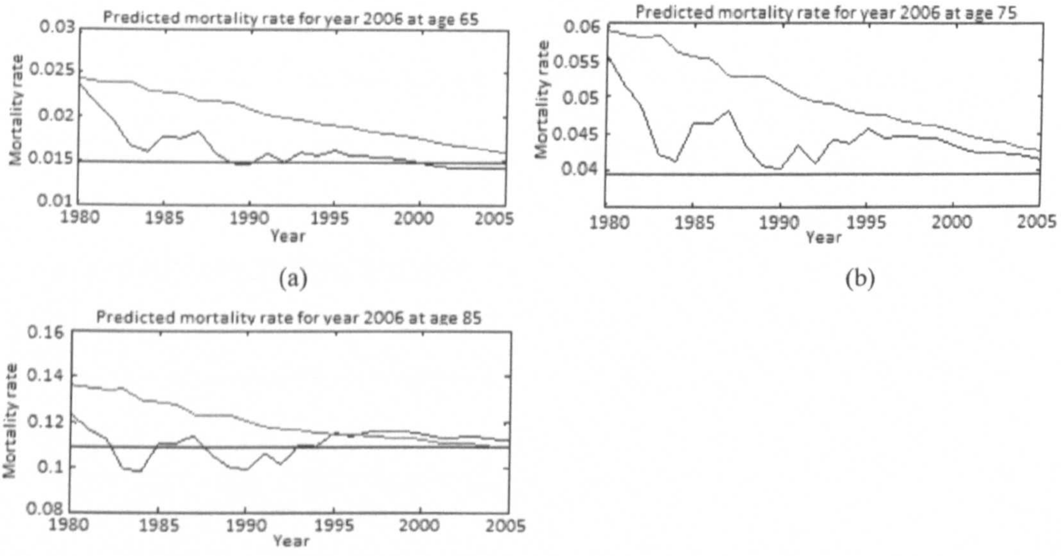
Age 65			Age 75			Age 85		
Year	Model (6.18)	CBD model	Year	Model (6.18)	CBD model	Year	Model (6.18)	CBD model
1981	-4.46	-5.89	1981	-0.12	0.24	1981	-0.31	1.56
1982	-3.73	-5.62	1982	0.52	0.91	1982	-0.46	1.81
1983	-3.43	-5.49	1983	-0.02	0.67	1983	-0.57	1.97
1984	-5.78	-7.98	1984	-3.07	-2.42	1984	-4.31	-1.71
1985	-4.92	-6.88	1985	0.30	0.69	1985	1.23	3.00
1986	-7.19	-9.50	1986	-1.19	-1.22	1986	-1.62	0.22
1987	-8.20	-10.75	1987	-4.52	-4.31	1987	-5.98	-4.50
1988	-7.61	-10.03	1988	-4.76	-4.62	1988	-5.01	-3.06
1989	-9.57	-12.75	1989	-5.03	-5.56	1989	-1.70	0.69
1990	-11.41	-14.10	1990	-8.17	-8.62	1990	-5.35	-2.72
1991	-12.08	-15.14	1991	-8.07	-8.22	1991	-4.23	-1.81
1992	-15.55	-18.53	1992	-9.42	-9.26	1992	-6.55	-5.20
1993	-14.18	-17.40	1993	-7.13	-7.44	1993	-2.09	-1.08
1994	-18.67	-22.00	1994	-11.21	-11.70	1994	-7.05	-5.93
1995	-19.23	-22.50	1995	-10.13	-10.70	1995	-4.15	-2.79
1996	-22.24	-25.52	1996	-12.30	-13.47	1996	-5.91	-4.69
1997	-25.39	-28.04	1997	-13.42	-14.72	1997	-6.60	-5.50
1998	-25.97	-29.19	1998	-14.20	-16.00	1998	-6.81	-5.37
1999	-27.67	-30.86	1999	-15.29	-16.86	1999	-6.51	-5.49
2000	-31.19	-34.07	2000	-18.07	-19.35	2000	-9.30	-8.84
2001	-33.93	-37.06	2001	-20.76	-22.21	2001	-10.00	-9.19
2002	-34.48	-38.04	2002	-21.80	-23.59	2002	-12.29	-11.37
2003	-34.94	-38.26	2003	-24.02	-25.18	2003	-12.39	-11.48
2004	-36.84	-40.19	2004	-28.42	-29.73	2004	-17.81	-16.62
2005	-38.15	-41.72	2005	-30.44	-31.61	2005	-19.33	-18.50

According to the Table 6.3, the selected model (6.18) has better prediction ability as CBD model. It is clearly in Figure 6.6 that model (6.15) produces similar prediction as model (6.18) which is still better than the prediction of CBD model. Therefore, both the selected model (6.18) and unselected model (6.15) can be used as alternative methods to model and predict the mortality rate.

Since the model (6.15) produced similar predictions as model (6.18), the issue rises that why the terms of the model need to be selected. Therefore, in order to test the prediction ability of the model (6.15) and selected model (6.18), the back testing techniques are used to distinguish the two models.

6.5.2 Back-testing for the forecast performance of the mortality models

We now selected the year 2006 is selected as the forecasts destination and the forecast of model (6.15) and (6.19) are based sequentially on estimates using observations up to 1980, estimates using observations up to 1981, and till up to 2005. Firstly, the historical data set within a rolling 20-year window are used to fit model (6.15) and (6.18) to give the prediction of year 2006. The corresponding prediction of model (6.15) and model (6.18) for age 65, 75 and 85 are drawn in Figure 6.7.



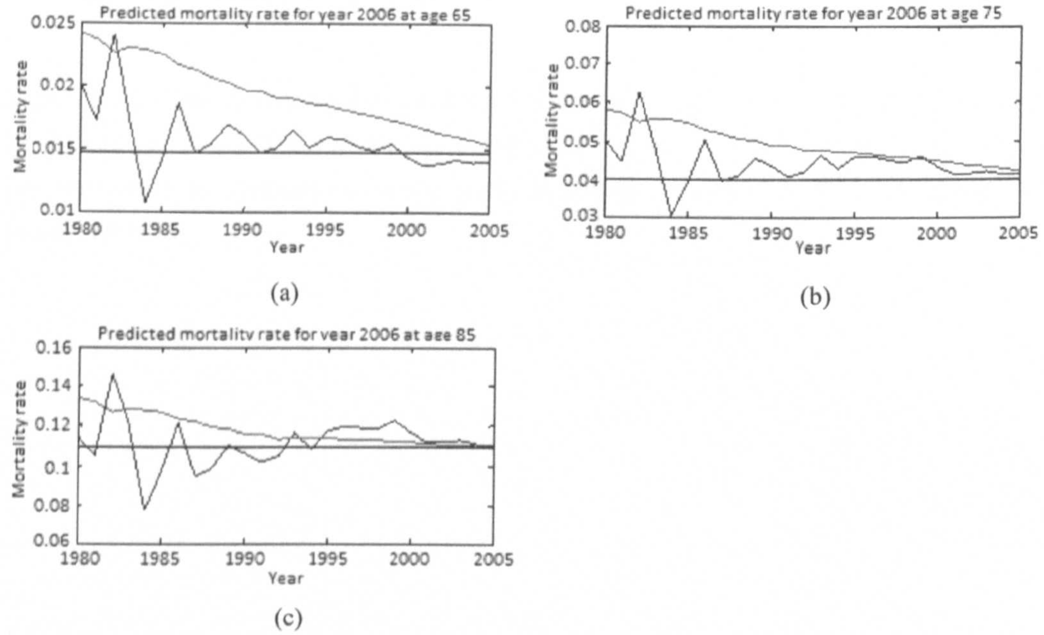
(a)

(b)

(c)

Figure 6.7 20 year rolling window prediction of mortality rate for year 2006 at age 65 (a), age 75 (b) and age 85 (c). The red line represents the prediction of selected model (6.18) and blue line represents the prediction of model (6.15). Black line represents the realized mortality rate at year 2006.

Then the rolling 15-year window and 10-year window are used to fit the model (6.15) and selected model (6.18) and the predicted mortality rate for year 2006 are drawn in Figure 6.8 and Figure 6.9. According to Figure 6.7, the prediction for year 2006 of model (6.15) converges faster than model (6.18). However, when forecasting for older ages like age 85, the model (6.15) begins to lose robustness. When the length of rolling window reduces, the prediction of model (6.15) loses more robustness as can be seen from Figure 6.8 and 6.9. Selected model (6.18) however gives more a consistent prediction for year 2006 with different rolling window length and at different ages always stay above the realized mortality rate of year 2006 and converges to the realized value. The ERR ranking results of the model (6.15) using different length of rolling window are listed in Table 6.4, Table 6.5 and Table 6.6.

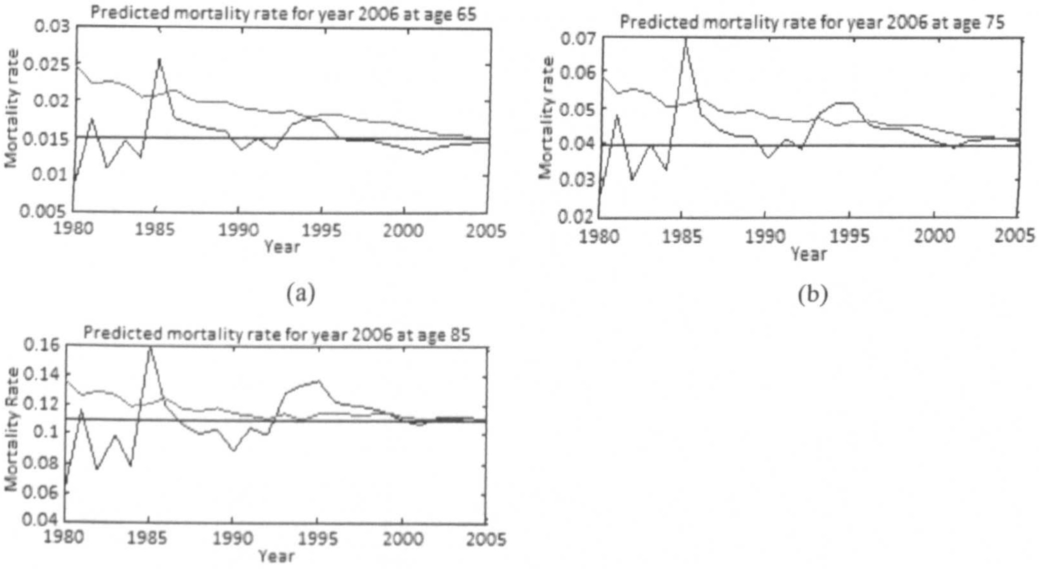


(a)

(b)

(c)

Figure 6.8 15 year rolling window prediction of mortality rate for year 2006 at age 65 (a), age 75 (b) and age 85 (c). The red line represents the prediction of selected model (6.18) and blue line represents the prediction of model (6.15). Black line represents the realized mortality rate at year 2006.



(a)

(b)

(c)

Figure 6.9 10 year rolling window prediction of mortality rate for year 2006 at age 65 (a), age 75 (b) and age 85 (c). The red line represents the prediction of selected model (6.18) and blue line represents the prediction of model (6.15). Black line represents the realized mortality rate at year 2006.

Table 6.4 ERR ranking results for model (6.15) using 20-year rolling window

Period\Rank	1st	2nd	3rd	4th	5th	6th
1961-1980	t^2	x	x^2	Constant	x*t	t
ERR	90.5013	9.4537	0.0002163	3.59E-05	4.95E-04	4.08E-05
1962-1981	t^2	x	x^2	Constant	t	x*t
ERR	90.5694	9.3862	0.0005794	9.83E-06	1.75E-04	1.75E-04
1963-1982	t^2	x	x^2	x*t	t	Constant
ERR	90.6407	9.3158	0.001135	1.58E-05	3.19E-07	4.09E-04
1964-1983	t^2	x	x^2	x*t	T	Constant
ERR	90.7230	9.2344	0.001348	1.31E-04	6.64E-05	1.48E-03
1965-1984	t^2	x	x^2	t	Constant	x*t
ERR	90.8087	9.1508	0.003610	7.49E-05	1.58E-03	2.48E-04
1966-1985	t^2	x	x^2	t	x*t	Constant
ERR	90.8566	9.1046	0.005274	1.26E-04	7.39E-04	6.66E-04
1967-1986	t^2	x	x^2	t	x*t	Constant
ERR	90.9122	9.0489	0.006387	2.41E-03	9.99E-04	6.94E-04
1968-1987	t^2	x	x^2	t	x*t	Constant
ERR	90.9626	8.9975	0.009948	1.59E-03	9.38E-04	2.89E-04
1969-1988	t^2	x	x^2	t	x*t	Constant
ERR	91.1108	8.8582	0.01251	0.001164	0.003048	0.001239
1970-1989	t^2	x	x^2	t	x*t	Constant
ERR	91.1420	8.8306	0.01291	0.0008831	0.003674	0.002399
1971-1990	t^2	x	x^2	t	x*t	Constant
ERR	91.1416	8.8288	0.01470	0.002809	0.002766	0.002443
1972-1991	t^2	x	x^2	t	x*t	Constant
ERR	91.1366	8.8307	0.01695	0.005614	0.002871	0.001067
1973-1992	t^2	x	t	x^2	x*t	Constant
ERR	91.1767	8.7882	0.01752	0.006977	0.003023	0.002069
1974-1993	t^2	x	t	x^2	x*t	Constant
ERR	91.1885	8.77440	0.01957	0.006593	0.003867	8.69E-04
1975-1994	t^2	x	t	x^2	x*t	Constant
ERR	91.2108	8.7463	0.02501	0.006286	0.004511	0.000897
1976-1995	t^2	x	t	x^2	x*t	Constant
ERR	91.2253	8.7285	0.02724	0.006151	0.006019	0.000366
1977-1996	t^2	x	t	x*t	x^2	Constant
ERR	91.2667	8.6848	0.02784	0.008168	0.005984	0.000786
1978-1997	t^2	x	t	x*t	x^2	Constant
ERR	91.2606	8.6846	0.03391	0.009333	0.005454	0.000516
1979-1998	t^2	x	t	x*t	x^2	Constant
ERR	91.2719	8.6700	0.03692	0.01025	0.004953	0.000556
1980-1999	t^2	x	t	x*t	x^2	Constant
ERR	91.2876	8.6515	0.03855	0.01186	0.004606	0.000688
1981-2000	t^2	x	t	x*t	x^2	Constant
ERR	91.3035	8.6275	0.04557	0.01286	0.004232	0.001051
1982-2001	t^2	x	t	x*t	x^2	Constant
ERR	91.3219	8.5987	0.05389	0.01465	0.003859	0.001531
1983-2002	t^2	x	t	x*t	x^2	Constant
ERR	91.3494	8.5632	0.06030	0.01654	0.00334	0.001847
1984-2003	t^2	x	t	x*t	x^2	Constant
ERR	91.3726	8.5352	0.06479	0.01713	0.00269	0.001934
1985-2004	t^2	x	t	x*t	x^2	Constant
ERR	91.3844	8.5116	0.07833	0.01543	0.00212	0.001868
1986-2005	t^2	x	t	x*t	Constant	x^2
ERR	91.4356	8.4535	0.08454	0.015087	0.003088	0.001522

Table 6.5 ERR ranking results for model (6.15) using 15-year rolling window

Period\Rank	1st	2nd	3rd	4th	5th	6th
1966-1980	t^2	x	x^2	x*t	t	Constant
ERR	90.7658	9.1891	0.003534	1.53E-04	1.01E-05	1.11E-04
1967-1981	t^2	x	x^2	x*t	t	Constant
ERR	90.8150	9.1401	0.004574	6.71E-05	1.86E-04	3.14E-04
1968-1982	t^2	x	x^2	Constant	x*t	T
ERR	90.8274	9.1302	0.007100	2.04E-04	5.44E-04	6.37E-05
1969-1983	t^2	x	x^2	x*t	T	Constant
ERR	90.9979	8.9740	0.009495	9.03E-05	0.003263	1.34E-04
1970-1984	t^2	x	x^2	x*t	T	Constant
ERR	91.0709	8.9055	0.01083	8.12E-05	0.002633	0.002413
1971-1985	t^2	x	x^2	t	x*t	Constant
ERR	91.0372	8.9420	0.01032	4.76E-05	0.001821	8.67E-04
1972-1986	t^2	x	x^2	t	x*t	Constant
ERR	91.0387	8.9414	0.01133	8.11E-04	0.001242	4.05E-05
1973-1987	t^2	x	x^2	T	Constant	x*t
ERR	91.1086	8.8712	0.01134	0.001315	9.01E-04	7.03E-04
1974-1988	t^2	x	x^2	T	Constant	x*t
ERR	91.1566	8.8219	0.01175	0.002775	6.51E-04	6.10E-04
1975-1989	t^2	x	x^2	T	x*t	Constant
ERR	91.1731	8.8047	0.01130	0.003942	9.85E-04	1.78E-04
1976-1990	t^2	x	t	x^2	x*t	Constant
ERR	91.2270	8.7469	0.01275	0.006001	0.001199	2.78E-04
1977-1991	t^2	x	t	x^2	x*t	Constant
ERR	91.2886	8.6852	0.01213	0.005838	0.002341	7.41E-04
1978-1992	t^2	x	t	x^2	x*t	Constant
ERR	91.3089	8.6609	0.01727	0.005382	0.002332	4.62E-04
1979-1993	t^2	x	t	x^2	x*t	Constant
ERR	91.3179	8.6527	0.01612	0.004918	0.002934	6.18E-05
1980-1994	t^2	x	t	x^2	x*t	Constant
ERR	91.3547	8.6126	0.01830	0.004895	0.003668	4.37E-04
1981-1995	t^2	x	t	x^2	x*t	Constant
ERR	91.3575	8.6078	0.01893	0.005031	0.004962	8.62E-05
1982-1996	t^2	x	t	x*t	x^2	Constant
ERR	91.3755	8.5860	0.02124	0.006640	0.005159	4.44E-05
1983-1997	t^2	x	t	x*t	x^2	Constant
ERR	91.3956	8.5611	0.02370	0.009208	0.005118	1.26E-04
1984-1998	t^2	x	t	x*t	x^2	Constant
ERR	91.4091	8.5454	0.02495	0.01055	0.004999	2.24E-04
1985-1999	t^2	x	t	x*t	x^2	Constant
ERR	91.3825	8.5685	0.02935	0.01071	0.004722	5.99E-07
1986-2000	t^2	x	t	x*t	x^2	Constant
ERR	91.4177	8.5302	0.03140	0.01171	0.004327	4.87E-04
1987-2001	t^2	x	t	x*t	x^2	Constant
ERR	91.4296	8.5131	0.03571	0.01201	0.003782	0.001637
1988-2002	t^2	x	t	x*t	x^2	Constant
ERR	91.4158	8.5218	0.04276	0.01061	0.002981	0.001385
1989-2003	t^2	x	t	x*t	x^2	Constant
ERR	91.4215	8.5140	0.04666	0.009147	0.002188	0.001022
1990-2004	t^2	x	t	x*t	Constant	x^2
ERR	91.4783	8.4507	0.05300	0.008260	0.002130	0.001649
1991-2005	t^2	x	t	x*t	Constant	x^2
ERR	91.5070	8.4137	0.06368	0.006337	0.001689	9.08E-04

Table 6.6 ERR ranking results for model (6.15) using 10-year rolling window

Period\Rank	1st	2nd	3rd	4th	5th	6th
1971-1980	t^2	x	x^2	x*t	t	Constant
ERR	90.0305	9.9008	0.01131	7.94E-04	7.32E-04	0.0007229
1972-1981	t^2	x	x^2	x*t	Constant	t
ERR	90.2048	9.7247	1.20E-02	9.26E-05	7.04E-04	0.0008740
1973-1982	t^2	x	x^2	x*t	t	Constant
ERR	90.3395	9.5893	0.01031	3.39E-06	8.40E-04	8.51E-05
1974-1983	t^2	x	x^2	t	x*t	Constant
ERR	90.4781	9.4520	9.71E-03	1.89E-04	7.30E-04	0.001458
1975-1984	t^2	x	x^2	t	Constant	x*t
ERR	90.5864	9.3509	9.82E-03	0.001688	3.57E-04	5.55E—04
1976-1985	t^2	x	x^2	Constant	x*t	t
ERR	90.6600	9.2814	8.76E-03	0.001392	4.70E-04	7.60E-05
1977-1986	t^2	x	x^2	t	x*t	Constant
ERR	90.6796	9.2610	0.007610	5.34E-04	8.71E-04	8.76E-06
1978-1987	t^2	x	x^2	t	x*t	Constant
ERR	90.7353	9.2087	0.007983	0.002773	2.44E-04	0.001237
1979-1988	t^2	x	x^2	t	Constant	x*t
ERR	90.9866	8.9796	0.007222	0.003511	8.41E-05	1.45E-04
1980-1989	t^2	x	x^2	t	x*t	Constant
ERR	91.0251	8.9502	0.006213	0.002735	3.90E-04	5.51E-04
1981-1990	t^2	x	t	x^2	x*t	Constant
ERR	90.9782	9.0001	0.007693	0.004291	4.27E-04	8.45E-04
1982-1991	t^2	x	t	x^2	x*t	Constant
ERR	90.9429	9.0376	0.008947	0.004551	0.001206	2.49E-05
1983-1992	t^2	x	t	x^2	x*t	Constant
ERR	90.9597	9.0232	0.01115	0.004839	0.002223	5.11E-04
1984-1993	t^2	x	Constant	x^2	x*t	t
ERR	90.9875	8.9965	0.008235	0.004967	3.66E-03	1.46E-04
1985-1994	t^2	x	Constant	x^2	x*t	t
ERR	91.0596	8.9228	0.01378	0.005216	0.003485	2.28E-04
1986-1995	t^2	x	Constant	x^2	x*t	t
ERR	91.0872	8.8959	0.009565	0.005593	5.58E-03	1.40E-04
1987-1996	t^2	x	x^2	t	x*t	Constant
ERR	91.1853	8.8004	0.008978	0.005349	6.41E-03	2.66E-05
1988-1997	t^2	x	t	x*t	x^2	Constant
ERR	91.2391	8.7449	0.01181	0.005975	0.005341	4.56E-05
1989-1998	t^2	x	t	x^2	x*t	Constant
ERR	91.2965	8.6883	0.01311	0.005146	0.005088	7.52E-05
1990-1999	t^2	x	t	x*t	x^2	Constant
ERR	91.3255	8.6605	0.01193	0.006045	5.04E-03	1.09E-04
1991-2000	t^2	x	t	x*t	x^2	Constant
ERR	91.3789	8.6038	0.01822	0.004762	0.004194	3.93E-04
1992-2001	t^2	x	t	x*t	x^2	Constant
ERR	91.4132	8.5670	0.02372	0.004733	0.003284	1.07E-04
1993-2002	t^2	x	t	x*t	x^2	Constant
ERR	91.4814	8.4955	0.03065	0.003640	0.002214	2.51E-04
1994-2003	t^2	x	t	x*t	x^2	Constant
ERR	91.4872	8.4903	0.02778	0.003165	0.001215	8.87E-07
1995-2004	t^2	x	t	x*t	x^2	Constant
ERR	91.4699	8.5027	0.03745	0.001755	4.89E-04	1.15E-04
1996-2005	t^2	x	t	x*t	Constant	Constant
ERR	91.4865	8.4883	0.03908	0.001239	2.35E-04	7.63E-05

It is obvious that in Table 6.4, 6.5 and 6.6 the terms $Year^2$ and Age in model (6.18) remain the most significant terms during rolling-windows estimation while the rank of other terms are changing. Therefore, the selected model (6.18) produces more robust prediction than the model (6.15).

6.5.3 Mortality model with smoking rate contained

In order to investigate the impact on the mortality rate of other possible variables apart from the year and age relationship, the smoking rate has also been involved into the reference model and the model can be written as

$$\log it\left(q\left(x,t\right)\right)=a_0+a_1x^2+a_2t^2+a_3xt+a_4x+a_5t+a_6S^2+a_7xS+a_8tS+a_9S+\varepsilon\left(x,t\right) \tag{6.20}$$

where S is the smoking rate of males of England and Wales at each year. The smoking rate data is then pre-processed similarly as the year input variables and the mortality rate data from year 1961 to 1980 between ages 60 to 89 is used to apply the OFR algorithm. The term selection results are then listed in Table 6.7.

Table 6.7 The rank of terms in model (6.20) after applying the OFR algorithm

Rank	Term	ERR in %
1	t^2	90.5013
2	x	9.4537
3	x^2	0.0002163
4	Constant	0.00003590
5	xt	0.0004953
6	S^2	0.0001177
7	xS	0.0003751
8	t	0.0005739
9	tS	0.000002570
10	S	0.00009757

According to the results listed in Table 6.7, the smoking rate variable is not obviously significant which means that the mortality rate of males of England and Wales from year 1961 to 1980 between ages 60 and 89 is not directly decided by the smoking rate according to OFR algorithm. Therefore, the selected model based on Table 6.7 is still model (6.18).

This exercise does however illustrate strength of the NARMAX model: it can easily incorporate additional exogenous variables that most conventional models (such as CBD) cannot.

6.6 Conclusions

In this chapter, some of the most recent mortality rate models have been reviewed and a quadratic polynomial mortality model has been proposed to fit the mortality rate surface. The NARMAX modelling method has been used to give a term selection from the proposed polynomial model. Long term prediction comparisons have been given between the proposed model, selected model and according to the comparison results, the proposed quadratic model is the best model to produce minimal prediction errors among the three models. A back testing technology was employed to indicate the importance of term selection of the polynomial mortality model and to compare the proposed model with the CBD model, one of the standard models in the literature. The long term prediction results indicated that the proposed quadratic models can give better long term prediction than the CBD model. But backtesting results indicated that care must be taken with the term selection of the quadratic model. The impact of the smoking rate was also considered, but results suggested that this does not have a significant effect on the mortality rates in our data set.

Chapter 7 Conclusions

Financial volatility forecasting is an important topic in financial risk management and option pricing. Many volatility models have been invented such as the Generalized AutoRegressive Conditional Heteroskedastic (GARCH) class of volatility models and some of them have achieved great success in the financial field. In the major GARCH literature, the mean process is usually fitted using the linear model and the volatility is calculated based on the residuals of the linear mean model. Some papers even treat mean process as a constant. However, much evidence suggests that the mean process should be nonlinear.

The Nonlinear AutoRegressive Moving Average with eXogenous inputs (NARMAX) (Leontaritis and Billings, 1985) modelling technique provides a powerful tool to approximate the nonlinear process with a selected structure. We find that an extension of a NARMAX methodology to fitting the nonlinear mean process can improve the prediction performance of both the mean and volatility models.

This thesis investigated the development of the financial volatility modelling in recent 20 years and the volatility models including major GARCH class of models were introduced. A new volatility model based on the asymmetric GARCH model was proposed. The parameter estimation methods for the GARCH class of models and the popular forecast evaluation methods of volatility models were investigated. Based on the NARMAX term selection algorithm, the impact of heteroskedastic noise on term selection was theoretically derived and a Weighted Orthogonal Forward Regression (WOFR) algorithm was proposed to correct this impact. As the weights in WOFR algorithm are usually unknown, an iterative refined parameter estimation procedure was proposed to improve simultaneously the parameter estimation of both the selected nonlinear mean model and volatility model.

The fitted models need to be validated in order to verify the model assumptions and check the model prediction performance. This thesis proposed to use the Cross

Validation (CV) method to validate the prediction performance of both the mean and volatility models. Since GARCH class of volatility models assume the standard mean residuals are distributed as i.i.d., the Brock-Dechert-Scheinkman (BDS) test was employed to test the standard one-step-ahead prediction during CV. The WOFR algorithm combined with CV method provides a systematic identification method for the nonlinear mean process in the context of financial volatility modelling.

A second application of NARMAX in mortality rate modelling was also provided and a forecast performance comparison between the commonly used Cairns-Blake-Dowd (CBD) mortality model was given to indicate the forecasting superiority of the selected polynomial mortality model.

7.1 Main Contributions of this thesis

This thesis proposed to use NARMAX modelling methodology in the nonlinear mean process during financial volatility modelling to give the term selection and parameter estimation for the nonlinear mean model. NARMAX techniques have been successfully proved to model many real world nonlinear systems. The extension of NARMAX model to financial volatility modelling opened a door in the financial area application of the NARMAX techniques and filled the gap between financial volatility mean process modelling and the nonlinear model term selection. The main contribution of this thesis can be summarised as follows.

(1) In this thesis, the commonly used volatility model such as GARCH class of models and the parameter estimation process of volatility model have been summarized. In some GARCH class of models, a regime switch like term is commonly used to approximate the asymmetry observed in realized data. However, during numerical estimation of the parameters in MLE, the partial differences subject to parameters of regime switch term sometimes jump at the switching point. The logistic STAR function provides a smooth transition for the process and therefore, this thesis proposed a new logistic STAR GARCH model based on the logistic STAR function to model the asymmetry of the volatilities.

(2) In this thesis, the commonly used mean models were investigated in the major GARCH class of models. The GARCH class of models have been developed very fast over the recent twenty years and there are almost a hundred of volatility models which are derived from the ARCH model. Meanwhile, mean process in the financial volatility modelling is typically fitted by ARMA models and almost not treated as nonlinear. Some literature even uses a constant mean model instead. However, this contradicts the evidence of nonlinearity observed in many empirical practices. Therefore, this thesis simulated a nonlinear mean process with time varying volatility derived by a General ARCH (GARCH) model and fitted the mean process with a linear model to estimate the volatility. The results successfully proved that inaccurate mean model could impact heavily on the volatility forecast even with the same volatility model structure.

(3) This thesis derived a new Weighted Orthogonal Forward Regression (WOFR) algorithm to compensate for the impact of the heteroskedastic noise on the term selection of the mean model in financial volatility modelling. NARMAX modelling techniques are based on the assumption of homoskedastic noise. However, heteroskedastic noise usually exists in financial return data and the homoskedasticity assumption of the OFR algorithm is usually violated; this in turn affects the term selection. To deal with this problem, a new WOFR algorithm and iterative refined parameter estimation were successfully applied to improve the mean model term selection and parameter estimation of both the mean and volatility models.

(4) Cross validation (CV) for the mean and volatility model was introduced to validate the prediction performance of the selected nonlinear mean model and the volatility model. In system identification, model validation is essential to verify the model assumptions and the goodness of fit for the underlying process. In the GARCH class of volatility models, the standard mean model residuals are assumed to be distributed as i.i.d. However, an inaccurate mean model or volatility model may cause the rejection of this assumption. This thesis proposed validation of both the mean and the volatility simultaneously by testing the i.i.d. assumption of the

standard mean model residuals and the one-step-ahead prediction errors. Simulations showed that the volatility model estimated from the inaccurate mean model produced larger forecast errors than that from the accurate mean model. Although both the linear and the nonlinear mean model passed the autocorrelation test, the CV method was very effective to reject the inaccurate linear mean model in testing the standard prediction errors.

(5) This thesis applied the NARMAX methodology to model the mortality rate and compared the selected model with CBD mortality model using a backtesting method. Mortality rate forecasting plays a key role in hedging the longevity risk for the pension providers and mortality rate modelling has therefore attracted much attention in recent years. Due to the fact that none of the existing mortality rate models is total satisfactory, the NARMAX modelling method was proposed in this thesis to fit the mortality rate surface. The selected model was mainly a quadratic polynomial model with both year and age factors. In order to compare the results with the statistic mortality rate model, a backtesting method was used to access the prediction performance of the selected nonlinear mortality model. The testing results showed that compared to the CBD mortality rate model, the selected mortality rate model produced better mortality rate predictions. The select quadratic model also had better robustness comparing with unselected quadratic model.

7.2 Suggestion for Further Research

Although a systematic identification method for nonlinear mean modelling of financial return data has been proposed and simulations has successfully proved the effectiveness of this method, the research in the application of the nonlinear modelling approaches to financial volatility is still at a very early stage. Further research may be worth carrying on in the following topics.

(1) In Chapter 4 and Chapter 5, the data used is simulated from a nonlinear mean model and a GARCH model. It is possible to apply the WOFR and CV methods to realized data in further research.

(2) Since different volatility models are based on the GARCH model, it is possible to extend the structure determination of the NARMAX method to the volatility process and so allow the volatility model structure to be determined using a NARMAX model based on the data set instead of an assumed GARCH model.

(3) As mentioned in Chapter 2, according to the transformation of standard GARCH model, the volatility is actually given by squared mean model residuals with some noise. This arises the using of least squared method to estimate the parameter of GARCH model. However, this noise is not distributed as a normal distribution and forecasts of squared residuals cannot be negative. Therefore, a cost function may be found to force the positivity of the forecasts and used to estimate the parameters of the GARCH model.

(4) Fan charts are now very popular in projecting the forecast uncertainties because it is more visually understandable than pure figures and numbers. Therefore, it would be useful to extend the fan chart projections to the NARMAX modelling method to indicate the prediction ability of selected models in a more visible manner.

References

Abhyankar, A. L., Copeland, L., and Wong, W. (1995). "Nonlinear Dynamics in Real- time Equity Martket Indices: Evidence from the U.K." *Economic Journal*(105), 864-880.

Ahlstedt, M. (1998). "Analysis of Financial Risks in a GARCH Framework." *Bank of Finland Studies*(11).

Akgiray, V. (1989). "Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts." *Journal of Business*(62), 55-79.

Andersen, T. G., and Bollerslev, T. (1997). "Answering the Critics: Yes, ARCH Models Do Provide Good Volatility Forecasts." *Working Paper #227, Department of Finance, Kellogg Graduate School of Management, Northwestern University.*

Andersen, T. G., Bollerslev, T. and Lange, S. (1999). "Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon." *Journal of Empirical Finance*(6), 457-477.

Andersen, T. G., Bollerslev T., Diebold, F. and Labys, P. (2001). "The Distribution of Realized Exchange Rate Volatility." *Journal of American Statistis Association*, 96(453), 42-57.

Andersen, T. G., Bollerslev, T. , Diebold, F.X. and Labys, P. (2002). "Modeling and Forecasting Realized Volatility." *Working paper, North Western University, Duke University and Pennsylvania University.*

Antoniadis, A., and Oppenheim, G. (1995). *Wavelets and statistics, in Lecture Notes in Statistics.* New york: Springer-Verlag.

Berndt, E. B., Hall, B., Hall, R., and Hausman, J. (1974). "Estimation and Inference in Nonlinear Structural Models." Annals of Economic and Social Measurement (3), 653-665.

Barnett, W., Gallant, R., and Hinich, M. (1993). "Detection of Nonlinearity and Chaos: Application to Money Stock." *Proceedings of the American Statistical Association: Business and Economics Section.*

Barnett, W., Gallant, R., Hinich, M., Kaplan, D.T., and Jensen, M.J. (1997). "A Single-Blind Controlled Competition among Tests for Nonlinearity and Chaos." *Journal of Econometrics*(82), 157-192.

Benjamin, B., and Soliman, A.S. (1993). "Mortality on the Move", *Institute of Actuaries*. Oxford.

Billings, S. A., and Leontaritis, I. J. (1981). "Idenitification of nonlinear systems using parameter estimation techniques.", *Proceedings of the I.E.E. Conference on Control and its Applications*. Warwich: Englands.

Billings, S. A., and Voon, W. S. F. (1984). "Least Squares Parameter Estimation Algorithms for Non-linear Systems." *International Journal of System Science*, 15(6), 601 - 615.

Billings, S. A., and Voon, W. S. F. (1986). "Correlation based model validity tests for nonlinear models." *Int. Journal of Control*, 44(1), 235-244.

Billings, S. A., Korenberg, M. and Chen, S. (1988). "Identification of nonlinear output-affine systems using an orthogonal least-squares algorithm." *Int. Journal of Systems Science*(19), 1559-1568.

Billings, S. A., Chen, S., and Korenberg, M. J. (1989). "Indentification of MIMO non-linear systems using a forward regression orghogonal estimator." *International Journal of Control*(49), 2157-2189.

Billings, S. A., and Zhu, Q. M. . (1994). "A structure detection algorithm for nonlinear rational models." *International Journal of Control*(59), 1439-1463.

Billings, S. A., and Coca, D. (2001). "Identification of NARMAX and Related Models." *Reseach Report. Department of Automatic Control and System Engineering, University of Sheffield.*(786).

Bjorck, A. (1996). "Numberical methods for least squares problems". Philadelphia: Society for Industrial and Applied Mathematics.

Black, F., and Scholes, M. (1973). "The Pricing of Options and Corporate Labilities." *The Journal of Political Economy.*, 81(3), 637-654.

Blake, D., Cairns, A.J.G., and Dowd, K.,. (2006). "Living with mortality: Longevity bonds and other mortality-linked securities." *Presented to the Faculty of Actuaries, 16 Jan.*

Bloomfield, P., and Watson, G. S. (1975). "The inefficiency of least squares." *Biometrika*, 62(1), 121.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* (31), 307-327.

Bollerslev, T. (1987). "A Conditional Heteroskedastic Time Series Model for Security Prices and Rates of Return Data." *Review of Economics and Statistics*(69), 542-547.

Bollerslev, T., Engle, R.F., and Wooldridge, J.M. (1988). "A capital asset pricing model with time-varying covariances." *The Journal of Politial Economy*, 96(1), 116-131.

Bollerslev, T., and Ghysels, E.,. (1996). "Periodic Autoregressive Conditional Heteroskedasticity." *Journal of Business and Economic Statistics*(14), 139-157.

Bollerslev, T., Engle, R.F., and Nelson, D.B.,. (2007). "ARCH models". *The Handbook of Econometrics*. Amsterdam: Elsevier.

Bollerslev, T. (2008). "Glossary to ARCH (GARCH)." *GREATS Research Paper* 49.

Box, G., and Jenkins, G. (1970). "Time series analysis: Forecasting and control". San Francisco: Holden-Day.

Brock, W. A., Dechert,W.D., and Sheinkman, J.A. (1987). "A Test for Independence Based on the Correlation Dimension." *unpublished manuscript*, Department of Economics,University of Wisconsin, Madison.

Brock, W. A., Hsieh, D. A., and LeBaron, B. (1991). "Nonlinear Dynamics, Choas, and Instability: Statistical Theory and Economic Evidence". London: The MIT Press.

Brock, W. A., Dechert, W.D., Sheinkman, J.A., and LeBaron, B. (1996). "A test for independence based on the correlation dimension." *Econometric Reviews*(15), 197-235.

Brouhns, N., Denuit, M., and Vermunt,J.K. (2002). "A Poisson log-bilinear regression approach to the construction of projected lifetables." *Insurance: Mathematics and Economics*, 31, 373-393.

Cairns, A. J. G., Blake, D., and Dowd, K. (2006). "A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration." *Journal of Risk and Insurance*, 73(687-718).

Cairns, A. J. G., Blake, D., and Dowd, K. (2007). "A quantitative comparison of stochastic mortality models using data from England & Wales and the United States." *Pensions Institute Discussion Paper*, 0701.

Caporale, G. M., Ntantamis, C., Pantelidis, T., and Pittis, N. (2004). "The BDS Test as a Test for the Adequacy of a GARCH (1,1) Specification: A Monte Carlo Study." *Economics Series*, Institute for Advanced Studies, Vienna.

Carroll, R. J., and Cline, D. B. H. (1988). "An asymptotic theory for weighted least-squares with weights estimated by replication." *Biometrika*, 75(1), 35-43.

Chavas, J., and Holt, M. (1991). "On Nolinear Dynamics: The Case of the Pork Cycle." *American Journal of Agricultural Economics*(73), 819-828.

Chen, J., and Shao, J. (1993). "Iterative weighted least squares estimators." *The Annals of Statistics*, 21(2), 1071-1092.

Chen, S., Billings, S. A., and Luo, W. (1989). "least squares methods and their application to non-linear system identification." *International Journal of Control*(50), 1873-1986.

Chen, S., Cowan, C. F. N., and Grant, P. M. (1991). "Orthogonal least-square learning algorithm for radial basis funciton networks." *IEEE Transactions on Neural Networks*(2), 302-309.

Chou, R. Y. (1987). "Volatility Persistence and Stock Returns-Some Empirical Evidence Using GARCH." *Journal of Applied Econometrics*, 3, 279-294.

Chow, W. S., Tan, H. Z, and Fang, Y. (2001). *Nonlinear System Represetation*: JoHn Wiley & Sons, Inc.

Cramer, H., and Wold, H. (1935). "Mortality variations in Sweden: a study in graduation and forecasting." *Scandinavian Actuarial Journal*(18), 161-241.

Devijver, P. A., and Kittler, J. (1982). "Pattern Recognition: A Statistical Approach." Prentice-Hall, London

Ding, Z. X., Granger, C. W. J., and Engle, R. F. (1993). "A long memory property of stock market returns and a new model." *Journal of Empirical Finance*(1), 83-106.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2008). "Backtesting Stochastic Mortality Model: An Ex-Post Evaluation of Multi-Period-Ahead Density Forecasts." *Pensions Institute Discussion Paper*, 0803.

Engle, R. F. (1983). "Estimates of the Variance of U. S. inflation Based upon the ARCH Model." *Journal o Money, Credit and Banking*, 15(3), 286-301.

Engle, R. F., and Bollerslev, T. (1986). "Modelling the Persistence of Conditional Variances." *Econometric Reviews*(5), 1-50.

Engle, R. F. (1989). "Stock Volatility and the Crash of 87: Discussion." *The Review of Financial Studies,*, 3(1), 103-106.

Engle, R. F., and Gonzalez-Rivera, G. (1991). "Semiparametric ARCH Models." *Journal of Business & Economic Statistics*, 9(4), 345-359.

Engle, R. F., and Kroner, K. F.,. (1995). "Multivariate Simultaneous Generalized ARCH." *Econometric Theory* 11(1), 122-150.

Figlewski, S. (1997). "Forecast Volatility". in Finan. Markets, Inst. Instruments. NYU, Salomon Center, pp. 1-88.

French, K. R., Schwert, G.W., and Stambaugh, R.F. (1987). "Expected Stock Returns and Volatility." *Journal of Financial Economics*, 19, 3-29.

Gallant, A. R., Hsieh, D., and Tauchen, G. (1989). "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate." *Working Paper from Chicago-Graduate School of Business*.

Gardner, E. S., Jr. (1985). "Exponential smoothing: the state of the art." *Journal of Forecasting*(4), 1-28.

Gray, S. F. (1992). "Modeling the conditional distribution of interest rates as a regime-switching process." *Journal o Financial Economics*(42), 27-62.

Hamilton, J. D., and Susmel, R. (1994). "Autoregressive conditional heteroskedasticity and changes in regime." *Journal of Econometrics*(64), 307-333.

Heathcote, C., and Higgins, T. (2001). *A Regression Model of Mortality, with Application to the Netherlands.*, Netherlands.

Higgins, M. L., and Bera, A.K. (1992). "A Class of Nonlinear Arch Models." *International Economic Review*, 33(1), 137-158.

Hinich, M. J., and Patterson, D. M. (1985). "Evidence of Nonlinearity in Daily Stock Returns." *Journal of Business & Economic Statistics*, 3(1), 69-77.

Hong, C. (1988). "Options, Volatilities and the Hedge Strategy." *Ph.D. dissertation, University of California, San Diego, Dept. of Economics.*

Hong, X., and Harris, C. J. (2001). "Variable selection algorithm for the construction of MIMO operating point dependent neurofuzzy networks." *IEEE Transactions on Fuzzy Systems*(8), 88-101.

Hsieh, D. A. (1993). "Implications of Nonlinear Dynamics for Financial Risk Management." *The Journal of Financial and Quantitative Analysis*, 28(1), 41-64.

Jarque, C. M., and Bera, A.K. (1980). "Efficient tests for normality, homoscedasticity and serial independence of regression residuals." *Economics Letters*(6), 255-259.

Kawakatsu, H. (2006). "Matrix exponential GARCH." *Journal of Econometrics*(134), 95-128.

Keyfitz, N. (1982). "Choice of funciton for mortality analysis: Effective forecasting depends on a minimum parameter representation." *Theoretical Population Biology*(21), 329-252.

Korenberg, M., Billings, S. A., Liu, Y. P., and Mcilroy, P. J. (1988). "Orthogonal parameter estimation algorithm for non-linear stochastic systems." *International Journal of Control*(48), 193-210.

Krane, S. A. (1963). "Analysis of Survival Data by Regression Techniques." *American Statistical Association and American Society for Quality*, 5(2), 161-174.

Lamoureux, C. G., and Lastrapes, W. D. (1990). "Persistence in Variance, Structural Change, and the GARCH Model." *American Statistical Association*(8), 225-234.

Lebaron, B. (1992). "Some relations between volatility and serial correlations in stock market returns." *The Journal of Business*, 65(2), 199-219.

Lee, K. Y. (1991). "Are the GARCH models best in out-of-sample performance?" *Economics Letters*(37), 305-308.

Lee, R. D., and Carter, L. (1992). "Modelling and forecasting the time series of US mortality." *Journal of the American Statistical Association.*(87), 659-671.

Leontaritis, I. J., and Billings, S. A. (1985). "Input-output parametric models for non-linear systems Part I: deterministic non-linear systems." *International Journal of Control*, 41(2), 303-328.

Leontaritis, I. J., and Billings, S. A. (1985). "Input-output parametric models for non-linear systems Part II: stochastic non-linear systems." *International Journal of Control*, 41(2), 329-344.

Lin, K. (1997). "The ABC's of BDS." *Journal of Computional Intelligence in Finance*(97), 23-26.

Lo, A., and Mackinlay, C. (1988). "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test." *Review of Financial Studies*(1), 41-66.

Lopez, J. A. (2001). "Evaluating the Predictive Accuracy of Volatility Models." *Journal of Forecast*, 20(2), 87-109.

Lumsdaine, R. L. (1991). "Asymptotic Properties of the Quasi-Maximum Likelihood Estimator in GARCH (1, 1) and IGARCH (1, 1) Models." *Mimeo*, Princeton University.

Mandelbrot, B. (1963). "The Variation of Certain Speculative Prices." *Journal of Business*(36), 394-419.

Mangani, R. (2009). "Macroeconomic effects on individual JSE Stocks: a GARCH representation." *Investment Analysts Journal*, 69, 47-57.

McMillan, D., Speight, A., and Apgwilym, O. (2000). "Forecasting UK stock market volatity." *Applied Financial Economics*(10), 435-448.

McNown, R., and Rogers, A. (1989). "Forecasting Mortality: A Parameterized Time Series Approach." *Demography*, 26, 645-660.

Mikosch, T., and Starica, C. (2000). "Limit theory for the sample autocorrelations and extremes of a GARCH (1,1) process." *The Annals of Statistics*, 28, 1427-1451.

Moody, J., and Darken, C. J. (1989). "Fast learning in networks of locally tuned processing units." *Neural Computation*(1), 281-294.

Nelson. (1996). "Modelling stock market volatility changes", in P. Rossi, (ed.), *Modelling Stock Market Volatility* Academic Press, pp. 3-15.

Ogunfunmi, T. (2007). "Adaptive Nonlinear System Identification: The Volterra and Wiener Model Approaches". Santa Clara: Springer.

Pagan, A. R., and Schwert, G. W. (1990). "Alternative Models for Conditional Stock Volatillity." *Journal of Econometrics*, 45(1-2), 267-290.

Renshaw, A. E., and Haberman, S. (2006). "A cohort-based extension to the Lee-Carter model for morality reduction factors." *Insurance: Mathematics and Economics*, 38, 556-570.

Rogers, A., and Gard, K. (1991). "Applications of Heligman/Pollard Model Mortality Schedule." *Population Bulletin of the United Nations*, 30, 70-105.

Rugh, W. J. (1981). *Nonlinear System Theory: The Volterra/Wiener Approach*, Baltimore: Johns Hopkins University Press.

Sentana, E. (1995). "Quadratic ARCH Models." *The Review of Economic Studies*, 62(4), 639-661.

Swain, A. K., and Billings, S. A. (1998). "Weighted complex orthogonal estimator for indentifying linear and non-linear continuous time models from generalized frequency response functions." *Mechanical System and Signal Processing*, 12(2), 269-292.

Tabeau, E., Jeths, A.V.D.B., and Heathcote, C. (2001). "Forecasting Mortality in Developed Countries.". Netherlands: Kluwer Academic Publishers.

Taylor, S. (1986). "Modelling financial time series.". New York: John Wiley & Sons.

Taylor, S. J. (2005). *Asset Price Dynamics, Volatility, and Prediction*, New Jersey: Princeton University Press.

Tsay, R. S. (2002). "Analysis of financial time series: financial econometrics.". New York: Wiley.

Tse, Y. K. (1991). "Stock returns volatility in the Tokyo stock exchange." *Japan and the World Economy*(3), 285-298.

Vilasuso, J. (2002). "Forecasting Exchange Rate Volatility." *Economic Letters*(76), 59-64.

Volterra, V. (1930). "Theory of functionals and of integral and integro-differential equations". Blackie & Son.

Wang, L. X., and Mendel, J. M. (1992). "Fuzzy basis funcitons, universal approximations, and orthogonal least squares learning." *IEEE Transactions on Neural Networks*(3), 807-814.

Wei, H. L., Billings, S. A. and Liu, J. (2004). "Term and variable selection for non-linear system identification." *Int. Journal of Control* 77(1), 86-110.

Weiss, A. A. (1986). "Asymptotic theory for ARCH models: Estimation and testing."*Econometric Theory 2.* City, pp. 107-131.

Wiener, N. (1958). "Nonlinear Problems in Random Theory". New York: Wiley.

Willey, T. (1992). "Testing for nonlinear dependence in daily stock indices." *Journal of Economics and Business*, 44, 63-74.

Zakoian, J. M. (1994). "Threshold Hetroskedastic Models." *JOurnal of Economic Dynamics and Control*, 18, 931-955.

Zhu, Q. M., and Billings, S. A. (1996). "Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks." *International Journal of Control*(64), 871-886.

Zumbach, G. (2002). "Volatility Processes and Volatility Forecast with Long Memory." *working Paper , Olsen Associates.*