University of the Aegean
School of Business Studies
Department of Financial Engineering & Management

New Trends in Financial Engineering: Combining Stochastic and Computational Intelligent Methodologies

By

Nikolaos Thomaidis, DipEng,
MSc in Mathematics and Finance

Submitted in Partial Fulfilment of
the Requirements for the Degree of
Doctor of Philosophy

March 2007

Supervising committee:

Giorgios Dounias	Aristophanis Dimakis	Giorgios Liagouras
Associate Professor	Professor	Assistant Professor

Contents

[1] This chapter also appears in Thomaidis et al. (2007).

Summary

The purpose of this thesis is to introduce a semi-parametric financial forecasting model that combines an intelligent learning technique, artificial neural networks, with common econometric GARCH models of volatility. We show how this flexible modelling framework can accommodate most of the stylised facts reported about financial prices or rates of return (nonlinear corrections, asymmetric GARCH effects and non-gaussian errors). We analytically discuss several strategies for the specification of the mean and variance components of the model by means of sequential statistical tests and propose variations of the standard testing framework that are *robust* to model misspecification, i.e. they preserve their asymptotic validity when the model is not correctly specified for the true conditional distribution. The finite-sample performance of testing procedures is investigated by means of Monte-Carlo simulations. To demonstrate various aspects of the model-building strategy, we present two empirical studies. In the first one, we apply NN-GARCH models to forecasting the conditional distribution of daily returns on three major international stock indexes (DAX, FTSE 100, S&P 500) and in the second one we compare the performance of the sequential testing procedure with other statistical and heuristic neural network model-selection strategies in accurately pricing options on the S&P 500 index.

Acknowledgements

The research presented in this thesis was carried out in the Decision & Management Engineering Laboratory at the Department of Financial Engineering & Management. It is the result of combining my past research interests together with new ideas from the field of computational intelligence.

I would especially like to thank my supervisor Associate Professor Giorgios Dounias for his constant encouragement and support in the development of this research. I am also thankful to Professor Aristophanis Dimakis and Assistant Professor Giorgios Liagouras for willingly accepting to evaluate my thesis and sit in the examination committee.

I am very grateful to Dr Nick Kondakis, CEO and Managing Director of Kepler Asset Management, not only for providing data and software but also for the many stimulating and instructive discussions we had on the practical aspects of my research. Many thanks also go to Associate Professors Jan Jantzen and Uzay Kaymak, research fellows from the EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems (EUNITE) and the Network On Nature-Inspired Smart Information Systems (NISIS), for their guidance and inspiration during the formative stages of this research and Dr Michael Michalopoulos, Director of the Bank of Greece, for providing access to data. I also wish to thank all the members of staff at the Department of Financial Engineering & Management, particularly Mrs Despina Amygdalou for secretarial support, Mr Panagiotis Leondaridis for technical support, and Dr Dimosthenis Drivaliaris and Lecturer Athanassios Koulakiotis for fruitful discussions. I am also thankful to two former undergraduate students of the Department, Vasilios Tzastoudis and Vasilios Vasiliadis, for their assistance during the progression of this research.

This PhD project was financially supported by the Public Benefit Foundation Alexander S. Onassis (http://www.onassis.gr), under the 2003-04 Scholarships Programme, and by a grant from Empirikion Foundation (http://www.empirikion.gr).

List of Figures

List of Tables

Abbreviation List

AIC	Akaike's information criterion
AP	Accumulated profit
APT	Average profit per trade
AR	Autoregressive
ARCH	Autoregressive conditional heteroskedasticity
ARIMA	Autoregressive integrated moving average
BS	Black-Scholes
CAC	The Paris Stock Exchange index
CI	Computational intelligent
DAX	The Frankfurt Stock Exchange index
EGARCH	Exponential GARCH model
ELL	Empirical log-likelihood
FE	Financial Engineering
FML	Full maximum likelihood
FTSE	The Financial Times London Stock Exchange index
GARCH	Generalised autoregressive conditional heteroskedasticity
GJR-GARCH	Glosten et al. (1993) GARCH model
HR	Hit rate
IC	Information criterion
LFEG	The forecasting model used in chapter 6 that comprises a linear AR model in mean with five lags and an EGARCH(1,1) model in variance
LM	Lagrange Multiplier
LS	Least squares
MAE	Mean absolute error
ML	Maximum likelihood
MLE	Maximum likelihood estimator
MSE	Mean squared error
MWSE	Mean weighted squared error
MWAE	Mean weighted absolute error
NASDAQ	A stock index of mainly industry equities
NL3F	The neural network forecasting model used in chapter 6 with three neurons in the hidden layer and no volatility component
NIKKEI	The Tokyo Stock Exchange financial index
NMAE	Normalised mean absolute error
NN	Neural network
OLS	Ordinary least squares

QML	Quasi maximum likelihood
QMLE	Quasi maximum likelihood estimator
RB-LM	The robustified LM test or the forecasting model designed by this type of test
RBV-LM	The robustified LM test with volatility estimates or the forecasting model designed by this type of test
RBV-LM1	A RBV-LM test in which the volatility model is separately estimated from the mean model
RBV-LM2	A RBV-LM test in which the volatility model is jointly estimated with the mean model, using the full maximum likelihood
S&P	The Standard and Poor's New York Stock Exchange index
SBIC	Schwarz's Bayesian information criterion
S-LM	The standard (non-robust) Lagrange Multiplier test or the forecasting model designed by this type of test
SP	Sequential pruning
SST	Sequential statistical testing
VaR	Value at Risk

Preface

I Motivation

In the past twenty years, both financial science and practice have changed in a truly fundamental way. These changes were fuelled in part by remarkable advances in financial theory and modelling techniques, the liberalisation of financial markets and the development of complex financial products (derivatives), the growth of information and computer technology and the overall increase in financial uncertainty followed some unprecedent financial disasters in the 90's. This emerging financial environment coupled with the advancing theory has led to a critical demand on sophisticated tools and methods for assessing, pricing and forecasting increasingly complex financial outcomes. This is the purpose of *financial engineering* (FE), a recently emerged cross-disciplinary field that combines financial theory, mathematical techniques, numerical methods and computer algorithms to solve financial problems. Utilizing a variety of techniques, practitioners of FE aim to derive methods for forecasting the market price of primarily and derivative securities, quantifying the risk associated with trading in these assets and designing portfolios that best meet the financial purposes and needs of the issuing authority.

Traditionally, probabilistic concepts/techniques such as expectations, distributions, stochastic processes and martingales were used to model the uncertainty in financial markets and to construct solutions to financial problems. On the other hand, recent advances in computer science have led to the development of computer algorithms that simulate elements of learning, adaptation and evolution to create programmes characterised by "intelligent" behaviour. *Computational intelligence* (CI) embraces a variety of modelling techniques, from Artificial Neural Networks, Fuzzy Inference Systems and Evolutionary Computation to Swarm Intelligent techniques and Artificial Immune Systems. The sophistication of these methodologies, the significant increase in computing power and the need for models that rely on more realistic assumptions inspired researchers to revise traditional solutions to FE problems. The result has been an explosive number of intelligent applications in financial engineering tasks with significant improvements in many cases.

Ever since computational intelligent techniques were used in application domains with a strong statistical culture, there appeared a tendency to apply well-known econometric principles in the construction and testing of CI models. The ultimate purpose of econometrics is to derive models that approximate the probabilistic relationship between the target and a set of explanatory variables, which is summarised by the *conditional probability distribution* or *probability density function.* This is a measure of the relative frequency of the target variable taking values in a specific interval given the values of the explanatory variables. Finding an acceptable approximation to the real density typically requires a *specification step*, by which the researcher restricts attention to a parametric family of densities, an *estimation step*, by which the researcher finds the member of the family that offers the optimal approximation to the real density, and a *diagnostic step*, in which the researcher judges the quality or the redundancy of the current approximation. Model-building iterates between the above steps until an acceptable specification is found.

Depending on the class of specifications the researcher restricts attention to, different aspects of the conditional distribution are modelled. Most computational learning techniques essentially employ a regression model that focuses on mean dependencies between the target and explanatory variables. This is because these algorithms have been initially designed for engineering application domains, such as signal processing, where either the conditional mean is the only object of interest to the researcher or the "noise" contained in measurements has no special statistical features resembling a memoryless process with constant variance. The empirical study of financial time-series, however, has revealed that asset price changes or rates of return follow more complex statistical laws, which highly deviate from the *"white-noise"* prototype. The empirical distribution of stock returns or exchange rates exhibit a substantial degree of asymmetry and kurtosis. Additionally, the conditional distribution is not invariant but changes in dependence to the recent history of price movements. In particular, large unanticipated fluctuations tend to increase on average the short-term variability of price changes, resulting in a wider conditional distribution. This effect is known in the literature as *autoregressive conditional heteroskedasticity* (ARCH)[2]. Quite often, the sign of returns is important to the determination of volatility, so that a sudden price drop has a different impact on future uncertainty about the realised return than an unexpected rise.

Such asymmetries in the short-term dynamic behaviour of asset prices are also observed when an asset is *mispriced*, i.e. its market price differs from the fundamental value. Classical asset-pricing models assume that securities traded in the market are on average correctly priced, so that the deviation between market and fundamental value does not contain any predictive information about future price movements. However, numerous empirical studies on market prices have confirmed large and persistent deviations from fundamental levels, which led scientists to review their interpretation of the price-formation mechanism. Modern theories of *behavioural finance* support that the trading activity of irrational "trend-chasing" investors drives prices away from fundamentals. Due to institutional restrictions and transaction costs, rational traders are not able to instantly "arbitrage away" pricing inefficiencies and thus bring prices back to fundamentals. The larger is the deviation the more likely is a future price movement in the direction that corrects the mispricing, although the speed of correction may depend on the level and/or the sign (direction) of mispricing. This fact results in a complex dynamic behaviour, in which the probability distribution of future returns may depend both on the size and sign of the mispricing as well as on recent price movements.

[2] A time-series is said to be heteroskedastic if its variance exhibits temporal changes.

One of the challenges of financial engineering is to develop models that are capable of capturing the main aspects of the statistical distribution of price movements and returns. Based however on the above discussion on the statistical properties of financial time-series, we realise that in order to obtain a faithful representation of the dynamic behaviour of financial instruments we need to consider a general class of models that could possibly accommodate nonlinear adjustments, symmetric or asymmetric ARCH effects and non-gaussian distributions. These models can in turn be used in common prediction or forecasting tasks or as simulators of realistic future market scenarios for pricing more advanced derivative products, such as options, futures and swaps.

II Main contribution

The purpose of this thesis is to present a new class of time-series models that combine a computational intelligent method, artificial neural networks, with econometric GARCH models of volatility. This flexible modelling framework can accommodate most of the stylised facts associated with financial time-series. The NN-GARCH models considered in this thesis belong to the general class of dynamic models that *jointly* parametrise the mean and the variance of the conditional distribution. They combine a neural network model for capturing nonlinear adjustments in the conditional mean of the distribution with GARCH parametrisations for the modelling of the conditional variance dynamics. By jointly modelling the conditional mean and volatility of the data-generating process, we extend the scope of NNs from function approximation to *density forecasting* tasks and we also reconsider the construction of neural network models under special statistical features existing in financial and economic data. This combination enables also the researcher to investigate interesting hypotheses concerning both the mean and the variance structure of the data-generating process. Joint mean-variance models are nowadays very popular in economic and financial time-series applications and their statistical properties are well-documented (see Bollerslev and Wooldridge (1992)).

In the application of any flexible semi-parametric methodology, the issue of choosing the *model specification* becomes of paramount importance. Not all data sets share the same features and therefore in any practical application it is important to determine the level of model complexity that is most appropriate for the particular data set. In this thesis, we devote much attention to this issue. Following the econometric tradition, we discuss several model-building strategies for NN-GARCH parametrisations that are based on solid statistical procedures rather than heuristic criteria. In this way, model-selection becomes a transparent and clearly-defined process that is applicable to a wide range of data sets.

In econometrics, there are generally two procedures for reaching the final model specification. One starts with a large, possibly "over-parametrised", model and simplifies the structure by "pruning" redundant or insignificant components. This is the so-called "*top-down*", *pruning* or *backward* approach to model specification, often encountered in computational intelligent techniques. When trying however to combine pruning with statistical inference, a puzzling theoretical issue arises that is related to the possibility that the large model is *redundant* for the particular data sets. Redundancy generally implies that a subset of model parameters can take any value without effecting its output and are thus *non-identifiable.* The non-identifiability of parameters has important consequences for the asymptotic distribution of common test statistics (Wald, Lagrange Multiplier, Information Criteria) and renders them inappropriate to use in significance tests.

These theoretical considerations led us to adopt the opposite ("*bottom-up*" or *forward*) route in model specification that starts with the simplest possible model and gradually complicates the structure in the direction indicated by special features existing in data[3]. If, for example, the price of an asset responds asymmetrically to previous movements, extra neurons are added to capture this effect. If nevertheless the residuals of the neural network model are heterogeneously distributed and the conditional volatility tends to cluster, an extra GARCH model is placed to parametrise the volatility dynamics. The model-specification strategy that we adopt in this thesis is based on *sequential Lagrange-Multiplier (LM) hypotheses tests* of neglected structure in the mean and the variance equation of the model. Roughly speaking, the model-building procedure begins with a linear model of the conditional mean. The residuals are then used to compute regression-based tests for additional nonlinearity, serial correlation or other misspecifications and further neurons are added according to the test results. Once the model for the conditional mean is deemed satisfactory, additional tests are carried out to detect strong features in the conditional variance of the distribution, such as ARCH heteroskedasticity. On the basis of these tests, GARCH models for the conditional variance are estimated in conjunction with the mean equation and tested against other effects.

The tests involved in each intermediate step of the above procedure are generally simple and inexpensive to construct compared to other computationally intensive specification procedures. LM tests involve only the computation of first derivatives and the asymptotic null distributions of the test statistics are standard (chi-square) and well-tabulated. Each time a new candidate model is tested against the existing specification, the researcher does not have to re-estimate the full model, as tests are directly computable from the model imposed by the null hypothesis by running a set of *auxiliary regressions.* This is a big advantage for highly nonlinear specifications, as it significantly lowers the computational burden that would be associated with estimating all candidate model.

While this "bottom-up" strategy follows a natural progression, its ability to successfully identify the data structure is very much determined by the type of specification tests employed. Generally, most of specification tests currently applied in the literature implicitly impose additional assumptions that are only tested on subsequent stages of the model-building cycle. For example, standard LM tests for additional nonlinearity or serial correlation in mean implicitly assume homogeneously distributed errors and are thus invalid in the presence of ARCH or other forms of heteroskedasticity. As simulation results show in this thesis, when applied to heteroskedastic data, they generally misinterpret consecutive large price movements, attributed to a temporal increase in the volatility level, as systematic nonlinearity

[3] This type of strategy in model selection is also termed as forward.

or serial correlation and hence spuriously indicate additional structure in the mean equation (more neurons or more lags) to capture these effects. As these price movements however are not due to systematic market reactions, the extended model is expected to have poorer performance on new unseen data. One strategy to avoid this deficiency of the testing procedure would be to continue the model-building cycle, specify a model for volatility and then go back and test the conditional mean equation. However, this procedure is not theoretically justified because the estimation of the volatility model is also based on the assumption that the conditional mean model is correctly specified.

Two remedies are applied in this thesis against the adverse effects of changing variance on the validity of mean specification tests. The first is to *ignore heteroskedasticity* and perform the LM test in way that is *robust* to general forms of heteroskedasticity. This approach is based on the general regression-based robust-testing framework, suggested by Wooldridge (1991), and also commonly applied in the context of other nonlinear econometric models. As simulations presented in this thesis show, robustified LM tests manage to closely follow the nominal type I error under heteroskedasticity, allowing thus the researcher to control the complexity of the mean model. However, this improvement comes occasionally with a loss of efficiency in detecting hidden nonlinearity or serial correlation in mean, which poses a problem if one is equally interested in modelling strong features in mean apart from heteroskedastic effects. The other approach is to *model heteroskedasticity*, i.e. incorporate information from the volatility structure during the testing procedure, which seems to offer a good choice between ignoring heteroskedasticity and using a non-robust test. To our knowledge, this strategy of model specification has not been previously applied in the context of neural networks and other popular nonlinear specifications (e.g. smooth transition or threshold models) encountered in the literature. The new test is carried out in a way that is robust to misspecification of the variance model, which means that the researcher does *not* need to explicitly model the volatility dynamics in order to obtain a valid testing procedure. However, as simulation results show, the more accurate is the model the higher is the efficiency of the LM test to identify neglected serial correlation or nonlinearity in the mean equation. Hence, when it comes to model construction, it always pays to put some effort on deriving a good approximation to both moments of the conditional distribution.

Besides the specification of a NN-GARCH model, a considerable part of this thesis is also devoted to derive procedures for in-sample statistical evaluation of the model. Based on the maximum likelihood theory, we device Wald-type tests for testing the *joint significance* of parameters of an estimated NN-GARCH model and thus offer the opportunity to the researcher to investigate hypotheses of interest regarding the data-generating process. We also present a series of *diagnostics*, based on LM tests, that examine the extend to which the derived specification is a faithful approximation to the conditional distribution. The distinguishing feature of these testing procedure is that they lead to valid inference regarding insignificance or structural misspecification even in the case where the distributional assumptions made by the model are *not* correct (i.e. the empirical density of standardised errors is fat-tailed or asymmetric). This is a quite useful feature as it permits statistical inference without having to explicitly model all aspects of the conditional distribution.

Once a NN-GARCH model has been specified for a particular data set, it can be used in deriving *density predictions* or *confidence intervals* on the future value of the variable of interest. The issue of forecasting with NN-GARCH models is also discussed in this thesis. We present a technique for obtaining one- as well as multi-step-ahead predictions, which avoids any restrictive assumption on the conditional density of errors (normality, symmetry, etc), and also discuss methods for evaluating model forecasts. Quite often in the econometric literature, forecasts are evaluated on a pure statistical basis by employing criteria that concentrate on the goodness-of-fit to the data. However, from a financial engineering perspective, the *economic significance* of forecasts is also an important issue. To this end, we give many ideas on how to design a trading strategy suitable for NN-GARCH or other density models that takes into account confidence bounds rather than point forecasts, as is customary in the NN literature. This is based on the detection of *"exceptional"* price movements.

The performance of the sequential statistical testing procedure, examined in this thesis, is also compared with other statistical and heuristic neural network model-selection techniques in a special application: the pricing of derivative securities. The purpose of this study is to derive "flexible" models that can be used for the pricing of option derivative contracts, relaxing the restrictive assumptions of parametric models, such as the famous Black-Scholes.

In summary, the main points of contribution of the thesis are the following:

1. We propose a new class of combined neural network GARCH models that can accommodate major properties of financial prices or rates of return (nonlinear correction, ARCH effects, leptokurtic error distribution).

2. NN-GARCH models provide *density predictions* in the form of *confidence intervals* on the future value of the target variable. Therefore, they give a better description of the underlying dynamics and, most important, of the risk associated with trading in the securities represented by the corresponding time-series data.

3. We examine various strategies for *model selection*, i.e. the determination of the number of hidden neurons in the mean and the level of complexity in the variance equation. The proposed tests are generally simple to implement and computationally inexpensive, as the model has been already estimated under the null hypothesis.

4. Model-selection strategies are also *robust*, meaning that they give the opportunity to the researcher to progressively build the model without referring to the properties of higher moments of the conditional distribution. Hence, when investigating mean relations, the researcher does not have to explicitly model the variance dynamics or when structuring the variance model, the researcher does not have to impose restrictive assumptions on the error distribution.

5. One strategy that we propose in this thesis for the specification of the neural network model uses information from the conditional variance structure. This procedure seems to increase the efficiency of the mean tests in separating nonlinearities in mean from heteroskedasticity.

6. We device Wald-type tests for testing the joint significance of the parameters of a pre-estimated NN-GARCH model and thus offer the opportunity to the researcher to investigate hypotheses of interest regarding the "true" data-generating process.

7. We also present a series of simple diagnostic tests that examine the extend to which the derived specification is a faithful approximation to the conditional distribution.

8. The distinguishing feature of these testing procedures is that they lead to valid inference even in the case where the distributional model is *inappropriate*, i.e. the empirical density of standardised errors is fat-tailed or even asymmetric.

9. We investigate the finite-sample performance of specification/diagnostic tests in a realistic simulation environment, which takes into account common statistical features of financial time-series data (heteroskedasticity, non-normality).

10. We present a technique for obtaining one- as well as multi-step-ahead predictions based on a NN-GARCH model, which avoids any restrictive assumption on the conditional density of errors (normality, symmetry, etc).

11. We design a trading strategy suitable for NN-GARCH or other density models that takes into account confidence bounds rather than point forecasts. This strategy is based on the detection of statistically *"exceptional"* price movements and hence attains a better control on the risk associated with trading positions.

12. We derive "flexible" semi-parametric neural network models that can be used for the pricing of option-type derivative contracts. We also present a *hybrid* methodology for estimating/forecasting the price of stock options that takes into account relevant parameters from established models and augments them using non-linear neural network techniques. The emphasis is on model selection strategies that can lead to optimal performance of the hybrid model.

III Organisation of the thesis

This thesis is organised into eight chapters:

Chapters 1 to 3 provide the essential background on the topics covered in our research. Chapter 1 discusses financial engineering, computational intelligence and artificial neural networks. Chapter 2 reviews characteristic features or *stylised facts* associated with the behaviour of financial prices (leptokurtosis, volatility clustering, leverage effects and nonlinear adjustments) and show how these features generally convey information about the probability distributions that govern future prices. This chapter provides also an introduction to the dynamic behaviour of asset prices, based on empirical evidence and modern theories of behavioural finance. Chapter 3 discusses the principles of econometric modelling with emphasis on the basic stages of the model-building cycle (specification, estimation, evaluation).

Chapter 4 introduces NN-GARCH models and discusses several statistical and numerical issues pertaining to the estimation and the specification of these models. To examine the extend to which our testing procedures can identify the right complexity of the data-generating process (number of hidden units, serial correlation), we perform two Monte-Carlo simulation studies, whose details are given in chapter 5.

The empirical properties of the modelling techniques presented in this thesis is investigated in chapters 6 and 7. Chapter 6 presents an empirical application of NN-GARCH models to predicting the conditional distribution of daily returns on three major international stock indexes (DAX, FTSE 100, S&P 500). This study gives us the opportunity to discuss several issues related to forecasting with NN-GARCH models. Special emphasis is given to the computation of one- and multi-period predictions, the construction of confidence intervals and the evaluation of the forecasting performance. In chapter 7 we provide another empirical study that compares the performance of the sequential testing procedure with other statistical and heuristic neural network model-selection strategies in accurately pricing options on the S&P 500 stock index.

This text concludes with chapter 8 that summarises the main findings and results. It also emphasises on the major contributions of this work to the research fields of financial engineering, computational intelligence and econometrics, and points to directions of future research.

Chapter 1

Essential background

1.1 Introduction

The purpose of this chapter is to provide the essential background on the topics covered in this thesis. Section 1.2 introduces the reader to financial engineering and discusses common areas of research in this discipline. Section 1.3 is devoted to the presentation of computational intelligent (CI) algorithms, a modern approach to extracting knowledge from data that has recently become popular in financial engineering. Popular intelligent learning models, such as artificial neural networks, neuro-fuzzy inference systems and genetically-evolved regressions, have been successfully applied to a wide range of complex financial tasks, with rather promising results (an up-to-date review of financial applications with CI models is provided). This apparent success has attracted the interest of many researchers in quantitative finance and econometrics, as most of the CI techniques have fundamental differences from conventional statistical methods. Section 1.4 discusses Artificial Neural Networks, a class of parallel semi-parametric computational models that have been mostly appreciated among all CI methods by practitioners and researchers in financial engineering. After a brief introduction to the topic, we concentrate the discussion on methods and techniques for determining the architecture of a NN model, an important issue in practical applications. Section 1.5 summarises and concludes the chapter.

1.2 Financial Engineering

1.2.1 The new trend

By the late 80's finance was changing in a truly fundamental way. This change was driven in large part by the globalisation and automation of financial markets, the development of complex financial products, the new regulatory framework concerning the management of financial risk and some remarkable advances in financial modelling. The challenges of this emerging financial environment, coupled with the advancing theory, gave rise to a new area of research and development characterised by the term *financial engineering*. Broadly speaking, financial engineering (FE) is the use of financial products or instruments, such as stocks, futures, swaps, options, to restructure or rearrange an existing financial profile in order to achieve certain financial goals, particularly the management of financial risk[1].

Just as any engineering field, FE is based upon fundamental sciences, such as economics or finance, that seek to understand the principles governing financial phenomena, and therefore they are unavoidably abstractive. Scientific theories rest on a series of assumptions or "*idealisations*" of the real world that, although they capture some of the essential features, they are not very accurate descriptions of the characteristic behavior of financial objects. Real markets are often filled with anomalies that disagree with standard economic theories. These anomalies lead to characteristic patterns of price evolution that cannot be successfully described by a physical model, such as the Brownian motion, or to explain by an economic theory of rational investment behaviour. Although the purpose of financial science is to "simplify and decompose in order to understand", engineering involves structuring solutions to *complex* problems by *creative* development and composition of modelling techniques, theories and financial instruments[2].

1.2.2 Some fundamental problems in Financial Engineering

Securities pricing

In most financial markets, one typically identifies two classes of securities: *primitive* and *derivative.* Primitive securities are tradable assets, such as stocks or bonds, whose price largely depends on the financial status of the issuer. Derivative securities, such as options, futures and swaps, derive their value from the performance of other underlying, mostly primitive, securities[3]. One of the main concerns of financial engineering is the development of theoretical or empirical models that describe the price evolution of both primitive and derivative securities. Most pricing models suggested in the financial literature typically assume a well-developed market with many tradable assets and enough rational agents who make investment decisions based on *fundamental* information regarding the prospects of each asset. All the information available to investors is almost instantly reflected to market asset prices and one cannot continuously make excessive trading profits based on publicly available information. This is the well-known *efficient markets hypothesis*, perhaps one of the most debated concepts in modern finance (see Fama (1970, 1991, 1998); Farmer and Lo (1999)).

Popular frameworks for the pricing of primitive assets such as common stocks, are the *Capital Asset Pricing Model* (CAPM), developed by Sharpe (1964); Litner (1965); Mossin (1966), or the *Arbitrage Pricing Theory*, which dates back to the seminar contributions of Ross (1976, 1977). The underlying idea of both models is to relate

[1] See e.g. the website of International Association of Financial Engineers (accessed from http://www.iafe.org at December 2006) and Galitz (1995) for relative definitions.

[2] The on-line magazine *Financial Engineering News* accessed from http://www.fenews.com/what-is-fe/what-is-fe.html [February 2007] provides a discussion on financial engineering and how this is related to other scientific or engineering fields. For some general references on financial engineering, the reader is referred to Mulvey et al. (1997); Zenios (2002) and the special issues of the *Annals of Operations Research* (Konno et al. (1993); Vladimirou (2007a,b)).

[3] See e.g. http://financial-dictionary.thefreedictionary.com [December 2006] for definitions of terms.

the expected rate of return on primitive securities to their exposure on a number of *fundamental factors* driving the market (stock indexes or other economic indicators). Based on a set of assumption discussed above, they suggest that what is important in the determination of the price of an asset is its exposure to fundamental factors, what is called the *fundamental* or *factor*-related *risk*, and not the risk associated with the issuing company. This is because in a well-developed market with many tradable assets, "idiosyncratic" risk can be diversified away by holding a proper portfolio of primitive securities.

The approach that is often adopted in the pricing of derivative assets is the no-*arbitrage* framework, initiated by Black and Scholes (1973) and developed by Merton (1973), Harrison and Kreps (1979) and Harrison and Pliska (1981). Arbitrage is the act of simultaneously buying and selling assets in order to create a riskless and profitable investment opportunity. The starting point in a no-arbitrage pricing theory is that a well-organised financial market should not permit the persistent gain of profits without any exposure to risk. Therefore, any arbitrage opportunities will be sooner or latter exploited by investors, ensuring that prices reach an equilibrium state where arbitrage is no more feasible. The assumption of arbitrage lies behind the famous Black and Scholes (1973) model for European-type options as well other parametric derivative pricing models, such as the Jump-Diffusion (Merton (1976)), Constant Elasticity of Variance (Cox et al. (1976)), and Hull and White (1987)'s Stochastic Volatility.

It is commonly believed that financial engineering is mainly concerned with pricing derivative instruments using arbitrage arguments, while the pricing of primitive assets is in the scope of financial economics. However, this distinction is illusive as the price evolution of primitive assets is an important determinant of the derivative pricing model.

Risk analysis

Following the overall increase in financial uncertainty in the 90's associated with a number of famous financial disasters, there has been an intensive research from financial institutions, regulators and academics to develop more sophisticated tools for properly measuring and managing financial risk[4]. The central concept was the *Value-at-risk* (VaR), a statistical measure of how the market value of an asset or a portfolio of assets is likely to decrease over a certain time period under usual market conditions. The calculation of a VaR statistic involves three parameters: the investment horizon, the confidence level and a loss amount (or loss percentage). All these components are important in the quantification of the risk of a trading position. To illustrate the above concept, assume that the predicted distribution of the price of an asset ten days ahead is given by figure 1.1. Let P_0 be the current value of the asset. Note that with probability $(1-\alpha)$ the maximum downfall is not expected to be below P_1, the "worse" $\alpha\%$ quantile of the distribution. Hence, the maximum loss, i.e. the VaR, at the given significance level is equal to the difference $P_0 - P_1$.

Note that the Value-at-Risk is an estimator of the tails of the empirical distribution of the portfolio value after a certain period of time. Hence, for accurate

[4] Dowd (1998); Jorion (1997); Ris (1996); Basle (1996).

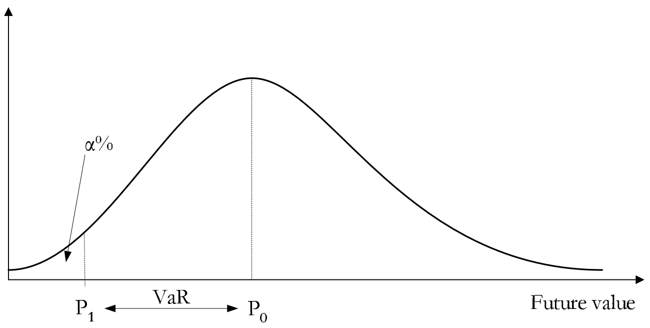

Figure 1.1: The calculation of Value-at-Risk.

quantification of risk one needs a model for the time-evolution of the joint distribution of the assets composing the portfolio. Various methods are available in the literature (see Jorion (1997); Sinha and Chamu (2000) for good surveys) each of which has its own advantages and disadvantages. It is important to note that if asset returns were normally distributed, VaR would be a function of the standard deviation of returns. However, it is widely documented that the empirical distribution of asset returns exhibit richer statistical properties, such as skewness and excess kurtosis, that raise the need for more sophisticated models than a gaussian distribution. These properties are very crucial in the modelling of financial asset prices and hence are analytically discussed in chapter 2.

Portfolio optimisation

The ultimate goal of financial engineering is the design of a trading strategy, i.e. a rule of allocating capital between different assets, that optimises the objectives posed by the investor under various institutional constraints. Depending on these objectives we obtain a number of portfolio construction methods. The traditional mean-variance portfolio analysis, originated with Markowitz (1952), attempts to simultaneously maximise the expected return on a portfolio and minimise its expected risk, as measured by the standard deviation of returns. To this end, historical quantitative analysis and forecasting are carried out in order to optimise the risk/return relationship. Other portfolio selection methods are not based on the maximisation of expected profit but rather concentrate on the avoidance of “bad outcomes”. Using the VaR framework, for example, portfolio managers allocate assets in a way that the VaR of a portfolio is not greater than a pre-specified amount, the maximum allowed loss that will only be exceeded with a small probability[5].

1.3 Models of Computational Intelligence

The last two decades have experienced a growing development of computer-based algorithms inspired by

[5] See Elton et al. (2003) and Michalopoulos et al. (2004), section 3, for a comprehensive discussion on portfolio selection methods.

- various aspects of physical human intelligence (knowledge acquisition and representation, approximate reasoning, etc.)
- biological structures or formations (cells, organs, human brain, ant colonies, etc.)
- and biological processes (evolution of species)

Popular examples of these methodologies are artificial neural networks, adaptive neuro-fuzzy inference systems (ANFIS) and genetically-evolved models. Although most of these intelligent-learning paradigms are in principle different in nature and have historically emerged from a variety of scientific fields, they are nowadays grouped under the umbrella of *computational intelligence* (CI). CI models manifest themselves in a great variety of forms and combinations and a detailed exposition of the topic would certainly go beyond the scope of our analysis. Therefore, in what follows we avoid getting into the details of each technique and rather focus on general characteristics which seem to be the unifying features of these intelligent methodologies. For a comprehensive and systematic introduction to computational intelligence methods, the reader is referred to Chen (2000); Engelbrecht (2003); Konar (2005). In section 1.4, we provide an analytical discussion on artificial neural networks (NN), the type of CI methodology employed in our thesis. Further references are given in the text when necessary.

Intelligent tools and techniques have been successfully applied to a wide range of complex application domains with promising modelling and forecasting results. The growing interest in this type of methodologies is generally justified by the fact that they are purely *empirical* or *data-driven* and offer very flexible model specifications that rest on few (if any) assumptions on the data-generating process. Many intelligent learning methods, such as feedforward NNs or Takagi-Sugeno fuzzy inference systems (Jang et al. (1996); Takagi and Sugeno (1985)), possess a *universal approximation property*, which means that under weak assumptions they are capable of approximating highly nonlinear mappings to an arbitrary degree of accuracy[6,7]. This is a quite desirable property as in many application domains (especially in economics and finance) theory can not fully guide the model-building process by suggesting the relevant input variables or the correct functional form that an empirical model should take.

[6] See Tikk et al. (2003) for a comprehensive review of the universal approximation properties of common intelligent techniques.

[7] As computational intelligent models do not make explicit assumptions on the type of relations inherent in data, they are characterised as *semi-* or *non-parametric* to discriminate them from other models that cannot accommodate any functional form (e.g. linear regressions). In this sense, they are comparable to other non-parametric techniques, such as wavelets, splines, projection pursuit or kernel smothers, often used in empirical statistics. The interested reader is refereed to Hardle et al. (1997) for a comprehensive review on non-parametric time-series analysis and Percival and Walden (2006); Gencay et al. (2001) for time-series and economic/financial applications of wavelets. Cherkassky et al. (1996) present a study whose purpose is to compare the predictive performance of various function estimation methods, including artificial neural networks, projection pursuit and adaptive regression splines. Based on artificially-generated data sets, they show that no single method is superior to the others and each method's performance depends significantly on the type of the target function and on the properties of the data used in model selection/estimation.

Another difference between CI and conventional statistical models lies in the representation of knowledge obtained from data. Common statistical empirical methods, such as regression models, are in fact mathematical equations. However, many intelligent paradigms provide alternative representations of the underlying data relationships, such as a system of logical rules or a decision tree, that are often easier to understand, interpret and validate by experts.

Intelligent modelling gives finally the researcher great flexibility in combining individual algorithms to create *hybrid* systems which share the advantages and minimise the disadvantages of individual schemes[8]. A neuro-fuzzy combination, for example, keeps the powerful approximating capability of NNs while it adds much to the comprehensibility of the induced model. An evolutionary algorithm, such as genetic programming, is often useful for performing a more consistent exploration of the space of possible network topologies (or else model specifications).

The last two decades have experienced an explosion in the number of applications of intelligent learning methodologies in time-series forecasting. A great majority of intelligent approaches employ a network learning technique, such as feedforward, radial basis function or recurrent NN (Zhang et al. (1998); Swanson and White (1997b)), although certain paradigms such as genetically-evolved regression models (Cortez et al. (2001); Farley and Jones (1996); Szpiro (1997); Koza (1991)) or inductive fuzzy inference systems (Fiordaliso (1998)) are also encountered in the literature. CI methodologies have been particularly popular in financial applications, including the prediction of stock prices and interest rates, the pricing of derivatives and the forecasting of foreign exchange rates and volatility. Table 1.1 gives a indicative list of references classified by the application task and the type of the learning technology (neural, fuzzy, evolutionary, etc.). Apparently, due to the rapidly growing literature it would be difficult to provide an exhaustive classification of intelligent applications in financial time-series. Besides, CI methods are often more difficult to categorise as they appear in many variations and different combinations (i.e. hybrid systems). Nevertheless, table 1.1 gives a good indication of the research trends in this area. More references of CI applications in financial engineering can be found in Abu-Mostafa et al. (2001); Trippi and Turban (1996); Chen (2002); Azoff (1994); Kingdon et al. (1997).

Computational intelligent algorithms are flexible semi-parametric models and this seems to be an important desirable property as concerns complex real-life applications. However, the "atheoretical" nature of CI models can also be the source of many problems in applied work. A too complex CI model may not only capture salient features of the relationship between variables of interest but also random effects pertaining to the particular data set. In this sense, the model is redundant for the particular data set and may *overfit* the data. Overfitting is a serious problem because although it reduces the approximation error in the *training* data set (i.e. the data used to specify the structure and estimate the values of the parameters of the model) it leads to bad performance on "*unseen*" data (i.e. *out-of-sample*). The possibility that a particular CI model be over-adequate for a particular data set raises the importance of a *model-building* strategy whose purpose is to determine the *optimal*

[8] See e.g. Tsakonas and Dounias (2002) for a review of this approach.

	Neural Network Architectures	Fuzzy Systems	Evolutionary Schemes
Stock markets	McCluskey (1993); Bergerson and Wunsch (1991); Kohara et al. (1997); Schoneburg (1990); Kimoto et al. (1990); White (1988a)	Fiordaliso (1998); Singh and Fieldsend (2000); Setnes and van Drempt (1999)	Chen and Yeh (1997); Nikolaev and Iba (2001); Kanungo (2004); McCluskey (1993)
Term Structure	Barucci and Landi (1993); Deboeck and Cader (1994)	Mohammadian and Kingham (2004)	Mohammadian and Kingham (2004)
Foreign Exchange	Weigend et al. (1992); Kuan and Liu (1995); Harm and Steurer (1996); Bolland et al. (1998); Tenti (1996); Nag and Mitra (2002)	Kim and Kim (1997); Li et al. (1995); Ghoshray (1996a,b)	Nag and Mitra (2002); Kim and Kim (1997)
Derivatives	Garcia and Gencay (2000); Hutchinson et al. (1994); Gencay and Qi (2001)		Chen et al. (1998); Chidambaran et al. (1999); Chen et al. (1999); Chidambaran et al. (1998); Schuster (2003)
Volatility	Donaldson and Kamstra (1997); Bartlmae and Rauscher (2000); Dunis and Huang (2001); Harrald and Kamstra (1997)		Harrald and Kamstra (1997); Neely and A.Weller (2002)

Table 1.1: An indicative list of computational intelligent applications in financial time-series as categorised by the application domain (rows) and the intelligent technology used (columns). For hybrid schemes that combine more that one technologies, we occasionally quote each intelligent component separately.

complexity of a CI model that guarantees good fit in the training set and reasonable out-of-sample performance. This systematic procedure is often referred to as *model selection*, *determination* or *identification*.

It is important to note that despite the huge amount of CI theories and the recognised success of these models in applied work, there are still no universal model-building methodologies that can be applied to a wide range of application data sets. Much of the model-building strategies are *ad hoc*, i.e. specially designed for the particular methodology and application, and mainly driven by heuristic criteria or "rules" of thumb. Hence in a typical application the success of any intelligent approach is the result of repeated and time-consuming experimentation involving much trail and error.

In financial engineering and time-series analysis the contribution of CI models is still questionable. This is mainly because in this type of application domain there is a strong culture for investigating not only the predictive power but also the statistical significance of various aspects of the derived model. Financial engineers are in many cases interested in knowing how many of the lags of a model to keep in the final specification, whether a particular variable is important in explaining future movements of the target variable or which is a 5% confidence interval for a particular set of model parameters. Such issues greatly enhance the modelling procedure and our understanding of the underlying phenomenon. In contrast to other computational intelligent paradigms, the investigation of the statistical properties of the final specification has been a very active research area in the neural networks literature, where nowadays solid theoretical procedures exist for testing the individual or joint irrelevance of network inputs and weights (see e.g. Zapranis and Refenes (1999); White (1989b); White and Racine (2001)).

1.4 Artificial Neural Networks

1.4.1 Introduction

Artificial Neural Networks are a class of parallel semi-parametric computational models inspired by the biology of human brain[9]. The most typical architecture for a NN is the so-called *single-hidden-layer feedforward network*, depicted in figure 1.2. In this topology, the values of explanatory variables are passed though links or "connections" (represented by solid lines) to the intermediate or *hidden* layer. Each intermediate processing unit, called *hidden neuron*, sums up the pre-weighted arriving signals and passes them though a nonlinear "activation" function $F : \mathbb{R} \rightarrow \mathbb{R}$. Common choices for F are the logistic, the tan-sigmoid and the radial basis function. The output of each neuron is also amplified and sent to the output layer. Note that the architecture depicted in figure 1.2 also performs a linear mapping between input variables (and a constant) to the output level through direct connections (dotted lines). The

[9] An introductory exposition to artificial neural networks can be found in Bishop (1995); Haykin (1999).

combined effect of the neural network can be expressed by the nonlinear function:

$$g(x_t;\mu) = \phi'\bar{x}_t + \sum_{j=1}^{h} \lambda_j F(w_j' x_j - c_j)$$

where $\bar{x}_t = (1, x_t')'$, $\phi \in \mathbb{R}^{n+1}$ is the vector of the parameters of the linear model, h is the number of hidden neurons, $w_j \in \mathbb{R}^n$ are the weights from input variables to neuron j, $\lambda_j \in \mathbb{R}$ is the weight from neuron j to the output level and $c_j \in \mathbb{R}$ is the bias term.

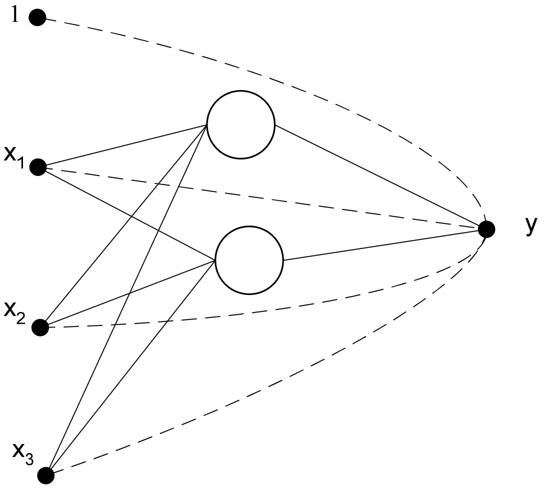

FIGURE 1.2: A single-hidden-layer feedforward neural network with three inputs and two hidden neurons.

As an example, consider a simple neural network with a single input x and an output y with two neurons in the hidden layer. The function relating x to y is defined as $y = \sum_{j=1}^{3} y_j$, where $y_1 = 0.5 + 0.2x$, $y_2 = F(-1.5x)$, $y_3 = 1.2F(11x + 21)$ and F is the logistic function. Figure 1.3 provides a plot of the output of the NN, measured in the vertical axis, with the input x measured on the horizontal axis. As seen, when $x \ll -2$, $y_2 = 1$ and $y_3 = 0$ hence y's behaviour is largely determined by the linear function $y_1 = 0.5{+}0.2x$ augmented by 1. As x rises past about -2, y_3 rapidly increases to its maximum ($y_3 = 1.2$) so that by the time x reaches roughly -1.4, y has increased from 1.1 to 2.3 approximately. As x continues towards zero, neuron y_2 is beginning to activate although its response is less abrupt due to the smaller coefficient of x (-1.5 compared to 11). By the time $x \approx 1$, $y_2 = 0$ and from that point onwards the output of the network is largely determined by the linear function.

NNs are widely used as function approximators. Given a sample of observations $\{x_t, y_t, t = 1, 2, \ldots, T\}$, where T is the sample size, $x \in \mathbb{R}^n$ is the set of input variables and $y \in \mathbb{R}$ is the target variable, the task is to determine the structure of the NN that can optimally approximate the relationship between x and y. The goodness of the approximation can be measured by a proper performance or error criterion $\pi_T(\mu)$, which is function of the parameters of the network and the data set. The definition of this criterion is in the researcher's judgement and largely depends on the particular

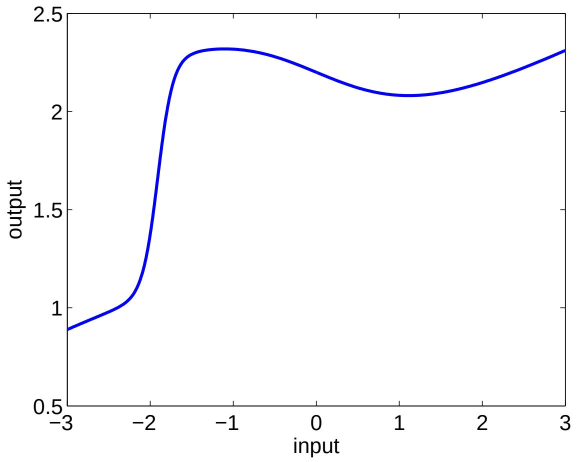

FIGURE 1.3: An example of a neural network mapping.

application, although common choices are the root mean squared or the root mean absolute error between the actual values of y and the response given by the network. Once an appropriate measure has been defined, the last stage is finding the values of the parameters that minimise the error or maximise the performance of the network. This stage is often called *training* or *estimation*.

Artificial neural networks are indisputably the most popular CI methodology within the area of financial engineering, especially in time-series forecasting and the pricing of stocks and options[10]. Due to the complex nature of economic and financial relationships, theory cannot fully guide the researcher as to which is the appropriate set of input variables or which is the appropriate functional form of the model that should be used for the particular data set. This difficulty makes it attractive to consider a flexible class of models that do not make explicit assumptions about the data-generating process. Artificial neural networks, as semi-parametric models, are well suited for this purpose. A single-hidden-layer feedforward NN is capable of approximating any Borel measurable function to an arbitrary degree of accuracy, by appropriately adjusting the number of neurons in the hidden layer. This is the so-called *universal approximation property* of NNs (see e.g. Hornik et al. (1989); Hornik (1991)).

1.4.2 Architecture selection

Model selection in artificial neural networks involves making a decision on how many neurons to include in the hidden layer and which subset of variables to connect to each neuron. Over the years of development and research in the neural network area, the number of model selection techniques that have been proposed in the relevant lit-

[10] See Zhang et al. (1998); Azoff (1994) for relatively recent surveys on the forecasting literature and table 1.1 for general financial engineering applications. Chapter 7 is exclusively concerned with the pricing of derivatives using NNs.

erature is great. Most of the approaches borrow their principles from disciplines such as statistics and information theory, however a great number of empirical heuristic procedures is also encountered. In what follows we attempt a brief review of the current practice in NN model selection. For further details, the interested reader is referred to Zapranis and Refenes (1999); Fine (1999)[11].

- *Regularisation*

  The idea behind regularisation is to compromise between the goodness-of-fit and the complexity of a neural network model, by penalising "over-parametrised" topologies. This is typically achieved by adding a complexity penalty term to the usual performance measure, thus minimising an objective function of the type
  $$l_T(\mu) = \pi_T(\mu) + ac_T(\mu)$$
  where $c_T(\mu)$ is the *regularisation* or *penalty* term, an increasing function of model parameters, and a is the *decay* or *smoothing* parameter, which represents the trade-off between model accuracy and complexity. A large value of a indicates a preference towards simpler topologies.

  In practice, various choices for the form of the penalty term are possible. Usually, $c_T(\mu)$ is taken as the sum of squared parameters. Another common formulation is $c_T(\mu) = (\sum_i \mu_i) \,/\, (1 + \sum_i \mu_i)$. The choice of the value of the smoothing parameter in an open issue. Weigend et al. (1991) proposed to iteratively increase or decrease the value of a during the training of the NN. However, this often requires much experimentation on behalf of the researcher and can also bias the model-selection procedure. One approach to determining the optimal value for the smoothing parameter has been proposed by MacKay (1992) in the context of Bayesian analysis.

- *Pruning*

  Pruning model-selection algorithms typically start with a complex network topology and gradually simplify the structure by removing insignificant neurons or connections. In this category, a vast number of techniques have been proposed, the majority of which are based on a simple heuristic or "rule-of-thumb". Optimal Brain Damage (Le Cun et al. (1990)) and Optimal Brain Surgeon (Hassibi and Stork (1993)) are two popular techniques of this sort that approximate the change of the error function when pruning a certain weight. A simple heuristic pruning strategy is also considered in chapter 7 for creating semi-parametric NN-based option pricing models.

- *Sequential statistical hypothesis testing*

  Ever since NNs were used in financial and economic problems, where well-established statistical and econometric procedures had already existed, there appeared a tendency to apply formal statistical procedures on NN model selection. These included among others statistical hypotheses tests and information

[11] Zapranis and Refenes (1999) provide a comprehensive survey on the application of statistical techniques in neural network modelling.

criteria. The popularity of statistical procedures in NNs was mainly because in this application domain there is a strong culture for investigating not only the predictive power but also the correct specification and the statistical significance of various aspects of the derived model.

Common statistical procedures in NN model selection typically follow a *bottom-up* direction, meaning that they start with a small model, mostly linear, and then gradually add hidden neurons if the data structure indicates so. The decision to further complicate the topology of the NN is often based on the value of some information criterion (AIC or SBIC) or the outcome of a statistical hypothesis test (see e.g. Anders and Korn (1999); Teräsvirta and Lin (1993); Medeiros et al. (2006); Swanson and White (1995, 1997a,b)). The latter is the approach that we adopt in this thesis for determining the structure of neural network models. In chapter 7, we compare bottom-up NN selection strategies guided by statistical hypothesis tests and information criteria with a top-down pruning algorithm on the basis of pricing S&P 500 stock index options.

1.5 Summary

The purpose of this chapter was to introduce the reader to the main topics covered in this thesis. We discussed the essence of financial engineering, a recently emerged multidisciplinary field that integrates financial theory with methods of engineering, statistical tools and computer algorithms, in order to structure sophisticated solutions to complex financial problems. We presented common research areas in financial engineering, such as the development of pricing models for primitive and derivative financial securities, the quantification of risk and the design of portfolios and trading strategies, and we reviewed the current practice in each of the above research directions.

The second part of this chapter was devoted to the presentation of computational intelligent algorithms, a modern approach to “learning-from-data” that has recently become popular in financial engineering. Over the years of development and research in this area, intelligent learning models such as artificial neural networks, neuro-fuzzy inference systems and genetically-evolved models have been successfully applied to a wide range of complex financial tasks, with rather promising results. The growing interest on these models is generally justified by their flexibility and their empirical or data-driven nature. A considerable part of this chapter is devoted to Artificial Neural Networks, a class of parallel semi-parametric computational models that are indisputably the most popular CI method among practitioners and researchers of financial engineering. After a brief introduction to the topic, we concentrate the discussion on methods and techniques for determining the architecture of a NN model. This is perhaps the most crucial stage in a practical application, as careful model selection helps to avoid overfitting and to find an adequate approximation of the true data-generating process.

The extensive use of NNs in financial problems, where well-established statistical and econometric procedures had already been used, led to an active research direction whose purpose was to apply formal statistical procedures on NN model estimation and selection. The popularity of statistical methods in NNs was in part due to the longstanding relationship between finance and stochastics but mainly came as a need to base model-building on clearly defined statistical principles that increase the transparency of the procedure. We will come back to the discussion on statistical modelling principles in chapter 3. The following chapter is about distinguishing properties of financial time-series.

Chapter 2

Some stylised facts on financial returns

2.1 Introduction

This chapter presents some stylised facts on the statistics and dynamic behaviour of financial prices and returns. Section 2.2 shows that the empirical distribution of returns is typically more "peaked" around the centre and has fatter tails than a normal probability density function. Section 2.3 shows that the average uncertainty and risk about the size of fluctuations in an asset's price typically changes over time, resulting in periods of high and low volatility. Section 2.4 discusses the dynamic behaviour of asset prices when the transaction price of an asset deviates from its fundamental value. Section 2.5 summarises the main findings and concludes with a discussion on the implications of the statistical properties for financial modelling.

2.2 Leptokurtosis

The empirical analysis of financial time-series has shown that price changes and rates of returns typically have "fatter tails" than are compatible with a normal distribution. To illustrate the property, we collect a sample of daily data from the French CAC 40 and German DAX stock indexes covering the period from July 3rd, 1987 to March 22nd, 2002[1]. We exclude weekends, public holidays and other non-trading days and we calculate the rates of return on each index by taking logarithmic differences between successive trading days, i.e. $r_t = \log(P_t) - \log(P_{t-1})$, where P_t denotes the index's closing value at time t and r_t the corresponding return between $t-1$ and t. The daily values and logreturns time-series are depicted in figure 2.1 and figure 2.2 respectively. Table 2.1 provides summary statistics for the logarithmic returns on CAC and DAX. We observe that the empirical distribution of returns is characterised by negative skewness and excess kurtosis (greater than three implied

[1] CAC (Compagnie des Agents de Change) 40 is a stock market index that tracks the forty largest French stocks based on the market capitalisation on the Paris Stock Exchange. DAX (Deutscher Aktienindex) is the leading index of the Frankfurt Stock Exchange and it measures the performance of the thirty largest German companies in terms of order book volume and market capitalisation.

by a normal distribution). This means that more probability is assigned to negative than positive returns and to extreme than moderate price movements. The Jarque-Bera test statistic in either case strongly rejects the hypothesis of normality. Figure 2.3 shows the histograms of logarithmic returns along with a normal density that provides the optimal fit to the data. As seen, the empirical distribution of returns largely deviates from the normal prototype, being sharply peaked around the mean and "heavy"-or "fat"-tailed. This leptokurtic shape of the empirical distribution is characteristic of most financial and economic time-series.

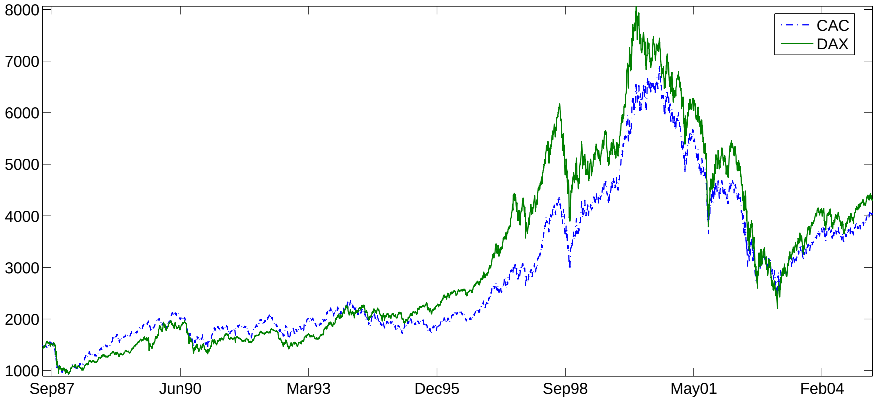

FIGURE 2.1: The CAC and DAX stock index values from July 3rd, 1987 to March 22nd, 2002.

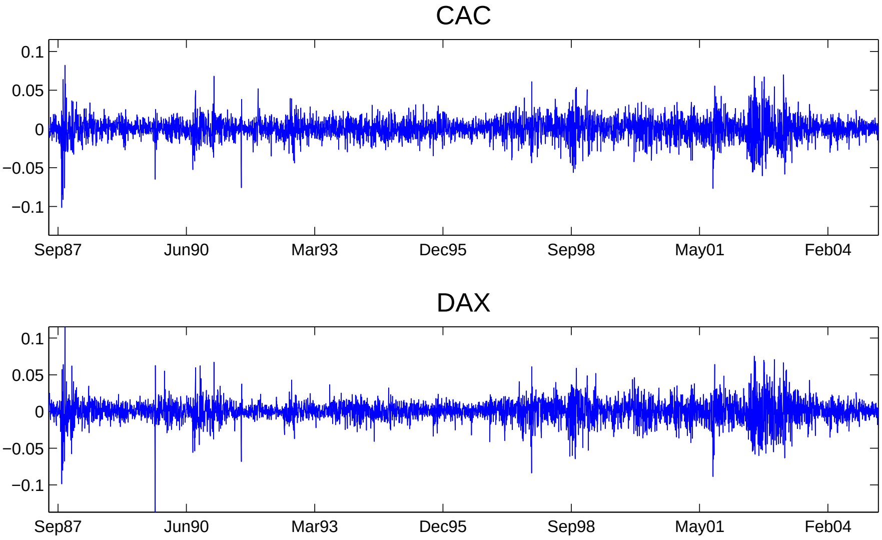

FIGURE 2.2: The CAC and DAX logreturns from July 3rd, 1987 to March 22nd, 2002.

Descriptive statistics		
	CAC	DAC
min	-0.101	-0.137
max	0.082	0.115
mean	2.283×10^{-4}	2.544×10^{-4}
standard deviation	0.014	0.015
skewness	-0.259	-0.364
kurtosis	7.251	9.154
Jarque-Bera normality test		
statistic	3.355×10^{3}	7.028×10^{3}
p-value	0	0

Table 2.1: The sample statistics of the CAC and DAX returns series. Panel A provides several distributional measures, while panel B shows the results of the Jarque-Bera normality test. The null hypothesis is that returns were sampled from a normal distribution. The p-value denotes the probability that one gets a value for the test statistic higher than the one tabulated above, given that the null hypothesis is true.

2.3 Predictability of the variance of returns

Another distinguishing property of the financial time-series refers to the variability of returns. Figure 2.4 shows the time evolution of the squared returns for both CAC and DAX. Observe that in both time-series the average uncertainty about the realised return is not constant but changes significantly with time. In particular, large price movements tend to be followed by large movements of either sign, resulting in succeeding periods of high and low volatility. On the contrary, the long-run variability of returns tends to settle down to a mean level. This pattern, common in many financial and economic time-series, was termed by Engle (1982) as *Autoregressive Conditional Heteroskedasticity* (ARCH).

Observable ARCH effects add an extra degree of short-term predictability of price movements that is related to the average uncertainty about future movements. It is important to note, however, that this source of predictability stems from the *second moment* of the conditional distribution of returns, i.e. the variance. Hence, although price changes or rates of return may be found uncorrelated, this does not mean that they are statistically independent, as the variability of returns may be largely determined by past extreme returns. Some models that are intended to capture this distinguishing pattern of volatility evolution are discussed in chapter 4.

Recent empirical studies on financial time-series have revealed additional features of volatility dynamics. French et al. (1987); Nelson (1990); Schwert (1990), among others, have confirmed that the short-term volatility does not often react in a symmetric way to past extreme price changes but the direction or sign of change is important to future volatility. Typically, a negative shock increases on average the short-term volatility more than an unexpected positive movement of the same magnitude, known as the *leverage effect* (see Engle and Ng (1993)). Inspired by the asymmetries and nonlinearities observed in data, many authors have proposed new models of volatility that are intended to capture these effects. Some examples are

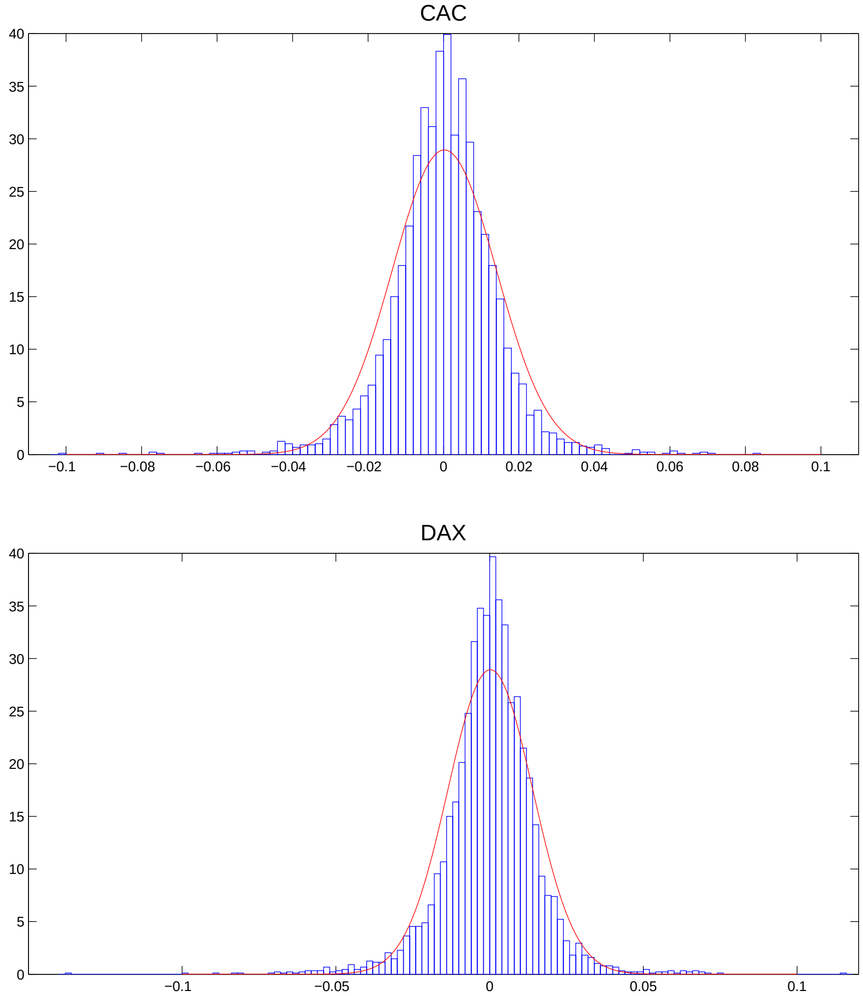

Figure 2.3: The empirical distribution of CAC and DAX logreturns. Along with the histograms we show the best fit from the family of normal densities.

discussed in chapter 4.

2.4 Nonlinearities in price adjustments

In high-frequency financial time-series, especially those that describe the intraday evolution of asset prices, different dynamic patterns arise when the market enters into exceptional regimes. These patterns are often the outcome of investors' actions in adjusting the price of the asset towards its *fundamental value*.

Due to the fact that nowadays most financial assets are traded in well-organised markets, where a great variety of securities exist, one can often derive a *theoretical* or *fair* value for these assets which is in accordance with the value of similar -in terms of payoffs- assets existing in the market. Typical examples are the Capital Asset

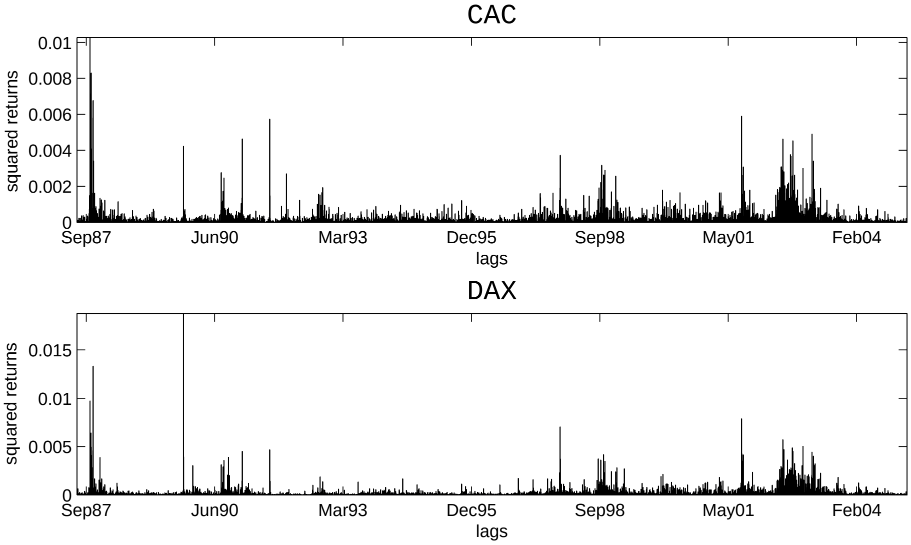

FIGURE 2.4: The short-run volatility of returns on CAC and DAX, as proxied by the square of logarithmic returns (r_t^2). Note that the time evolution of volatility shows a distinguishing pattern of heteroskedasticity (i.e. changing variance), characterised by successive periods of high uncertainty and relative "calm" (i.e. *volatility clusters*). This pattern, common in many financial time-series, is known in the literature as *Autoregressive Conditional Heteroskedasticity* (ARCH).

Pricing Model (CAPM), which links the price of a common stock with the value of the market portfolio, and the famous Black-Scholes option pricing model, which derives a fair value for a call option in relation to the price of the underlying security and the risk-free interest rate prevailing in the market.

In an effort to derive a predictive model of asset prices, most fundamental valuation theories start with the assumption that the actual price at which an asset is traded in a market is approximately equal to its fundamental value. The difference between these two does not convey any useful information from a modelers' point of view. This assertion is typically based on the standard financial argument that if the price is significantly different from its fundamental value, there appears a low- or even zero-risk profit opportunity through the tool of *arbitrage*. This is realised by buying and selling differently priced items of the same value and thus profiting from the difference. As this investment opportunity will attract many profit-seeking investors, one expects that the trading activity will almost instantly adjust the market price so that arbitrage is no more feasible. At this equilibrium state, the price of the asset is equal to its fundamental value.

This dogma has received extensive criticism by many researchers in financial economics, especially in the light of growing empirical evidence confirming large and persistent deviations from fundamental levels. These observations have led to the development of a new theory that views the relationship between price and fundamentals from a new perspective. The so-called *limits to arbitrage* have become a central issue in the recently emerged field of behavioural finance that tries to explain a variety of anomalies observed in the formation of market prices[2].

Arbitrage theories are typically based on the assumption that securities can be “substituted” or “replicated” by proper portfolios of other securities traded in the market. They also assume an ideal market environment with no transaction costs and enough well-informed rational traders. The common argument is that as soon as there is a deviation from the fundamental value an attractive investment opportunity arises, which rational traders will immediately snap up thereby correcting the mispricing. However, practical implementation issues pose limits to the profitability and feasibility of arbitrage strategies. Hence, even when an asset is persistently mispriced, strategies designed to correct the mispricing can be both *risky* and *costly*, hence leaving prices in a non-equilibrium state for protracted periods of time.

The theory of limits to arbitrage suggests that transaction costs, liquidity concerns, margin payments and limitations to short selling - which are often summarised by the term *market frictions* - are very important as concerns the practical implementation of a trading strategy. Empirical research conducted in various market environments revealed that these imperfections are responsible for statistically significant deviations from the no-arbitrage situation. Some studies concluded, for example, that the futures contract is selling at a discount relative to its theoretical price (Cornell and French (1983a,b); Figlewski (1984)). Jawadi (2005) reports similar results for prices of common stocks. Apparently, due to implementation costs, arbitrage strategies are profitable only when the benefits from arbitrage well exceed implementation costs.

Even when implementation costs are insignificant, arbitrage strategies may still be unattractive due to the existence of uninformed “noise” traders that engage in “trend-chasing” and drive prices away from fundamentals. Although fundamental valuation theories claim that the activity of arbitrageurs will sooner or later lead this type of investors out of the market, in the short-run arbitrageurs run the risk that the trading of noise investors will cause further deviations from the fundamental value. This is what is often called as *noise-trading risk.* As most arbitrageurs are restrained by short-term investment horizons, the activity of noise traders makes them less aggressive in combating any mispricing in the first place. On the other hand, there is extensive evidence that noise trading is not always a “bad signal” for rational investors but also a profitable opportunity. Occasionally, arbitrageurs may find it more profitable to trade in the direction of noise traders rather than in the direction that corrects the mispricing. This is the well-known “feeding the bubble” strategy (see e.g. Samuels et al. (1998), ch. 8).

The fact that arbitrage can be of limited affectiveness in real markets has important implications to the short-run dynamics of financial prices and often suggests a new approach to the modelling of securities prices. Thomaidis and Dounias (2006a); Thomaidis (2006) give an extensive discussion on the relationship between behavioural finance and common econometric models and Thomaidis and Dounias (2005) propose a new pricing methodology based on computational intelligent methodologies. To illustrate the ideas, consider figure 2.5 that depicts the contemporaneous deviation of the actual price of an asset from its fundamental value. Note that when

[2] For recent surveys on the behavioural finance literature see Shleifer (2000); Barberis and Thaler (2001).

mispricings are in the order of transaction costs the price may meander without a tendency to settle down to an equilibrium level, as perhaps none investor will find it advantageous to correct the mispricing. However, the larger the deviation from the fundamental value is the more investors are expected to engage in arbitrage trading, thereby exerting a greater pressure on the price to return to the zero level (fundamental value). This market behaviour suggests a certain type of nonlinear relationship between the mispricing occurring at some time and the correction of mispricing that is expected to take place in the near future. This relationship is depicted in figure 2.6. We observe that the average speed of correction attains its minimum for values of the mispricing around zero. However, the impact of mispricing increases as we move away from zero, possibly in an asymmetric way (i.e. negative mispricings may be corrected more abruptly than positive ones). Of course, this figures gives a rough approximation to reality. The mispricing-correction mechanism is very much depended upon the level of transaction costs and other market frictions, the number of investors engaged in arbitrage activity and the security under consideration. For example, in liquid markets with low transaction costs, the area of inaction around the fundamental value may be considerable smaller compared to illiquid securities. Therefore, the exact pattern of price reversions has necessarily to be determined on a case-to-case basis.

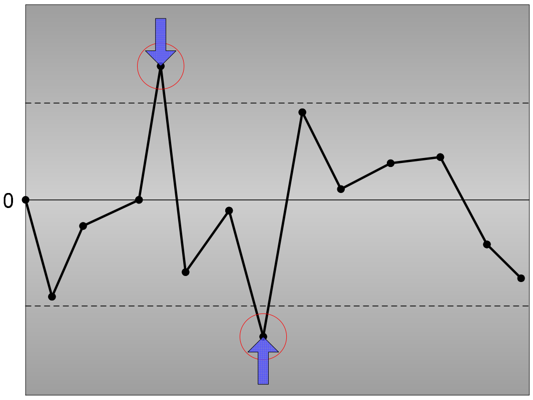

FIGURE 2.5: Nonlinear mean-reversion dynamics.

2.5 Summary

The purpose of this chapter was to discuss some distinguishable properties of financial time-series. We showed that price changes or rates of return typically follow nontrivial statistical laws that highly deviate from the "gaussian white noise" prototype, assumed in most engineering applications of NNs. Empirical distributions are more "peaked" around the mean and have heavier tails than a normal probability density. The average uncertainty about price movements is not constant but depends on the history of returns. In particular, large unanticipated price movements tend to increase on average the short-term variability of the series and quite often a sudden

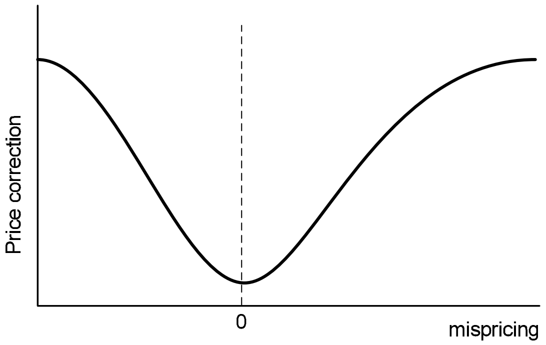

Figure 2.6: Average price correction vs mispricing.

price drop has a different impact from an unexpected rise. Hence, although most financial securities go through period of high and low volatility any shift in volatility levels is largely determined by the size and sign of past unexpected movements and hence is predictable.

A considerable part of this chapter was also devoted to a discussion on the short-run dynamic behaviour of prices when an asset is mispriced, i.e. its market price deviates from the fundamental value. Classical asset pricing models assume that most securities traded in the market are priced according to their fundamental value, because as soon as there is any deviation from fundamentals there appears an opportunity for arbitrage, which rational investors will instantly exploit thereby correcting the mispricing. However, quite often practical implementation issues pose limits to the profitability and feasibility of arbitrage strategies, resulting in persistent deviations from fundamental levels. When mispricings are in the order of transaction costs, the price may meander without a tendency to revert to an equilibrium level, as perhaps none investor has any benefit from correcting the mispricing. However, the larger is the deviation from the fundamental value the more investors are expected to engage in arbitrage trading thereby exerting a greater pressure on the price to return to the fundamental level. This means that the probability distribution of future price changes may depend on the size/sign of the mispricing and on recent price movements.

The empirical properties of financial time-series have important implications for the modelling techniques used in financial engineering. Obtaining a faithful representation of the dynamic behaviour of financial instruments requires considering a general class of models that could possibly accommodate nonlinear adjustment mechanisms, symmetric or asymmetric ARCH effects and non-gaussian distributions. However, not all data sets share the same features. In high-frequency price series, for example, asymmetric responses may arise more often than in a time-series representing daily data where heteroskedasticity may be the dominant effect. From a practical point of view, it is therefore absolutely important to have a model-selection strategy that can adjust the level of complexity of the predictive model according to the particular features existing in data.

However, these features may often invalidate the procedure used to determine the optimal structure of the model leading to unpleasant results. For example, as we show later in chapters 4 and 5, standard statistical procedures that are used to determine the number of neurons in a NN model may misinterpret heteroskedasticity as systematic nonlinearity in data and hence increase the complexity of NN in order to capture this effect. As heteroskedasticity, however, is a property of the error process this can lead to overfitting and bad out-of-sample performance.

Chapter 3

Principles of statistical and econometric modelling

3.1 Introduction

This chapter reviews the basic principles in the construction of statistical-econometric models with financial applications. Section 3.2 discusses the main econometric assumptions about the data-generating processes encountered in financial markets and section 3.3 presents all stages of the econometric model-building cycle (specification, estimation, evaluation). Section 3.4 summarises and concludes the chapter with a discussion on the statistical properties of intelligent learning algorithms and the interaction between statistics and computational intelligence.

3.2 The general setting

Most computational intelligent techniques are traditionally treated as approximators to the functional relationship between a set of explanatory variables and another target variable of interest to the researcher. However, this approach should be treated with care in application domains where the researcher has incomplete control over the process that generates the data. This situation arises when the "nature" or "chance" has a hand in generating measurements[1]. Nature' s involvement may be partial, as for example in an experimental setting where observations are "noisy" i.e. their precise value is also determined by chance, or complete, as in financial markets where the modeler is a sole observer and has no control on the values that economic variables take. In such an inherently uncertain environment, it is no longer possible to express an exact functional relationship between target and explanatory variables; however it is always possible to find a relationship in terms of probability.

Suppose the researcher is interested in explaining the behaviour of an observable variable and for that purpose he assembles a sample of observations $\{(y_t, z_t), \quad t = 1, 2, \ldots, T\}$, where y_t is the realisation of the *target* or *dependent* variable and z_t includes other "exogenous" variables that are believed to determine y_t. For example, z_t could be measurements of various economic indexes up to a particular point in

[1] See White (1989b) for a comprehensive discussion.

time and y_t could be the closing value of the American Standard and Poor's stock index on the next day. The purpose of the analysis is to predict and test hypotheses about the behaviour of y_t given a set of predetermined *conditioning* variables x_t. The set of variables included in x_t depends on the application domain and the type of information the analyst wants to incorporate in his model. If, for example, y_t and z_t represent independent observations that are not indexed by time, the set of conditioning variables may be simply $x_t = z_t$. In a dynamic model, it is common to view y_t and z_t as being contemporaneous variables and then x_t is any subsequence of $(z_t, y_{t-1}, z_{t-1}, \ldots, y_1, z_1)$. If one wants to condition only on information observed before t, z_t can be excluded from x_t. The probabilistic relationship between x_t and y_t is completely summarised by *the conditional probability law* of y_t, $P(y_t \in A|x_t)$, which essentially describes the relative frequency of y_t taking values in A given the values of x_t. In this probabilistic context, the focus of interest is shifted from an exact functional relationship to the probabilistic relationship between x_t and y_t. It is this conditional probability that embodies everything there is to know about the effect of x_t on y_t.

Assuming that the conditional probability law is well behaved, one can define the *conditional probability density function* $p(y_t|x_t)$, which is the probability of the target variable taking a value close to y_t given x_t. Several features of the conditional density are of particular interest in financial applications:

- *The conditional expectation*

  The *first moment* or *conditional expectation* $E(y_t|x_t)$ is a measure of the value of y_t that will be realised on average given x_t. $E(y_t|x_t)$ is a deterministic quantity and its value is only determined by x_t, hence $E(y_t|x_t) = g(x_t)$ for some nonlinear mapping $g : \mathbb{R}^n \to \mathbb{R}$. Since y_t is stochastic, its actual realisation will (almost certainly) differ from $E(y_t|x_t)$, so there will be an *expectational error*

  $$\epsilon_t \equiv y_t - E(y_t|x_t)$$

  By definition of ϵ_t and by the properties of conditional expectation, it follows that $E(\epsilon_t|x_t) = 0$, i.e. the average expectation of ϵ_t given x_t is zero. This means that ϵ_t is unpredictable by x_t, which is why it is often called the *innovation* or *"surprise"* term. In finance and economics, it is common to treat ϵ_t as a collective measure of news or events that occur at time t, through were not expected by market participants and hence have not been discounted into current prices. Because $g(x_t) = E(y_t|x_t)$ we can also write

  $$y_t = g(x_t) + \epsilon_t, \tag{3.2.1}$$

  The alternative representation (3.2.1) allows us to discuss some fundamental differences in the meaning assigned to ϵ_t among different applications of NNs. In signal processing, for example, y_t is thought of being the result of a deterministic signal $g(x_t)$ and some exogenous noise source ϵ_t. Hence, equation (3.2.1) is essentially read from right to left and y_t is the outcome of x_t and ϵ_t. In econometrics, however, a fundamentally different interpretation is given to

(3.2.1): the process that generates y_t can be decomposed into a part which can be explained though x_t and the remainder or unexplained part ${\epsilon_t}^2$. Of course, the properties of ϵ_t are largely determined by x_t and the quality of approximation. If, for example, x_t contains everything there is to know about y_t then ϵ_t should behave like white noise, i.e. an *independent identically distributed* (iid) process with zero mean and constant variance. One the other hand, if x_t does not influence y_t then all properties of y_t are inherited by ϵ_t. In stochastic environments where the researcher is a sole observer of variables, ϵ_t does not simply enter (3.2.1) in the form of "exogenous" noise. On the contrary, it contains important information about the conditional distribution of y_t which may deserve further investigation. One such feature is the conditional variance of y_t discussed below.

- *The conditional variance*

  The *second moment* or *conditional variance* $\text{Var}(y_t|x_t)$ of y_t given x_t is a measure of the average dispersion of y_t around $E(y_t|x_t)$. Hence it shows the uncertainty or "risk" about the realisation of y_t given the information included in x_t. In financial markets, the accurate modelling of the conditional variance becomes a very important task for risk analysis and portfolio management. This is because, as we showed in the previous chapter, the second moment of economic variables typically changes with time, following characteristic patterns of evolution.

3.3 The construction of econometric models

The task of econometrics is to specify models that approximate the real conditional distribution set by nature or certain features of it (conditional mean, conditional variance). Finding these approximations typically requires a specification step, in which one restricts attention to a class of candidate models believed to contain the real distribution, and an estimation step that picks out one candidate model from the class that by some criterion seems to be closest to the true density. The last stage is concerned with judging the quality of model specification as an approximation to reality and the plausibility of assumptions set by the model. The process of econometric model-building is schematically presented in figure 3.1 and analytically discussed below.

3.3.1 Specification

The first task of the researcher is to specify a set of explanatory variables x_t and a parametric family of specifications that is believed to encompass the true conditional density. This parametric family is written as:

$$\mathcal{P}(\delta) = \{\rho(y_t|x_t;\delta), \quad \delta \in \mathbb{R}^l\}$$

where $\rho(.;\delta)$ is a candidate density model and δ is a vector of free parameters that index each member of the family. Probably the most popular and commonly used

[2] See Hendry (1995) for a discussion on the topic.

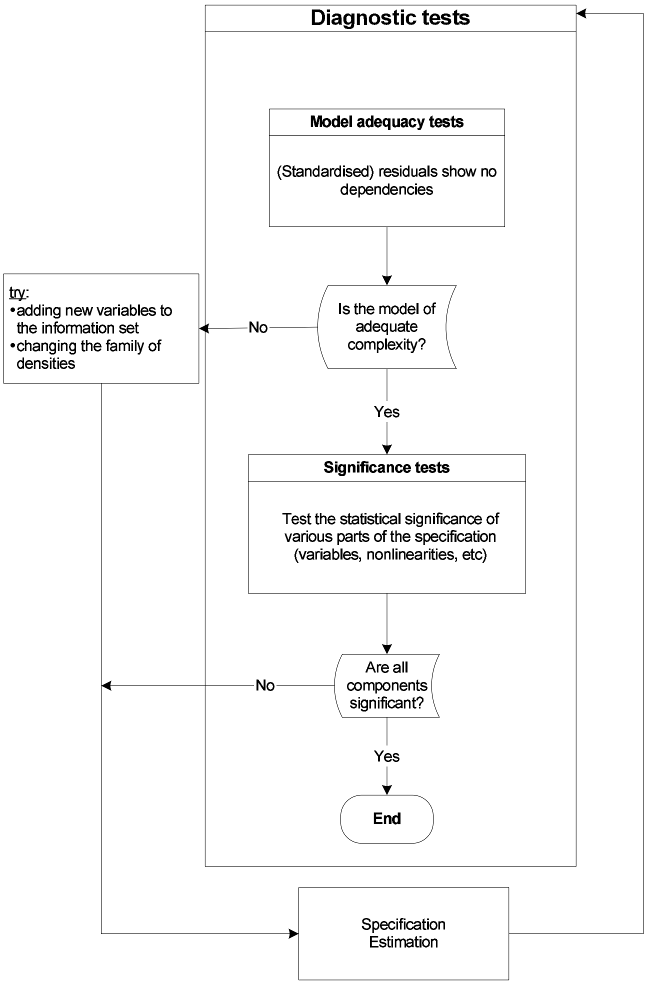
Diagnostic tests
Model adequacy tests
(Standardised) residuals show no dependencies
Is the model of adequate complexity?
No
try:
•adding new variables to the information set
•changing the family of densities
Yes
Significance tests
Test the statistical significance of various parts of the specification (variables, nonlinearities, etc)
Are all components significant?
No
Yes
End
Specification
Estimation

FIGURE 3.1: The process of constructing an econometric model (adapted from Zapranis and Refenes (1999), p. 15).

econometric specifications are the *regression* or *expectational* models that are exclusively focused on capturing first-moment relationships between y_t and x_t. A typical regression model takes the form

$$y_t = m(x_t; \theta) + \epsilon_t$$

where $m(.; \theta)$ is a possibly nonlinear function of x_t and ϵ_t's follow a certain distribution (normal or student) with constant variance. Apart from the classical regression model where $m(.)$ is specified as a linear function, nonlinear regression models are also very popular in time-series applications, especially in finance and economics (see Granger and Teräsvirta (1998) for a comprehensive discussion). Most computational intelligent algorithms encountered in the literature, such as artificial neural networks, genetically evolved models, adaptive neuron-fuzzy inference systems, etc, essentially employ a regression model. However, there also exist methods, especially in the neural network area, for directly approximating the conditional probability density of y_t given x_t. Popular examples are the mixture density (Bishop (1994)) and the recurrent mixture density networks (Schittenkopf et al. (2000)). Dynamic models that jointly parametrise the mean *and* the variance of the conditional distribution are also often encountered in financial time-series analysis and forecasting. Perhaps, the most popular examples of this class is the family of GARCH models and their variations. These are examined in the next chapter.

3.3.2 Estimation

Once the parametric family of models is specified the next step is determine the value of δ^*, i.e. to choose the member of $\mathcal{P}$, that give the best approximation to the true density. Different estimation methods are available in the literature, depending on the type of specification employed. *Least-squares* is the most commonly used method for expectational models; *maximum likelihood* is more relevant to models that parametrise higher moments of the distribution.

3.3.3 Evaluation

The application of most econometric models is typically followed by a third stage of model evaluation, in which the predictive ability of the model is judged by means of several goodness-of-fit measures. In econometrics, there is also a tradition to apply in-sample evaluation tests, which are mainly concerned with the adequacy of the model for the particular data set as well as the statistical significance of its components. Those are examined below.

- *Adequacy* or *diagnostic* tests form an integral part of a model's post-evaluation stage. They generally investigate whether the specified model structure is a faithful representation of $p(y_t|x_t)$. Diagnostic testing is based on the general principle that if the model is well specified for the underlying data-generating process, what is left unmodelled should not contain interesting features of the conditional density. If, for example, a regression model is an adequate structure for the data, the residuals (what is left unmodelled in this case) should resemble a memoryless process with constant variance. The existence of "strong" statistical properties in the error term indicates some sort of specification bias.

- *Model significance tests.* For a model that passes the adequacy test, a second type of diagnostics is employed which concerns evaluating the statistical significance of the various parts thereof, especially the explanatory variables. The common trend in econometric modelling is to identify models with minimal complexity, enough to capture the salient features or "*driving forces*" of the data-generating process. This is the well-known principle of *parsimony* or *Occam's razor*, often found in econometrics textbooks (see for example Box et al. (1994)). Parsimonious models are easier to handle and typically have better forecasting performance on unseen data.

  Depending on the type of the specification employed by the researcher, significance tests can change in nature. In linear regressions, model significance is equivalent to testing the coefficients of explanatory variables. In nonlinear regression models, including neural networks, significance tests are also directed to the parameters that contribute to nonlinearity (e.g. the weights connecting input variables to hidden neurons and to the output level). In second- or higher-moment density models, significance tests are typically directed to each submodel corresponding to a specific moment of the distribution.

3.4 Summary and discussion

The purpose of this chapter was to review current trends in statistical-econometric model-building. The ultimate purpose of econometrics is to specify models that approximate the probabilistic relationship between the target variable and the set of explanatory variables, as summarised by the conditional probability distribution or density function. Finding this approximation typically requires a *specification* step, in which one restricts attention to a class of density models, an *estimation* step, which chooses the member of the class that seems to offer the best approximation to the true density and a *diagnostic* step, in which the quality/redundancy of the approximation is evaluated. These steps are repeated until an acceptable specification is found.

The majority of computational intelligent applications in financial engineering, including neural networks, essentially employ a regression model, although various methods for approximating the entire density are also available in the literature. However, the discussion in chapter 2 on stylised facts on financial returns showed that returns series typically follow no trivial statistical laws characterised by mean and higher-order dependencies in data. For example, the existence of ARCH effects in a time-series implies that a large price movement of either sign is expected to temporarily increase the uncertainty about the size of future price changes, naturally resulting in a wider conditional distribution.

In the light of this empirical evidence, a single regression model that focuses on mean dependencies is not an adequate description of the data and hence is expected to have worse forecasting performance compared to other models that are directed to higher moments of the conditional distribution. The natural extension is to consider models that modify the shape of the conditional distribution in dependence of past data. Chapter 4 proposes a new class of intelligent models that have this property.

From an econometric point of view, computational intelligent models can be seen as nonlinear semi-parametric approximations to the conditional density of y_t given x_t. From this perspective, various model specification problems can be solved by applying proper statistical inference. Statistical theory gives us also the tools to construct confidence intervals on the values of the parameters of the model and test hypotheses regarding the obtained specifications. Taking a statistical perspective is especially important for semi-parametric CI models, like neural networks, because the reason for applying them is the lack of knowledge about an adequate functional form. When based on a clearly defined statistical decision rule, model-selection becomes more transparent and easy to reconstruct. These issues are further discussed in the next chapter that shows how artificial neural networks can be combined with econometric models of volatility into an integrated modelling framework.

Chapter 4

Neural Network GARCH models: a hybrid approach

4.1 Introduction

This chapter introduces a class of hybrid semi-parametric models that combine neural networks with econometric GARCH parametrisations of volatility. Following the main principles underlying the construction of econometric models, we propose a flexible modelling framework that is intended to capture interesting features of the *entire* conditional distribution. We show how the class of NN-GARCH models can accommodate most of the stylised facts on financial returns discussed in chapter 2 (nonlinear adjustments, symmetric/asymmetric GARCH effects and non-gaussian errors).

The structure of this chapter is as follows: Section 4.2 reviews current trends in the modelling of volatility with GARCH-type models and section 4.3 introduces the class of NN-GARCH models and extensively discusses several statistical and numerical issues that arise in the estimation of parameters. In section 4.4 we propose a statistical procedure for judging the significance of the parameters of the model that also gives the opportunity to the researcher to test hypotheses of interest regarding the mean and variance structure of the data-generating process. Section 4.5 proposes a "bottom-up" model-building strategy for the family of NN-GARCH models that is based on sequential statistical tests of additional structure in the mean/variance of the model. The specification of the mean and the variance components are discussed in detail and several hints are given. In section 4.6 we present a general framework for diagnostic checking on an estimated NN-GARCH model, whose general purpose is to judge the quality of approximation to the real density. All tests considered here do not pose restrictive assumptions holding in addition to null hypothesis (normality, heteroskedasticity), are simple to construct and involve much less computation compared to other approaches proposed in the literature. The chapter concludes with section 4.8 which summarises the main points.

Joint NN-GARCH models were first introduced in Thomaidis et al. (2005a) and Thomaidis and Dounias (2006b) under the assumption that the variance follows a symmetric GARCH model. For the determination of the neural network model, we proposed the misspecification - robust LM test of White (1996) which has the desirable property of leading to correct inference in the presence of general forms of heteroskedasticity in errors. However, this test significantly increases the computational burden of model specification as it involves inversion of matrices and the computation of second derivatives of the log-likelihood. In this work, we extend the framework of NN-GARCH to incorporate asymmetric parametrisations for the volatility and we also examine a new set of methods for determining the number of hidden units in the neural network part, based on the robustified auxiliary-regression testing framework of Wooldridge (1990, 1991).

4.2 Generalised autoregressive heteroskedastic models

Nowadays, the family of generalised autoregressive heteroskedastic (GARCH) parametrisations are no doubt the most popular models for volatility clustering. GARCH parametrisations belong to the general class of dynamic models with time-dependent volatility, whose form is:

$$y_t = m(x_t;\delta) + \epsilon_t \tag{4.2.1a}$$

$$\epsilon_t | x_t \sim D(0, h_t(x_t;\delta)) \tag{4.2.1b}$$

where $y_t \in \mathbb{R}$ is the target variable, $x_t \in \mathbb{R}^n$ is the vector of explanatory variables, δ is a vector of parameters and $m(x_t;\delta)$ is the mean or expectational model of y_t given x_t. The unpredictable component ϵ_t is assumed to follow a certain distribution D with zero mean and conditional variance h_t, which is generally a function of x_t and δ. Typical choices for D are the normal and the Student t-distribution. Note that (4.2.1) is implicitly a conditional density model for y_t as $y_t|x_t \sim D(m_t, h_t)$.

Depending on the type of model employed in the expectational or the variance part, we obtain a variety of linear and nonlinear parametrisations for the first two moments of the conditional density. Possibly, most practical applications of GARCH models assume a linear-in-mean model, where $m_t \equiv \phi' \bar{x}_t,\ \phi \in \mathbb{R}^{n+1}$ and $\bar{x}_t = (1, x_t')'$, although various forms of nonlinear regressions, such as threshold (Li and Li (1996)) or smooth transition (Lundbergh and Teräsvirta (1998)), have also been used in the parametrisation of m_t.

As a first attempt to model volatility clustering, Engle (1982) introduced the ARCH(q) model expressing the conditional variance as a linear function of the past q squared innovations, i.e.

$$h_t = a_0 + \sum_{i=1}^{q} a_i \epsilon_{t-i}^2 \tag{4.2.2}$$

Later, Bollershev (1986) proposed a generalisation of ARCH, the GARCH(p,q) model, which achieves simpler and more parsimonious parametrisations of volatility dynamics. The general form of a GARCH(p,q) model is

$$h_t = a_0 + \sum_{i=1}^{p} a_i h_{t-i} + \sum_{j=1}^{q} b_j \epsilon_{t-j}^2 \tag{4.2.3}$$

where h_{t-i}'s and ϵ^2_{t-j}'s are often called the GARCH and ARCH terms. In an ARCH or GARCH model, the effect of a shock on current volatility declines geometrically with time. Empirically, the family of GARCH models, and especially the simplest GARCH(1,1), has been very successful in practice[1].

Despite the apparent success and simplicity of ARCH and GARCH models, these parametrisations cannot sufficiently describe all empirical features of volatility dynamics. Note that equations (4.2.2) and (4.2.3) impose a symmetry in the response of the conditional variance to past shocks ϵ_{t-j}, $j = 1, 2, \ldots, q$. It is thus implicitly assumed that the volatility depends on the size but *not* on the sign of the shock, i.e. the amount of volatility following "bad news" is the same as the amount of volatility following "good news". Many authors have empirically discovered, however, that the direction of news is important to future volatility. French et al. (1987); Nelson (1990); Schwert (1990) among others, report that an unexpected drop in price (negative shock) seems to increase on average the short-term volatility more than an unexpected increase in price (positive shock) of the same magnitude. In order to capture the asymmetry observed in data, Nelson (1990) has introduced the family of *exponential* GARCH, or EGARCH(p, q), models

$$h_t = \exp\left(a_0 + \sum_{i=1}^{p} a_i \log(h_{t-i}) + \sum_{j=1}^{q} b_j(|u_{t-j}| - E) + \sum_{j=1}^{q} l_j u_{t-j}\right) \tag{4.2.4}$$

where $u_t \equiv \epsilon_t/\sqrt{h_t}$ are the standardised residuals and $E = \sqrt{2/\pi}$ if u_t is gaussianly distributed and $E = \sqrt{\frac{\nu-2}{\pi}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$ if u_t is t-distributed (with degrees of freedom $\nu > 2$). The parameters l_j of the model are the *leverage coefficients*, allowing for asymmetric responses of h_t to u_{t-j}. If l_j's are identically equal to zero the EGARCH becomes a symmetric volatility model in which a positive surprise ($u_{t-1} > 0$) has the same effect on volatility as a negative surprise ($u_{t-1} < 0$). In the EGARCH(1,1), a negative value of l_1 implies that past unanticipated bad news ($u_{t-1} < 0$) has a greater (lower) impact on future volatility than good news $u_{t-1} > 0$.

Apart from EGARCH models, a great number of alternative formulations of asymmetric volatility have been proposed in the literature, motivated by empirical research on financial data (see Engle and Ng (1993) for a good survey). Among them is the GJR-GARCH(p, q) model proposed by Glosten et al. (1993):

$$h_t = a_0 + \sum_{i=1}^{p} a_i h_{t-i} + \sum_{j=1}^{q} b_j \epsilon^2_{t-j} + \sum_{j=1}^{q} d_j S^-_{t-j}\epsilon^2_{t-j} \tag{4.2.5}$$

where $S^-_{t-j}, j = 1, \ldots, q$ is an indicator variable taking value 1 if $\epsilon_{t-j} < 0$ and 0 otherwise. By means of S^-_{t-j}'s, the GJR-GARCH model allows for different response of h_t to positive and negative values of ϵ_{t-j}.

[1] See Bollerslev et al. (1992); Bera and Higgins (1993) for comprehensive surveys.

4.3 A general class of neural network GARCH models

The NN-GARCH models, introduced in this dissertation, are members of the general class (4.2.1) of dynamic models with GARCH heteroskedasticity, in which a neural network is used in the conditional mean equation. In particular, we assume that

$$y_t = \phi' \bar{x}_t + f(x_t;\theta) + \epsilon_t \tag{4.3.1a}$$

$$\epsilon_t | x_t \sim D(0, h_t(x_t;\delta)) \tag{4.3.1b}$$

where $\bar{x}_t = (1, x_t')'$, $\phi \in \mathbb{R}^{n+1}$ is the vector of the parameters of the linear model (or weights directly connecting input variables to the output level) and $f(x_t;\theta)$ is a single-layer feedforward neural network with h hidden neurons, i.e.

$$f(x_t;\theta) = \sum_{j=1}^{h} \lambda_j F(w_j' x_j - c_j) \tag{4.3.2}$$

where $F(z) = 1/(1+e^{-z})$ is the logistic function, $w_j \in \mathbb{R}^n$ are the weights from input variables to neuron j, $\lambda_j \in \mathbb{R}$ is the weight from neuron j to the output level and $c_j \in \mathbb{R}$ is the bias term. The conditional variance of (4.3.1) is modelled using any of the parametrisation (4.2.2) to (4.2.5) discussed in the previous chapter.

The vector of free parameters of the model is $\delta = (\phi', \theta', \alpha')' \in \mathbb{R}^m$, where $\phi \in \mathbb{R}^{n+1}$ and

$$\theta = (\lambda_1, \ldots, \lambda_h, w_1', \ldots, w_h', c_1, \ldots, c_h)' \in \mathbb{R}^{(n+2)h}$$

are the parameters of the conditional mean equation and α are the parameters of the conditional variance model. The dimensionality of this vector depends on the type of the model employed.

If errors are assumed conditionally normal, the density function induced by the family of NN-GARCH specifications is

$$\rho(y_t|x_t;\delta) = \frac{1}{\sqrt{2\pi h_t}} \exp\left(-\frac{\epsilon_t^2}{2h_t}\right)$$

However, Bollerslev (1987) noted that ARCH or GARCH models do not fully account for leptokurtosis in financial series and quite often the unconditional error distribution corresponding to ARCH and GARCH models has fatter tails than predicted by a normal distribution. Therefore, he proposed an extension of the original GARCH model where errors follow a student-t distribution. In this case, the conditional density function is given by

$$\rho(y_t|x_t;\delta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}} \left(1 + \frac{u_t^2}{(\nu-2)}\right)^{-(\nu+1)/2}$$

where $\Gamma(.)$ is the gamma function, $u_t \equiv \epsilon_t/\sqrt{h_t}$ are the standardised residuals and ν are the degrees of freedom. The above density function is also symmetric around zero and has an additional factor that controls the "heaviness" of the distribution tails. For $\nu > 4$ the conditional kurtosis equals $3(\nu-2)/(\nu-4)$, which is greater than that of the normal's. For large values of ν, the density converges to a standard normal.

Note that formula (4.3.1) provides the researcher with a quite general class of models that incorporate linear/nonlinear-in-mean processes, linear/nonlinear GARCH effects in variance and possibly non-gaussian errors. GARCH specifications with linear expectational models as well as pure neural network regression models are encompassed by the above framework.

4.3.1 Estimation of parameters

The parameters of the model are estimated by maximising the log-likelihood function $l(\delta) = T^{-1}\sum_{t=1}^{T} l_t(\delta)$ where

$$l_t(\delta) = -0.5\log(2\pi h_t) - 0.5\,\epsilon_t^2/h_t \tag{4.3.3}$$

assuming normally distributed disturbances or

$$\begin{aligned} l_t(\delta) = \log\left(\frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)}(\pi(\nu-2))^{-0.5}\right) &- 0.5\log(h_t) \\ &- 0.5(\nu+1)\,\log\left(1+\frac{\epsilon_t^2}{h_t(\nu-2)}\right) \end{aligned} \tag{4.3.4}$$

assuming t-distributed residuals. Typically, several numerical problems in the maximisation of the log-likelihood are avoided if analytical formulae for the gradient and the hessian are used. Those are given in appendix A for the case of a normal density function. The value $\hat{\delta}_T$ that maximises the log-likelihood is called the *maximum likelihood estimator* (MLE).

It is important to note that the MLE of the parameters of a neural network, and hence of the full model, is generally not unique, unless the NN is *identifiable*. In the general case, a density model is called identifiable if the mapping from the parameter vector δ to $p(y_t|x_t;\delta)$ is one-to-one, in other words the model does not produce the same output for different configurations of its parameters (Watanabe (2001)).

There are three characteristics of the neural network specification that cause non-identifiability (see e.g. Anders and Korn (1999); Medeiros et al. (2006); Hwang and Ding (1997) and their references). Note that in the topology presented in figure 1.2 the hidden neurons can be permuted resulting in different, though discrete, maxima of the log-likelihood. Equivalent models are also obtained due to the property of the logistic function $F(x) = 1 - F(-x)$. The third and most important source of non-identifiability is the possibility that the network is overly parametrised for the conditional expectation, in which case some of its neurons are *redundant.* Note that if a neuron j is redundant then either λ_j or w_j are zero and the output is constant for all sample observations. But if λ_j equals zero, the corresponding w_j weights leading into that neuron can take any value and are thus not identifiable. Similarly, if the w_j weights are all zero, the corresponding value of the λ_j weight does not have any effect on the value of the log-likelihood. In either case, the set of optimum parameter solutions to the maximization problem corresponds to flat regions of the log-likelihood function across several directions of the parameter space.

The first two sources of non-identifiability can be easily remedied by applying proper parameter constraints. This is done as in Trapletti et al. (2000); Medeiros et al.

(2006)) by requiring that a) $c_1 < \ldots < c_h$, which precludes permutation of hidden neurons and b) $w_{1j} > 0, j = 1, 2, \ldots, h$, which remedies any possible misidentification due to the symmetry of the logistic function. The existence of redundant neurons however cannot be handled in the same way and requires careful selection of the architecture of the neural network. This issue is further discussed in section 4.5.1.

Several restrictions also apply to the volatility part and depend on the conditional variance model employed. Those generally guarantee positivity and non-explosive behaviour of the variance process. In standard GARCH models, for example, positivity of the conditional variance parameters requires that $a_0 > 0$ and $a_i, b_j \geq 0$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, q$. In addition, if $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$ we also have stability and second order stationarity of the variance process (i.e. stationarity in mean and variance), although weaker conditions could be used instead to guarantee non-explosiveness of the volatility process (see e.g. Bougerol and Picard (1992)). In contrast to GARCH specifications, no positivity restrictions need to be imposed in the estimation of an EGARCH model, since the logarithmic transformation ensures that the forecasts of the variance are non-negative. The stationarity constraint for this model is satisfied by ensuring that the roots of the characteristic polynomial

$$\lambda^p - a_1\lambda^{p-1} - a_2\lambda^{p-2} - \ldots - a_p$$

are inside the unit circle. In the simplest EGARCH$(1, 1)$ this condition simplifies to $|a_1| < 1$. Finally, for the GJR-GARCH(p, q) model the corresponding conditions are $a_0 > 0$ and $a_i, b_j, b_j + d_j \geq 0$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, q$ for positivity and $\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j + 0.5 \sum_{j=1}^{q} d_j < 1$ for stationarity.

4.3.2 Consistency and asymptotic normality of the MLE

Consistency and asymptotic normality for the maximum likelihood estimator of a NN-GARCH model is guaranteed by imposing some further conditions regarding the adequacy of the mean and variance specifications. Generally, a NN-GARCH model is said to be *structurally correct* if the mean and the volatility model are correctly specified for the corresponding moments of the conditional distribution. This means that for some δ_0 in the corresponding parameter space

$$E(y_t|x_t) = m(x_t; \delta_0), \text{ correct specification of the mean} \tag{4.3.5a}$$

$$V(y_t|x_t) = h(x_t; \delta_0), \text{ correct specification of the variance} \tag{4.3.5b}$$

where $E(y_t|x_t)$ and $V(y_t|x_t)$ are the conditional expectation and variance of y_t given x_t. This conditions can be equivalently written in terms of the error process $\{\epsilon_t,\ t = 1, 2 \ldots, T\}$ as

$$E(\epsilon_{0t}|x_t) = 0 \tag{4.3.6a}$$

$$E\left[\epsilon_{0t}^2 \middle| x_t\right] = h(x_t; \delta_0) \tag{4.3.6b}$$

where ϵ_{0t} denotes the error of the model evaluated under the true parameter vector δ_0. Assume that the procedure used to estimate the parameters of the NN-GARCH model is maximisation of the log-likelihood function under the assumption that y_t

given x_t is normally distributed. If in addition to 4.3.6 the conditional distribution of y_t happens to be normal, the NN-GARCH model is said to be *correctly specified for the entire density.* Otherwise, the model suffers from *distributional misspecification* and the value of $\hat{\delta}_T$ that maximises the log-likelihood function is called the *quasi maximum likelihood estimator* (QMLE)(see White (1996)).

Consistency and asymptotic normality of the QMLE in the particular class of NN-GARCH models can be shown along the lines of Bollerslev and Wooldridge (1992), who study the properties of the QMLE in general dynamic models that jointly parametrise the mean and variance of the conditional distribution. Let $\nabla l(\delta)$, $\nabla^2 l(\delta)$ denote the gradient and the hessian of the log-likelihood evaluated at δ, $s_t(\delta)$ denote the gradient of the t-contribution to the log-likelihood l_t and $d_t(\delta) = -E\left(\nabla s_t(\delta)|x_t\right)$, the negative conditional expectation of the gradient of s_t. If the NN-GARCH model is structurally correct and under some additional regularity and moment conditions, the QMLE is consistent for δ_0 and also asymptotically normally distributed around δ_0 with variance-covariance matrix

$$C = A^{-1} I A^{-1} \tag{4.3.7a}$$

$$A = E\left(-\nabla^2 l(\delta_0)\right) = -T^{-1}\sum_{t=1}^{T} E\left(\nabla s_t(\delta_0)\right) = T^{-1}\sum_{t=1}^{T} d_t(\delta_0) \tag{4.3.7b}$$

$$I = V\left(T^{1/2}\nabla l(\delta_0)\nabla' l(\delta_0)\right) = T^{-1}\sum_{t=1}^{T} E\left(s_t(\delta_0)' s_t(\delta_0)\right) \tag{4.3.7c}$$

where the equality in (4.3.7c) follows from conditions 4.3.6, the formulae for the gradient and the hessian of the normal likelihood presented in the appendix, and the properties of the conditional expectation. The matrix C is called the *robust asymptotic variance-covariance matrix* because it is valid under non-gaussianly distributed errors (White (1982, 1996)). Following appendix A, it can further be shown that in the case where the actual density is normal, $E\left(\nabla s_t(\delta_0)\right) = E\left(s_t(\delta_0)' s_t(\delta_0)\right)$, and the variance-covariance matrix simplifies to $C = A^{-1} = -I^{-1}$.

Note that the computation of C involves taking expectations under the true parameter vector δ_0. Hence, from a practical point of view, it would useful to have a consistent estimator of C that can be exclusively computed from sample quantities. According to Bollerslev and Wooldridge (1992), such a sample estimator $\hat{C}_T$ exists and is given by

$$\hat{C}_T = \hat{A}_T^{-1} \hat{I}_T \hat{A}_T^{-1} \tag{4.3.8a}$$

where

$$\hat{A}_T = T^{-1}\sum_{t=1}^{T} d_t(\hat{\delta}_T) \tag{4.3.8b}$$

$$\hat{I}_T = T^{-1}\sum_{t=1}^{T} s_t(\hat{\delta}_T)' s_t(\hat{\delta}_T) \tag{4.3.8c}$$

are the corresponding sample estimators of A and I[2]. Matrix $\hat{C}_T$ is practically useful in obtaining confidence intervals on the estimates $\hat{\delta}_T$ and testing hypothesis involving the true parameter vector δ_0. This is possible by treating $\hat{\delta}_T$ as approximately normally distributed with mean value δ_0 and variance $\hat{A}_T^{-1}\hat{I}_T\hat{A}_T^{-1}/T$.

The estimators $\hat{A}_T$ and $\hat{I}_T$ have the convenient property of being positive definite, or at least positive semi-definite, which remedies several problems that arise in the calculation of confidence intervals. Moreover, as the calculations in the appendix show, they are computable entirely by the residuals, the mean and the variance equations and the derivatives of the mean and the variance functions. Apart from the computation of confidence intervals, $\hat{A}_T$ can be also used in the optimisation of the log-likelihood, as an estimate of the negative hessian which is much easier to compute as it does not involve the calculation of second derivatives of the log-likelihood. It can also be shown, along the lines of Engle (1982), that for a symmetric volatility model (ARCH or GARCH) $\hat{A}_T$ is *block-diagonal* between the parameters of the condition mean and variance. Block-diagonality gives us the opportunity to use an alternative estimation procedure for the NN-GARCH model, which splits the full optimisation problem into two problems of lower dimensionality. This is as follows:

1. Start with estimating the parameters of the neural network using nonlinear least squares and calculate the residuals $\hat{\epsilon}_t$.

2. Use the residuals of the mean model to estimate the parameters of the variance model, by means of maximum likelihood. Calculate the variance estimates $\hat{h}_t$.

3. Use $\hat{h}_t$ in place of h_t and re-estimate the parameters of the mean part maximising the full log-likelihood function. Repeat step 2 until a convergence criterion for the parameter values is satisfied.

The iterative estimation procedure, presented above, is asymptotically efficient to full maximum likelihood estimation and thus offers an alternative easier way to obtain the optimal parameters of the NN-GARCH model (see Cox and Hinkley (1974), p. 308). Note, however, that such iterative estimation is theoretically justified *only* under symmetry of the volatility model. In any case, where an EGARCH, GJR-GARCH or any other asymmetric specification is used, splitting the optimisation problem between the mean and variance parameters results in less efficient estimates and loss of information.

Note that the correct specification of both conditional moments of the distribution is a necessary requirement for the consistency and asymptotic normality of the QMLE. Hence, when estimating the parameters of the neural network by jointly estimating the variance, there is the danger that the mean estimator is *not* consistent unless the variance model is correctly specified for the conditional variance, and vice versa. This is a major difference between estimation of NN-GARCH models and estimation of neural networks using nonlinear least squares, where consistency of NN parameters is independent from the variance specification. Additional regularity conditions require that the likelihood is maximised in the interior of the parameter

[2] The summands in the formula of $\hat{A}_T$ are calculated after applying the conditional expectation operator

space and the true parameter vector δ_0 is uniquely identified. The first condition precludes cases where the log-likelihood estimator falls into the boundary of the parameter space, e.g when a GARCH model is used under constant-variance errors. The second condition is responsible for asymptotic normality of the QMLE. Indeed, if convergence occurs in a flat region of the log-likelihood, the QMLE is no longer asymptotically normally distributed. Identifiability would be invalidated, for example, when the neural network contains inactive hidden neurons.

The implications of these conditions are very important for practical applications of NN-GARCH models and signify the importance of the model-building procedure employed by the researcher. Simply speaking, one cannot combine any neural network with any model of volatility and expect this combination to be successful, unless special attention is paid to identifying the right structure for the mean and variance model. This issue takes a significant part of our thesis, where various model-building procedures are proposed that take into account the interaction of the first two moments of the condition distribution. This discussion is given in sections 4.5.1 and 4.5.2.

4.4 Testing hypotheses about the parameters of a NN-GARCH model

Most models used in empirical research do not typically cover all aspects of the data-generating process. For instance, regression models leave the variance or higher moments of the distribution unspecified. It is therefore an issue whether an a priori misspecified model can serve as a valid device for statistical inference.

Hypotheses in the framework of maximum likelihood are typically formulated as restrictions on the "true" parameter vector and tested by means of a Wald statistic. A broad range of interesting hypotheses can be stated as

$$H_0 : S\delta_0 = 0,$$

against the alternative

$$H_1 : S\delta_0 \neq 0,$$

where S is a $p \times m$ indicator matrix ($p \leq m$) that picks certain elements of δ. If δ_0 belongs to the interior of the parameter space, then under conditions that guarantee consistency and asymptotic normality of the QMLE, the Wald statistic

$$W_T = T\,\hat{\delta}_T' S' \left(S\hat{C}_T S'\right)^{-1} S\hat{\delta}_T$$

follows asymptotically under H_0 the χ^2_p distribution, with p degrees of freedom. This version of the Wald statistic uses the robust asymptotic variance-covariance matrix and hence preserves its asymptotic chi-square distribution under non-normal errors.

Note that in practical applications, practitioners quite often use $\hat{I}_T$ in place of $\hat{C}_T$ in the formula of the Wald statistic and thus implicitly employ a non-robust version of this test. However, the validity of this choice depends on the assumption that the parametric family of models is correct for the *entire* density. For regression models, this would imply that errors are homoskedastic and follow a gaussian distribution. In the context of NN-GARCH models, the assumption requires that the conditionally distribution of standardised errors is correctly specified (normal or t, in our case). Under severe misspecification, the asymptotic distribution of the non-robust Wald statistic is no longer normal and hence becomes impossible to control the theoretical *type I error* or *size* of the test (i.e. the probability of falsely rejecting the null hypothesis).

4.5 Model-building strategy

Assuming a specific structure for the NN-GARCH model (4.3.1), it is always possible that a simpler submodel nested within the former can adequately describe the data under consideration. For example, it is by no means impossible that the conditional mean be a linear function of x_t or the conditional variance be constant over time. In this case, the addition of an extra neuron or the inclusion of a GARCH model in the volatility model is redundant and unnecessarily increases the complexity of the model. Therefore, it would be much more reasonable to start with the simplest possible specification and complicate the model structure in the direction indicated by special features existing in data.

This strategy takes also precautions against various statistical considerations that arise from over-identifiable models and parameter redundancy. As discussed in previous sections, standard statistical inference using maximum likelihood is not possible in NN-GARCH models that include inactive neurons in the mean part or a GARCH volatility model when errors are homoskedastic. Non-identifiability renders also impossible to follow the opposite route in model specification, i.e. to start with a large, possibly over-parametrised, model and then remove insignificant parts by testing hypotheses about certain subsets of parameters. Other model selection approaches such as Bayesian regularization or information criteria (AIC, SBIC) implicitly assume that the model is identifiable.

Finally, a simple-to-complex model-building strategy is also totally justifiable when applying a NN-in-mean (and in fact any highly nonlinear) model to heteroskedastic data. As neural networks have a high approximating ability, it is very likely that some affects due to the presence of conditional heteroskedasticity in the data-generating process be misinterpreted as neglected nonlinearity and hence "absorbed" by the parameters of an over-parametrised NN. Simulations presented in chapter 5 shed more light on this issue and show that standard statistical methodologies can lead to over-identifiable networks in the presence of strong properties in the second-moment of the distribution. It is thus absolutely necessary that the methodology used in specifying the mean part offer some kind of *robustification* against these adverse affects. Robustified tests are hopefully available within the QML theory and discussed below.

In the specification of a NN-GARCH model, the general rule is to first specify the conditional mean equation, using linear or nonlinear least-squares, and then the conditional variance. The theoretical motivation for this choice lies in the well-documented result that if the expectational model m_t is correctly specified for the conditional expectation $E(y_t|x_t)$, the parameters of the mean equation can be consistently estimated without reference to the conditional variance (see e.g. White (1996), ch 5). On the contrary, it is not possible to consistently estimate the parameters of the variance model if the conditional mean is badly specified. This is because common tests on the parameters of the variance equation implicitly impose under the null correct specification of the mean part. Hence, rejecting the null gives no clear indication to the researcher on how to proceed.

Our model-building procedure is schematically presented in figures 4.1 and 4.2 and analytically discussed in the following sections.

4.5.1 Specification of the mean part

The specification of the mean follows the sequential statistical testing procedure of Medeiros et al. (2006). As seen from figure 4.1, this starts with specifying a linear regression model using ordinary least-squares (OLS). The optimal number of variables in the linear model is selected by means of an information criterion (AIC or SBIC) or by an autocorrelation test. The researcher then chooses a significance level (say $\alpha\%$) and tests the null hypothesis of linearity against a neural network model with a single hidden neuron. This test for neglected nonlinearity is repeated for all combinations of explanatory variables. If linearity is not rejected at the given significance level, the final model is linear in mean. Otherwise, a NN model with a single hidden neuron is estimated using nonlinear least squares. The set of variables attached to the neuron are the ones for which the lowest p-value of the test is reported. The single-neuron model is now tested against a NN with an additional neuron in the hidden layer. If the null is rejected, the above procedure is repeated for NN models with $h = 2, \ldots,$ neurons until first acceptance of the null. To favour parsimonious models, Medeiros et al. (2006) proposed to half the significance level at each stage of subsequent test.

The neglected nonlinearity test employed by Medeiros et al. (2006) is essentially an application of a *Lagrange Multiplier* (LM) test on additional structure in the mean model. Assume that at some stage of the procedure described above the analyst has detected an additive NN model with h neurons in the hidden part:

$$y_t = \phi'\bar{x}_t + f(x_t;\theta) + \epsilon_t \tag{4.5.1}$$

where $f(x_t;\theta)$ is given by (4.3.2). Starting with model (4.5.1), the question is whether the current approximation to the conditional expectation can be improved by adding more hidden neurons that capture neglected nonlinearities. If the answer is yes, the data can be described more accurately by adding one more neuron to the mean model

$$y_t = \phi'\bar{x}_t + f(x_t;\theta) + \lambda_{h+1}F(w_{h+1}x_t - c_{h+1}) + \epsilon_t^* \tag{4.5.2}$$

The appropriate test on the additional neuron is the *test against neglected nonlinearity* or simply *nonlinearity test.* If $f(x_t;\theta) = 0$ the neglected nonlinearity test becomes a test of model linearity against model nonlinearity. Note that if the extra neuron is redundant then either λ_{h+1} or w_{h+1} are identically equal to zero in which case the output of the neuron is constant and merges with the intercept of the linear part. Hence, the appropriate null hypothesis for the neglected nonlinearity test can be formulated as $H_0 : \lambda_{h+1} = 0$ or $H_0 : w_{h+1} = 0$. However if either of the nulls hold the corresponding w_{h+1}- or λ_{h+1}-parameters can take any values without effecting the output of the neuron; in other words, they are *non-identifiable*. This is a problem in statistical inference where some of the parameters of the model are only identifiable under the alternative hypothesis (see e.g. Davies (1977)). Non-identifiability has severe consequences for the properties of the least-squares estimators and renders inappropriate the use of classical statistics to test restrictions on the parameters imposed by the null hypothesis.

Two ways to carry out the nonlinearity test have been proposed in the literature that bypass the problem of non-identifiability. One is due to White (1989a), who proposed testing the hypothesis $H_0 : \lambda_{h+1} = 0$ by assigning random values to the weights w_{h+1} of the extra neuron. The finite sample properties of this test were investigated by Lee et al. (1993). Another technique with better performance was proposed by Teräsvirta et al. (1993). In this approach, the identification problem is solved in the spirit of Luukkonen et al. (1988); Davies (1977) by using a third-order Taylor approximation to the hypothetical additional neuron, in which case the augmented model (4.5.2) takes the form

$$y_t = \phi'\bar{x}_t + f(x_t;\theta) + \xi' z_t + R_3(x_t) + \epsilon_t^* \tag{4.5.3}$$

where ξ is a vector of parameters,

$$z_t = (x_{1t}^2, x_{1t}x_{2t}, \ldots, x_{it}x_{jt}, \ldots, x_{nt}^2, x_{1t}^3, x_{1t}^2x_{2t}, \ldots, x_{it}x_{jt}x_{kt}, \ldots, x_{nt}^3) \in \mathbb{R}^l$$

is the vector of the extra regressors of the mean model, with $l = n(n+1)/2 + n(n+1)(n+2)/6$, and $R_3(x_t)$ is the Taylor remainder. A test for an extra neuron is now equivalent to testing the hypothesis $H_0 : \xi = 0$ against the alternative $H_1 : \xi \neq 0$. Note that if H_0 is true, $R_3(x_t)$ vanishes also, so the error process remains affectively unchanged.

Teräsvirta et al. (1993) proposed testing the null hypothesis using an LM test on the additional ξ parameters. Let

$$\hat{\epsilon}_t = y_t - \hat{\phi}'\bar{x}_t + f(x_t;\hat{\theta})$$

where $\hat{\phi}$ and $\hat{\theta}$ are the least-squares estimates of the parameters of the restricted model (4.5.1). Let also $\nabla\hat{f}_t$ be the gradient of $f(x_t;\hat{\theta})$ evaluated at $\phi = \hat{\phi}$, $\theta = \hat{\theta}$ and $\xi = 0$. The LM test is carried out in three steps:

Procedure 4.5.1.1

1. Regress $\hat{\epsilon}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$ and compute the residuals $\hat{\varepsilon}_t$
2. Regress $\hat{\varepsilon}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$ and z_t and compute R^2, the coefficient of determination of the regression
3. The test statistic is calculated as $LM_T = TR^2$, where T is the sample size

The asymptotic (large-sample) distribution of LM is taken as χ^2 with l degrees of freedom. So, if the researcher wants the probability of falsely rejecting the null not to exceed α, the significance level, he should reject the hypothesis of no extra neuron whenever $LM_T > \chi^2_{\alpha,l}$, where $\chi^2_{\alpha,l}$ denotes the $100\alpha\%$ critical value of the χ^2 distribution with l degrees of freedom.

Typically, in applying least-squares optimisation to highly nonlinear neural networks convergence problems arise, so that $\hat{\epsilon}_t$ may not be precisely orthogonal to $\bar{x}_t$ and $\nabla\hat{f}_t$, which is the first-order optimality condition for nonlinear least squares. This adversely affects the size of the test, leading to over-rejections of the null hypothesis. To circumvent numerical convergence problems in the computation of the test statistic, it is typical to perform step 1 in the procedure above, where $\hat{\epsilon}_t$'s are regressed on $\bar{x}_t$, $\nabla\hat{f}_t$ and the new residuals $\hat{\varepsilon}_t$ of the regression are used in place of $\hat{\epsilon}_t$. By the properties of the regression, $\hat{\varepsilon}_t$ is orthogonal to $\bar{x}_t$ and $\nabla\hat{f}_t$.

Generally, the validity of the above procedure, henceforth called *standard* LM or simply S-LM test, depends on the assumption that errors ϵ_t are independent identically distributed. In particular, the S-LM statistic no more follows an asymptotic chi-square distribution in the presence of ARCH or other types of heteroskedasticity (Wooldridge (1990)). This means that using chi-square critical values does not result in a test with theoretical type I error equal to α, in which case the test suffers from *size distortions* as is commonly said in the statistical literature. Simulations presented in chapter 5 show that in the presence of ARCH effects in the disturbances of the mean model, the S-LM test tends to overreject the hypothesis of no additional neuron, occasionally producing excessively nonlinear NN models.

To overcome the problems of the S-LM test, Wooldridge (1990, 1991) proposed a modification of the test that is *robust* to changes in the variance of the errors. The robustified LM test, henceforth called RB-LM, is carried out in a similar way by running a set of auxiliary regressions:

Procedure 4.5.1.2

1. Regress $\hat{\epsilon}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$ and compute the residuals $\hat{\varepsilon}_t$
2. Regress z_t on $\bar{x}_t$, $\nabla\hat{f}_t$ and take the $1 \times l$ residuals vector $\hat{u}_t$
3. Run the regression 1 on $\hat{\varepsilon}_t\hat{u}_t$ and compute the sum of squared residuals SSR_1
4. Compute the statistic as $LM_T = T - SSR_1$

The modified LM_T statistic preserves its asymptotic χ^2 distribution under heterogeneously distributed errors and is the one recommended in Granger and Teräsvirta (1998); Medeiros et al. (2006) to resolve the problems of the standard nonlinearity test. Simulations presented in chapter 5 show that although the null rejection rate of the RB-LM test is close to the nominal size under various forms of variance misspecification, the test can have poor power in detecting hidden nonlinearity in the residuals. This is a problem if one is also interested in modelling nonlinearity in mean apart from heteroskedasticity.

Another version of the RB-LM test arises if one initially standardises the quantities $\hat{\epsilon}_t$, $\bar{x}_t$, $\nabla\hat{f}_t$ and z_t that enter in the procedure above by $\hat{h}_t$, an estimate of the time t conditional variance of ϵ_t. In this way, one tests the hypothesis of additional nonlinearity in mean by incorporating information from the volatility structure of the time-series[3]. According to Wooldridge (1991), the validity of this test, henceforth denoted by RBV-LM, is not affected by bad specification of the variance process, however a good guess of the true variance h_t may increase the power of the robustified test in detecting hidden nonlinearity in mean. Since in this thesis we are concerned with economic and financial applications of NNs, a natural option for obtaining a good initial guess is fitting a GARCH(1,1) model to the residuals of the NN model $\{\hat{\epsilon}_t, t = 1, 2, \ldots, T\}$, or any model the researcher believes that it adequately describes the volatility structure of errors (apart, of course, from a model of constant variance in which case the RBV-LM test is equivalent to RB-LM).

This modification of the LM test procedure reveals a new dimension in testing hypotheses about the parameters of a regression model. It is an attempt to incorporate information from the second-moment structure of the conditional distribution while testing for the adequacy of the mean specification. The main advantage of this testing procedure, henceforth called RBV-LM, is that it is *robust* to failure of the research's initial assumption about the underlying volatility process, but it might also have *more power* than the ordinary RB-LM if $\hat{h}_t$ offers a better approximation to the conditional variance of y_t given x_t than a model of constant variance does (Wooldridge (1991)). In this thesis, we propose the following procedure to carry out the RBV-LM test:

Procedure 4.5.1.3

1. Estimate the parameters of the restricted NN model using nonlinear least-squares
2. Compute the residuals $\hat{\epsilon}_t$ and estimate a GARCH(1,1) model using maximum likelihood.
3. Based on the GARCH specification, compute the time t variance estimates $\hat{h}_t$ and perform the testing procedure 4.5.1.2 by using $\tilde{\epsilon}_t \equiv \hat{\epsilon}_t/\sqrt{\hat{h}_t}$, $\tilde{x}_t \equiv \bar{x}_t/\sqrt{\hat{h}_t}$, $\nabla\tilde{f}_t \equiv \nabla\hat{f}_t/\sqrt{\hat{h}_t}$, $\tilde{z}_t \equiv z_t/\sqrt{\hat{h}_t}$ in place of the corresponding quantities.

Chapter 5 investigates the finite-sample performance of the RBV-LM under various types of volatility dynamics (homoskedasticity, GARCH or EGARCH) and error distributions (normal or t). Results show that the RBV-LM test with GARCH(1,1) volatility estimates has the right size in detecting nonlinearity under heteroskedasticity, even when the variance process is not correctly specified (i.e errors are homoskedastic or EGARCH). However, the RBV-LM test leads to a substantial increase in power if the volatility model is closely approximated. In any case, the RBV-LM is at least as powerful as RB-LM and sometimes more powerful in detecting hidden nonlinearity.

[3] To our knowledge, this version of the robustified LM test has not been previously applied in other relevant studies considering the specification of either neural networks or common nonlinear econometric models (see e.g. Becker and Hurn (2006); Medeiros and Veiga (2003); Medeiros et al. (2006)). This is possibly because in these studies the emphasis is put on modelling mean dependencies.

4.5.2 Specification of the variance part

Once the appropriate complexity for the mean model has been identified, the next step is to approximate the variance structure. As seen from figure 4.2, the model-building procedure carries on with testing the hypothesis of homoskedasticity in errors against ARCH effects. This can be done using Engle (1982)'s LM test. It should be noted here however that the validity of this test depends on the assumption of conditional homokurtosis and normality of errors (Wooldridge (1990)), hence Engle's procedure may lead to wrong inference under leptokurtic or asymmetric error distributions. A robust version of the ARCH LM test can be derived along the lines of the general robustification strategy proposed in Wooldridge (1991, 1990) (further details are given in section 4.6.4). This test is also performed by running a set of auxiliary regressions and, although robust against non-normality, it loses nothing in terms of asymptotic efficiency if the normality assumption happens to hold. In this sense, Wooldridge's LM test dominates Engle's and is the one we adopt in testing for heteroskedasticity.

If the null hypothesis of homoskedasticity in errors cannot be rejected, then no model for volatility is specified. Otherwise, a joint NN(h)-GARCH$(1, 1)$ model is estimated, with h being the number of neurons specified by the nonlinearity test ($h = 0, 1, 2, \ldots$). Another set of diagnostics, tests the hypothesis of GARCH$(1, 1)$ against additional ARCH structure in the squared standardised residuals. This is a test for *neglected autoregressive heteroskedasticity in the residuals* (see section 4.6.4)). If the null hypothesis is rejected, a NN(h)-GARCH(p, q) model is estimated, where p, q are determined accordingly. Once the appropriate symmetric volatility model is specified, additional diagnostics are performed that examine *asymmetric effects* in the variance process. Several alternatives are discussed in Engle and Ng (1993) and include a test for *sign*, *negative* and *positive size bias*. As the Engle's ARCH test, however, these diagnostics are also sensitive to the normality assumption and a rejection of the null hypothesis is not a clear indication of asymmetric effects in variance. In section 4.6.4 we propose a robustification procedure of these tests along the lines of Wooldridge (1991, 1990). If the null hypothesis of a symmetric variance process is rejected against sign or size bias, the next step is to jointly estimate a NN with an asymmetric variance model (EGARCH, GJR, etc).

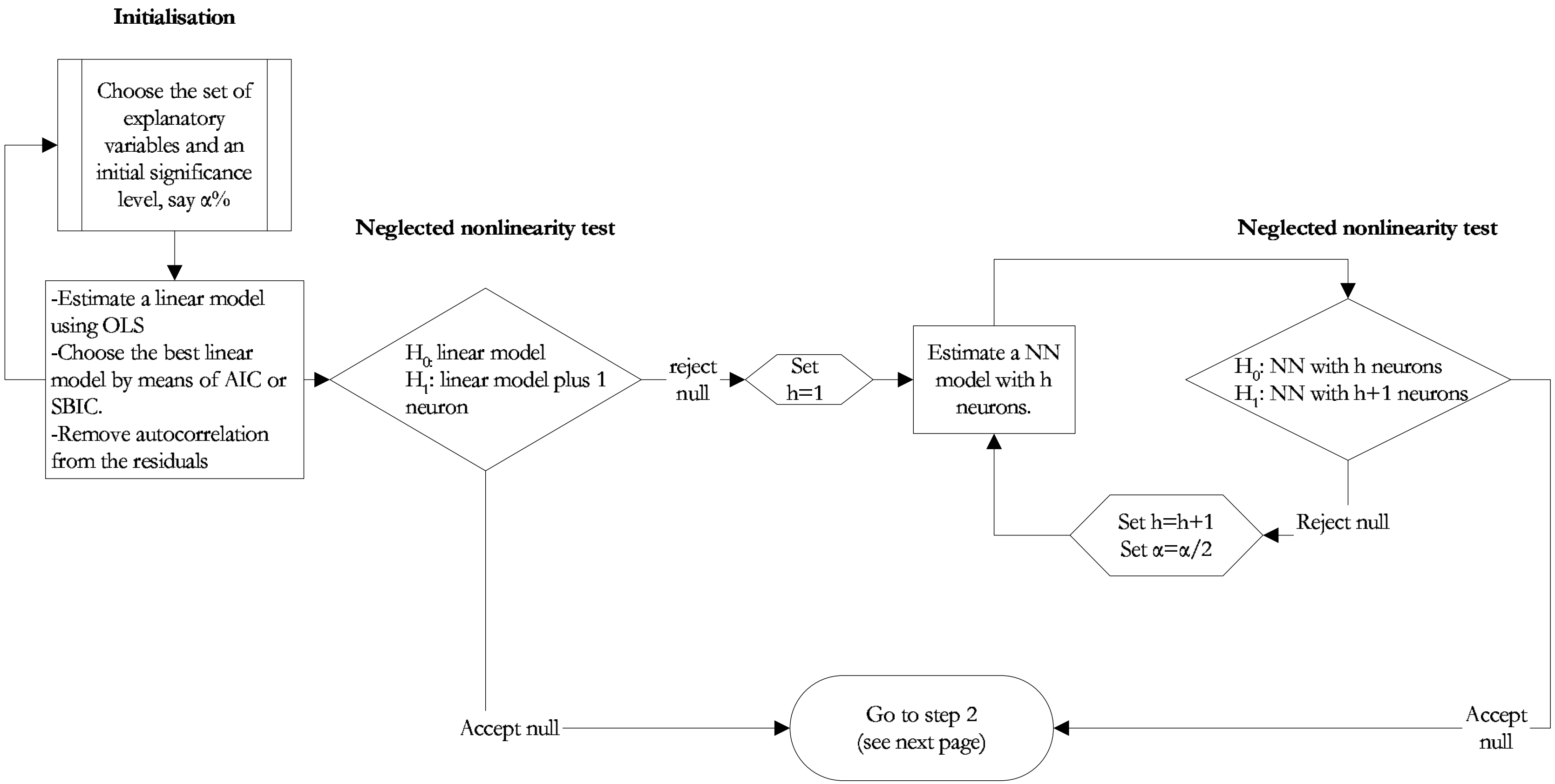
Initialisation
Choose the set of explanatory variables and an initial significance level, say α%
-Estimate a linear model using OLS
-Choose the best linear model by means of AIC or SBIC.
-Remove autocorrelation from the residuals
Neglected nonlinearity test
H0: linear model
H1: linear model plus 1 neuron
reject null
Set h=1
Estimate a NN model with h neurons.
Neglected nonlinearity test
H0: NN with h neurons
H1: NN with h+1 neurons
Reject null
Set h=h+1
Set α=α/2
Accept null
Go to step 2 (see next page)
Accept null

FIGURE 4.1: The strategy for specifying the conditional mean part of a NN-GARCH model. The neglected nonlinearity test is performed for all possible combinations of explanatory variables. Upon each rejection of the null, we estimate an additive linear and neural network model with $h + 1$ number of neurons. The variables attached to the extra neuron are the ones which produce the lowest p-value of the test.

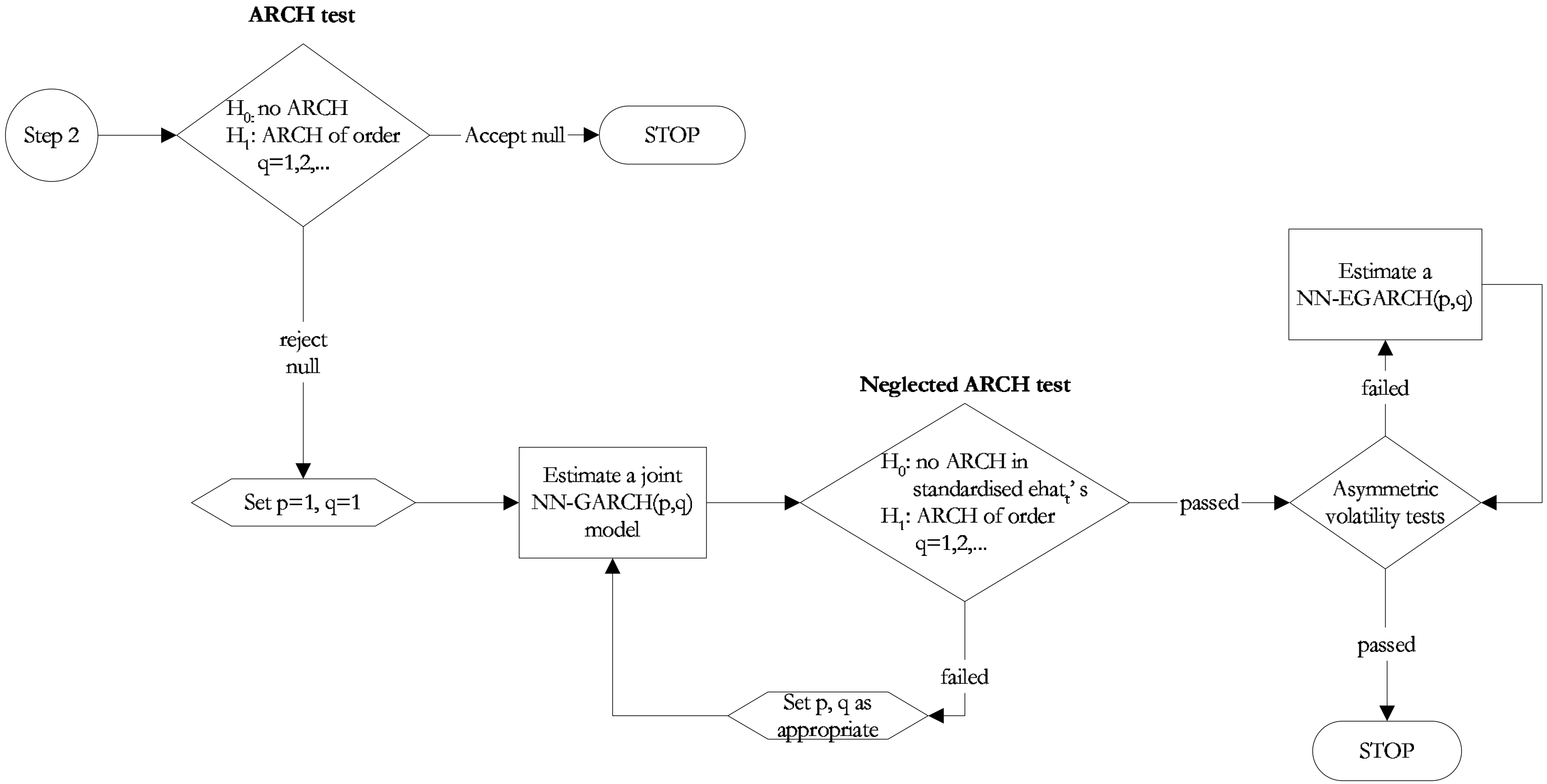
ARCH test
Step 2
H_0: no ARCH
H_1: ARCH of order q=1,2,...
Accept null
STOP
reject null
Set p=1, q=1
Estimate a joint NN-GARCH(p,q) model
Neglected ARCH test
H_0: no ARCH in standardised ehat$_t$' s
H_1: ARCH of order q=1,2,...
passed
Asymmetric volatility tests
failed
Estimate a NN-EGARCH(p,q)
passed
STOP
failed
Set p, q as appropriate

FIGURE 4.2: The strategy for specifying the conditional variance part of a NN-GARCH model.

4.6 Evaluation of the model specification

After constructing a NN-GARCH model it is important to test whether the hypotheses implicitly imposed by the specification conform with the data, i.e. whether

1. errors contain no forecastable structure in mean, i.e. there is no autocorrelation or neglected nonlinearity in the residuals
2. errors contain no forecastable structure in variance, i.e. there is no serial correlation or asymmetric effects in the squared standardised residuals
3. errors are conditionally distributed according to the density model assumed by the specification (gaussian or student)

From the requirements set above, 1 & 2 ensure that the NN-GARCH model is structurally correct, i.e. conditions (4.3.6) hold, which guarantees that the QMLE estimator is consistent and asymptotically normally distributed. Asymptotic normality is important for the chi-square distribution of the Wald statistic that tests constraints on the parameters on the model, and this is why significance tests on the final model specification can be safely performed *after* the diagnostic stage. Requirement 3 is concerned with the correct specification of the conditional distribution that determines whether the researcher should use a standard or a robust estimator of the variance-covariance matrix in the computation of the test statistic.

From a practical point of view, requirements 1& 2 are perhaps the most important as they are directly linked to the adequacy of the model for the particular data-generating process. Higher-order distributional properties of the data-generating process, such as asymmetry or extra kurtosis, are of interest in some cases, especially as concerns the ability of the model to predict realistic future scenarios on the target variable or to estimate the risk associated with extreme price movements. Even in this case, however, there are techniques for estimating the unconditional distribution of errors without resting on a particular distributional model. A discussion is given in chapter 6. Therefore, in what follows we shall mainly concentrate on diagnostics that evaluate 1 & 2 and resort to common non-parametric procedures, such as the Jarque-Bera, the Kolmogorov-Smirnov or Chi-Square test, for judging the goodness-of-fit of the distributional model.

The importance of a subsequent stage of diagnostic checking as a safeguard against specification bias has only recently been recognised by the computational intelligent society. Most practitioners employ heuristic procedures or popular econometric diagnostics, such as the Ljung - Box - Pierce test for autocorrelation in the residuals or the Breysch - Pagan test for heteroskedasticity. However, these tests have been developed in the framework of linear models and cannot be directly applied to our case. The reason is that the asymptotic distribution of the test statistics is not known under a NN-GARCH process and hence it is not possible to access the theoretical type I error of the test. In this section, we propose a unifying diagnostic framework for NN-GARCH models that is based on LM tests. We particularly consider tests of autocorrelation (or serial dependence in mean), heteroskedasticity (or serial dependence in variance), neglected nonlinearity in mean and asymmetric effects in variance. Similar in nature tests have been applied to econometric nonlinear models that jointly parametrise the conditional mean and variance (see e.g. Lundbergh and Teräsvirta (1998, 2002)).

4.6.1 The general framework

To derive a general framework for diagnostic checking on mean and variance, we consider the following extended NN-GARCH model:

$$y_t = \phi' \bar{x}_t + f(x_t;\theta) + g(z_t;\xi) + \epsilon_t^* \tag{4.6.1a}$$

$$\epsilon_t^* = u_t \sqrt{h(x_t;\alpha)\,\omega(z_t;\beta)} \tag{4.6.1b}$$

where ξ, β are vectors of parameters and $g(z_t;\xi)$, $\omega(z_t;\beta)$ are assumed continuous and twice differentiable for all ξ, β and (almost) everywhere in the sample space of z_t. z_t is a set of variables derived from the information set of the researcher that does not necessarily share common elements with x_t. Assume without loss of generality that $g(z_t;0) = 0$ and $\omega(z_t;0) = 1$. Note that contrary to the mean model, in the variance part we assume a multiplicative alternative volatility model. This structure allows us to compute a greater variety of model diagnostics, including diagnostics on an EGARCH model. The hypothesis of no additional structure in the mean *and* the variance equation can be stated as

$$H_0 : \xi = 0 \text{ and } \beta = 0$$

Under the null, $\epsilon_t^* = \epsilon_t$ and the error structure remains unchanged. Let $\hat{\delta}_T = (\hat{\phi}_T, \hat{\theta}_T, 0, \hat{\alpha}_T, 0)$ be the QMLE of the *restricted* NN-GARCH, i.e. the model with the restrictions imposed under the null imbedded, and

$$\hat{\epsilon}_t = y_t - \hat{\phi}' \bar{x}_t - f(x_t;\hat{\theta})$$

be the estimated residuals. A common way to investigate this hypothesis is by means of an LM test on the parameters ξ and β of the joint NN-GARCH model. Let the inverse of the fisher information matrix be

$$I^{-1} = \begin{pmatrix} J_{\theta\theta'} & J_{\theta\xi'} & J_{\theta\alpha'} & J_{\theta\beta'} \\ J_{\xi\theta'} & J_{\xi\xi'} & J_{\xi\alpha'} & J_{\xi\beta'} \\ J_{\alpha\theta'} & J_{\alpha\xi'} & J_{\alpha\alpha'} & J_{\alpha\beta'} \\ J_{\beta\theta'} & J_{\beta\xi'} & J_{\beta\alpha'} & J_{\beta\beta'} \end{pmatrix}$$

where $J_{xy'}$ denotes the xy block of I^{-1}. The LM test statistic is computed as

$$LM_T = T \begin{pmatrix} \nabla_\xi \hat{l} \\ \nabla_\beta \hat{l} \end{pmatrix}' \begin{pmatrix} \hat{J}_{\xi\xi'} & \hat{J}_{\xi\beta'} \\ \hat{J}_{\beta\xi'} & \hat{J}_{\beta\beta'} \end{pmatrix} \begin{pmatrix} \nabla_\xi \hat{l} \\ \nabla_\beta \hat{l} \end{pmatrix}$$

where hatted quantities are evaluated under the restricted QMLE. The above statistic can be equivalently computed from a set of auxiliary regressions. Let

$$\hat{\eta}_t \equiv \begin{pmatrix} \hat{\epsilon}_t \\ \hat{\epsilon}_t^2 - \hat{h}_t \end{pmatrix}$$

$$\hat{\Sigma}_t \equiv \begin{pmatrix} \hat{h}_t & 0 \\ 0 & 2\hat{h}_t^2 \end{pmatrix}$$

$$\hat{\Lambda}_t \equiv \begin{pmatrix} \nabla'_\mu \hat{m}_t & \nabla'_\xi \hat{g}_t & 0 & 0 \\ \nabla'_\mu \hat{\nu}_t & \nabla'_\xi \hat{\nu}_t & \nabla'_\alpha \hat{\nu}_t & \nabla'_\beta \hat{\nu}_t \end{pmatrix}$$

and

$$\hat{\Psi}_t \equiv \begin{pmatrix} \nabla'_\mu \hat{m}_t & 0 \\ \nabla'_\mu \hat{h}_t & \nabla'_\alpha \hat{h}_t \end{pmatrix}$$

where $m(.)$, $\nu(.)$ are the extended mean, volatility models and $\mu = (\phi', \theta')'$ is the vector of the mean parameters of the restricted model. The following procedure delivers the LM statistic (see Bollerslev and Wooldridge (1992), p. 15):

Procedure 4.6.1.1

1. Regress $\hat{\Sigma}_t^{-1/2}\hat{\eta}_t$ on $\hat{\Sigma}_t^{-1/2}\hat{\Lambda}_t$ and save the vector residuals $\hat{r}_t$
2. Compute the LM_T test statistic as $2TR^2$, where R^2 is the correlation coefficient of the regression

In a highly nonlinear NN-GARCH specification, convergence problems arise so that the matrix of generalised residuals is not precisely orthogonal to the matrix of $\hat{\Sigma}_t^{-1/2}\hat{\Psi}_t$, which is the first-order condition for the optimality of the QMLE. This may adversely affect the size of the test. To circumvent this problem, we can regress $\hat{\Sigma}_t^{-1/2}\hat{\eta}_t$ on $\hat{\Sigma}_t^{-1/2}\hat{\Psi}_t$ and use the residuals of the regression $\breve{\eta}_t$ in place of $\hat{\Sigma}_t^{-1/2}\hat{\eta}_t$ in the above procedure.

Assuming normality of errors, LM_T follows an asymptotic χ^2 with $l_\xi + l_\beta$ degrees of freedom, where l_* is the dimension of the corresponding vector. In order to derive a test that is robust to distributional misspecification of the NN-GARCH model, we can follow the procedure:

Procedure 4.6.1.2

1. Run the matrix regression

$$\hat{\Sigma}_t^{-1/2}\hat{\eta}_t \text{ on } \hat{\Sigma}_t^{-1/2}\hat{\Psi}_t, \quad t = 1, 2, \ldots, T$$

   and save the matrix residuals $\breve{\eta}_t$.

2. Run the matrix regression

$$\hat{\Sigma}_t^{-1/2}\hat{\Lambda}_t \text{ on } \hat{\Sigma}_t^{-1/2}\hat{\Psi}_t, \quad t = 1, 2, \ldots, T$$

   and save the matrix residuals $\breve{\Lambda}_t$.

3. Run the OLS regression

$$1 \text{ on } \breve{\eta}_t' \breve{\Lambda}_t, \quad t = 1, 2, \ldots, T$$

and compute LM as $TR^2 = T - SSR$, where SSR is the sum of squared residuals of the last regression.

Step 2 is effectively responsible for the robustification of the test. This form of the LM test has several attractive features. First, it is valid under non-normality of errors and also asymptotically equivalent to the standard LM test when the normality assumption happens to hold. In addition, the computation of the test statistic involves only first derivatives and is based on a set of regressions, whose implementation is straightforward. Based on the above general testing framework, we can derive a series of diagnostics on the NN-GARCH model by giving different forms to g and ω. Several cases are discussed below:

4.6.2 Testing the conditional mean model

In conditional mean tests, function ω can be taken trivially equal to one also under the alternative hypothesis, so it does not enter the formulae above. In this case, the matrix of gradients of the unrestricted model, evaluated under the null, takes the form:

$$\hat{\Lambda}_t \equiv \begin{pmatrix} \nabla_\mu' \hat{m}_t & \nabla_\xi' \hat{g}_t & 0 \\ \nabla_\mu' \hat{h}_t & \nabla_\xi' \hat{h}_t & \nabla_\alpha' \hat{h}_t \end{pmatrix}$$

Notice that $\hat{\Lambda}_t$ has the first and the third column common with $\hat{\Psi}_t$. Hence these two columns may be omitted from $\hat{\Lambda}_t$, when carrying out the test, as they do not affect the value of the test statistic.

Testing for serial correlation in mean

Testing serial independence of errors is a typical step towards the evaluation of the final model. Rejecting this hypothesis suggests that the model cannot adequately describe the dynamic structure of y_t. This is the case e.g. when more lags of y_t have to be included in x_t.

Let us assume that the errors of the restricted NN-GARCH model follow an autoregressive process of order l_ξ,

$$\epsilon_t = \xi' z_t + \epsilon_t^*$$

where $z_t = (\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-l_\xi})'$ and ϵ_t^* is an iid error process. To test linear independence in the residuals, we set $g(z_t; \xi) = \xi' z_t$ and take the null to be $H_0 : \xi = 0$ against the alternative hypothesis $H_1 : \xi \neq 0$. The gradient $\nabla_\xi \hat{g}_t$ of g_t evaluated at null is simply $\hat{z}_t$, where

$$\hat{z}_t = (\hat{\epsilon}_{t-1}, \hat{\epsilon}_{t-2}, \ldots, \hat{\epsilon}_{t-l_\xi})'$$

for $t = l_\xi + 1, m_\xi + 2, \ldots, T$.

Testing for omitted variables

An important case of misspecification arises, when the researcher has omitted from the set of explanatory variables x_t some exogenous variables that might influence the target variable y_t. Similarly to the test presented in the previous section, we can test the hypothesis of omitted variables in the linear part of the model, by stating the alternative as $g(z_t; \xi) = \xi' z_t$, with z_t being the set of additional exogenous regressors. Note that in this case $\nabla_\xi \hat{g}_t$ is simply z_t. We can also test the NN-GARCH specification against the alternative that important variables have been excluded from the neural network model. In this case, z_t could be specified as a vector of second- or higher-order cross product terms of the exogenous variables. This case is examined below.

Testing for additional nonlinearity in mean

We can test neglected nonlinearity in the mean part, generalising the procedure of Teräsvirta et al. (1993) for NN-GARCH specifications. In particular, we take $g(z_t; \xi)$ to be equal to the third-order Taylor polynomial $\xi' z_t$, where

$$z_t = (x_{1,t}^2, x_{1,t}x_{2,t}, \ldots, x_{i,t}x_{j,t}, \ldots, x_{1,t}^3, \ldots, x_{i,t}x_{j,t}x_{k,t}, \ldots, x_{n,t}^3)$$

Under the null hypothesis, $\xi = 0$ and the LM statistic follows asymptotically the χ^2 distribution with $l_\xi = n(n+1)/2 + n(n+1)(n+2)/6$ degrees of freedom, where n is the number of the explanatory variables.

4.6.3 Testing the conditional variance structure

The general testing framework on volatility models is based on investigating whether we can predict the squared normalised residuals by some variables observed in the past but not included in the volatility model being used. If these variables have predictive ability on the squared normalised residuals then the volatility model is misspecified. In evaluating the conditional variance, we treat function g as being trivially equal to zero also under the alternative hypothesis. In this case, the matrix of gradients of the unrestricted model takes the form:

$$\hat{\Lambda}_t = \begin{pmatrix} \nabla'_\mu \hat{m}_t & 0 & 0 \\ \nabla'_\mu \hat{\nu}_t & \nabla'_\alpha \hat{\nu}_t & \nabla'_\beta \hat{\nu}_t \end{pmatrix}$$

which, for the range of hypotheses considered here, becomes

$$\hat{\Lambda}_t = \begin{pmatrix} \nabla'_\mu \hat{m}_t & 0 & 0 \\ \nabla'_\mu \hat{h}_t & \nabla'_\alpha \hat{h}_t & \hat{h}_t \nabla'_\beta \hat{\omega}_t \end{pmatrix}$$

Notice that the gradient matrix of the restricted model $\hat{\Psi}_t$ shares the first two columns with $\hat{\Lambda}_t$. Hence, these columns may be omitted from $\hat{\Lambda}_t$ when performing the matrix regressions.

Testing for remaining serial correlation in the standardised squared residuals

Similarly to Lundbergh and Teräsvirta (1998), we can test the hypothesis of remaining ARCH effects in the standardised residuals of a NN-GARCH model by setting $\omega(z_t;\beta) = 1 + \beta' z_t$, with

$$z_t = (\hat{u}^2_{t-1}, \hat{u}^2_{t-2}, \ldots, \hat{u}^2_{t-l_\beta})$$

and $\hat{u}_t \equiv \hat{\epsilon}_t \Big/ \sqrt{\hat{h}_t}$. Note that if the alternative hypothesis is true in this case, then from the volatility part of the extended model it follows that $E(\epsilon^2_t) = h_t\omega_t$, which becomes

$$E\left(\frac{\epsilon^2_t}{h_t}\right) = 1 + \xi_1 \frac{\epsilon^2_{t-1}}{h_{t-1}} + \xi_2 \frac{\epsilon^2_{t-2}}{h_{t-2}} + \ldots + \xi_{l_\beta} \frac{\epsilon^2_{t-l_\beta}}{h_{t-l_\beta}}$$

This means that standardised errors are serially correlated and thus past $\hat{u}^2_{t-j}$'s contain useful information for the conditional variance.

The null hypothesis of no remaining heteroskedasticity is equivalent to testing $\beta = 0$. Under this null, the LM statistic is asymptotically χ^2-distributed with l_β degrees of freedom.

Testing for asymmetries in the volatility model

The final set of diagnostics on the volatility model examine unmodelled asymmetries in the variance of the squared errors. Following Engle and Ng (1993) we can derive tests for *sign*, *positive size* and a *negative size bias*[4]. The first of these tests examines whether positive or negative shocks can (further) predict future variance. The last two tests examine whether large and small negative (positive) shocks have an impact on volatility, not captured by the current conditional variance model.

Asymmetric variance effects can be tested by setting $\omega_t(\beta) = e^{\beta' z_t}$, where z_t equals

- S^-_{t-1}, for the sign bias test
- $S^-_{t-1}\epsilon_{t-1}$, for the negative size bias and
- $S^+_{t-1}\epsilon_{t-1}$, for the positive size bias test

S^-_t (S^+_t) are dummy variables taking the value 1 if $\hat{\epsilon}_t < 0$ ($\hat{\epsilon}_t > 0$) and 0 otherwise. One can also device a joint test against all these effects by taking $z_t = (S^-_{t-1},\, S^-_{t-1}\hat{\epsilon}_{t-1},\, S^+_{t-1}\hat{\epsilon}_{t-1})'$. Critical values are taken from a $\chi^2_{l_\beta}$ distribution where the degrees of freedom l_β are adjusted according to the number of terms included in z_t.

[4] In calculating the above test statistics, Engle and Ng (1993) assume a mean model that is equal to a constant. However, these formulae do not directly apply to our case, where the mean is parametrised as a neural network. In this section, we propose an extension of the testing procedure that accommodates nonlinearities in mean.

4.6.4 Some diagnostics for symmetric volatility models

Recall from section 4.3.2, that if the volatility formulation of the NN-GARCH model is homoskedastic or symmetric, as in the case of an ARCH/GARCH model, the off-diagonal blocks of the expected hessian are zero. In this case, the standard LM test statistic simplifies to

$$LM_T = T\,\nabla'_\beta \hat{l}(\hat{J}_{\beta\beta'})\nabla_\beta \hat{l}$$

so that the volatility specification can be tested *without* reference to the mean model[5]. It is important to bear in mind, however, that this simplification is only valid for symmetric volatility models; in cases where the conditional variance is paramatrised as an EGARCH or a GJR-GARCH process, one necessarily has to follow the procedure described in section 4.6.3.

If we let $\eta_t \equiv \epsilon_t^2/h_t - 1$, the standard test statistic can be computed from the following procedure

Procedure 4.6.4.1

1. Regress $\hat{\eta}_t$ on $\nabla_\alpha \hat{h}_t/\hat{h}_t$ for $t = 1, 2, \ldots, T$ and save the residuals $\breve{\eta}_t$.
2. Regress $\breve{\eta}_t$ on $\nabla_\alpha \hat{h}_t/\hat{h}_t$ and $\nabla_\beta \hat{\omega}_t$ for $t = 1, 2, \ldots, T$ and compute LM_T as TR^2 from the regression.

LM_T has an asymptotic chi-square distribution with l_β degrees of freedom assuming normality and conditional homokurtosis of errors. A version of the LM that is robust to non-normality can be computed in analogy to procedure 4.6.1.2 as follows:

Procedure 4.6.4.2

1. Regress $\hat{\eta}_t$ on $\nabla_\alpha \hat{h}_t/\hat{h}_t$ for $t = 1, 2, \ldots, T$ and save the residuals $\breve{\eta}_t$.
2. Run the vector regression

   $$\nabla_\beta \hat{\omega}_t \text{ on } \nabla_\alpha \hat{h}_t/\hat{h}_t, \quad t = 1, 2, \ldots, T$$

   and compute the residuals $\hat{r}_t$.
3. Run the ordinary least-squares regression

   $$1 \text{ on } \breve{\eta}_t \hat{r}_t, \quad t = 1, 2, \ldots, T$$

   and compute LM_T as $TR^2 = T - SSR$, where SSR is the sum of squared residuals.

Under a symmetric volatility process, the diagnostics discussed in the previous section coincide with various variance diagnostics presented in the literature.

[5] $\hat{J}_{\beta\beta'}$ denotes the $\beta\beta'$ block of the inverse of I_ν, the part of the Fisher information matrix corresponding to the parameters of the volatility model.

Testing for ARCH

If we take the extended volatility model in (4.6.1b) to be a constant σ_0^2 under the null and an ARCH(q) model under the alternative, then we have a test for heteroskedasticity in the residuals of the mean model, in the spirit of Engle's LM test. In this case, $\hat{\eta}_t^2 = \hat{\epsilon}_t^2/\hat{\sigma}^2 - 1$, where $\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}\hat{\epsilon}_t^2$ is the sample sum of squared errors, and

$$\hat{\omega}_t = 1 + \beta'\hat{z}_t$$

where $\hat{z}_t = (\hat{\epsilon}_{t-1}^2/\hat{\sigma}^2, \hat{\epsilon}_{t-2}^2/\hat{\sigma}^2, \ldots, \hat{\epsilon}_{t-q}^2/\hat{\sigma}^2)$. The standard version of the LM test, as computed from procedure 4.6.4.1, produces the Engle (1982) 's test for ARCH-type heteroskedasticity[6]. We can also derive an improved version of Engle's test that is robust to non-normal (standardised) residuals (see Bollerslev and Wooldridge (1992)). Let $\tilde{\epsilon}_t^2 = \hat{\epsilon}_t^2 - \hat{\sigma}^2$ be the squared demeaned residual at time t. The test statistic is computed as TR^2 from the regression

$$1 \text{ on } \tilde{\epsilon}_t^2\tilde{\epsilon}_{t-1}^2, \tilde{\epsilon}_t^2\tilde{\epsilon}_{t-2}^2, \ldots, \tilde{\epsilon}_t^2\tilde{\epsilon}_{t-l_\beta}^2, \quad t = 1, 2, \ldots, T$$

and follows a $\chi_{l_\beta}^2$ asymptotic distribution.

Testing for GARCH

Testing for remaining serial correlation in the squared normalised residuals when the volatility model is a GARCH(p, q) is considered by Lundbergh and Teräsvirta (1998). They also provide simulation evidence regarding the finite-sample properties of the robust and the non-robust test statistics. A common finding in all misspecification examples is that the robust version of the test proves superior to the non-robust one when errors follow a non-gaussian distribution and loses nothing in terms of power when errors happen to be normal.

Testing for sign bias, negative/positive size bias

The non-robust version of LM tests for asymmetric effects in the variance process is investigated by Engle and Ng (1993). A robustification of these tests is straightforward following procedure 4.6.4.2. In this case, $\hat{\omega}_t = \hat{z}_t$, where $\hat{z}_t$ is either a sign dummy variable or a product of a dummy with the past error.

4.7 Relevance to other approaches

At this point, it would be useful to compare the NN-GARCH modelling framework, proposed in this thesis, with other relevant approaches appeared in the literature. It has probably so far become clear that NN-GARCH specifications attempt to model the entire conditional density, by jointly parametrising the mean, the variance structure and the density of errors. This practice to model the first two moments of the distribution was very popular in econometrics in the 90's. Although the majority of GARCH models that have been proposed in the literature ignore the possibility of

[6] In this case step 1 is not necessary.

nonlinearities in mean, various nonlinear regression models, such as threshold (Li and Li (1996)) or smooth transition (Lundbergh and Teräsvirta (1998)) have appeared in combination with parametric or semi-parametric models of volatility.

In the framework of computational intelligent models, NN-GARCH models draw analogy with recurrent mixture density networks (RMDN) proposed by Schittenkopf et al. (2000). This approach to time-series analysis attempts to directly model the entire conditional probability density and shares several features with ours, mainly in that it encompasses nonlinearities in mean, nonlinear GARCH effects in variance and non-gaussian errors. However, Schittenkopf et al. (2000) discuss no particular model specification strategy for recurrent mixture density networks and the number of hidden units is chosen ad hoc. This makes unsafe the application of statistical inference, as simple heuristics do not exclude the possibility of obtaining an over-identifiable network with all the unpleasant consequences discussed in section 4.3.2. Contrary to this approach, in our thesis we pay special attention to identifying the right model complexity given the available data and avoiding over-parametrised models by means of sequential statistical tests.

Another approach that is related to our work is Donaldson and Kamstra (1997), who developed a NN-based model for conditional volatility that is linear in mean and nonlinear in variance. This was mainly intended to capture asymmetric responses of stock index volatility to past innovations. In Donaldson and Kamstra (1997)'s approach, nonlinearity enters directly into the variance part while in our approach we also consider the possibility of nonlinearity in the mean[7]. As concerns the specification of the model, Donaldson and Kamstra (1997) parameterize the mean as a linear AR process and estimate a variety of neural network-type GARCH models with growing complexity. The number of hidden units is determined by means of an information criterion. Despite its statistical foundations, this approach is more computationally demanding and makes it difficult to control the probability of producing over-identifiable models. On the contrary, in a sequential testing procedure, the percentage of selected models being over-parameterized is bounded by the *size* of the test, i.e. the probability of falsely rejecting the null hypothesis. This is an argument for the use of sequential tests in model selection instead of information criteria (see also Medeiros et al. (2006) for a discussion).

4.8 Summary and discussion

In this chapter, we showed how artificial neural networks and GARCH parametrisations can be combined into a flexible modelling framework that can accommodate a variety of features observed in financial time-series: nonlinearities in mean, nonlinear GARCH effects in variance and possibly non-gaussian errors. By jointly modelling the conditional mean and volatility of the data-generating process, we extend the scope of NNs from function approximation to *density forecasting* tasks and bring NNs right in the centre of current econometric research. Besides detecting a functional relationships between target and explanatory variables, with our combined

[7] Still, our methodological framework can be easily extended to incorporate a neural network in variance. See the discussion at the end of the thesis.

NN-GARCH model we can now capture other interesting features of the conditional distribution.

Our goal was also to provide a complete model-building cycle for the family of NN-GARCH specifications, which comprises all stages of econometric modelling: specification, estimation and evaluation. As with every flexible class of models, the issue of carefully selecting the final specification becomes of paramount importance. The analysis of the statistical properties of NN-GARCH models revealed that any combination of neural networks and GARCH parametrisation is not guaranteed to be successful unless special attention is paid to the specification of the mean equation. In fact, if nonlinear dependencies in data are due to GARCH effects extra hidden neurons in the nonlinear model are redundant leading to inconsistency of parameter estimates and possibly poor out-of-sample performance.

The specification of the NN-GARCH model follows a “bottom-up” procedure which avoids many statistical and numerical problems arising from non-identifiability of parameters. The specification of the mean structure is based on sequential Lagrange Multiplier (LM) tests of neglected nonlinearity that are *robustified*, i.e. preserve their asymptotic validity in the presence of heteroskedasticity. One of the testing procedure that we propose in this thesis for detecting the structure of the neural network gives the opportunity to the researcher to incorporate information from the variance structure of the distribution of errors while testing for extra neurons in the data-generating process.

Based on the quasi maximum likelihood theory, we device in-sample robustified diagnostics on the estimated model that investigate whether the derived model is a faithful approximation to the data-generating process. The distinguishing feature of these diagnostics is that they lead to valid inference regarding structural misspecification despite the fact that the distributional assumptions made by the model may not be correct (i.e. the empirical density of standardised errors is fat-tailed or asymmetric). We particularly consider LM tests for remaining autocorrelation in mean, remaining autocorrelation in variance, asymmetric variance effects and nonlinearity. In general, these tests are simple and inexpensive to construct, as the model has already been estimated under the null, and only require the computation of first derivatives and a set of auxiliary regressions to determine whether the residuals (or the standardised residuals) can be further explained by the conjectured alternative hypothesis. This is a big advantage for highly nonlinear specifications, where the numerical estimation of parameters becomes an issue. The empirical performance of some of the testing procedures discussed above is investigated in the next chapter by means of Monte-Carlo simulation.

Chapter 5

Monte Carlo simulation studies

5.1 Introduction

The testing procedures discussed in chapter 4 for detecting hidden nonlinearity in data, asymmetric effects in variance or checking the adequacy of a NN-GARCH model rest on *asymptotic*, large-sample size, theory. Therefore, it would be essential from a practical point of view to investigate their performance in sample sizes relative to financial applications, taking into account common statistical features of these application data (heteroskedasticity, non-normality). Section 5.2 details a Monte-Carlo simulation study on the sequential LM nonlinearity testing framework for detecting the number of hidden units and section 5.3 presents a similar exercise related to the diagnostic checking on a neural network regression model. Experimental results are discussed in section 5.4 which concludes the chapter.

5.2 Detecting the number of hidden units under heteroskedasticity in errors

Monte Carlo simulation experiments presented in Anders and Korn (1999); Medeiros et al. (2006); Teräsvirta et al. (1993) investigate the properties of the standard neural network nonlinearity test assuming a neural network regression model, in which errors follow a normal distribution with constant variance. In financial applications, however, it is highly possible that a pure NN model is not enough to capture the statistical features of the underlying probability-generating model (e.g. heteroskedasticity in errors). In this section we present a variety of simulations that examine the properties of the standard sequential LM testing procedure under ARCH-type heteroskedasticity and non-normality of errors and compare its performance with the two robustified sequential procedures presented in section 4.5.1. All subsequent experiments were conducted on Matlab$^{©}$ version 7.

5.2.1 Design of experiments

For the empirical investigation of the size and power of LM tests, we consider two data-generating processes for the conditional mean whose exact specification is given in table 5.1, Panel A. Model 0 is a typical linear autoregressive model with one lag-dependent variable and model 1 is an additive AR(1)-NN process with one hidden neuron. The set of conditioning variables in this case is (y_{t-1}) for model 0 and (y_{t-1}, y_{t-3}) for model 1. Errors are assumed to follow the general specification

$$\epsilon_t = \zeta_t \sqrt{h_t}$$

where $\{\zeta_t\}$ are drawn from a N(0,1) or a Student $t(\nu)$ distribution with ν degrees of freedom. In the second case, ζ_t's are normalised to have unit sample variance. As discussed in the section 4.5.1, the validity of LM nonlinearity tests does not depend on the normality of errors and this seems a big advantage in the light of empirical findings suggesting that the empirical distribution of economic time-series is heavy-tailed and leptokurtic. It is therefore interesting to investigate the properties of these tests when the distribution of residuals deviates from the normal prototype. The density function of a student distribution is also symmetric around zero and the additional factor ν controls the thickness of the distribution tails. In our experiments we set $\nu = 5$, which is also a common choice in other simulation studies. The use of a student distribution as a model for the error process is motivated by numerous empirical surveys supporting that the distribution of financial returns is often more heavy-tailed than would be predicted by an heteroskedastic model (Bollerslev (1987)).

Panel A: Expectational models

Model 0	$y_t = 0.001 + 0.6y_{t-1} + \epsilon_t$
Model 1	$y_t = 0.25 + 0.45y_{t-1} + 0.5\,F[2.5(y_{t-1} - 1.3y_{t-3} - 0.1)] + \epsilon_t$

Panel B: Volatility models

VGP0	$h_t = 10^{-3}$
VGP1	$h_t = 10^{-3} + 0.85h_{t-1} + 0.05\epsilon_{t-1}^2$
VGP2	$h_t = 10^{-3} + 0.90h_{t-1} + 0.07\epsilon_{t-1}^2$
VGP3	$\log(h_t) = -0.008 + 0.93\log(h_{t-1}) + 0.17(\lvert u_{t-1}\rvert - E) - 0.085u_{t-1}$

Table 5.1: Specification of the models used in simulation.

The empirical performance of tests is investigated under a variety of volatility processes, presented in table 5.1, Panel B. The first process (VGP0) imposes constant variance, while VGP1 and VGP2 assume that the conditional variance follows a GARCH(1,1) model. The difference between VGP2 and VGP1 is that volatility shocks are on average more persistent for the former than the latter. The last model for the conditional volatility is Nelson (1990)'s exponential GARCH (EGARCH), which is also very popular in financial time-series analysis. In this simulation exercise, motivated by empirical evidence presented in section 2.3, we assume that a negative ϵ_{t-1} has on average a greater impact on h_t than a positive one and hence assign a negative value to the coefficient of u_{t-1}.

For each combination of mean and variance models, we generated 1000 sample paths excluding the first 500 observations to eliminate the effect of initial values. The sequential testing procedure described in section 4.5.1 was then applied to decide the number of hidden neurons in the mean equation. In order to investigate the effects of heteroskedasticity in the specification of the mean part, we implemented the above procedure using all versions of nonlinearity tests: the standard LM (S-LM), the robustified LM (RB-LM) and the robustified LM test (RBV-LM) with volatility estimates obtained from a GARCH(1,1) model (independently of the true volatility generating process). In all experiments, the correct set of variables was given to the model a priori as a way to obtain an idea of the behaviour of various statistics free from the effects of an incorrectly selected set of variables. Perhaps, the extent to which each test is able to determine the appropriate set of variables could be the objective of another simulation study. SBIC was used to determine the set of variables composing the linear model.

5.2.2 Simulation results

Tables 5.2 to 5.5 show the empirical test performance under the hypothesis that the true expectational model is linear. We report results for three initial significance levels (1%, 5% and 10%) and two sample size of 700 and 2000 observations. Each cell shows the percentage of paths for which the corresponding testing procedure indicated a number of hidden neurons equal to the value of $\hat{h}$ given in the second column. A general conclusion drawn from the first set of experiments is that a larger initial significance level α unavoidably leads to more false rejections of the linearity hypothesis. When errors are homogeneously distributed (VGP0) the empirical rejection rate of all tests closely follows the nominal type I error. However, in the presence of heteroskedasticity the S-LM test tends to overreject the correct hypothesis of linearity. The size distortions are especially dramatic for the processes characterised by a strong or an asymmetric ARCH component, like VGP2 and VGP3, or a heavy-tailed distribution of errors (t distribution). The situation does seem to improve with an increasing number of observations. Note e.g. from tables 5.4 and 5.5 that at 10% significance 2000 observations and t-distributed errors the probability that the S-LM test wrongly indicate nonlinearity in mean is more than 30% under a persistent GARCH or an EGARCH volatility process. It seems that under strong heteroskedasticity the test misinterprets changes in the levels of variance as neglected nonlinearity in the residuals of the linear model and hence indicates additional neurons to capture these effects. In this way various features of the error component are transferred into the mean model and since these features are not due to systematic movements in y_t the resulting NN is expected to have poor performance on unseen data. Generally, experiments on the S-LM test give a cautionary remark against the use of non-heteroskedasticity-robust statistical procedures in the determination of the NN architecture. On the other hand, simulations show that sequential testing procedures based on robustified versions of the LM test are efficient in controlling the empirical type I error of the testing procedure under various forms of heteroskedasticity in errors. Of particular interest are the results for the RBV-LM test, whose main difference with RB-LM is that the researcher chooses to explicitly model heteroskedasticity rather to ignore it. Tables 5.2 and 5.5 represent two cases where the researcher has nevertheless incorrectly specified the form of heteroskedasticity. In table 5.2 the researcher erroneously employs an heteroskedastic model when the conditional variance of errors is in fact constant. Although in practical applications graphical inspection of the times-series or some preliminary data analysis would probably prevent the analyst from doing so, we see that the application of an RBV-LM test with GARCH(1,1) volatility estimates does not distort the empirical size. The same is true for the experiments reported in table 5.5, where in this case the GARCH model employed by the researcher misses an important feature of volatility dynamics, i.e. the asymmetric response of the conditional variance to past errors.

Tables 5.6 to 5.9 show the empirical rejection frequencies when the expectational model includes one neuron in the hidden layer. This part of simulation experiments allows us to investigate the ability of tests (or else the power) to detect hidden nonlinearity when various types of heteroskedasticity may also be present in the data-generating process. Table 5.6 shows that under homogeneously distributed errors all tests have similar power performance and tend to underreject the hypothesis of nonlinearity indicating a linear expectational process. Still, the percentage of rejections increases with the sample size, which is somewhat expected as nonlinearity becomes more apparent in large samples. Tables 5.7 to 5.9 show that all tests tend to detect more often nonlinear structure in mean under heteroskedasticity and a fat-tailed t distribution, the rejection rate being increased with the persistence of the volatility process. A possible explanation for this finding stems from the fact that an increase in the short-term volatility and the kurtosis of the unconditional distribution makes more likely the appearance of extreme movements of y_t in a specific sample path. As the overall sample variability of y_t is increased, the hidden neuron is activated more often making thus nonlinearity more apparent in the data-generating process. A common finding in this part of experiments is that the standard version of the LM test is adversely affected by the presence of autoregressive heteroskedasticity. The stronger is heteroskedasticity the more likely is S-LM to produce an excessive number of neurons. The RB-LM test is much well behaved against the adverse effects of heteroskedasticity. However it tends to be overly conservative especially at low significance levels, although its power performance is improved when more observations are used. Note also that the tendency of the RB-LM test to be conservative is slightly increased when $t(5)$ random variates are used in the simulation. Similar findings have also been reported in simulation studies that employ the RB-LM test to detect nonlinearity in other nonlinear regression models (see e.g. Becker and Hurn (2006); Lundbergh and Teräsvirta (1998)). The robustified LM test that uses information from the variance structure seems more efficient in isolating nonlinearity in mean from heteroskedasticity and also has a good control over the empirical type I error. The power of RBV-LM rapidly picks up with the persistence of the volatility process (table 5.8) and also seems quite satisfactory in the case of an EGARCH volatility process (table 5.9). The out-performance of the RBV-LM test with GARCH(1,1) volatility estimates can be attributed to the fact that a GARCH model is a better approximation to an EGARCH volatility-generating process than a model of constant variance, implicitly assumed by RB-LM. Hence, even if the volatility model is not the primary concern of the researcher, it pays to put some effort on deriving a good approximation to the underlying volatility structure.

Model 0, VGP0

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
700 observations										
Normal	$\hat{h} = 0$	0.995	0.965	0.920	0.992	0.953	0.919	0.994	0.956	0.915
	$\hat{h} = 1$	0.005	0.035	0.079	0.008	0.047	0.081	0.006	0.043	0.084
	$\hat{h} = 2$	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.001
Student	$\hat{h} = 0$	0.992	0.961	0.919	0.996	0.970	0.928	0.995	0.967	0.924
	$\hat{h} = 1$	0.008	0.039	0.081	0.004	0.029	0.071	0.005	0.033	0.074
	$\hat{h} = 2$	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.002
2000 observations										
Normal	$\hat{h} = 0$	0.991	0.952	0.917	0.991	0.959	0.908	0.993	0.960	0.911
	$\hat{h} = 1$	0.009	0.047	0.079	0.009	0.041	0.090	0.007	0.039	0.086
	$\hat{h} = 2$	0.000	0.001	0.004	0.000	0.000	0.002	0.000	0.001	0.003
Student	$\hat{h} = 0$	0.989	0.959	0.930	0.994	0.958	0.912	0.994	0.964	0.930
	$\hat{h} = 1$	0.011	0.041	0.070	0.006	0.041	0.087	0.006	0.036	0.069
	$\hat{h} = 2$	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.001

TABLE 5.2: The extend of overfitting in the case of the expectational model 0 and the variance-generating process VGP0. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 0, VGP1

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
700 observations										
Normal	$\hat{h} = 0$	0.979	0.928	0.857	0.991	0.956	0.903	0.990	0.950	0.896
	$\hat{h} = 1$	0.021	0.069	0.140	0.009	0.042	0.095	0.010	0.049	0.103
	$\hat{h} = 2$	0.000	0.003	0.003	0.000	0.002	0.002	0.000	0.001	0.001
Student	$\hat{h} = 0$	0.942	0.865	0.805	0.996	0.982	0.940	0.995	0.975	0.930
	$\hat{h} = 1$	0.056	0.132	0.189	0.004	0.018	0.060	0.005	0.025	0.070
	$\hat{h} = 2$	0.002	0.003	0.006	0.000	0.000	0.000	0.000	0.000	0.000
2000 observations										
Normal	$\hat{h} = 0$	0.945	0.821	0.686	0.998	0.972	0.923	0.997	0.970	0.924
	$\hat{h} = 1$	0.054	0.169	0.294	0.002	0.027	0.074	0.003	0.030	0.076
	$\hat{h} = 2$	0.001	0.010	0.019	0.000	0.001	0.003	0.000	0.000	0.000
Student	$\hat{h} = 0$	0.909	0.820	0.742	0.996	0.964	0.925	0.996	0.963	0.916
	$\hat{h} = 1$	0.079	0.164	0.236	0.004	0.036	0.075	0.004	0.037	0.082
	$\hat{h} = 2$	0.012	0.016	0.022	0.000	0.000	0.000	0.000	0.000	0.002

TABLE 5.3: The extend of overfitting in the case of the expectational model 0 and the variance-generating process VGP1. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 0, VGP2

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
					700 observations					
Normal	$\hat{h} = 0$	0.964	0.876	0.810	0.997	0.968	0.920	0.993	0.959	0.910
	$\hat{h} = 1$	0.035	0.116	0.177	0.003	0.032	0.080	0.007	0.041	0.087
	$\hat{h} = 2$	0.001	0.008	0.013	0.000	0.000	0.000	0.000	0.000	0.003
Student	$\hat{h} = 0$	0.894	0.810	0.739	0.991	0.957	0.908	0.992	0.953	0.906
	$\hat{h} = 1$	0.102	0.174	0.233	0.009	0.043	0.088	0.008	0.044	0.090
	$\hat{h} \geq 2$	0.004	0.016	0.028	0.000	0.000	0.004	0.000	0.003	0.004
					2000 observations					
Normal	$\hat{h} = 0$	0.930	0.840	0.774	0.987	0.953	0.898	0.991	0.947	0.900
	$\hat{h} = 1$	0.067	0.153	0.209	0.013	0.045	0.100	0.009	0.053	0.099
	$\hat{h} = 2$	0.003	0.007	0.017	0.000	0.002	0.002	0.000	0.000	0.001
Student	$\hat{h} = 0$	0.834	0.722	0.641	0.996	0.969	0.926	0.992	0.939	0.901
	$\hat{h} = 1$	0.148	0.245	0.308	0.004	0.031	0.074	0.008	0.060	0.095
	$\hat{h} = 2$	0.017	0.030	0.048	0.000	0.000	0.000	0.000	0.001	0.004
	$\hat{h} = 3$	0.001	0.003	0.003	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.4: The extend of overfitting in the case of the expectational model 0 and the variance-generating process VGP2. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 0, VGP3

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
					700 observations					
Normal	$\hat{h} = 0$	0.951	0.862	0.785	0.990	0.962	0.915	0.989	0.952	0.908
	$\hat{h} = 1$	0.049	0.135	0.204	0.010	0.038	0.080	0.011	0.046	0.088
	$\hat{h} = 2$	0.000	0.003	0.011	0.000	0.000	0.005	0.000	0.002	0.004
Student	$\hat{h} = 0$	0.914	0.821	0.750	0.995	0.965	0.916	0.991	0.961	0.901
	$\hat{h} = 1$	0.082	0.172	0.237	0.004	0.034	0.082	0.009	0.038	0.093
	$\hat{h} = 2$	0.004	0.007	0.013	0.001	0.001	0.002	0.000	0.001	0.006
					2000 observations					
Normal	$\hat{h} = 0$	0.924	0.825	0.760	0.994	0.955	0.910	0.990	0.947	0.902
	$\hat{h} = 1$	0.076	0.173	0.236	0.006	0.045	0.089	0.010	0.051	0.093
	$\hat{h} = 2$	0.000	0.002	0.004	0.000	0.000	0.001	0.000	0.002	0.005
Student	$\hat{h} = 0$	0.849	0.730	0.653	0.994	0.953	0.906	0.991	0.952	0.899
	$\hat{h} = 1$	0.145	0.253	0.318	0.006	0.046	0.090	0.009	0.046	0.098
	$\hat{h} = 2$	0.006	0.017	0.027	0.000	0.001	0.004	0.000	0.002	0.003

Table 5.5: The extend of overfitting in the case of the expectational model 1 and the variance-generating process VGP3. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 1, VGP0

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
700 observations										
Normal	$\hat{h} = 0$	0.972	0.865	0.774	0.977	0.888	0.769	0.979	0.876	0.770
	$\hat{h} = 1$	0.028	0.133	0.219	0.023	0.108	0.222	0.021	0.121	0.220
	$\hat{h} \geq 2$	0.000	0.002	0.007	0.000	0.004	0.009	0.000	0.003	0.010
Student	$\hat{h} = 0$	0.962	0.861	0.758	0.983	0.897	0.775	0.984	0.901	0.790
	$\hat{h} = 1$	0.037	0.136	0.227	0.017	0.102	0.220	0.016	0.098	0.204
	$\hat{h} \geq 2$	0.001	0.003	0.015	0.000	0.001	0.005	0.000	0.001	0.006
2000 observations										
Normal	$\hat{h} = 0$	0.951	0.827	0.691	0.958	0.827	0.707	0.960	0.835	0.707
	$\hat{h} = 1$	0.048	0.169	0.300	0.041	0.170	0.279	0.040	0.164	0.285
	$\hat{h} \geq 2$	0.001	0.004	0.009	0.001	0.003	0.013	0.000	0.001	0.008
Student	$\hat{h} = 0$	0.943	0.812	0.672	0.961	0.848	0.706	0.967	0.869	0.749
	$\hat{h} = 1$	0.057	0.186	0.322	0.039	0.147	0.284	0.033	0.127	0.243
	$\hat{h} \geq 2$	0.000	0.002	0.006	0.000	0.005	0.010	0.000	0.004	0.008

Table 5.6: The extend of overfitting in the case of the expectational model 1 and the variance-generating process VGP0. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 1, VGP1

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
700 observations										
Normal	$\hat{h} = 0$	0.840	0.635	0.509	0.940	0.788	0.646	0.936	0.763	0.633
	$\hat{h} = 1$	0.157	0.348	0.452	0.060	0.207	0.335	0.064	0.227	0.342
	$\hat{h} = 2$	0.003	0.017	0.039	0.000	0.005	0.019	0.000	0.010	0.025
Student	$\hat{h} = 0$	0.720	0.515	0.382	0.965	0.797	0.648	0.941	0.763	0.615
	$\hat{h} = 1$	0.263	0.430	0.534	0.035	0.200	0.338	0.059	0.233	0.370
	$\hat{h} = 2$	0.014	0.050	0.078	0.000	0.003	0.014	0.000	0.004	0.015
	$\hat{h} \geq 3$	0.003	0.005	0.006	0.000	0.000	0.000	0.000	0.000	0.000
2000 observations										
Normal	$\hat{h} = 0$	0.553	0.313	0.204	0.777	0.481	0.332	0.747	0.433	0.303
	$\hat{h} = 1$	0.435	0.643	0.702	0.223	0.508	0.641	0.253	0.553	0.648
	$\hat{h} = 2$	0.012	0.043	0.094	0.000	0.011	0.027	0.000	0.014	0.049
	$\hat{h} = 3$	0.000	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000
Student	$\hat{h} = 0$	0.331	0.150	0.105	0.813	0.501	0.348	0.730	0.411	0.275
	$\hat{h} = 1$	0.611	0.718	0.698	0.183	0.487	0.609	0.264	0.559	0.667
	$\hat{h} = 2$	0.054	0.121	0.182	0.004	0.012	0.043	0.006	0.030	0.058
	$\hat{h} \geq 3$	0.003	0.010	0.012	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.7: The extend of overfitting in the case of the expectational model 1 and the variance-generating process VGP1. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 1, VGP2

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
700 observations										
Normal	$\hat{h} = 0$	0.620	0.393	0.267	0.916	0.713	0.554	0.875	0.638	0.495
	$\hat{h} = 1$	0.353	0.533	0.621	0.081	0.275	0.423	0.123	0.347	0.479
	$\hat{h} = 2$	0.026	0.070	0.104	0.003	0.012	0.023	0.002	0.015	0.025
	$\hat{h} \geq 3$	0.001	0.004	0.008	0.000	0.000	0.000	0.000	0.000	0.001
Student	$\hat{h} = 0$	0.458	0.281	0.196	0.936	0.740	0.580	0.871	0.634	0.466
	$\hat{h} = 1$	0.467	0.579	0.607	0.064	0.257	0.407	0.126	0.349	0.498
	$\hat{h} = 2$	0.061	0.117	0.166	0.000	0.003	0.013	0.003	0.017	0.036
	$\hat{h} \geq 3$	0.014	0.023	0.031	0.000	0.000	0.000	0.000	0.000	0.000
2000 observations										
Normal	$\hat{h} = 0$	0.171	0.051	0.028	0.616	0.312	0.188	0.425	0.180	0.096
	$\hat{h} = 1$	0.754	0.777	0.728	0.381	0.657	0.748	0.566	0.777	0.824
	$\hat{h} = 2$	0.071	0.159	0.225	0.003	0.031	0.064	0.009	0.043	0.080
	$\hat{h} \geq 3$	0.004	0.012	0.019	0.000	0.000	0.000	0.000	0.000	0.000
Student	$\hat{h} = 0$	0.089	0.030	0.018	0.722	0.401	0.264	0.420	0.164	0.086
	$\hat{h} = 1$	0.701	0.629	0.544	0.274	0.575	0.673	0.558	0.782	0.808
	$\hat{h} = 2$	0.173	0.278	0.356	0.004	0.023	0.061	0.022	0.054	0.104
	$\hat{h} \geq 3$	0.037	0.063	0.082	0.000	0.001	0.002	0.000	0.000	0.002

Table 5.8: The extend of overfitting in the case of the expectational model 1 and the variance-generating process VGP2. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

Model 1, VGP3

Test type		**S-LM**			**RB-LM**			**RBV-LM**		
Significance level		1%	5%	10%	1%	5%	10%	1%	5%	10%
700 observations										
Normal	$\hat{h} = 0$	0.476	0.253	0.158	0.805	0.514	0.369	0.628	0.367	0.246
	$\hat{h} = 1$	0.506	0.681	0.727	0.194	0.475	0.611	0.371	0.612	0.714
	$\hat{h} = 2$	0.018	0.063	0.105	0.001	0.011	0.020	0.001	0.021	0.040
	$\hat{h} = 3$	0.000	0.003	0.010	0.000	0.000	0.000	0.000	0.000	0.000
Student	$\hat{h} = 0$	0.512	0.310	0.213	0.901	0.682	0.530	0.761	0.468	0.339
	$\hat{h} = 1$	0.444	0.581	0.632	0.099	0.312	0.456	0.237	0.515	0.620
	$\hat{h} = 2$	0.042	0.102	0.143	0.000	0.006	0.014	0.002	0.017	0.039
	$\hat{h} = 3$	0.001	0.007	0.012	0.000	0.000	0.000	0.000	0.000	0.002
2000 observations										
Normal	$\hat{h} = 0$	0.401	0.197	0.137	0.813	0.559	0.406	0.587	0.334	0.217
	$\hat{h} = 1$	0.541	0.659	0.653	0.182	0.426	0.558	0.407	0.628	0.705
	$\hat{h} = 2$	0.057	0.139	0.199	0.005	0.015	0.035	0.006	0.038	0.078
	$\hat{h} = 3$	0.001	0.005	0.011	0.000	0.000	0.001	0.000	0.000	0.000
Student	$\hat{h} = 0$	0.315	0.159	0.108	0.874	0.622	0.444	0.569	0.312	0.203
	$\hat{h} = 1$	0.564	0.610	0.601	0.124	0.358	0.518	0.416	0.622	0.690
	$\hat{h} = 2$	0.109	0.202	0.251	0.002	0.020	0.036	0.015	0.065	0.106
	$\hat{h} = 3$	0.012	0.029	0.040	0.000	0.000	0.002	0.000	0.001	0.001

Table 5.9: The extend of overfitting in the case of the expectational model 1 and the variance-generating process VGP3. We report results for all versions of the (non)linearity LM-based test. Each cell corresponds to the percentage of replications for which the number of neurons indicated by the test was equal to $\hat{h}$.

5.3 Testing the adequacy of a neural network regression model

Assume that the researcher is solely interested in conditional mean relations and does not want to explicitly model the structure of higher moments of the distribution (e.g. the variance). However, he would like to know whether the employed architecture of a NN adequately captures the main elements of the conditional expectation. This is a common problem one is typically faced with when applying a NN regression model in real data. It is common sense that if a regression model is of adequate structure it should not leave interesting features of the conditional expectation, such as autocorrelation, nonlinear dependencies, or omitted variables in the residuals. Therefore, most diagnostic procedures on a NN regression model are concentrated on testing the residuals for "strong" properties.

In the great majority of NN applications, authors employ heuristic procedures, such as sample autocorrelation diagrams or normality plots, to judge the randomness of the residuals of the model. Zapranis and Refenes (1999) were among the first to propose an integrated framework for neural network evaluation, which is essentially an application of the Box-Jenkins diagnostic checking principles for linear ARMA models (Box et al. (1994)). However, Medeiros et al. (2006) noticed that many popular portmanteau tests, such as the Ljung-Box or Box-Pierce, are inapplicable in the neural network case, because their asymptotic null distribution is unknown if the test is based on the estimated residuals of a neural network model. Therefore, they derived a series of Lagrange Multiplier (LM) test statistics specially designed for the diagnostic checking of a neural network regression model. NN specification testing based on maximum-likelihood statistics is also discussed in Kuan and White (1994).

In this chapter, we re-examine the issue of diagnostic checking on a neural network regression model under the existence of strong properties in the distribution of errors. We present the standard LM testing framework of Medeiros et al. (2006), suitable for homogeneously distributed errors, and we propose two new diagnostic procedures for neural network models that apply to heteroskedastic errors. These essentially employ the Wooldridge (1991)'s modifications of the classical LM statistic, also considered in previous sections. After presenting the general testing framework in section 5.3.1, in section 5.3.2 we study some special cases of model adequacy tests, such as serial correlation in errors and omitted variables. The finite-sample performance of the LM tests is investigated in sections 5.3.3 and 5.3.4 by means of a Monte Carlo study, in the special case of detecting error autocorrelation (of a fixed order) in a NN model when GARCH heteroskedasticity is also present.

5.3.1 The general framework

A general diagnostic framework for NN-GARCH models can be based on the following additive extension of (4.3.1a):

$$y_t = \phi' \bar{x}_t + f(x_t; \theta) + g(z_t; \xi) + \epsilon_t^* \tag{5.3.1}$$

where $g(z_t; \xi)$ is continuous and twice differentiable for all $\xi \in \mathbb{R}^l$ and (almost) everywhere in the corresponding sample space. Assume without loss of generality that

$g(z_t;0) = 0$. If the restricted model is adequate then $g(z_t;0)$ should be zero and $\epsilon_t^* = \epsilon_t$. Hence a general hypothesis of adequacy can be stated as

$$H_0 : \xi = 0$$

Failing to reject this hypothesis, suggests that errors contain additional structure, which is generally a problem if the NN model is intended for statistical inference or forecasting.

The diagnostic testing procedure proposed by Medeiros et al. (2006) is essentially an application of an LM test to the parameter vector ξ. The derivation of test statistics was built on early works of Eitrheim and Teräsvirta (1996) and Teräsvirta (1994). Let

$$\hat{\epsilon}_t = y_t - \hat{\phi}'\bar{x}_t - f(x_t;\hat{\theta})$$

where $\hat{\phi}$ and $\hat{\theta}$ are the least-squares estimates of the restricted model. Let also $\nabla\hat{f}_t$, $\nabla\hat{g}_t$ denote the gradients of $f(x_t;\theta)$, $g(x_t;\xi)$ evaluated at $\theta = \hat{\theta}$ and $\xi = 0$. The standard LM test is carried out in the following three steps:

Procedure 5.3.1.1

1. Regress $\hat{\epsilon}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$ and compute the new residual vector $\hat{\varepsilon}_t$
2. Regress $\hat{\varepsilon}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$, $\nabla\hat{g}_t$ and compute the sum of squared residuals $SSR_2 = T^{-1}\sum_{t=1}^{T} u_t^2$
3. Compute the test statistic as $LM_T = TR^2$, where R^2 is the coefficient of determination of the regression.

The asymptotic distribution of LM_T follows a χ^2 distribution with l degrees of freedom. However, the validity of the standard LM test depends on the correct specification of the volatility process of errors. Generally, the LM test statistic does not follow a chi-square asymptotic distribution in the presence of ARCH or other forms of heteroskedasticity. Hence using critical values from a chi-square distribution may lead to wrong inference regarding the presence of additional structure in the mean of the distribution. A robustified LM test that ignores the effect of heteroskedasticity can be computed from the following procedure:

Procedure 5.3.1.2

1. Regress $\hat{\epsilon}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$ and compute the residuals $\hat{\varepsilon}_t$
2. Regress $\nabla\hat{g}_t$ on $\bar{x}_t$, $\nabla\hat{f}_t$ and take the $l \times 1$ residuals vector $\hat{u}_t$
3. Run the regression 1 on $\hat{\varepsilon}_t\hat{u}_t$ and compute the sum of squared residuals SSR_1
4. Compute the statistic as $LM_T = T - SSR_1$

If one can also obtain an estimate of the conditional variance $\hat{h}_t$, it is possible to carry out the above testing procedure by using standardised quantities in place of the corresponding quantities in 5.3.1.2. The robustified LM statistics follow an asymptotic χ^2_l distribution under heterogeneously distributed errors and are also asymptotically efficient under homoskedastic errors, i.e. no efficiency is lost in carrying out the test.

In his discussion on the properties of robustified tests, Wooldridge (1991) assumes that the parameters $\mu = (\phi',\, \theta')'$ of the NN model are estimated using weighted least-squares, where model's error ϵ_t is weighted by $1\sqrt{\hat{h}_t}$. However, estimators obtained by other estimation procedures can be also used in the computation of the test without affecting its asymptotic size. The key feature of this approach is that tests are always asymptotically efficient, *no matter* which estimator is used for the variance or, more importantly, the mean (subject of course that it is $\sqrt{T}$-consistent for the mean parameters).

Although general statements can be made regarding asymptotic efficiency, one has little knowledge as to whether these properties carry over to finite samples. Using a more efficient estimation procedures for the parameters may lead to better power performance of the test in different sample sizes. This issue is further discussed in section 5.3.3. Among the many alternatives, we examine two possible ways to carry out the test. The first is the one we followed in nonlinearity tests and amounts to estimating the parameters of the NN model using nonlinear least-squares, computing the residuals $\hat{\epsilon}_t$ and estimating a GARCH(p, q) model using maximum likelihood. The other possibility is to *jointly* estimate the parameters of the mean and variance models by forming the full maximum likelihood (FML) function:

$$l(\mu, \alpha) = (1/T) \sum_{t=1}^{T} l_t(\mu, \alpha)$$

where

$$l_t(\mu, \alpha) = -0.5 \log(2\pi h_t) - 0.5\, \epsilon_t^2 / h_t$$

FML is the most efficient estimation procedure whenever the parameters of the NN model cannot be estimated without reference to the parameters of the variance model, and vice versa. This case emerges when one combines a NN with a GARCH model for volatility, as the conditional variance h_t is implicitly a function of the parameters of the mean model. However, a cautionary remark is in order here about the use of a full maximum likelihood estimation method. As we pointed out in section 4.3.2, where we discussed the conditions for consistency of the MLE of a NN-GARCH model, when jointly estimating the mean and the variance parameters of a model, there is a danger that the mean estimator is not consistent unless the variance model is correctly specified for the true conditional variance. This possibility does not theoretically justify the use of FML in the application of the RBV-LM test, however it would be insightful to see the practical consequences of this choice. Section 5.3.4 sheds light on this issue by means of simulation.

5.3.2 Adequacy tests

In this section, we apply the general testing framework presented above to derive two special model adequacy tests: a test for serial correlation in the residuals and a test for omitted variables.

Testing for serial correlation in the residuals

Let us assume that the errors of the a NN model follow an autoregressive process of order l,

$$\epsilon_t = \xi' z_t + \epsilon_t^*$$

where $z_t = (\epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-l})'$ and ϵ_t^* is an iid error process. To test serial independence in the residuals, we set the additive extension $g(z_t; \xi)$ equal to $\xi' z_t$ and take the null to be $H_0 : \xi = 0$ against the alternative hypothesis $H_1 : \xi \neq 0$. The gradient $\nabla \hat{g}_t$ of g_t evaluated at null is simply $\hat{z}_t$, where

$$\hat{z}_t = (\hat{\epsilon}_{t-1}, \hat{\epsilon}_{t-2}, ..., \hat{\epsilon}_{t-l})'$$

for $t = l + 1, l + 2, ..., T$.

Testing for omitted variables

The general framework presented above can be also used to test other specification errors in the NN model. One important case is the omission of other exogenous variables not included in x_t, though they might influence the target variable y_t. Similarly to the test presented in the previous section, we can test the hypothesis of omitted variables in the linear part of the model, by stating the alternative as $g(z_t; \xi) = \xi' z_t$, with z_t being the set of additional exogenous regressors. Note that in this case $\nabla \hat{g}_t$ is simply z_t. Testing for additional nonlinearity in the NN model can be also treated as an omitted variable test, where in this case z_t is a sufficiently high-order polynomial of x_t that is thought to well approximate the remaining nonlinear structure (see section 4.5.1 for more details).

5.3.3 Design of experiments

The performance of the LM diagnostic procedures proposed above was investigated in the special case of detecting error autocorrelation (of a fixed order) in a NN model when GARCH heteroskedasticity is also present in errors. It is important to note that in relevant simulation exercises, Medeiros and Veiga (2003); Medeiros et al. (2006) provide evidence regarding the empirical performance of the standard LM test under the assumption that the disturbances of the neural network are normally distributed with constant variance. This simulation model however does not adequately resemble the statistical properties of most economic and financial data, whose empirical distribution is characterised by heavy-tails and changing variance (heteroskedasticity). To evaluate the effectiveness of the procedures proposed in this section on a more realistic basis, we compute the empirical size and power of LM tests under GARCH-type heteroskedasticity and a fatter-tailed distribution.

Simulations are based on the following single-hidden-neuron autoregressive NN process[1]:

$$y_t = 0.7y_{t-1} - 0.9F(2.2y_{t-1} + 1.4y_{t-2} - 5.1) + \epsilon_t$$

with $y_{-1} = y_0 = 0$. Errors ϵ_t follow the general specification

$$\epsilon_t = u_t\sqrt{h_t}$$

where $u_t = \rho u_{t-2} + \zeta_t$, h_t is the conditional variance and $\{\zeta_t\}$ is a sequence of iid random variables drawn from a N(0,1) or a t distribution with 5 degrees of freedom. In the second case, ζ_t's were normalised to have unit variance. Varying the value of ρ allows us to investigate the size and the power of tests in detecting serial correlation of order 2 in the residuals. To investigate the performance of the LM tests under conditional heteroskedasticity, we assume that h_t follows a GARCH(1,1) model

$$h_t = 0.001 + 0.9h_{t-1} + 0.08\epsilon_{t-1}^2$$

or an EGARCH(1,1) model

$$\log(h_t) = -0.03 + 0.5\log(h_{t-1}) + 0.25(|u_{t-1}| - E) - 0.15u_{t-1}$$

The initial values of h_{t-1}, ϵ_{t-1}^2 are set equal to the sample mean of the squared innovations. The first model represents a symmetric volatility process, with persistence factor 0.9+0.08=0.98 very close to unity. This means that one average short-term disturbances of h_t die out at a slow rate. The EGARCH model incorporates asymmetric responses to volatility. In the specification presented above, negative ϵ_{t-1}'s have a greater impact on h_t than positive ones due to the negative value of the leverage coefficient (-0.15).

We generated 5000 sample paths from each data-generating process and removed the first 500 observations to eliminate the effect of initial values. For each replication, we conduct the following three versions of the LM serial correlation test: the standard LM (S-LM), the robustified LM (RB-LM) and the robustified LM test with volatility estimates (RBV-LM). Those were obtained by fitting a GARCH(1,1) model, independently of the true volatility generating process. The parameters of the NN and the GARCH(1,1) model are estimated by two procedures: a) using nonlinear least-squares for NN and maximum likelihood for the GARCH model and b) maximising the joint likelihood function of the mean and variance model. We denote the corresponding test statistics by RBV-LM1 and RBV-LM2, respectively.

5.3.4 Simulation results

Tables 5.10 and 5.11 show the empirical rejection probabilities for the tests mentioned before under the assumption that errors are no serially dependent ($\rho = 0$)

[1] Although a vast variety of design models could be used in this study, experiments conducted with more complex neural network models produced similar results. This makes us believe that the simulation results are not specific to the particular choice of the parameters or the architecture of the neural network.

though heteroskedastic. This part of Monte Carlo experiments allows us to draw conclusions on the size of the tests, i.e. the rate at which each test falsely rejects the hypothesis of no serial correlation in the residuals. We report results for three nominal significance levels 1%, 5% and 10% and three sample sizes (700, 1000 and 2000 observations). As seen, the S-LM test for the adequacy of a NN is strongly effected by the heteroskedasticity of errors and produces more false rejections as the sample size increases (especially for non-normal errors). This result is generally in accordance with the findings of other simulation studies on linear models (see e.g. Furno (2000); Hafner and Herwartz (2000); Whang (1998)). On the other hand, robustified LM tests seem to preserve their finite-sample validity under heteroskedasticity. Table 5.10 shows that under GARCH errors (i.e. correct specification of the volatility process) the empirical rejection probabilities are very close to the nominal significance levels for all robustified versions (RB-LM, RBV-LM1 RBV-LM2). The results of Table 5.11 have some interesting implications for the two versions of RBV test. In particular, they show that the size of the RBV-LM1 test in not distorted if the assumed volatility model does not adequately describe the heteroskedastic pattern of errors. This is important for practical applications of the test. Interestingly, no significant size distortions are also observed for RBV-LM2 - at least for the range of sample sizes studied here - despite the fact that for an EGARCH volatility process the mean estimator employed by this statistic might be inconsistent for the parameters of the NN model.

Tables 5.12-5.15 present the power performance of tests, i.e. the rejection probabilities when errors are autocorrelated ($\rho = \{0.2, 0.3\}$). Each cell reports both the actual and the *size-corrected* rejection probability (in parentheses). As seen, the S-LM has the highest actual finite-sample power but, after size correction, its power performance is similar or even inferior to RB-LM and RBV-LM. Tables 5.12 and 5.14 represent the cases where the researcher has correctly identified the volatility process of errors. Under this condition, both versions of the RBV-LM test clearly have the highest (size-corrected) power among all tests and also better power performance than the RB-LM test in terms of actual and size-corrected rejection probabilities. The RBV-LM2 statistic, computed by the joint maximum likelihood estimator, seems to outperform RBV-LM1 in all sample sizes and distributions. Hence, using a more efficient estimator of the parameters of the NN model seems to increase the finite-sample efficiency of the robustified test with volatility information. In Tables 5.13 and 5.15 power results are reported under the assumption that errors follow the EGARCH volatility process. Note that in contrast to GARCH errors, in this case the power of all tests rapidly picks up as the sample size and the value of ρ increases. Under the assumption of EGARCH conditional variance, the volatility model implicit in all robustified tests is *a priori* misspecified for the conditional variance of errors. Nevertheless, RBV-LM tests (especially RBV-LM2) typically lead to higher rejection probabilities - especially when the sample size gets smaller - although the difference is not as pronounced as in the previous case. In any case, the power of the RBV-LM tests is not lower than RB-LM's. The slight outperformance of the robustified test with GARCH volatility estimates can be attributed to the fact that a GARCH model is a better approximation to an EGARCH volatility-generating process than a model of constant variance.

$\rho = 0$, volatility generating process GARCH

Distribution	**Significance**	**S-LM**	**RB-LM**	**RBV-LM1**	**RBV-LM2**
Normal	700 observations				
	1.0%	0.036	0.008	0.008	0.008
	5.0%	0.113	0.047	0.049	0.048
	10.0%	0.184	0.102	0.096	0.097
	1000 observations				
	1.0%	0.041	0.008	0.010	0.010
	5.0%	0.115	0.046	0.049	0.049
	10.0%	0.183	0.093	0.092	0.095
	2000 observations				
	1.0%	0.052	0.010	0.008	0.008
	5.0%	0.140	0.049	0.042	0.042
	10.0%	0.219	0.100	0.092	0.090
Student	700 observations				
	1.0%	0.085	0.006	0.008	0.008
	5.0%	0.186	0.046	0.044	0.044
	10.0%	0.261	0.099	0.100	0.096
	1000 observations				
	1.0%	0.112	0.008	0.011	0.010
	5.0%	0.216	0.046	0.049	0.050
	10.0%	0.303	0.100	0.100	0.099
	2000 observations				
	1.0%	0.166	0.006	0.009	0.009
	5.0%	0.293	0.043	0.052	0.053
	10.0%	0.379	0.097	0.106	0.107

Table 5.10: The empirical rejection frequencies for the S-LM, RB-LM, RBV-LM1 and RBV-LM2 serial correlation tests under the assumption that errors have no serial correlation ($\rho = 0$) and follow a GARCH volatility model.

$\rho = 0$, volatility generating process EGARCH

Distribution	**Significance**	**S-LM**	**RB-LM**	**RBV-LM1**	**RBV-LM2**
Normal	700 observations				
	1.0%	0.023	0.012	0.011	0.013
	5.0%	0.076	0.055	0.052	0.053
	10.0%	0.141	0.103	0.107	0.108
	1000 observations				
	1.0%	0.022	0.011	0.011	0.011
	5.0%	0.079	0.054	0.053	0.054
	10.0%	0.145	0.104	0.107	0.108
	2000 observations				
	1.0%	0.020	0.009	0.012	0.012
	5.0%	0.071	0.046	0.054	0.055
	10.0%	0.136	0.098	0.108	0.110
Student	700 observations				
	1.0%	0.038	0.006	0.007	0.007
	5.0%	0.106	0.041	0.041	0.042
	10.0%	0.179	0.088	0.097	0.096
	1000 observations				
	1.0%	0.045	0.006	0.007	0.008
	5.0%	0.127	0.047	0.047	0.049
	10.0%	0.198	0.100	0.104	0.103
	2000 observations				
	1.0%	0.060	0.011	0.012	0.014
	5.0%	0.146	0.056	0.053	0.063
	10.0%	0.222	0.106	0.104	0.116

TABLE 5.11: The empirical rejection frequencies for the S-LM, RB-LM, RBV-LM1 and RBV-LM2 serial correlation tests under the hypothesis that errors have no serial correlation ($\rho = 0$) and follow an EGARCH volatility model.

$\rho = 0.2$, volatility generating process GARCH

Distribution	**Significance**	**S-LM**	**RB-LM**	**RBV-LM1**	**RBV-LM2**
Normal	700 observations				
	1.0%	0.213 (0.107)	0.058 (0.067)	0.102 (0.108)	0.116 (0.130)
	5.0%	0.385 (0.256)	0.178 (0.182)	0.256 (0.261)	0.281 (0.285)
	10.0%	0.484 (0.365)	0.276 (0.273)	0.379 (0.389)	0.408 (0.414)
	1000 observations				
	1.0%	0.318 (0.162)	0.086 (0.095)	0.173 (0.178)	0.199 (0.199)
	5.0%	0.502 (0.344)	0.234 (0.242)	0.386 (0.389)	0.417 (0.423)
	10.0%	0.601 (0.476)	0.357 (0.370)	0.514 (0.525)	0.544 (0.554)
	2000 observations				
	1.0%	0.605 (0.323)	0.235 (0.235)	0.466 (0.483)	0.519 (0.545)
	5.0%	0.747 (0.598)	0.454 (0.456)	0.698 (0.719)	0.742 (0.769)
	10.0%	0.818 (0.691)	0.575 (0.575)	0.796 (0.808)	0.831 (0.843)
Student	700 observations				
	1.0%	0.363 (0.108)	0.057 (0.074)	0.115 (0.135)	0.131 (0.148)
	5.0%	0.522 (0.267)	0.182 (0.190)	0.310 (0.333)	0.333 (0.353)
	10.0%	0.612 (0.393)	0.298 (0.300)	0.451 (0.451)	0.470 (0.479)
	1000 observations				
	1.0%	0.505 (0.121)	0.093 (0.111)	0.214 (0.204)	0.242 (0.236)
	5.0%	0.651 (0.338)	0.259 (0.270)	0.455 (0.461)	0.492 (0.494)
	10.0%	0.731 (0.476)	0.387 (0.385)	0.586 (0.586)	0.627 (0.628)
	2000 observations				
	1.0%	0.793 (0.209)	0.243 (0.286)	0.574 (0.587	0.629 (0.646)
	5.0%	0.877 (0.535)	0.489 (0.514)	0.787 (0.781)	0.833 (0.827)
	10.0%	0.912 (0.703)	0.631 (0.637)	0.868 (0.862)	0.903 (0.899)

Table 5.12: The empirical rejection frequencies for the S-LM, RB-LM, RBV-LM1 and RBV-LM2 serial correlation tests under the assumption that errors have serial correlation of order 2 ($\rho = 0.2$) and follow a GARCH volatility model. Values in parentheses show size-corrected rejection probabilities.

$\rho = 0.2$, volatility generating process EGARCH

Distribution	**Significance**	**S-LM**	**RB-LM**	**RBV-LM1**	**RBV-LM2**
Normal	700 observations				
	1.0%	0.570 (0.460)	0.473 (0.462)	0.487 (0.475)	0.510 (0.479)
	5.0%	0.765 (0.700)	0.700 (0.681)	0.707 (0.705)	0.722 (0.715)
	10.0%	0.838 (0.796)	0.797 (0.794)	0.800 (0.794)	0.810 (0.798)
	1000 observations				
	1.0%	0.786 (0.720)	0.706 (0.699)	0.725 (0.718)	0.747 (0.730)
	5.0%	0.902 (0.871)	0.873 (0.868)	0.872 (0.868)	0.888 (0.882)
	10.0%	0.942 (0.918)	0.920 (0.917)	0.924 (0.919)	0.933 (0.928)
	2000 observations				
	1.0%	0.988 (0.982)	0.980 (0.980)	0.984 (0.983)	0.986 (0.986)
	5.0%	0.996 (0.994)	0.994 (0.994)	0.996 (0.996)	0.997 (0.996)
	10.0%	0.999 (0.998)	0.997 (0.997)	0.998 (0.998)	0.999 (0.999)
Student	700 observations				
	1.0%	0.697 (0.448)	0.504 (0.557)	0.530 (0.570)	0.549 (0.572)
	5.0%	0.844 (0.744)	0.729 (0.752)	0.741 (0.759)	0.762 (0.779)
	10.0%	0.900 (0.838)	0.826 (0.840)	0.836 (0.839)	0.852 (0.859)
	1000 observations				
	1.0%	0.872 (0.656)	0.727 (0.773)	0.769 (0.801)	0.793 (0.802)
	5.0%	0.950 (0.881)	0.885 (0.889)	0.904 (0.907)	0.918 (0.919)
	10.0%	0.969 (0.935)	0.932 (0.931)	0.942 (0.940)	0.954 (0.953)
	2000 observations				
	1.0%	0.995 (0.941)	0.966 (0.965)	0.987 (0.986)	0.989 (0.987)
	5.0%	0.999 (0.994)	0.987 (0.986)	0.997 (0.997)	0.998 (0.997)
	10.0%	0.999 (0.997)	0.994 (0.993)	0.999 (0.999)	0.999 (0.999)

Table 5.13: The empirical rejection frequencies for the S-LM, RB-LM, RBV-LM1 and RBV-LM2 serial correlation tests under the hypothesis that errors have serial correlation of order 2 ($\rho = 0.2$) and follow an EGARCH volatility model. Values in parentheses show size-corrected rejection probabilities.

$\rho = 0.3$, volatility generating process GARCH

Distribution	**Significance**	**S-LM**	**RB-LM**	**RBV-LM1**	**RBV-LM2**
Normal	700 observations				
	1.0%	0.330 (0.204)	0.105 (0.119)	0.159 (0.166)	0.210 (0.224)
	5.0%	0.497 (0.371)	0.267 (0.275)	0.341 (0.347)	0.408 (0.413)
	10.0%	0.587 (0.477)	0.382 (0.380)	0.459 (0.465	0.527 (0.533)
	1000 observations				
	1.0%	0.472 (0.306)	0.168 (0.181)	0.261 (0.267)	0.347 (0.347)
	5.0%	0.629 (0.496)	0.363 (0.372)	0.484 (0.488)	0.579 (0.585)
	10.0%	0.709 (0.606)	0.491 (0.501)	0.600 (0.612)	0.686 (0.696)
	2000 observations				
	1.0%	0.771 (0.567)	0.383 (0.383)	0.594 (0.613)	0.742 (0.760)
	5.0%	0.866 (0.767)	0.618 (0.620)	0.780 (0.796)	0.883 (0.896)
	10.0%	0.905 (0.833)	0.720 (0.720)	0.854 (0.863)	0.932 (0.937)
Student	700 observations				
	1.0%	0.488 (0.192)	0.101 (0.131)	0.180 (0.203)	0.232 (0.254)
	5.0%	0.643 (0.392)	0.270 (0.282)	0.392 (0.411)	0.471 (0.489)
	10.0%	0.720 (0.519)	0.399 (0.401)	0.522 (0.521)	0.598 (0.605)
	1000 observations				
	1.0%	0.615 (0.236)	0.149 (0.176)	0.287 (0.274)	0.389 (0.383)
	5.0%	0.741 (0.468)	0.357 (0.370)	0.527 (0.533)	0.642 (0.643)
	10.0%	0.800 (0.596)	0.489 (0.487)	0.658 (0.658)	0.751 (0.753)
	2000 observations				
	1.0%	0.876 (0.413)	0.361 (0.409)	0.650 (0.659)	0.804 (0.815)
	5.0%	0.899 (0.678)	0.606 (0.630)	0.831 (0.827)	0.926 (0.924)
	10.0%	0.925 (0.783)	0.729 (0.734)	0.898 (0.893)	0.961 (0.957)

TABLE 5.14: The empirical rejection frequencies for the S-LM, RB-LM, RBV-LM1 and RBV-LM2 serial correlation tests under the assumption that errors have serial correlation of order 2 ($\rho = 0.3$) and follow a GARCH volatility model. Values in parentheses show size-corrected rejection probabilities.

$\rho = 0.3$, volatility generating process EGARCH

Distribution	**Significance**	**S-LM**	**RB-LM**	**RBV-LM1**	**RBV-LM2**
normal	700 observations				
	1.0%	0.864 (0.800)	0.816 (0.809)	0.809 (0.799)	0.820 (0.801)
	5.0%	0.947 (0.922)	0.922 (0.914)	0.916 (0.915)	0.925 (0.920)
	10.0%	0.967 (0.955)	0.957 (0.955)	0.953 (0.951)	0.956 (0.952)
	1000 observations				
	1.0%	0.956 (0.930)	0.930 (0.927)	0.932 (0.929)	0.943 (0.937)
	5.0%	0.986 (0.977)	0.980 (0.978)	0.981 (0.980)	0.984 (0.983)
	10.0%	0.992 (0.989)	0.990 (0.989)	0.988 (0.988)	0.991 (0.990)
	2000 observations				
	1.0%	0.999 (0.998)	0.998 (0.998)	0.999 (0.999)	0.999 (0.999)
	5.0%	1.000 (1.000)	0.999 (0.999)	0.999 (0.999)	1.000 (1.000)
	10.0%	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
Student	700 observations				
	1.0%	0.904 (0.755)	0.789 (0.820)	0.805 (0.831)	0.823 (0.835)
	5.0%	0.964 (0.927)	0.919 (0.928)	0.922 (0.927)	0.928 (0.935)
	10.0%	0.979 (0.960)	0.957 (0.962)	0.954 (0.956)	0.959 (0.959)
	1000 observations				
	1.0%	0.975 (0.895)	0.912 (0.932)	0.932 (0.945)	0.946 (0.950)
	5.0%	0.993 (0.979)	0.973 (0.973)	0.980 (0.982)	0.986 (0.986)
	10.0%	0.997 (0.991)	0.989 (0.989)	0.993 (0.992)	0.994 (0.994)
	2000 observations				
	1.0%	0.999 (0.991)	0.989 (0.989)	0.999 (0.998)	0.999 (0.998)
	5.0%	1.000 (0.999)	0.994 (0.994)	1.000 (1.000)	1.000 (1.000)
	10.0%	1.000 (1.000)	0.996 (0.996)	1.000 (1.000)	1.000 (1.000)

TABLE 5.15: The empirical rejection frequencies for the S-LM, RB-LM, RBV-LM1 and RBV-LM2 serial correlation tests under the hypothesis that errors have serial correlation of order 2 ($\rho = 0.3$) and follow an EGARCH volatility model. Values in parentheses show size-corrected rejection probabilities.

5.4 Summary and discussion

This chapter sheds light on the empirical performance of testing procedures proposed in this thesis by means of Monte-Carlo simulations. We present two experiments whose main issue is the ability of the standard and robustified LM testing procedures to detect the number of hidden units and error autocorrelation in a neural network regression model. Our attempt was to create a more realistic compared to other studies simulation environment that closely resembles the statistical properties of economic and financial data (heavy-tails and changing variance).

Simulations show that the non-robust version of the LM test cannot distinguish between mean dependencies and changing variance levels. The empirical size of the test is generally distorted under ARCH heteroskedasticity, leading to excessively false indications of nonlinearity or serial correlation in errors. This is a cautionary remark against using non-robust statistical procedures for testing NN specifications. Robustified LM tests, on the other hand, closely follow the nominal type I error under heteroskedasticity, allowing thus the researcher to control the complexity of the neural network *without* having to explicitly model variance and higher-order dependencies in the data-generating process. This is particularly desirable for practical applications of NNs given the empirical properties of economic and financial time-series. In addition, adopting robustified versions of the test when data are not in fact heteroskedastic does not result in a power decrease.

Comparing the two robustified versions of the LM test, we find that the RB-LM test that ignores heteroskedasticity in errors manages to restore the size of the test but this improvement comes occasionally with a loss in power when data are *also* characterised by additional structure in mean. Incorporating estimates of the conditional volatility of errors into the testing procedure, gives the researcher a more efficient device for testing the adequacy of a given NN parametrisation. Simulations show that the validity of this test is not affected by misspecification of the volatility process; in other words the test is robust to a failure of the researcher' prior guess on the true volatility dependencies. However, the finite-sample performance depends on the choice of the volatility model and more power is gained the closer are the prior estimates to the conditional volatility. Hence, although the primary concern of the researcher may be to investigate first-moment relations between the variables of interest (i.e. relations in mean), it pays to put some effort on deriving a good approximation to the underlying volatility structure. This is another argument for extending modelling to higher-order dependencies of conditional distributions.

The experiments performed in this study can be extended in various directions. First of all, more network configurations can be tested with different distributions, volatility models and signal-to-noise ratios. Further research is necessary to compare the empirical performance of the robustified testing procedures, examined in this thesis, with that of other statistical or heuristic methods encountered in the literature for selecting the structure of a neural network. A third important aspect is to see how the selection strategies work when the "true" data-generating model is not nested in the class of neural networks. In this case, the neural network is only an approximator to the underlying functional relation and therefore a priori misspecified for the data-generating process. Simulation experiments presented in Anders and Korn (1999) show that the sequence of hypothesis tests employing a Taylor expansion leads to reliable results compared to information criteria or cross-validation, when the “true” model in not encompassed by the class of neural network architectures. However, in their experiments random variates are drawn from a normal distribution with constant variance. It would be thus interesting to evaluate the effectiveness of the above procedures under the presence of strong statistical features in disturbances, such as heteroskedasticity or non-normality.

Chapter 6

Short-term predictability of daily stock index returns

6.1 Introduction

This chapter presents an empirical application of NN-GARCH models to forecasting the conditional distribution of daily returns on three major international stock indexes (DAX, FTSE 100, S&P 500)[1]. By means of this study, we try to address another important issue in modern finance, the predictability of asset prices. Although supporters of market efficiency claim that stock prices follow random walks and are generally unpredictable, recent empirical finance research has revealed several "nonlinearities" in the dynamics of prices that enhance their predictability. These were discussed in chapter 2.

While researches believe that financial markets are nonlinear, there is generally disagreement on the sources of nonlinearity and hence on the proper type of models that should be employed in each case to capture price dependencies. Some authors believe that nonlinearities enter through the conditional mean, mostly as a result of asymmetric market responses, while others believe that they enter through the conditional variance, in the form of autoregressive conditional heteroskedasticity (ARCH) or leverage effects. In any case, it is important to distinguish between these types of nonlinearities as each type has different implications for portfolio selection and the design of risk management strategies (see Hsieh (1993)). Note however that both types of dynamic behaviour may as well coexist, as the existence of asymmetric and threshold affects in one moment of the distribution does not imply anything for the other. In a process exhibiting autoregressive conditional heteroskedasticity, nonlinear adjustments may also be present, because ARCH does interact with the conditional mean.

By means of this study, we take the opportunity to discuss several issues in a forecasting application of a NN-GARCH model. Special emphasis is given to the

[1] FTSE 100, is a financial Stock Exchange index of the 100 most highly capitalised companies listed on the London Stock Exchange. The S&P (Standard and Poor's) 500 index is a market weighted value index consisting of 500 stocks, chosen for their market size, liquidity and industry group representation.

computation of one- and multi-period predictions, the construction of confidence intervals and the evaluation of the forecasting performance. Besides statistical criteria, we introduce a new framework for judging the economic significance of forecasts provided by a NN-GARCH or a general dynamic model that parametrises conditional densities.

The plan of this chapter is as follows: section 6.2 provides the motivation for this study and reviews the current literature on predictability of asset prices. Section 6.3 presents the data set employed in our empirical application and section 6.4 details the construction of forecasting models. As we are interested in real-time forecasting, we re-estimate the parameters of models on a daily basis each time a new observation becomes available and occasionally re-specify the structure of forecasting models. In section 6.4.2 we discuss a method for computing one- and multi-step-ahead prediction densities, which is based on the simulation of future market scenarios. Section 6.5 is devoted to the evaluation of forecasting models and proposes a set of criteria suitably adjusted to density predictions. Quite often, in the literature, forecasting models are evaluated by purely statistical measures that concentrate on the goodness-of-fit of the model. However, from a financial engineering perspective, the economic significance of forecasts is also an important issue. We propose a trading strategy that takes into account confidence bounds rather than point forecasts, as is customary in the NN literature, and is based on the detection of "exceptional" price movements. Section 6.6 discusses the results of the forecasting exercise and section 6.7 summarises and concludes the chapter.

6.2 Motivation and empirical evidence

The issue of short-term predictability of financial prices has a long history in empirical finance. Although most of the theoretical literature traditionally supports market efficiency and thus unpredictability of returns, recent empirical research has reported short-term predictability and nonlinear dynamics. In most empirical studies, the issue of predictability of financial prices has been investigated using linear regression models with possible searches for macroeconomic variables that further explain abnormal returns. The evidence from this empirical research is mixed: some authors report predictability while others claim that stock prices follow a (geometric) random walk possibly with a drift[2]. There is a strong objection however that any evidence of unpredictability may in reality be a statistical bias introduced by the non-adequacy of linear-in-mean models to capture all features of the return-generating process.

A good deal of research addressed the same issues employing flexible or less "tightly" parametrised specifications as a means of accommodating generic nonlinearity in the conditional mean. Among them neural networks seemed to offer an ideal choice for flexible nonlinear modelling, mostly due to their universal approximation property. The empirical results where though inconclusive. White (1988b), for example, found that neural networks have poor out-of-sample forecasting ability when applied to IBM daily stock prices. For monthly New York stock-index returns, Chuah (1992) found that neural networks could not beat the benchmark linear model

[2] See e.g. Fama (1970, 1991, 1998); Farmer and Lo (1999) for related literature.

in terms of forecasting accuracy. On the other hand, Wu and Zheng (2003) showed that the daily closing prices of several American stock indexes (S&P 500, Dow Jones, and NASDAQ 100[3]) are highly forecastable and quite often a recurrent NN outperformed linear ARIMA models in terms of profit rates. Phua et al. (2003) confirmed that for European markets such as DAX, FTSE 100.

Short-term predictability of financial time-series has also been investigated by means of parametric nonlinear econometric models, such as threshold and smooth-transition autoregressions. These empirical models are often consistent with a behavioural interpretation of the price formation process, discussed in section 2.4. Uninformed noise traders typically engage in "trade-chasing" which in the short-run drives prices away from fundamentals, with abnormal changes in returns, but with the result that in the long-run arbitrage traders will cause a reversion of prices back to fundamentals. Applying smooth-transition autoregressive models to the returns on indexes of six international stock markets (France, Germany, Hong Kong, Japan, Malaysia, Singapore), McMillan (2005) showed that these models are likely to improve in- and out-of-sample fit compared to linear alternatives.

Among the various forecasting exercises involving nonlinear models, Teräsvirta et al. (2005) deserves special attention as it addresses the issue of in-sample model specification. Based on a number of monthly macroeconomic time-series, Teräsvirta et al. (2005) compare the forecasting accuracy of linear models against smooth transition autoregressions (STAR) and neural networks. They found that, as far as point forecasts are concerned, a carefully specified STAR model generally outperforms its linear counterpart. The results for the NN models are mixed: typically there is not much gain in using a NN instead of a simple linear autoregressive model, although a NN obtained by Bayesian regularisation produces more accurate forecasts that a corresponding model specified using a specific-to-general methodology. In the last approach, Teräsvirta et al. (2005) adopted the non-robust version of the neglected nonlinearity test, proposed by Teräsvirta et al. (1993) and also examined in section 4.5.1, to judge the significance of extra neurons. They empirically show that careful in-sample specification of NN models is important in order to obtain acceptable and successful out-of-sample forecasting results. Arbitrarily determined NN architectures containing redundant neurons, can cause instability and explosive forecasting behaviour, especially when applied to multi-step-ahead predictions.

A number of authors claim that nonlinear dependencies in daily prices are more the result of GARCH effects and volatility clustering rather than mean interactions. Hsieh (1989), after conducting a systematic specification of the conditional mean and variance parts, concludes that reported nonlinearity in daily exchange rates is more likely to enter through variances rather than through means and a GARCH(1,1) model with a fat-tailed distribution can account for most of the nonlinear dependence. Hong and White (2005), applying entropy measures of serial dependence, give mixed evidence of serial dependence in S&P 500 log-returns after removing the persistent effects of GARCH volatility (using an AR-GARCH). Their results support that a linear AR-GARCH model cannot fully capture the dynamics of S&P 500 daily returns,

[3] A stock index including 100 of the largest domestic and international securities listed on the NASDAQ Stock Market based on market capitalisation. The index mainly reflects companies across major industry groups and does not contain securities of financial or investment companies.

although they claim that this may be due to misspecification of the conditional mean and variance.

In the empirical studies presented above, some authors focus exclusively on mean and others on variance dependencies, often ignoring the interaction between these two moments. It is therefore possible that nonlinear dependencies detected in mean are in fact the result of ARCH effects, and vice versa. McMillan (2005); Teräsvirta et al. (2005), for example, employ the non-robust version of the neglected nonlinearity test to build nonlinear regression models, although both report substantial excess kurtosis in data. Whether this excess kurtosis is the results of homogeneously distributed heavy-tailed errors or ARCH effects in the variance of errors is not clear. In the second case, reported nonlinearities are questionable, because, as simulation results presented in chapter 5 show, the empirical size of the standard version of the nonlinearity test is distorted in the presence of ARCH effects leading to excessively false indications of nonlinearity in mean.

When investigating the source of dependencies in financial data, there is a need for jointly determining the first two moments of the distribution and also employ model specifications that can account for generic nonlinearity in the mean and the variance of the data-generating process. Experience gained from other forecasting exercises suggests that introducing much flexibility in modelling can have adverse effects as concerns the long-run performance of forecasting models. Over-parametrised mean structures, such as neural networks, can absorb some affects due to the presence of conditional heteroskedasticity in the data-generating process. Therefore, it is absolutely necessary that the methodology used to determine the right complexity of the mean model offer some kind of robustification against the adverse affects of conditional heteroskedasticity. The common practice to remove persistent heteroskedasticity, by applying for example a linear GARCH model, is probably not the proper way to deal with nonlinearities in variance, because as we discussed in section 4.3.2, bad specification of the mean process can yield inconsistency for the variance model, and vice versa.

In this chapter, we investigate the predictability of daily returns on three major stock indexes (DAX, S&P 500, FTSE 100) using NN-GARCH models. Due to the dual nature of this class of parametrisations, it is possible to investigate whether past returns or return "surprises" contain useful information as regards future returns and the volatility of returns.

6.3 Data set

Our sample data comprise daily closing values of DAX, FTSE 100 and S&P 500 from 30/11/1994 to 10/09/2001, a total of 1700 approximately trading days. The rates of return on each index are calculated by taking logarithmic differences between successive trading days, i.e. $r_t = \log(P_t) - \log(P_{t-1})$, where P_t denotes the index's closing value at time t and r_t the corresponding return between $t-1$ and t.

Many authors have noted systematic effects both in the mean and variance of price movements that are related to calendar, weekday or exchange holiday influences. These price anomalies are likely to cause spurious nonlinearity in the return-generating process and hence have to be filtered-out in advance. In order to adjust for irregular shifts, we apply the two-stage adjustment process described in Gallant et al. (1992) in which systematic effects are first removed from the mean and then from the variance. This is done by running two linear regressions against a set of dummy variables capturing calendar, weekday and exchange holiday effects.

Figure 6.1 shows the resulting log-return time-series and table 6.1 provides summary statistics. The empirical distribution of returns is generally characterised by negative skewness and excess kurtosis. The Jarque-Bera test statistic strongly rejects the hypothesis of normality in each case. The last two rows of table 6.1 report the Box-Pierce Q and Q^2 test for up to fifth order autocorrelation in returns and squared returns. Test statistics are corrected for heteroskedasticity or higher-moment dependencies according to Diebold (1988) (see also Lobato et al. (2001)). The Q-test rejects the hypothesis of no correlation in mean returns for all indexes except DAX. The non-modified Q test (not reported here) indicated serial dependence in mean in all indexes. The modified Q^2 statistic for autocorrelation in the squared returns is highly significant in all cases indicating strong dependencies in the volatility of returns. Similar results are obtained using the non-modified Q^2 statistic.

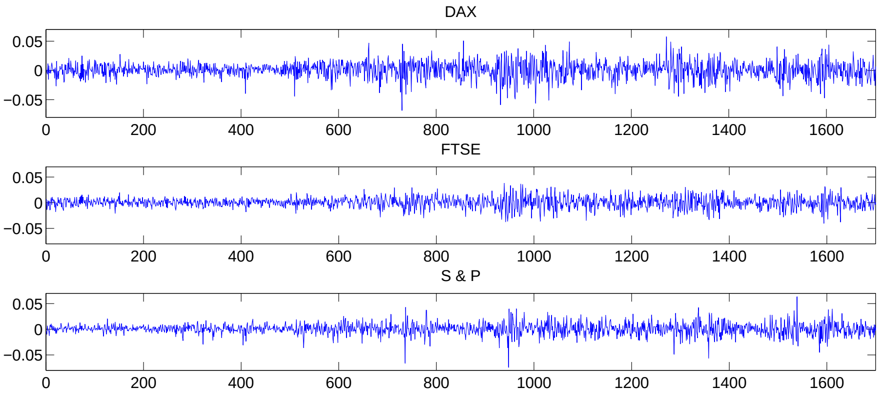

Figure 6.1: Daily log-returns on the DAX, CAC and S&P indexes from 30/11/1994 to 10/09/2001 after adjusting for calendar, week-of-the-day and exchange holiday effects.

6.4 Forecasting models

6.4.1 Model building

NN-GARCH specifications are used to derive out-of-sample predictions on the one-day ahead DAX, FTSE and S&P returns conditionally on the returns observed in the last five consecutive trading days. As we are interested in real-time forecasting, we re-estimate the parameters of models on a daily basis each time a new observation becomes available. In flexible forecasting structures, such as NN-GARCH models, that encompass a variety of specifications, attention should also be paid to the specification of the model. Owning to the time-varying nature of economic and trading activity, one expects that the return-generating process has undergone many changes

	DAX	FTSE	S&P
min	-0.068	-0.040	-0.074
max	0.058	0.038	0.063
mean	5.956×10^{-4}	3.160×10^{-4}	5.033×10^{-4}
std	0.014	0.010	0.011
skewness	-0.297	-0.193	-0.281
kurtosis	4.517	4.130	6.907
JB test	187.478 (0.000)	100.904 (0.000)	1.106×10^{3} (0.000)
$Q(5)$	4.034 (0.545)	24.114 (0.000)	8.981 (0.110)
$Q^2(5)$	86.095 (0.000)	90.070 (0.000)	39.410 (0.000)

Table 6.1: Summary statistics on DAX, FTSE and S&P 500 logarithmic returns after adjusting for seasonality, calendar and week-of-the-day effects. JB is the Jarque-Bera test of normality and $Q(5)$, $Q^2(5)$ are the Box-Pierce tests for up to fifth order autocorrelation in returns and squared returns, respectively. Test statistics are corrected for heteroskedasticity. The significance is given in parentheses.

in the almost seven-year period spanned by our sample data. Hence, it is possible that the expectational law of returns switches from nonlinear to linear or that the returns series is locally homoskedastic. But in the former case, pre-detected hidden neurons are not identifiable which yields inconsistency for the parameters of the NN-GARCH model. In order to obtain more parsimonious specifications and thus decrease the possibility of parameter redundancy, we chose to periodically re-specify the structure of the model.

Specification, estimation, and forecasting were carried out in rolling samples of 700 observations The beginning of the first sample window is placed at 30/11/1994 and the last window ends at 10/09/2001. All models are re-estimated each day and re-specified only once every fifty days (or two trading months approximately) to reduce the computational burden. At each re-specification stage, the bottom-up procedure described in section 4.5 was used to determine the optimal structure of the conditional mean and variance parts of the NN-GARCH model. The variables of the linear model where determined using SBIC. Additional lags where added when necessary to remove serial correlation of order 1 to 10 in the residuals of the linear model. In order to investigate the effects of heteroskedasticity or higher-moment model misspecification in the performance of the forecasting models, we applied all three versions of the LM neglected nonlinearity test to determine the number of neurons in the hidden layer: the standard LM (S-LM), the robustified LM (RB-LM) and the robustified LM with GARCH(1,1) volatility estimates (RBV-LM)[4]. For simplicity, we shall henceforth refer to the models resulting from the above procedures as S-LM, RB-LM and RBV-LM. After the specification of the mean part, we tested the residuals for autoregressive conditional heteroskedasticity. Whenever the test indicated the presence of ARCH effects, we jointly estimated a NN-GARCH(1,1) model. The adequacy of the GARCH(1,1) model structure was then tested against neglected

[4] All models were estimated and specified in Matlab© version 7. The code that implements the neural network model-building strategy with the standard LM test was downloaded from http://swopec.hhs.se/hastef/abs/hastef0508.htm on January 2005.

ARCH and asymmetric GARCH effects, using either the standard or the robust testing procedures described in section 4.5.2, in accordance to the type of the nonlinearity test employed in the mean. Whenever tests indicated the presence of asymmetric effects in the conditional variance, we jointly estimated a NN-EGARCH(1,1) model and then further tested its adequacy[5]. All intermediate models were estimated using QML under a normal density assumption. Return data were standardised before presented to the NN-GARCH model. This has been found to significantly improve the numerical optimisation of the empirical log-likelihood function.

In order to compare the relative advantage of careful in-sample model selection as opposed to simple re-estimation of parameters, we also considered two other fixed-specification forecasting structures: a linear-in-mean AR(5)-EGARCH(1,1) model, with five lags in the mean part, and a fully-interconnected neural network with three neurons in the hidden layer and no volatility component. The latter model helps us to clarify whether there is any advantage in modelling time-varying conditional heteroskedasticity against a model that assumes steady conditional distribution. The predefined forecasting models will be henceforth referred to as LFEG and NL3F, respectively.

6.4.2 Computing multi-step density forecasts

All models are compared on the basis of a set of criteria discussed below and on a forecasting horizon of one, three and five days ahead. Comparison on multiple horizons helps us to determine whether forecasting models have captured the salient features of the underlying data-generating probability model and not merely over-fitted the data. Multi-step-ahead forecasts were computed from the same (one-step ahead) model, presented in the form of confidence bounds on the future return on the stock index. In the context of NN-GARCH models, and in most nonlinear models of mean and variance, there is no analytical formulae for the j-step ahead conditional density $\rho(y_{t+j}|x_t)$ and the estimation has to be done either by numeric approximation of nested integrals or by Monte-Carlo simulation[6]. In the present work, we adopt the second approach. In order to derive the j-step-ahead conditional distribution of daily returns, we simulate $N = 2000$ return scenarios (paths) of length j. For each scenario i, a set of j disturbances is sampled with replacement from model's in-sample *standardised* residuals (the residuals divided by the estimated standard deviation). These are denoted by $\left\{\hat{u}_t^{(i)},\, t = 1, \ldots, j\right\}$. The error scenario is then computed as $\left\{\sqrt{\hat{h}_t^{(i)}}\hat{u}_t^{(i)},\, t = 1, \ldots, j\right\}$, where $\hat{h}_t^{(i)}$ is the model's volatility forecast at time t based on the history of errors. For homoskedastic models, where no dependence in conditional volatilities is assumed, this is set equal to the in-sample variance $\hat{\sigma}^2$. Note that bootstrapping standardised residuals from the in-sample distribution of errors, effectively by-passes any restrictive assumption on the conditional density of errors

[5] In order to reduce the computational burden due to the large number of estimated models, we did not apply a posteriori tests for parameter and variable significance. Because of this omission, more complicated structures for the forecasting models may have been occasionally obtained.

[6] Granger and Teräsvirta (1998); Tay and Wallis (2002) give an analytical discussion on techniques for density forecasting with nonlinear econometric models.

		S-LM	RB-LM	RBV-LM
DAX	$\hat{h} = 0$	0.007	0.901	1.000
	$\hat{h} = 1$	0.695	0.099	0.000
	$\hat{h} = 2$	0.199	0.000	0.000
	$\hat{h} \geq 3$	0.099	0.000	0.000
FTSE	$\hat{h} = 0$	0.099	0.596	0.249
	$\hat{h} = 1$	0.255	0.404	0.701
	$\hat{h} = 2$	0.299	0.000	0.050
	$\hat{h} = 3$	0.150	0.000	0.000
	$\hat{h} = 4$	0.099	0.000	0.000
	$\hat{h} \geq 5$	0.099	0.000	0.000
S&P	$\hat{h} = 0$	0.000	0.304	0.249
	$\hat{h} = 1$	0.304	0.646	0.701
	$\hat{h} = 2$	0.497	0.050	0.050
	$\hat{h} \geq 3$	0.199	0.000	0.000

Table 6.2: Model specifications (a): percentage of samples where each test indicated a number of neurons equal to the value of $\hat{h}$ given in the second column.

imposed by the model specification (normality, symmetry, etc).

6.4.3 Final specifications

Due to space limitations, we are confined to a short description of the specification of models that were derived for each index and sample period, though further details are available from the author upon request. Table 6.2 shows the percentage of samples where each test indicated a number of neurons equal to the value of $\hat{h}$ given in the second column. As seen, the S-LM test was generally supportive of nonlinearity in mean and in many cases the number of neurons in the hidden layer exceeded 2. On average, more nonlinearity was detected in FTSE than in other indexes. The RB-LM and the RBV-LM tests on the other hand produced overall simpler models. An entirely linear-in-mean return-generating process was suggested by RBV-LM for DAX returns. An analysis of the in-sample residuals of the S-LM models derived for each index showed that the addition of extra neurons and lag variables did not significantly reduce the skewness and kurtosis of the empirical distribution. The adjusted Q^2 was in most cases significant indicating strong heteroskedasticity in errors. It is therefore likely that the extra neurons placed by the S-LM test were the result of size distortions rather than actual nonlinear dynamics in index returns.

Table 6.3 provides details about the specification of the variance part. Each cell reports the percentage of samples where the (standard or robustified) volatility tests detected heteroskedasticity and asymmetric variance effects in errors. Note that the ARCH LM tests reject on average more often the hypothesis of homoskedasticity for the residuals of the RB-LM and RBV-LM than the residuals of the S-LM model. This difference could be attributed to the inclusion of extra neurons in the mean part of S-LM models, whose result is to absorb part of the heteroskedasticity. Strong asymmetric responses of conditional variance to past shocks was detected in the residuals of most rolling models, hence an EGARCH model was estimated. In all cases, the simplest EGARCH(1,1) volatility model was found adequate to capture

		S-LM	RB-LM	RBV-LM
DAX	NO ARCH	0.047	0.000	0.000
	GARCH	0.140	0.233	0.187
	EGARCH	0.813	0.767	0.813
FTSE	NO ARCH	0.234	0.047	0.093
	GARCH	0.094	0.374	0.374
	EGARCH	0.672	0.579	0.533
S&P	No ARCH	0.354	0.188	0.235
	GARCH	0.000	0.141	0.141
	EGARCH	0.646	0.671	0.624

Table 6.3: Model specifications (b): percentage of samples where ARCH effects and asymmetric volatility were detected in errors.

volatility clustering in unexpected returns and there was no need to add further ARCH or GARCH terms.

Generally, the in-sample specification of forecasting models reveals strong asymmetries and nonlinear dependencies in daily returns on the three major stock indexes, though there is generally disagreement on the source of these nonlinearities. Some models indicate that nonlinearities enter through the mean and others through the variance of the return-generating process. Since these results may be due to model specification bias and the deficiency of the statistical procedures to discriminate between these two effects, this issue has necessarily to be addressed out-of-sample.

6.5 Performance measures

In our study, density forecasts are obtained from models that focus on different aspects of the conditional density. Some parametrise the mean and assume homoskedasticity while others jointly parametrise the mean and the variance of the conditional distribution. The different nature of these models generally makes it difficult to find a widely acceptable set of criteria to compare alternative density forecasts. Besides, many popular out-of-sample criteria that are often applied in the literature, such as the root mean squared error or the confusion rate (Swanson and White (1997a)), are mostly applicable to point forecasts. The criteria we adopt in this study are an attempt to generalise the notions of forecasting error and goodness-of-fit to densities. For the evaluation of the out-of-sample forecasting performance of the models specified above, we use both *statistical* and *non-statistical* (economic) criteria. Those are described below in detail.

6.5.1 Statistical criteria

The most natural indicator of the out-of-sample "fitness" of a model is the *empirical log-likelihood* (ELL), the performance criterion used at the estimation stage. This addresses the question of how likely model forecasts given the new data. Apart from the ELL, we also used the Schwarz's Bayesian information criterion (SBIC), which adds a complexity penalty term to the empirical log likelihood, and the model that minimises this penalized log-likelihood is preferred. The penalty term is an increasing function of m, the total number of model's parameters, which poses a problem to our case since most models are not of fixed specification and thus the number of the parameters changes over time. To calculate the SBIC of flexible specifications, we use in place of m the average number of parameters in the corresponding samples. Although in this case the information theoretic interpretation of the criterion is not clear, SBIC may still serve as a good indicator of the balance between complexity and goodness-of-fit achieved by each specification.

As our models are in fact approximations to the true underlying conditional density $p(y_t|x_t)$, they can be used to calculate confidence bounds on future returns. The *empirical confidence level* (ECL), i.e. the percentage of observations that fall outside the confidence bounds, is another indicator of how well the model approximates the underlying statistical process. In a well-specified model one expects that the percentage of sample returns lying outside the confidence bounds is close to the nominal significance level.

The last two criteria of density forecasting performance are the *normalised mean absolute error* (NMAE) and the *hit rate* (HR), both adapted from Schittenkopf et al. (2000). These are two indicators of the ability of models to forecast short-term changes in conditional volatility. NMAE compares model's conditional volatility estimates with those obtained by a "naive" predictor. HR is the relative frequency of correctly predicted increases and decreases of volatility, i.e. it measures how often the model gives the correct direction of change of volatility. In order to apply both criteria, one needs a measure for the true volatility of unexpected index returns on each day. In our application, we used the simplest possible proxy, the squared model's error at time t ($\hat{\epsilon}_t^2$). As a naive predictor, we used an average of the past squared shocks over a window of 30 observations. We thus define NMAE and HR as follows:

$$\text{NMAE} = \frac{\sum_t \left|\hat{\epsilon}_{t+j}^2 - \hat{h}_{t+j}\right|}{\sum_t \left|\hat{\epsilon}_{t+j}^2 - \hat{n}_{t+j}\right|}$$

$$\text{HR} = \frac{1}{T}\sum_t \delta_t, \quad \delta_t = \begin{cases} 1, & (\hat{\epsilon}_{t+j}^2 - \hat{\epsilon}_t^2)(\hat{h}_{t+j} - \hat{\epsilon}_t^2) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\hat{h}_{t+j}$ denotes the $t+j$ model's volatility forecast given the information available up to time t and $\hat{n}_{t+j}$ denotes the naive predictor. Note that NMAE generally takes positive values, though in order for a model to beat the naive predictor, it should attain a NMAE less than 1. The HR lies between 0 and 1. A value of 0.5 indicates that the model is not better than a random predictor of ups and downs.

6.5.2 Non-statistical criteria

Quite often in the literature forecasting models are evaluated on a statistical, goodness-of-fit basis. However, when comes to financial engineering applications, the economic significance of forecasts is also an important issue. To investigate this, one typically converts the predictions provided by the model into a suitable trading strategy that involves buying and selling orders and then calculates the distribution of profits/losses at the end of the investment period.

In most empirical studies with NN models, predictions are given in the form of point forecasts hence trading strategies usually evolve buying or selling the asset if the predicted value is higher or lower than the current price. In our application, we deviate from the main trend and design a trading strategy that takes into account confidence bounds rather than point forecasts. The economic performance of the trading strategy becomes another means of testing the credibility of confidence bounds provided by the forecasting model. The underlying idea is to buy (sell) the asset when its value becomes "exceptionally" low (high). Note that forecasting models are estimated based on the log-returns time-series, as this is mean-reverting, while the tradable "quantity" is the index[7]. However, experiments conducted in this thesis have shown that a model that is successful in predicting shifts in returns does not necessarily have good forecasting performance on an index level, when comes to multi-period predictions. As the ultimate objective of this study is not to create a trading strategy for daily stock indexes but rather to evaluate the predictive performance of different models, we assume that index returns form a "hypothetical" tradable asset whose price evolution is described by the corresponding time-series.

The trading strategy adopted is this study takes a buying (long) position on the asset when $r_t < \hat{r}^{L,\alpha}_{t+j}$ and a selling (short) position when $r_t > \hat{r}^{H,\alpha}_{t+j}$, where r_t is the observed return at time t and $\hat{r}^{L,\alpha}_{t+j}$ $\hat{r}^{H,\alpha}_{t+j}$ denote the $(1-\alpha)\%$ low and high j-step-ahead confidence bounds. The above trading strategy was first introduced in Thomaidis et al. (2006) for exploiting price discrepancies in a pair of related stocks. A trade entered at some time t closes after j trading days ahead, although this is not necessary as in order to avoid placing too many orders, one could leave the position open unless the model predicts a movement in the opposite direction (e.g. if for some $t_1 > t$, $r_{t_1} \geq \hat{r}^{L,\alpha}_{t_1+h}$ while $r_t < \hat{r}^{L,\alpha}_{t+h}$, and so on)).

Based on the profit/loss distribution of the aforementioned trading strategy, we report three performance measures: a) the *accumulated profit* (AP) gained by the end of the forecasting period, b) the *number of trades* (NT) placed by the model and c) the *average profit per trade* (APT), i.e. the total profit divided by the number of trades. Note that in a real market environment with transaction costs, the AP can be a misleading index of economic performance as it does not explicitly specify the number of trades been placed to achieve the final wealth. Note that a trading system that puts an excessive number of orders may eventually experience gross losses, if trades are on average not profitable enough to overbalance costs. From this perspective, the APT is an equally important to the AP performance measure that is often overlooked in the forecasting literature.

6.6 Empirical results

The out-of-sample performance of forecasting models is summarised in tables 6.4-6.9, at the end of the chapter. It should be noted here that occasionally the numerical estimation of highly nonlinear models (especially S-LM or NL3F) resulted in instable values for the parameters and explosive out-of-sample behaviour at long prediction

[7] In fact, stock indexes per se are not tradable securities. However, one can track index movements by holding a portfolio of stocks composing the index or by trading on the index future contract.

horizons. In order to create more realistic forecasting devices and to make model comparison fair, we decided to truncate implausible out-of-sample index forecasts by implying a filter, in the spirit of Swanson and White (1995); Teräsvirta et al. (2005). For each trading day t, we computed the average and the standard deviation of past returns in a window of 30 days. If the upper or lower 10% confidence bound was beyond plus/minus three standard deviations from the average, model's forecast was replaced by this naive predictor.

For an illustration of the forecasting differences between models, figure 6.2 plots the one-day-ahead conditional density estimates (excluding RB-LM's, whose predicted density is very similar to RBV-LM) for two specific days of the DAX sample set: October 1st, 1998 (bottom figure) and May 18th, 1999 (top figure). The first date signals the last day of a series of consecutive large negative returns and the 18th of May 1999 represents a normal period in the German Stock Exchange with returns fluctuating around zero. For these particular days, the structure of the flexible forecasting models is as follows: RBV-LM, RB-LM are linear in mean (with no lags) while S-LM includes one neuron in the hidden layer. The volatility part of all models is parametrised as an EGARCH(1,1) process.

The top plot of figure 6.2 shows that under normal market conditions, all models deliver approximately the same forecast in the form of a leptokurtic distribution. The large unexpected price drop that took place on the 1st of October 1998 increased on average the short-term uncertainty about the next-day price movement, resulting in wider density forecasts for the heteroskedastic models (see bottom plot). The distribution of the pure NN model however remains largely unaffected. Note that on October 1st, 1998 the one-step-ahead prediction densities are skewed to the right, assigning higher probability to a subsequent positive than negative return. The density corresponding to the S-LM forecasting structure has its peak further shifted to the right, which is mainly the result of nonlinearity existing in the mean model. We thus observe that different parametrisations of the conditional mean and variance result in a different response to the same market conditions.

Figures 6.3 to 6.6 show models' density forecasts based on the information available up to October, 1st 1998 for one, three and five days ahead. Note that multi-day-ahead densities are generally less skewed that the following-day's density forecast, meaning that the large negative shock on DAX returns that took place on the 1st of October is absorbed and the market moves into a steady state. The deviation between density forecasts is more pronounced in models that contain extra nonlinearity in mean, such as the S-LM and NL3F, and exhibit richer forecasting behaviour.

Table 6.4 shows the normalised mean absolute error across different models and for various forecasting horizons. We observe that RB-LM, RBV-LM and LFEG models appear slightly better than a naive volatility predictor, although in some cases the value of the normalised mean absolute error is above 1. Highly parametrised models like S-LM or NL3F, one the other hand, have worse performance, especially at longer forecasting horizons. The values of the hit rate in table 6.5 are generally greater than 0.5 for all models indicating an advantage over a random predictor of volatility increases and decreases. Overall, the results for NMAE and HR are inconclusive and also questionable, as the ability of both criteria to identify a good forecasting model relies on a reliable estimate of the true conditional volatility, an otherwise ill-defined

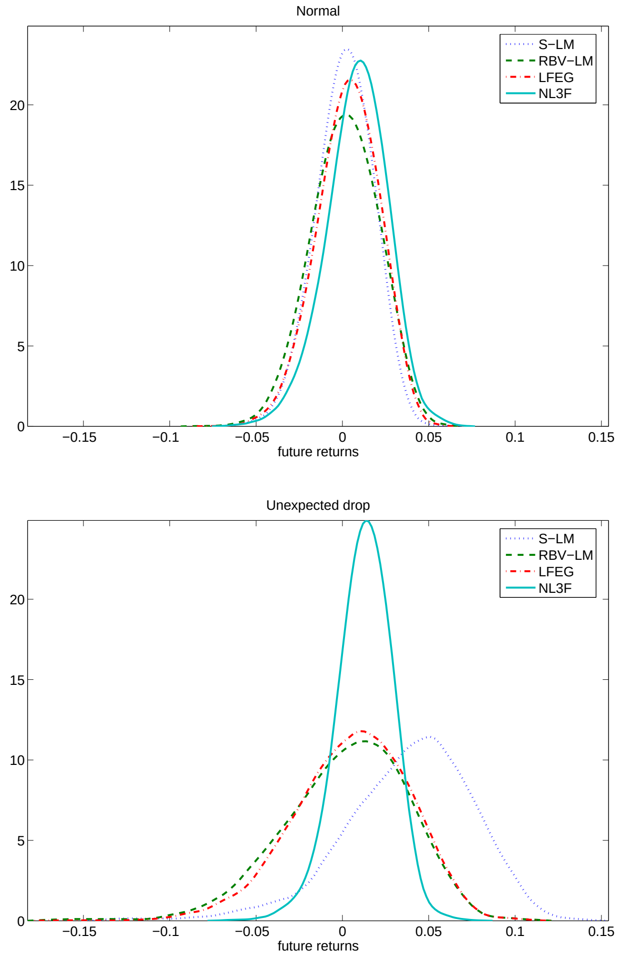

FIGURE 6.2: The estimated one-day-ahead conditional densities for DAX log-returns on 18/05/99 (top figure), 01/10/98 (bottom figure). Solid lines show the density forecast of the pure neural network model (NL3F), dashed line shows the density of RBV-LM, dash-dotted line shows the density of LFEG and dotted line shows the density of S-LM. Densities estimates were obtained using the Epanechnikov kernel-smoothing method.

concept (see Schittenkopf et al. (2000) for a discussion).

Tables 6.6 and 6.7 show the empirical likelihood and the SBIC for the forecasting models. Although differences are not generally significant, the empirical likelihood of the S-LM and NL3F models tends to be lower. The complexity penalised log-likelihood mostly favours flexible-specification models that also take into account the volatility structure. It seems that robustified LM tests result in parsimonious specifications that combine reasonable fit with minimum complexity. On the contrary, the extra complexity indicated by the standard LM test, or included in LFEG and NL3F specifications, does not add much to the out-of-sample fit of the models. Table 6.7 shows a slight advantage of flexible specifications versus a linear fixed AR(5)-EGARCH(1,1) model, possibly because in the latter the number of parameters is not

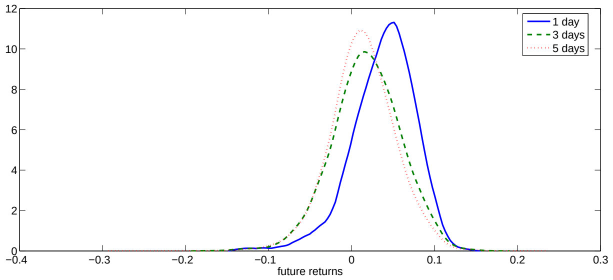

FIGURE 6.3: The 1,3 and 5-day-ahead distribution of returns predicted by the S-LM model on the 1st of October 1998.

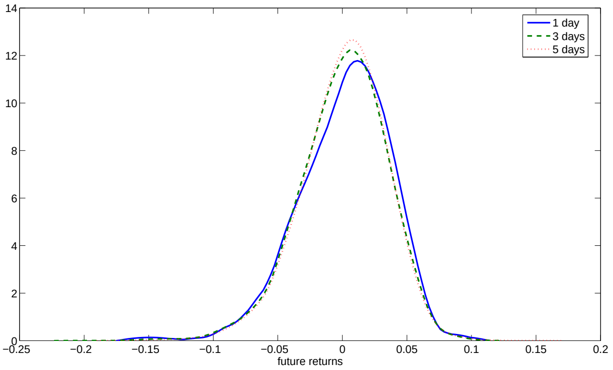

FIGURE 6.4: The 1,3 and 5-day-ahead distribution of returns predicted by the RBV-LM model on the 1st of October 1998.

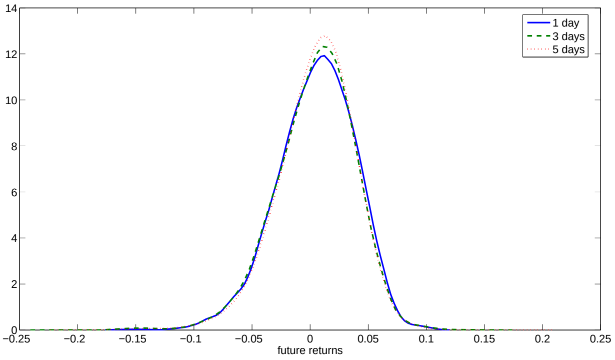

FIGURE 6.5: The 1,3 and 5-day-ahead distribution of returns predicted by the LFEG model on the 1st of October 1998.

adjusted according to data complexity.

As pertains to the empirical confidence, the results from table 6.8 clearly advo-

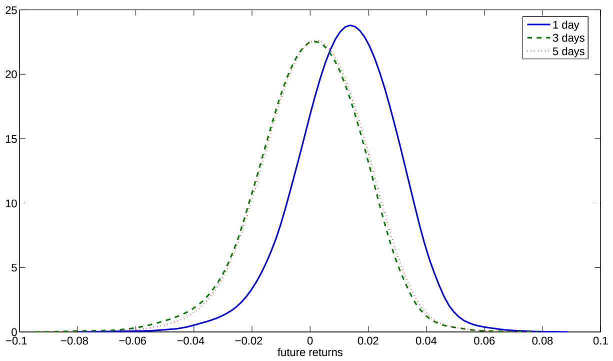

FIGURE 6.6: The 1,3 and 5-day-ahead distribution of returns predicted by the pure NN model (NL3F) on the 1st of October 1998.

cate the credibility of the confidence bounds provided by models that specify the conditional variance (RB-, RBV, S-LM and LFEG). For this class of models, the percentage of observations falling outside the confidence areas closely follows the corresponding nominal rates for all forecasting horizons. The bounds provided by the standard neural network model, on the other hand, are often too narrow missing a greater percentage of observations. Hence, a pure nonlinear-in-mean model cannot fully account for the variability of index returns. The results of table 6.8 do not show significant differences between a flexible specification NN-GARCH and a fixed specification AR(5)-EGARCH model, meaning that in terms of empirical confidence the two classes of models are almost operationally equivalent.

The results for the non-statistical criteria have some interesting implications for the economic significance of forecasting models. Tables 6.9, 6.10 and 6.11 show the accumulated profit, the number of trades placed by the models in the investment period and the average profit per trade. A general remark is that the higher is the significance level α, the more trades are placed by the trading system and the average profit per trade tends to go down. The reason is that as α increases confidence bounds become narrower and more points are characterised as "exceptional". Note that on average the S-LM and the pure neural network model give systematically more buy and sell signals than the others, possibly because nonlinearity in mean results in narrower confidence bounds. However, the trades indicated by these models are not more profitable than GARCH ones, and in many cases (mainly for longer forecasting horizons) they tend to be less profitable. We observe that GARCH models are on average more conservative in placing orders, hence the tendency of the average profit per trade to increase over samples. Interestingly, models indicated by robustified tests very often outperform S-LMs in terms of average profit per trade, rendering thus the addition model complexity that the standard test induces non economically significant.

6.7 Summary and discussion

This chapter presented an empirical application of NN-GARCH models on three major stock indexes (DAX, S&P 500, FTSE 100) with the purpose of investigating the predictability of daily returns. The joint nature of these models permitted us to investigate whether past return shocks contain useful information as regards both future returns and the volatility of returns.

On the occasion of this study, we discussed several issues related to forecasting with NN-GARCH models, such as the computation of one- and multi-period predictions, the construction of confidence intervals and the evaluation of the forecasting performance. Quite often, in the literature, forecasting models are evaluated by pure statistical criteria that concentrate on the goodness-of-fit of the model. However, from a financial engineering perspective, the economic significance of forecasts is also an important issue. Judging the economic significance is possible by designing a trading strategy that incorporates models' forecasts. In the class of models studied in this thesis, forecasts are presented in the form of density predictions or confidence bounds. For this reason we introduce a trading strategy that explicitly takes into account this information and focuses on the detection of *abnormal* or "*exceptional*" price movements. In this way, we deviate from the main trend in forecasting applications of NNs, where trading strategies are commonly based on point forecasts.

Owing to the time-varying nature of economic activity, we applied an adaptive forecasting scheme which amounts to re-estimating the parameters of the models each time a new observation becomes available and occasionally re-specify their structure. The model-building strategy analysed in section 4.5.1 was used to derive the optimal specification of the mean and variance equations of forecasting models. In order to investigate the effects of heteroskedasticity or higher-moment model misspecification in the performance of forecasting models, we applied robust as well as non-robust versions of the tests for additional structure in mean and variance.

The results of the forecasting exercise have some interesting implications regarding both the nature of predictability of asset prices and the performance of testing procedures presented in this thesis. Many were the cases where robustified tests simplified the structure of the forecasting device, sometimes as much as to a linear in mean model. The standard LM nonlinearity test, on the other hand, produced on average more parametrised neural network models that "absorbed" much of the heteroskedasticity in errors. However, the additional nonlinearity introduced by the S-LM typically brought no gain in forecasting accuracy and occasionally led to unstable performance on unseen data. This makes us believe that nonlinearities reported in daily stock index movements are more the result of GARCH effects and asymmetric volatility response to past shocks rather than mean interactions. However, this result is strongly depended upon the market and time-period under study as well as on the frequency at which data were sampled. As most of the stocks composing the German, English and U.S. stock index are heavily traded, it is very likely that exceptional price movements and nonlinear adjustments disappear within hours, minutes or even seconds, and hence cannot be seen on daily data.

It should be noted that the computational burden imposed by the large number of forecasting models, did not allow us to perform detailed in-sample evaluation of estimated models before applying them to out-of-sample forecasting. We believe that a combination of diagnostics with significance tests would generally improve the parsimony and performance of forecasting models.

The most important lesson gained from this forecasting exercise is that when one considers choosing a forecasting structure from a large family of models, carefully determining the right model complexity (i.e. the right member of the family) improves out-of sample forecasting performance. The existence of heteroskedasticity and higher-moment properties in the distribution of errors may often invalidate standard model-selection procedures leading to over-parametrised models with poor generalising ability. Robustified tests seem unaffected by the presence of strong heteroskedasticity in the disturbances and also result in parsimonious specifications that combine reasonable fit with minimal complexity. It is thus absolutely necessary that the procedures used to specify the complexity of the mean part offer some kind of robustification against the adverse effects of heteroskedasticity. Of course, further experimentation with financial and economic time-series is necessary to reveal the real benefits of robustification, but this is left to future research.

Forecasting Horizon	**1**	**3**	**5**
	DAX		
S-LM	0.983	1.096	1.431
RB-LM	0.986	0.998	0.995
RBV-LM	0.990	0.989	0.979
LFEG	0.970	0.962	0.964
NL3F	0.955	0.973	0.964
	FTSE		
S-LM	0.940	0.989	1.525
RB-LM	1.108	1.159	0.995
RBV-LM	0.953	0.989	0.999
LFEG	0.957	0.946	0.936
NL3F	0.958	1.172	1.426
	S&P		
S-LM	0.899	3.328	3.597
RB-LM	0.920	0.940	1.069
RBV-LM	0.930	1.076	1.154
LFEG	1.576	2.235	3.848
NL3F	0.879	1.266	2.086

Table 6.4: The normalised mean absolute error.

Forecasting Horizon	**1**	**3**	**5**
	DAX		
S-LM	0.709	0.697	0.666
RB-LM	0.715	0.706	0.684
RBV-LM	0.720	0.705	0.694
LFEG	0.707	0.706	0.691
NL3F	0.699	0.687	0.684
	FTSE		
S-LM	0.711	0.705	0.691
RB-LM	0.715	0.689	0.695
RBV-LM	0.719	0.702	0.711
LFEG	0.718	0.716	0.723
NL3F	0.722	0.674	0.654
	S&P		
S-LM	0.725	0.708	0.685
RB-LM	0.723	0.716	0.708
RBV-LM	0.717	0.716	0.703
LFEG	0.682	0.576	0.717
NL3F	0.731	0.654	0.630

Table 6.5: The hit rate.

Forecasting Horizon	**1**	**3**	**5**
	DAX		
S-LM	2.691	2.704	2.692
RB-LM	2.723	2.717	2.707
RBV-LM	2.723	2.715	2.709
LFEG	2.722	2.713	2.713
NL3F	2.578	2.638	2.638
	FTSE		
S-LM	2.930	2.936	2.909
RB-LM	3.071	2.970	2.995
RBV-LM	3.007	2.961	2.943
LFEG	2.937	2.979	2.982
NL3F	2.864	2.882	2.841
	S&P		
S-LM	2.878	2.844	2.805
RB-LM	2.949	2.904	2.882
RBV-LM	2.951	2.888	2.873
LFEG	2.663	2.736	2.564
NL3F	2.837	2.632	2.542

Table 6.6: The empirical log-likelihood.

Forecasting Horizon	**1**	**3**	**5**
	DAX		
S-LM	-5.284	-5.311	-5.288
RB-LM	-5.412	-5.399	-5.379
RBV-LM	-5.411	-5.396	-5.384
LFEG	-5.374	-5.357	-5.357
NL3F	-4.970	-5.090	-5.090
	FTSE		
S-LM	-5.722	-5.733	-5.676
RB-LM	-5.885	-5.884	-5.809
RBV-LM	-5.860	-5.860	-5.824
LFEG	-5.945	-5.888	-5.894
NL3F	-5.541	-5.577	-5.495
	S&P		
S-LM	-5.659	-5.591	-5.603
RB-LM	-5.836	-5.745	-5.702
RBV-LM	-5.846	-5.720	-5.690
LFEG	-5.257	-5.403	-5.059
NL3F	-5.488	-5.078	-4.896

Table 6.7: The Schwarz's Bayesian information criterion.

	α	**0.01**	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
	DAX						
S-LM	*1 day*	0.028	0.069	0.128	0.228	0.506	0.809
	3 days	0.013	0.051	0.111	0.217	0.506	0.795
	5 days	0.012	0.048	0.112	0.211	0.505	0.797
RB-LM	*1 day*	0.014	0.059	0.120	0.220	0.504	0.802
	3 days	0.015	0.060	0.114	0.231	0.517	0.804
	5 days	0.016	0.063	0.116	0.235	0.517	0.806
RBV-LM	*1 day*	0.013	0.060	0.116	0.219	0.505	0.817
	3 days	0.016	0.060	0.120	0.228	0.517	0.807
	5 days	0.013	0.063	0.120	0.228	0.522	0.810
LFEG	*1 day*	0.014	0.061	0.124	0.226	0.509	0.818
	3 days	0.013	0.063	0.127	0.235	0.529	0.817
	5 days	0.015	0.063	0.125	0.237	0.532	0.809
NL3F	*1 day*	0.023	0.097	0.167	0.285	0.550	0.839
	3 days	0.025	0.081	0.144	0.275	0.553	0.825
	5 days	0.023	0.081	0.151	0.276	0.546	0.825
	FTSE						
S-LM	*1 day*	0.022	0.093	0.156	0.265	0.542	0.804
	3 days	0.024	0.093	0.161	0.265	0.555	0.826
	5 days	0.039	0.111	0.172	0.295	0.556	0.823
RB-LM	*1 day*	0.014	0.066	0.129	0.235	0.525	0.787
	3 days	0.016	0.060	0.121	0.230	0.533	0.816
	5 days	0.016	0.064	0.123	0.238	0.542	0.824
RBV-LM	*1 day*	0.016	0.076	0.139	0.246	0.538	0.811
	3 days	0.022	0.075	0.144	0.244	0.545	0.815
	5 days	0.027	0.076	0.144	0.253	0.547	0.815
LFEG	*1 day*	0.011	0.066	0.123	0.226	0.518	0.789
	3 days	0.017	0.072	0.139	0.242	0.542	0.818
	5 days	0.021	0.081	0.133	0.241	0.540	0.816
NL3F	*1 day*	0.032	0.103	0.170	0.292	0.587	0.828
	3 days	0.045	0.111	0.159	0.256	0.495	0.840
	5 days	0.069	0.200	0.257	0.368	0.623	0.843
	S&P						
S-LM	*1 day*	0.023	0.072	0.129	0.270	0.573	0.822
	3 days	0.020	0.060	0.132	0.267	0.560	0.826
	5 days	0.025	0.074	0.134	0.261	0.562	0.828
RB-LM	*1 day*	0.016	0.053	0.119	0.241	0.536	0.817
	3 days	0.019	0.057	0.123	0.242	0.549	0.813
	5 days	0.020	0.059	0.119	0.245	0.541	0.821
RBV-LM	*1 day*	0.017	0.053	0.121	0.232	0.538	0.812
	3 days	0.017	0.055	0.114	0.241	0.542	0.810
	5 days	0.019	0.061	0.112	0.242	0.542	0.816
LFEG	*1 day*	0.007	0.041	0.074	0.199	0.453	0.709
	3 days	0.003	0.003	0.006	0.082	0.309	0.439
	5 days	0.005	0.035	0.085	0.143	0.449	0.761
NL3F	*1 day*	0.029	0.085	0.152	0.284	0.590	0.833
	3 days	0.102	0.245	0.346	0.452	0.691	0.878
	5 days	0.107	0.246	0.321	0.422	0.651	0.885

TABLE 6.8: The empirical confidence levels.

	α	**0.01**	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
	DAX						
S-LM	*1 day*	0.158	0.785	2.686	5.066	9.865	11.090
	3 days	0.142	0.946	2.220	5.129	10.071	11.821
	5 days	0.090	0.716	2.251	5.049	9.612	11.751
RB-LM	*1 day*	0.134	0.948	2.678	5.244	9.998	11.704
	3 days	0.113	1.100	2.796	5.682	10.140	12.198
	5 days	0.101	1.114	3.189	5.226	10.055	11.971
RBV-LM	*1 day*	0.167	0.962	2.710	5.199	10.298	11.681
	3 days	0.062	1.045	2.913	5.823	10.325	12.131
	5 days	0.122	0.966	3.239	5.453	9.983	11.898
LFEG	*1 day*	0.094	0.962	2.808	5.163	9.835	11.761
	3 days	0.173	1.057	2.940	5.817	10.321	12.080
	5 days	0.158	1.152	3.208	5.458	9.834	12.012
NL3F	*1 day*	0.781	2.890	4.659	6.706	10.010	11.495
	3 days	0.797	2.813	4.596	6.956	10.754	12.056
	5 days	0.761	2.668	4.523	6.506	10.106	11.826
	FTSE						
S-LM	*1 day*	0.426	1.301	2.366	4.085	6.835	8.137
	3 days	0.409	1.848	3.449	5.321	7.942	9.673
	5 days	0.409	1.326	2.826	4.704	7.734	9.078
RB-LM	*1 day*	0.021	0.729	1.837	3.373	6.989	8.294
	3 days	0.025	1.129	2.505	4.460	8.139	10.142
	5 days	0.000	0.909	2.322	3.917	7.770	9.363
RBV-LM	*1 day*	0.109	1.037	2.206	3.752	6.913	8.158
	3 days	0.086	1.370	2.846	4.834	8.411	10.003
	5 days	0.131	1.533	3.226	4.911	8.024	9.475
LFEG	*1 day*	0.025	0.669	1.978	3.538	7.095	8.217
	3 days	0.025	0.717	2.172	3.982	7.877	10.047
	5 days	0.033	0.856	2.229	3.870	7.353	9.296
NL3F	*1 day*	0.673	1.715	2.874	4.380	7.223	8.324
	3 days	1.082	2.430	3.280	4.446	7.082	8.433
	5 days	0.746	1.838	2.642	4.159	6.510	7.507
	S&P						
S-LM	*1 day*	0.751	1.852	2.802	4.586	7.388	8.691
	3 days	0.616	1.563	2.682	5.051	8.773	10.187
	5 days	0.485	1.549	2.827	4.797	8.456	9.537
RB-LM	*1 day*	0.347	1.285	2.203	4.229	7.445	8.697
	3 days	0.375	1.322	2.586	4.987	8.919	10.518
	5 days	0.185	1.626	2.620	4.994	8.618	9.979
RBV-LM	*1 day*	0.524	1.235	2.283	4.185	7.549	8.629
	3 days	0.273	1.316	2.468	5.037	8.969	10.486
	5 days	0.274	1.364	2.640	5.048	8.716	9.935
LFEG	*1 day*	0.000	0.266	0.470	0.979	2.321	3.796
	3 days	0.000	0.000	0.012	0.693	3.659	4.874
	5 days	0.000	0.000	0.000	0.104	3.940	4.494
NL3F	*1 day*	1.035	2.399	3.877	5.148	7.290	8.859
	3 days	1.532	2.948	3.230	4.000	6.244	7.192
	5 days	1.725	2.236	2.818	3.841	6.127	7.045

TABLE 6.9: The accumulated profit at the end of the forecasting period.

	α	**0.01**	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
	DAX						
S-LM	*1 day*	9	36	95	206	522	817
	3 days	3	23	69	180	513	814
	5 days	1	18	70	184	502	804
RB-LM	*1 day*	3	27	79	195	521	821
	3 days	2	27	84	198	519	824
	5 days	2	26	91	203	524	823
RBV-LM	*1 day*	4	26	83	195	529	819
	3 days	1	26	90	204	525	823
	5 days	2	25	92	207	524	819
LFEG	*1 day*	2	27	82	192	519	821
	3 days	3	27	91	207	525	821
	5 days	3	27	91	208	523	818
NL3F	*1 day*	15	87	172	284	544	825
	3 days	19	79	136	268	541	815
	5 days	19	72	143	264	537	815
	FTSE						
S-LM	*1 day*	18	56	112	211	509	785
	3 days	16	75	150	273	562	823
	5 days	27	78	158	272	565	826
RB-LM	*1 day*	1	25	66	156	493	795
	3 days	2	36	91	190	523	818
	5 days	0	31	87	184	527	812
RBV-LM	*1 day*	5	40	95	192	508	798
	3 days	5	52	114	225	549	805
	5 days	4	55	134	239	547	807
LFEG	*1 day*	1	24	73	170	484	793
	3 days	1	25	79	166	488	787
	5 days	1	27	77	174	479	798
NL3F	*1 day*	24	70	137	240	540	811
	3 days	71	185	247	356	611	830
	5 days	35	98	150	253	490	831
	S&P						
S-LM	*1 day*	18	57	104	221	528	804
	3 days	16	51	104	242	560	824
	5 days	19	62	124	257	567	822
RB-LM	*1 day*	8	35	74	192	516	807
	3 days	10	40	92	226	541	815
	5 days	7	43	93	234	544	819
RBV-LM	*1 day*	11	35	80	184	513	798
	3 days	9	39	85	228	542	807
	5 days	7	36	97	232	545	813
LFEG	*1 day*	0	18	32	80	225	432
	3 days	0	0	1	42	303	447
	5 days	0	0	0	7	349	459
NL3F	*1 day*	27	79	154	249	539	824
	3 days	103	240	327	450	684	871
	5 days	111	242	317	421	639	874

Table 6.10: The number of trades.

	α	**0.01**	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
	DAX						
S-LM	*1 day*	0.018	0.022	0.028	0.025	0.019	0.014
	3 days	0.047	0.041	0.032	0.028	0.020	0.015
	5 days	0.090	0.040	0.032	0.027	0.019	0.015
RB-LM	*1 day*	0.045	0.035	0.034	0.027	0.019	0.014
	3 days	0.057	0.041	0.033	0.029	0.020	0.015
	5 days	0.050	0.043	0.035	0.026	0.019	0.015
RBV-LM	*1 day*	0.042	0.037	0.033	0.027	0.019	0.014
	3 days	0.062	0.040	0.032	0.029	0.020	0.015
	5 days	0.061	0.039	0.035	0.026	0.019	0.015
LFEG	*1 day*	0.047	0.036	0.034	0.027	0.019	0.014
	3 days	0.058	0.039	0.032	0.028	0.020	0.015
	5 days	0.053	0.043	0.035	0.026	0.019	0.015
NL3F	*1 day*	0.052	0.033	0.027	0.024	0.018	0.014
	3 days	0.042	0.031	0.026	0.024	0.020	0.015
	5 days	0.040	0.030	0.032	0.025	0.019	0.015
	FTSE						
S-LM	*1 day*	0.024	0.023	0.021	0.019	0.013	0.010
	3 days	0.026	0.025	0.023	0.019	0.014	0.012
	5 days	0.015	0.017	0.018	0.017	0.014	0.011
RB-LM	*1 day*	0.021	0.029	0.028	0.022	0.014	0.010
	3 days	0.013	0.031	0.028	0.023	0.016	0.012
	5 days	-	0.029	0.027	0.021	0.015	0.012
RBV-LM	*1 day*	0.022	0.026	0.023	0.020	0.014	0.010
	3 days	0.017	0.026	0.025	0.021	0.015	0.012
	5 days	0.033	0.028	0.024	0.021	0.015	0.012
LFEG	*1 day*	0.025	0.028	0.027	0.021	0.015	0.010
	3 days	0.025	0.029	0.027	0.024	0.016	0.013
	5 days	0.033	0.032	0.029	0.022	0.015	0.012
NL3F	*1 day*	0.028	0.025	0.021	0.018	0.013	0.010
	3 days	0.015	0.013	0.013	0.012	0.012	0.010
	5 days	0.021	0.019	0.018	0.016	0.013	0.009
	S&P						
S-LM	*1 day*	0.042	0.032	0.027	0.021	0.014	0.011
	3 days	0.038	0.031	0.026	0.021	0.016	0.012
	5 days	0.026	0.025	0.023	0.019	0.015	0.012
RB-LM	*1 day*	0.043	0.037	0.030	0.022	0.014	0.011
	3 days	0.038	0.033	0.028	0.022	0.016	0.013
	5 days	0.026	0.038	0.028	0.021	0.016	0.012
RBV-LM	*1 day*	0.048	0.035	0.029	0.023	0.015	0.011
	3 days	0.030	0.034	0.029	0.022	0.017	0.013
	5 days	0.039	0.038	0.027	0.022	0.016	0.012
LFEG	*1 day*	-	0.015	0.015	0.012	0.010	0.009
	3 days	-	-	0.012	0.016	0.012	0.011
	5 days	-	-	-	0.015	0.011	0.010
NL3F	*1 day*	0.038	0.030	0.025	0.021	0.014	0.011
	3 days	0.015	0.012	0.010	0.009	0.009	0.008
	5 days	0.016	0.009	0.009	0.009	0.010	0.008

Table 6.11: Average profit per trade. Dashes denote cases where no trades were placed.

Chapter 7

A comparison of neural network model-selection strategies for the pricing of S&P 500 stock index options[1]

7.1 Introduction

An option is a type of tradeable financial contract whose price depends on a number of factors such as the exercise price, the time to maturity, the price and the volatility of the underlying asset and the risk-free interest rate prevailing in the market. The growing interest on option pricing was stimulated by the seminal work of Black and Scholes (1973) who managed to derive a “fair” value for a European-type call option based on dynamic hedging and arbitrage arguments. Since the publication of the Black-Scholes model in 1973, considerable research effort has been made on deriving parametric models that relax some of the restrictive assumptions underlying the Black-Scholes formula (normality of log-returns, constant volatility, etc.). Despite the elegance of functional forms and their theoretical appeal, most of these models are difficult to implement, have poor out-of-sample performance and are sometimes inconsistent with market data.

Semi-parametric computational intelligent models, especially neural networks, are more flexible in relaxing the restrictive assumptions of parametric models with a potential for improvement on out-of-sample pricing performance. The success of NN-based option pricing models is now well documented (see e.g. Hutchinson et al. (1994); Gencay and Qi (2001)). Essentially, a neural network offers to the researcher an estimation technique that puts up a flexible pricing formula with a set of unknown parameters and lets the optimisation routine search for the values of parameters that provide the optimal fit to the data.

In Tzastoudis et al. (2006), we applied a series of neural network models, determined by heuristic criteria, to the pricing of call options on the S&P 500 index.

[1] This chapter also appears in Thomaidis et al. (2007).

We attempted to reduce the overall number of inputs to the model by taking advantage of the no-arbitrage value of a forward contract, written on the same underlying asset. In addition, we experimented with *hybrid* intelligent schemes that combine a semi-parametric NN model with theoretical option-pricing formulae. Theoretical arguments were used in "problematic" data regions and NN learning was directed to more actively traded areas of the option surface. This combination seemed to increase the efficiency of the learning process and also reduce the overall in-sample and out-of-sample error.

In this chapter, we present an extension to our previous work, which contrary to many approaches on option pricing pays special attention to model selection. For the choice of the optimal architecture of the neural net, we experiment with both an iterative "top-down" pruning techniques as well as two "bottom-up" strategies that start with simple models and gradually complicate the architecture if data indicate so. Apart from heuristics, we also employ methods that base model selection on solid statistical techniques, such as statistical hypothesis tests and information criteria, and we compare their performance in accurately pricing and forecasting stock index options. We start with fitting the entire surface using a single NN and then examine fitting restricted areas of the option matrix using hybrid intelligent models, in the spirit of Tzastoudis et al. (2006).

The structure of this chapter is as follows: in section 7.2 we provide the necessary financial background on option contracts and we present the famous Black & Scholes and other parametric approaches to option pricing. In section 7.3 we review the literature on option pricing with neural networks. Section 7.4 presents the procedures employed in our study for the selection of the optimal NN architecture; section 7.4.1 details bottom-up model selection based on sequential statistical hypothesis tests and information criteria and section 7.4.2 reviews an iterative "top-down" pruning technique. Section 7.5 discusses the application data used in comparing the performance of different model selection strategies. Section 7.6 presents a first approach to option pricing, fitting the entire option matrix, and section 7.7 considers hybrid intelligent NN models directed to restricted areas of the option surface. Section 7.8 summarises the main findings and proposes future research directions.

7.2 Financial background

7.2.1 Options theory

An option is a certain type of financial contract that gives the *right* to the owner to buy or sell certain quantities of an asset, also called the *underlying asset* or simply the *underlying*, at some future date (*expiration* or *maturity date*) and at a price that is agreed in advance (*exercise* or *strike price*). In this way, options allow investors to bet on future market scenarios and also reduce financial risk. *Call* options give one the right to buy while *put* options give the right to sell the underlying. An investor who makes use of this right in expiration is said to have *exercised* the option. In standard option terminology, a call option is said to be *in-the-money* if the current market price of the underlying is greater than the strike price, *out-of-the-money* if the underlying price is lower than the strike price and *at-the-money* if the above two are close to each other. "Deep" in-the-money calls are generally highly priced, especially those that are near to expiration, as the holder can buy the underlying from the counter-party and make profit by selling it at the market at a higher price. For the same reason, far out-of-the money calls are almost worthless as the holder has no benefit from exercising the contract.

7.2.2 Black & Scholes option pricing

As soon as option trading started in organised markets, there came the problem of discovering a fair price to be paid by an investor who enters an option contract. This is known in the literature as the *option pricing problem.* The Black and Scholes (1973) model is considered to be the first successful attempt to obtain a fair value for a call option which is based on the fundamental idea of arbitrage (see sections 1.2.2 & 2.4). The original Black & Scholes mathematical formula for European-style call options, which was later modified by Merton (1973) for a dividend-paying underlying asset, is as follows:

$$C = Se^{-\delta \mathrm{T}} N(d_1) - Ke^{-rT} N(d_2) \tag{7.2.1}$$

$$d_1 = \frac{ln(S/K) + [(r-\delta) + (1/2)\sigma^2]T}{\sigma\sqrt{T}} \tag{7.2.2}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{7.2.3}$$

where C is the fair value of a call option at some time T before expiration, S is the current price of the underlying, K is the strike price of the option, r is the risk-free interest rate, δ is the dividend yield, σ is the volatility of the underlying and $N(.)$ denotes the standard normal cumulative density function.

The BS model links the price of a call option with a number of factors that affect its value. Intuitively, a rise in the underlying price S or in the volatility σ has a positive impact on the value of the call, as it increases the probability that the option will expire in the in-the-money area (i.e. above K). Dividends decrease the price of the underlying (i.e. stock) and hence the value of the call option. The effect of time-to-maturity and risk-free interest rate is less clear.

If we take into account the no-arbitrage value of a forward contract written on the same index, $F = Se^{(r-\delta)\mathrm{T}}$, we can reformulate the BS formula as follows[2]

$$C = DF\,(F\,N(d_1) - K\;N(d_2)) \tag{7.2.4}$$

$$d_1 = \frac{ln(F/K) + (1/2)\sigma^2 T}{\sigma\sqrt{\mathrm{T}}} \tag{7.2.5}$$

$$d_2 = d_1 - \sigma\sqrt{\mathrm{T}} \tag{7.2.6}$$

where DF is the discounting term, representing the amount of money that has to be invested in the risk-free interest rate r in order to obtain 1$ after time T. This transformation proves to be more convenient than the original BS formula, as the forward price is a *well-defined* tradable quantity that incorporates all information

[2] See e.g. Hull (1998).

about the prevailing interest rate and the dividends. Formula 7.7.3 forms our basis for setting-up a neural network mapping from (F, K, DF, σ) to the target value C.

Although pioneering in its conception, the BS model has been empirically shown to suffer from systematic biases when compared to market prices (Bakshi et al. (1997); Cont and Forseca (1997)). Most of the biases steam from the fact that the development of the BS formula has been based on a set of assumptions that fail to hold true in practice. In an attempt to relax the BS assumptions, researchers have come up with a variety of other parametric option pricing models, such as the jump-diffusion (Merton (1976)), constant elasticity of variance (Cox et al. (1976)), and Hull and White (1987)'s stochastic volatility. Despite their analytical tractability, the majority of these models are often too complex to implement, have poor out-of-sample pricing performance and sometimes inconsistent with implied parameters (Bakshi et al. (1997)). In addition, they are often based on restrictive assumptions concerning the market infrastructure and/or investors' attitude, which are often questionable from a theoretical or empirical point of view.

7.3 Option pricing with semi-parametric neural network models

Computational intelligent models, like neural networks or genetic programming, seem to offer a promising semi-parametric alternative to option pricing. This is mainly due to their ability to approximate highly nonlinear relationships without relying on the restrictive assumptions concerning the time-evolution of the underlying price, the efficiency of the market, the rationality of agents, etc. Option pricing with neural networks, in particular, has attracted the interest of many practitioners and researchers worldwide. Hutchinson et al. (1994) were among the first to apply a neural network model to the pricing of S&P 500 futures options. In order to reduce the number of inputs to the neural net they applied the so called "homogeneity property" of the BS formula. Their research shows that the resulting network model can be used successfully out-of-sample for the pricing and delta-hedging of options. An approach analogous to the above was followed by Garcia and Gencay (2000) in pricing European S&P 500 index options for various periods between 1987 and 1994. Bennell and Sutcliffe (2004) used a neural network to price options on FTSE 100 index and concluded that this nonparametric model is superior to the BS formula for out-the-money options. Lajbcygier et al. (1996) price options on futures using a two-input (S/K,T) and a four-input (S/K, T, r, σ) model. Their results suggest that the four-input model outperforms both the two-input one and the BS model and it works extremely well for a reduced data region (i.e. for options near the money and short maturity). Yao et al. (2000) use neural nets to forecast option prices on NIKKEI 225 index futures[3]. Their work also shows that a neural network model can outperform the BS model in volatile markets, even when the parameter of volatility is not feeded as an input into the neural network.

[3] NIKKEI 225 is an unweighted index of the largest 225 shares traded on the Tokyo Stock Exchange.

7.4 Neural network identification strategies

One major problem with semi-parametric approaches is the selection of model architecture. Many studies on option pricing typically employ a heuristic criterion (such as cross validation or sequential pruning) to determine the optimal architecture of the neural net. This however raises doubts as to whether the model has managed to reproduce the pricing formula assumed by the market or simply overfitted the data.

Many popular NN specification techniques follow a “general-to-specific” or “top-down” approach in which one starts with a large model and applies appropriate techniques to remove “redundant” components, i.e. hidden neurons and variables, that do not contribute much to the prediction accuracy of the model. Another strategy is the so-called “bottom-up” or “simple-to-complex”. The idea is to start with the simplest (linear) model and gradually complicate the structure by adding neurons if nonlinearity exists in data. The procedure is sequential so that at each step two decisions are made as regards the specification of the neural network model: first whether to add an (extra) neuron and second which input variables to append to this neuron.

The neural network models considered in this work belong to the general class of single-hidden-layer feedforward NNs with a linear component. The performance measure used in the estimation of these models is either the mean squared error (MSE)

$$MSE = (1/T)\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \tag{7.4.1}$$

or the mean weighted squared error (MWSE)

$$MWSE = (1/T)\sum_{t=1}^{T}\omega_t(y_t - \hat{y}_t)^2 \tag{7.4.2}$$

where $\hat{y}_t$ denotes the neural network's forecast for y_t and $\omega = (\omega_{1,}\omega_2, \ldots, \omega_T)$ is a vector of weights with $\omega_t \geq 0$. In what follows, we shall discuss various model selection strategies assuming a *MWSE* criterion, as the *MSE* is a special case of *MWSE* with ω_t being equal to one for all observations.

7.4.1 Simple-to-complex

The first two approaches employed in our work use a *simple-to-complex* strategy to determine the optimal architecture for a neural net. Model selection in this case is guided by statistical procedures, such as sequential statistical hypothesis tests and information criteria.

Neural network identification based on sequential statistical tests

The first approach is the sequential testing procedure presented in section 4.5.1, suitably modified for the case where the network is estimated using weighted nonlinear least-squares. As the data of our option pricing problem are not chronologically ordered, certain statistical properties, such as heteroskedasticity, are not likely to occur here. Hence, in choosing the number of units in the NN model we used the standard version of the LM test of neglected nonlinearity, although robustified versions could be used instead. It should be noted here that the LM testing procedures presented in section 4.5.1 implicitly assumes a least-squares estimation method for the NN. A modification of the tests for weighted LS is possible if one initially multiplies $\hat{\epsilon}_t$, $\bar{x}_t$, $\nabla \hat{f}_t$ and z_t by $\sqrt{\omega_t}$, where ω_t is the weight applying to observation t.

Neural network identification based on information criteria

Information criteria, such as the AIC or SBIC, are often used in choosing between two models with different degrees of complexity. These are typically a decreasing function of the goodness-of-fit of a model and an increasing function of the number of parameters. They thus favour *parsimonious* models, models that attain a good fit with minimum complexity.

Our model selection strategy based on information criteria is similar to the statistical testing procedure described above instead that it decides on the grounds of an information criterion whether the model can be improved by adding more hidden neurons. The strategy was adapted from Anders and Korn (1999), although in our case we followed the "bottom-up" route to determine both the number of hidden units of the network *and* the optimal combination of hidden neurons and variables. Problems in applying information criteria to neural network model selection also arise if the model includes redundant neurons. Therefore information criteria can only be combined with a simple-to-complex strategy. In the comparison between NN models the extra neuron is approximated by a third-order Taylor polynomial, as in the sequential testing procedure.

The NN model-specification strategy based on information criteria is described below:

1. Start with estimating a linear model $y_t = \phi' \bar{x}_t$ by weighted ordinary least squares and choose the number of variables by means of an information criterion (i.e. AIC, SBIC).

2. Calculate the errors of the linear model and regress them on a third-order Taylor expansion of an additional hidden neuron for all combinations of explanatory variables. Compute the value of the information criterion for the errors of the linear model, assuming a zero number of parameters (an "empty"model), and compare it to that obtained from the new regressions.

3. If the value of either AIC or SBIC shows no improvement over the "empty" model stop. Otherwise, estimate a NN model with one hidden unit and the combination of variables showing the lowest value for the information criterion connected to it.

4. Repeat steps 2-3: compute new errors, judge the relevance of an additional hidden neuron on the basis of an information criterion and if necessary estimate enlarged models. The procedure stops when an additional hidden neuron does not lead to further improvements.

Model selection based on information criteria involves approximately the same amount of computation with statistical sequential testing, as only auxiliary regressions are run at each step of the process.

7.4.2 Neural network identification based on sequential pruning

The opposite route to the afore described methods is sequential pruning. According to this strategy, a model with a large number of hidden neurons is first estimated and the size of the model is subsequently reduced by applying an appropriate technique. In this paper we adopt a simple network-pruning heuristic proposed by Kaashoek and van Dijk (1998). It is based on the concept of *incremental contribution*, originally proposed by Theil (1971) as a method for choosing variables in a linear model. The main idea is to find how much of the variance of the target variable y_t is explained by inclusion of an additional neuron or explanatory variable, holding all other parts of the model constant. This approach is different in two aspects from the ones listed above. First, it follows the opposite direction in the specification of the NN model going from larger to smaller networks, which as noted in the previous section is not possible if one intends to apply statistical procedures. Second, it employs a simple heuristic rather than a solid statistical criterion to decide whether to remove a redundant component (neuron/variable) of the model. In this sense, it serves as a good benchmark to judge the performance of more sophisticated approaches.

Let $\hat{\epsilon}_t$ be the error of the full network and $\hat{\epsilon}_t^j$ the error of the network with the j-th neuron excluded. The *incremental contribution of hidden neuron* j is defined by $R^2_{full} - R^2_j$ where

$$R^2_{full} = 1 - \sum_{t=1}^{T} \hat{\epsilon}_t^2 \Bigg/ \sum_{t=1}^{T} \bar{y}_t^2 \tag{7.4.3}$$

$$R^2_j = 1 - \sum_{t=1}^{T} \left(\hat{\epsilon}_t^j\right)^2 \Bigg/ \sum_{t=1}^{T} \bar{y}_t^2 \tag{7.4.4}$$

and $\bar{y}_t$ denotes the demeaned value of y_t. Based on this definition, the network pruning method is implemented as follows:

1. Start with a large NN model, including h neurons.

2. Remove the i-th neuron, $i = 1, \ldots, h$, re-estimate the restricted models and compute the value of the percentage incremental contribution

$$\Delta R^2_j = \frac{R^2_{full} - R^2_j}{R^2_{full}} \tag{7.4.5}$$

1. If ΔR^2_j is smaller than a threshold TH, say 1%, remove the corresponding neuron from the model.

2. Repeat the above step until the incremental contribution of each neuron is significant.

3. Continue with variable selection by removing one input variable each time and comparing the correlation coefficients. Finally, remove all variables for which ΔR_k^2, $k = 1, ..., n$, is smaller than TH and re-estimate the restricted model.

Contrary to model selection based on sequential testing and information criteria, this method is much more computationally demanding, as for each comparison of the correlation coefficients the restricted model has to be re-estimated. Another drawback of pruning is that it estimates NN models with a large number of inputs and hidden units, which is problematic as quite often the optimisation algorithm converges to a local optimum.

7.5 Sample data

For the application and testing of the proposed methodologies, we used equity option contracts on the S&P 500 index of the New York Stock Exchange. S&P 500 index options are European-style options and considered to be among the most tradable index options worldwide in terms of liquidity. We obtained two "snapshots" of the option matrix, quoted on May, 17th 2002 and July, 29th 2002. Figure 7.1 depicts the market price of options quoted on May, 17th 2002 as a function of the time to maturity T and the strike level K. Both strike and option prices are given as a percentage of the underlying and maturities range from 0.083 to 5 years. Note that in the option surface the value of the call is a characteristic U-shaped function of the strike price that is more skewed for short maturities. Furthermore, the value is monotonically increasing with time-to-maturity, the slope being more pronounced for around-the-money options. These features are related to the "smile" and "skew" of the implied volatility surface as often mentioned in the literature (see e.g. Hull (1998); Jackwerth and Rubinstein (1996)).

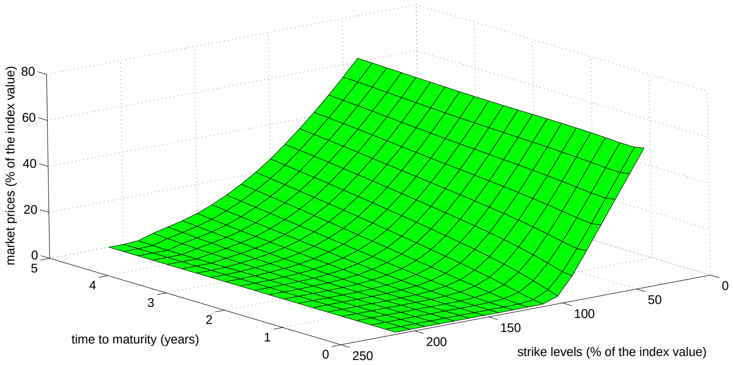

FIGURE 7.1: The S&P 500 call option surface on 17/05/2002.

7.6 Study A: Fitting the entire option surface

In the first set of experiments, we employed neural network models to fit the entire option surface. The call option matrix quoted on 17/05/2002 (a total of 378 observations) was used as a training set and the same amount of data corresponding to the option matrix on 29/07/2002 were used in out-of-sample evaluation of the NN models. Our purpose is to determine the value C of the call in terms of the forward price F, the strike level K, the discounting factor DF, and the volatility σ on that day, e.g

$$C_{NN} = f_{NN}(F, K, DF, \sigma) \tag{7.6.1}$$

Note that among all inputs, the volatility depends exclusively on the price of the underlying asset and it is thus constant for a given day. This poses a problem to our network approximation problem because one input variable takes the same value over the whole training set and is perfectly correlated with the constant of the linear model. As our focus is on capturing the shape of the option surface, volatility can be safely excluded from the set of input variables. Of course, if we possessed more option surfaces at different time-frames and the task was to forecast option prices in a time-series context then the dynamics of volatility would be an important explanatory factor.

All three model-building strategies were applied to the selection of the NN architecture. As a performance measure, we adopted the mean squared error between the call option market price and the network output. In order to avoid constantly firing hidden neurons at the saturation area, input data were standardised so that they have mean and standard deviation equal to zero and one, respectively. The Schwarz's Bayesian information criterion was used in the selection of parsimonious NN architectures. In the literature, SBIC has been found to deliver more conservative (i.e. less complex) models than the Akaike's information criterion. Due to the fact that more parsimonious models often outperform more complicated models when used in forecasting tasks and because the SBIC has been found to perform well in selecting forecasting models in other contexts (see e.g. Engle and Brown (1986)), we also adopt it in the present pricing task as a method for penalising the mean squared error.

The specification of the NN models indicated by each methodology is summarised in Table 7.1. Note that all models include from one to three neurons in the hidden layer and K and/or DF in the linear part. Interestingly, both sequential statistical testing (SST) and SBIC have chosen the same network architecture. Sequential pruning (SP), on the other hand, indicated a much simpler one-neuron NN model with only the strike price K contributing to nonlinearity. Table 7.2 shows the performance of SST and SP selection methods (IC is excluded as it indicated the same model as SST) for the training and test data sets. As measures of performance, we report the mean squared error, the mean absolute error and Schwarz's information criterion. Although the latter is commonly used in model selection tasks, we also adopted here as an indicator of the *parsimony* of the model (i.e. how much complexity was necessary to achieve the given fitting accuracy). Table 7.2 shows that the neural network specified by SST is better than SP in terms of in-sample and output-of-sample fit. The extra two neurons that sequential statistical tests indicated significantly increase the forecasting ability of the NN model both in the training and test set. In addition, and much more important, this decrease in model's error is enough to compensate for the additional complexity, as shown by the significantly lower value of SBIC in both sets. Hence, "bottom-up" approaches seem to deliver more parsimonious architectures than the heuristic-driven "top-down" pruning.

Table 7.2 also reports the corresponding fitting measures for the Black-Scholes (BS) model. This is mainly to get an idea of the relative advantage of modelling with a semi-parametric approach in comparison to a parametric market standard. Note that the BS formula requires an estimate for the volatility of the stock index at the day corresponding to the training and test option surface. Volatility forecasts were obtained by fitting a NN-GARCH model in the daily S&P 500 time-series. For comparison purposes we also employed a *exponential weighted moving average* (EWMA) model, suggested by J.P. Morgan in RiskMetrics^{TM} (Ris (1996)). This model tries also to incorporate the empirical observation that the time evolution of stock indexes is characterised by succeeding periods of relatively high and low variability. In this study, we used a time-period of almost 500 days preceding 17/05/2002 and 29/07/2002 to estimate the two models and then obtained one-day-ahead volatility forecasts. Both models gave similar results. The volatility estimates were $\sigma^2_{EWMA} = 0.107$ and $\sigma^2_{GARCH} = 0.127$ for 17/05/2002 and $\sigma^2_{EWMA} = 0.301$ $\sigma^2_{GARCH} = 0.281$ for 29/07/2002. We chose the volatility forecasts of the NN-GARCH model as input to the BS model, although similar results were obtained when feeding EWMA estimates.

Architecture		**Sequential Statistical Testing**	**Information Criterion**	**Sequential Pruning**
Linear Part		K, DF	K, DF	DF
Nonlinear Part	**Neuron 1**	F, K, DF	F, K, DF	K
	Neuron 2	K, DF	K, DF	-
	Neuron 3	K, DF	K, DF	-

Table 7.1: The architecture of neural network models obtained by sequential statistical tests (SST), Schwarz's information criterion (SBIC) and sequential pruning (SP). All models include from one to three neurons in the nonlinear part. Row two shows the variables of the linear part and rows three-to-five shows the variables connected to each neuron. Note that both SST and SBIC have chosen the same architecture for the neural network and SP has indicated a one-neuron feedforward NN model with only the strike price K contributing to nonlinearity.

Table 7.2, last column, shows the various measures of performance for the BS model. The value of the SBIC in this particular case was computed assuming a zero number of parameters; it is thus expected that our comparison will be slightly biased towards the BS model. Nevertheless, all NN models clearly outperform the BS model both in terms of goodness-of-fit and complexity-penalised mean squared error. It would be instructive to illustrate the quality of fitting obtained by a NN compared to the BS model. Figure 7.2 shows various intersections of the test option surface corresponding to a short (0.083 years), medium (2 years) and long (3 years) maturity. Solid lines denote market option prices, squared lines show the forecast of the SST-NN model and circled lines correspond to the forecast of the BS model. Observe that the BS model tends to underprice in-the-money options, especially at shorter maturities, and to overprice out-of-the money options at longer maturities. The NN model, on the other hand, manages to closely follow the gradual decrease in the slope of the skew as maturity gets longer.

Data Set	**Measure**	**SST**	**SP**	**BS**
Training Set	MSE	0.331	5.215	30.281
	MAE	0.319	1.761	4.250
	SBIC	-1.219	1.730	3.411
Test Set	MSE	0.440	4.427	24.353
	MAE	0.491	1.621	3.839
	SBIC	-0.570	1.566	3.193

Table 7.2: The performance of the SST and SP selection methods in the training and testing data set. We report three measures of the goodness-of-fit and parsimony of the models: the mean squared error, the mean absolute squared error and Schwatz's information criterion.

7.7 Study B: Fitting restricted areas of the option matrix

When applying a neural net to option pricing tasks, it should be noted that not all areas of the option surface are equally "dense", in the sense that most liquid options are traded in certain strike-maturity pairs. These are mainly characterised by short-to-medium maturity and strike prices around 100. The absence of an active market usually gives rise to unpleasant phenomena, such as bid/ask spreads and longer average times between consecutive transactions. This means that reliable training data are available for certain areas of the option surface, which poses a problem to the application of a NN, and in fact of any semi-parametric technique. As the out-of-sample performance of semi-parametric methods is very much determined by the quality of the training data, it is expected that a NN would have a lower generalisation ability and higher prediction error for non-liquid areas of the option matrix.

In addition to liquidity concerns, it should be noted that not all areas of the option matrix are equally interesting to traders. Typically, the value of deep out- or deep in-the-money options is more or less predictable, especially at short maturities, and the majority of trading activity mainly takes place in around-the-money options not so close to expiration. Using standard financial arguments of arbitrage, we can approximately derive the value of a call option that is "deep" in- or deep out-of-the money and close to expiration (i.e. $|K - 100| > 50$ and $T < 1$). Starting with in-the-money options, if $S \gg K$ the holder of the option can benefit from paying $K\$$ to buy the underlying from the counter-party and then sell it at the market at

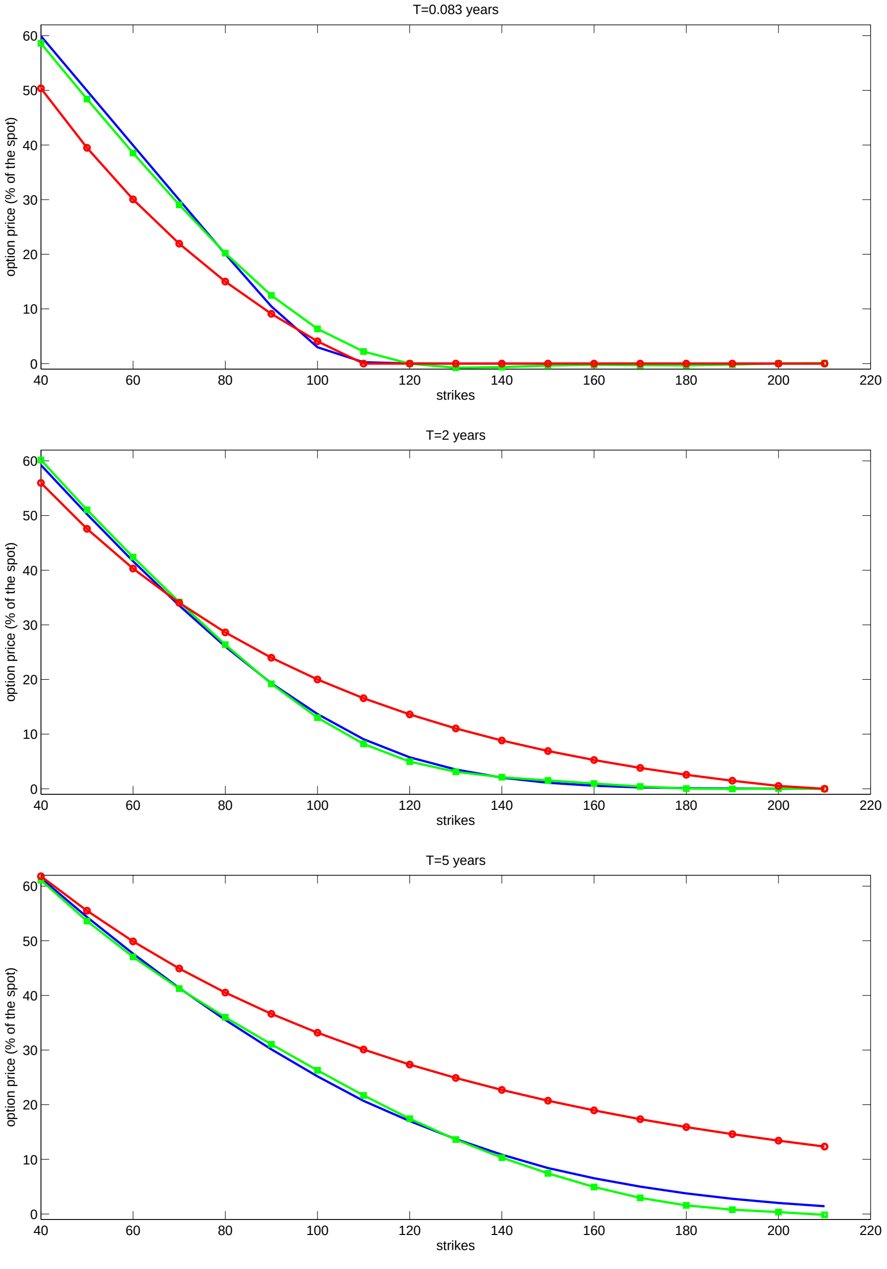

FIGURE 7.2: The quality of fitting in the test option matrix. We present three intersections of the option surface corresponding to a short (0.083 years), medium (2 years) and long (5 years) maturity. Pure line shows the option market prices, squares show the forecast of the SST-NN model and circles show the forecast of the BS model.

a higher price. This results in an immediate profit of $S - K$. But, $S = Fe^{(\delta - r)\mathrm{T}}$and for options with short time to maturity ($T \ll 1$), $e^{(\delta-r)T} \approx 1$. Hence, the value of a deep in-the-money option is approximately equal to $F - K$.

Using a similar argument, we can show that the value of a deep out-of-the money option is close to 0. If $S \ll K$ it is not worth exercising the option, as the underlying is sold in the market at a much lower price. Furthermore, none will be interested in buying a deep out-of-the-money option, rendering its value close to zero. For options that are around-the-money ($S \approx K$), their value cannot be easily determined as it depends on the short-term volatility of the asset. If S is currently slightly below K and the underlying is highly volatile, then it is likely that S will cross the K-threshold and the option become valuable.

From the above discussion, we infer that it makes more sense to direct a NN model to short-maturity at-the-money options and gradually increase the range of strikes as maturity gets longer. This was actually the pricing approach adopted in Tzastoudis et al. (2006). In this case, the price of a call option is defined as the outcome of the hybrid model:

$$C = \begin{cases} F - K, & S \gg K,\, T \in [T_1, T_2] \\ f_{NN}(F, K, DF), & S \approx K,\, T \in [T_1, T_2] \\ 0, & S \ll K,\, T \in [T_1, T_2] \end{cases} \tag{7.7.1}$$

where f_{NN} denotes the NN component. One way to implement the above model is to employ a mean weighted squared error objective function for the NN

$$MWSE = \sum_T \sum_K \omega_{K,T} \left(C^{K,T}_{market} - C^{K,T} \right) \tag{7.7.2}$$

where each $\omega_{K,T}$ denotes the weight assigned to the strike-maturity pair (K, T). In our research, we experimented with three matrices of weights, W_1, W_2 and W_3 that cover a range of 60-150% of strike levels and specific ranges of maturities aiming at short-, middle- and long-horizon fitting. The three weighting schemes are presented in Table 7.3. The graduation of weights reveals the importance assigned by experts to each pair of the option matrix.

The three methods described in section 7.4 were applied to the selection of the NN architecture using a mean weighted squared error criterion. For the bottom-up model selection approach based on information criteria, we adopted the following formula for the SBIC applying to weighted least squares

$$SBIC = \log(MWSE) + p\log(T)/T \tag{7.7.3}$$

where $MWSE$ is defined as above and p is the number of free parameters in the model.

Table 7.4 presents the architecture of the NN indicated by each methodology for the three weighting schemes. Column three shows the variables of the linear part and columns four to six the number of hidden neurons and the variables connected to each of them. Note that directing neural network to restricted areas of the option matrix generally results in a simplification of the architecture of the model. Fewer neurons are needed in this case than those employed in study A to represent the entire

Strike Levels	**Weights**											
60	0	0	0	0	0	0.5	0.5	0.7	0.7	0.7	0.7	0.7
70	0	0	0.5	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
80	1.0	1.0	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
90	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
100	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
110	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2.0	2.0	2.0	2.0
120	1.0	1.0	1.2	1.2	1.2	1.2	1.2	1.2	1.5	1.5	1.5	1.5
130	0	0.5	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
140	0	0	0.5	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
150	0	0	0	0	0	0.5	0.5	0.7	0.7	0.7	0.7	0.7
Ranges of Maturities												
R$_1$	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00
R$_2$	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75
R$_3$	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00

Table 7.3: The three matrices of weights corresponding to a short-, middle- and long-maturity fitting (R$_1$, R$_2$ & R$_3$). Pairs that do not appear in the table are assigned a weight of zero.

option matrix. The only exception is models generated by SBIC. The specification of models at all maturity ranges are very similar, possibly because the shape of the fitted surface for around-the-money options does not significantly alter from short to long maturities.

Weighting scheme	**Models**	**Linear part**	**Nonlinear part**		
			Neuron 1	**Neuron 2**	**Neuron 3**
W$_1$	**SST**	K	K, DF	-	-
	SBIC	K	F, K	F, K	F, K
	SP	K, DF	K	-	-
W$_2$	**SST**	K	F, K, DF	-	-
	SBIC	K	F, K, DF	K, DF	F, K, DF
	SP	DF	K	-	-
W$_3$	**SST**	F, K, DF	DF	-	-
	SBIC	K	K, DF	K, DF	K, DF
	SP	K, DF	K	-	-

Table 7.4: The architecture of the hybrid NN models used in fitting restricted areas of the option matrix, designated by the three weighting schemes W$_1$, W$_2$ and W$_3$ (see also table 7.3). Column three shows the variables of the linear part and columns four to six the number of hidden neurons and the variables connected to each of them. Statistical sequential testing (SST) and sequential pruning (SP) choose one hidden neuron for all weighting schemes, while the number of neurons specified by the SBIC is three.

Table 7.5 reports goodness-of-fit and parsimony measures for the various NN selection methods in the training and test data set. In restricted areas fitting, all models are of almost equal performance and none seems to clearly excel the others. Only SST-NN models seem to score better in the second range of maturities (W_2). Experiments conducted with various weighting schemes have shown that the performance of the NN models is sensitive to the placement of weights and a different weighting scheme typically results in different values for MWSE, MWAE and SBIC. The architecture of models, on the other hand, is partially unaffected by shifts in weights except when the area where the NN is effective (i.e. the area with higher weights) gets wider. All NN model-selection strategies tend to indicate more complex pricing formulae as higher weights are assigned to far from-the money areas with previously low or zero importance. This is also because the effective number of observations presented to the network (i.e. those with non-zero weight) increases and nonlinearity becomes more apparent with a larger sample of data. Note that for the weighting schemes presented in table 7.3, only 105 out of the 378, i.e. less than half, input data are used in training and testing.

7.8 Summary and further research

This chapter compared a number of neural network model selection approaches to pricing options on the S&P 500 stock index. Generally, the accurate forecasting of an option matrix is a difficult task though very important to traders and investors in option markets. The price of an option is determined by multiple factors through highly nonlinear relationships. Thus, the application of a semi-parametric intelligent technique with high approximation ability, such as a neural network, is very much recommended. However, most of these techniques have an open architecture and much effort is needed to determine the structure of the model that is appropriate for the particular application domain or data set.

In our approach, we attempted to treat model selection in a more systematic way. For the choice of the optimal architecture of the neural network model, we experimented with a "top-down" pruning technique and two "bottom-up" specification strategies that start with simple models and gradually complicate the architecture, if data indicate so. Apart from heuristic-driven approaches, we also employed methods that base model selection on statistical hypothesis tests and information criteria. In the first set of experiments, NN models were employed to fit the entire option surface and in the second one they were used as parts of a hybrid scheme directed to capture certain areas of the surface.

Results from both experiments generally indicate that "bottom-up" approaches outperform the "top-down" heuristic both in terms of in-sample error and out-of-sample forecasting accuracy. In particular, architecture selection based on sequential statistical tests seems to deliver the most parsimonious structures that combine good out-of-sample performance with minimum model complexity. Both simple-to-complex strategies considered in this paper are also much more preferred in terms of the computational burden needed to reach the final specification. In each method, subsequent training of enlarged NN models is avoided and the decision of whether to add an extra neuron is based on auxiliary regressions whose implementation is straightforward.

The experiments performed in this study can be extended in various directions. First of all, additional model selection approaches could be employed and more comparative results be obtained. A more interesting task would be to employ different NN models to learn a generalised option pricing formula in a time-series context. For this purpose, one needs various snapshots of the option surface taken at different time-frames. It is our belief that in this case the estimated volatility would be an important factor in determining the time-evolution of option prices.

Weighting Scheme $\mathbf{W}_1$					
		SST	**IC**	**SP**	**BS**
Training Set	**MWSE**	1.205	0.913	1.034	4.016
	MWAE	0.601	0.519	0.584	1.001
	SBIC	0.336	0.257	0.183	1.390
Test Set	**MWSE**	1.383	1.566	1.221	10.242
	MWAE	0.660	0.695	0.650	1.846
	SBIC	0.473	0.797	0.349	2.327
Weighting Scheme $\mathbf{W}_2$					
		SST	**IC**	**SP**	**BS**
Training Set	**MWSE**	1.814	1.837	1.871	3.822
	MWAE	0.795	0.769	0.793	1.010
	SBIC	0.770	1.006	0.751	1.341
Test Set	**MWSE**	1.636	1.925	2.051	14.960
	MWAE	0.769	0.790	0.841	2.322
	SBIC	0.667	1.053	0.843	2.705
Weighting Scheme $\mathbf{W}_3$					
		SST	**IC**	**SP**	**BS**
Training Set	**MWSE**	4.674	4.725	4.741	4.027
	MWAE	1.243	1.220	1.237	1.105
	SBIC	1.692	1.901	1.706	1.393
Test Set	**MWSE**	4.882	4.926	4.936	20.78
	MWAE	1.283	1.258	1.263	2.816
	SBIC	1.735	1.943	1.746	3.034

Table 7.5: The performance of the hybrid NN models used in fitting restricted areas of the training and test option matrix, designated by the three weighting schemes W_1, W_2 and W_3 (see also table 7.4).

Chapter 8

Conclusions and further research

In this thesis, we showed how computational intelligent models (artificial neural networks) and econometrics GARCH parametrisations can be combined into a flexible modelling framework that can accommodate most of the stylised facts associated with financial time-series (nonlinearities in mean, asymmetric GARCH effects in variance and possibly non-gaussian errors). By jointly modelling the conditional mean and volatility of the data-generating process, we manage to extend the scope of NNs from function approximation to *density forecasting* tasks and thus investigate interesting dependencies on higher-moments of the conditional distribution.

As with every flexible class of models, the issue of carefully selecting the final specification becomes of paramount importance. The analysis of the statistical properties of NN-GARCH models revealed that any combination of neural networks with GARCH parametrisations is not guaranteed to be successful unless special attention is paid to the specification of the mean and the variance equation. In fact, if nonlinear dependencies in data are due to GARCH effects the placement of extra hidden neurons in the mean model renders some of the parameters of the model redundant, which destroys the asymptotic normality of the maximum likelihood estimator and leads to poor out-of-sample performance. On the other hand, it is not possible to consistently estimate a GARCH model unless the specification of the mean equation is adequate for the conditional mean. Therefore, in a practical application it is important for the researcher to have a systematic procedure that can decide on the level of complexity to be placed on each side of the model, according to special features existing in data.

Following the principles underlying the construction of econometric models, we propose a complete model-building cycle that comprises specification, estimation and evaluation of the model. For the determination of the number of hidden neurons in the mean and the level of complexity in the variance equation, we follow a sequential testing procedure which avoids many statistical and numerical problems arising from the non-identifiability of neural networks. Based on the maximum likelihood theory, we device Wald-type tests for testing the *joint significance* of parameters of an estimated NN-GARCH model and thus offer the opportunity to the researcher to investigate hypotheses of interest regarding the nature of the underlying statistical process. We also present a series of in-sample *diagnostics* for the mean and variance equation that examine the extent to which the derived specification is a faithful approximation to the conditional distribution. These are very important as far as the quality of approximation as well as the forecasting performance of the model are concerned. Among the various cases of structural misspecification that we consider are remaining autocorrelation in mean, remaining autocorrelation in variance, nonlinearity in mean and asymmetric variance effects. The distinguishing feature of significance or diagnostic tests is that they lead to valid inference even in the case where the distributional assumptions made by the model are *not* correct (i.e. the empirical density of standardised errors is fat-tailed or asymmetric). This is a quite useful feature as it permits statistical inference without having to explicitly model all aspects of the conditional distribution.

The statistical procedures considered in this thesis are generally simple and inexpensive to construct, as they directly apply to a pre-estimated model, and only require the computation of first derivatives and a set of auxiliary regressions to determine whether the residuals (or the standardised residuals) contain additional features conjectured by the alternative hypothesis. Most important, the validity of tests does *not* depend on restrictive assumptions, such as homoskedasticity or normality of errors, holding in addition to the null hypothesis being investigated. This makes possible to control the empirical type I error of the test (i.e. the probability that it mistakenly rejects the null hypothesis) without having to explicitly model all aspects of the conditional distribution.

A considerable part of this thesis is devoted to investigating the finite-sample-size performance of testing procedures in a simulation environment that resembles most of the statistical features observed in financial data (nonlinearity, heteroskedasticity, non-normality). The results of simulation studies show that the non-robust version of the LM test cannot distinguish between mean dependencies and changing variance levels. The empirical size of the test is generally distorted under ARCH heteroskedasticity, leading to excessively false indications of nonlinearity or serial correlation in errors. This is a cautionary remark against using non-robust statistical procedures for testing NN specifications. Similar problems quite probably arise in other neural network specification or diagnostic procedures proposed in the literature (information criteria, pruning heuristics, etc), although general statements about the magnitude of the size and power distortions that can be expected in each case are difficult to come by. This issue can be investigated by means of further Monte Carlo simulations.

The robustification of testing procedures considered in this thesis, along the lines of Wooldridge (1990, 1991), allows the researcher to closely follow the nominal type I error when investigating certain hypotheses on the conditional mean, without having to explicitly model the variance structure or the distribution of standardised errors. Incorporating estimates of the conditional volatility of errors into the testing procedure, gives the researcher an option for increasing the efficiency (or else the power) of the conditional mean test in detecting hidden nonlinearity or serial correlation in the residuals. The forecasting exercise presented in chapter 6 shows that robustified tests are generally able to distinguish between nonlinearities in mean and ARCH effects, something which is not feasible with the non-robust LM test.

After establishing a complete model-building cycle for the family of NN-GARCH specifications, our next goal is to compare the performance of sequential statistical tests with other statistical or non-statistical procedures used in the design of neural network GARCH models. This comparison can be made on the basis of simulated as well as real data In parallel, work has to be done on the asymptotic statistical theory of the model, including conditions for the existence of moments and for consistency and asymptotic normality of the QML estimator. These are very important as far as the validity of diagnostic tests is concerned. Some specialised domains of application are discussed below.

Statistical arbitrage

An interesting application area for NN-GARCH models is related to the detection of statistical mispricings in a group of assets. Several authors have suggested approaches that attempt to take advantage of price discrepancies by taking proper transformations of financial time-series and creating "synthetic" assets; see e.g. Burgess (2002, 2000); Towers (2002) for stocks of FTSE 100 Burgess and Refenes (1996); Garrett and Taylor (2001) for equity index futures and Steurer and Hann (1996) for exchange rates. Amongst them, Burgess and Refenes (1996); Burgess (2002); Steurer and Hann (1996) note that the correction of statistical mispricings is often characterised by strong nonlinearity and hence employ neural network autoregressions to model the dynamics of statistical mispricings. This is to obtain an idea of how mispricings of different size and sign (positive/negative) are on average corrected over time.

In Thomaidis et al. (2006) an intelligent NN-based system was also used in the detection of mispricings in a pair of two closely related stocks. The mispricing series $\{z_t,\, t = 1, 2, ...T\}$ was constructed by comparing the price of each stock to a fundamental value indicator based on the price of the other stock. A statistical analysis of mispricings at a high (intra-day) sampling-frequency revealed to us that the volatility of z_t is not constant over time but strongly depends, in an ARCH fashion, on the history of mispricings. Any changes in the short-term volatility of the synthetic asset have important implications for the risk control of statistical arbitrages; hence a combined neural network-GARCH autoregressive model was used to model both nonlinearities in the correction of z_t as well as volatility clustering. The trading strategy employed in that work is similar to the one presented in section 6.5 and is based on the idea of taking proper positions on the constituent assets when mispricings become exceptionally high or low.

At the moment, our statistical arbitrage detector comprises only symmetric GARCH parametrisations for the modelling of the conditional variance, although it shows an improvement over a pure NN or a linear AR-GARCH model. An immediate extension would be to introduce asymmetric models for the volatility, such as EGARCH, GJR-GARCH, etc. In this way, the model could track more efficiently short-term changes in the uncertainty associated with any deviation of the synthetic from the mean and hence have a better control over the risk associated with the trading strategy.

Semi-parametric estimation of volatility

Although the initial version of the proposed NN-GARCH model includes parametric variance models, such GARCH or EGARCH, it can be easily extended to include more flexible parametrisations of the variance equation that do not make explicit assumptions on the nature of asymmetries in the volatility-generating process. One of our future research directions is to extend the neural network model from the mean to the variance part, thus creating a combined NN-NNGARCH model. This is also the tendency in other nonlinear time-series methodologies (see e.g. Lundbergh and Teräsvirta (1998) for smooth transition mean-variance models). To avoid problems of non-identifiability, an incremental strategy based on sequential LM tests could also be used to determine the number of hidden neurons in the volatility process. First experiments conducted on this type of models showed that a bottom-up approach similar to that used in the specification of the mean is advisable, as the inclusion of many neurons in the volatility equation quite often leads to numerical problems in the optimisation of the likelihood as well as instability of the variance process. However, an immediate application of mean nonlinearity tests based on third-order Taylor series approximations of extra neurons may not be appropriate in this case, as further conditions on the parameters need to be imposed to guarantee positive and non-explosive variance estimates.

Interdependencies between financial markets

In recent years, global markets tend to become more integrated as a result of a broad tendency toward liberalisation and deregulation in the capital markets of developed as well as developing countries. Empirical research documents a significant level of interdependence among international markets, with the U.S. market being one of the most influential in the world (see e.g. Hamao et al. (1990); Koch and Koch (1991); Masih and Masih (1997); Berben and Jansen (2005)). Hence, in order to capture the whole picture of the price formation mechanism, it seems more appropriate to depart from pure time-series models and include other exogenous economic variables that could possibly help in forecasting the target variable.

Interdependencies between domestic and international markets are often complex and can manifest themselves in various ways (see e.g. Thomaidis et al. (2005b) for a discussion). Short-term return transmissions are perhaps the most common type of dependencies been investigated in the literature. Transmission of returns takes place between two markets when past returns on one index can one average forecast future returns on the other. In this case, a *causality* relationship exists between the two indexes and we say that the one index "causes" the other. Causality may or may not be bi-directional and can vary substantially in time.

Apart from short-term dependencies, markets are often tied-up in long-run relations that act as an *attractor* or *equilibrium* to individual series. Whenever short-run deviations are observed, markets are expected to react in the direction that restores the equilibrium. Detecting equilibrium relations may be useful in predicting the long-run behavior of a group of markets, hence long-run relations can be combined with short-term causalities to better describe the dynamic response paths of individual series.

Perhaps, the most common type of models used to investigate the short-run dynamics are the so-called *equilibrium correction* models (see e.g. Engle and White (1999); RSAS (2003); Thomaidis and Dounias (2005)). Assume two variables x_t and y_t, possibly representing the value of two stock indexes at time t, and another variable z_t that measures temporal deviations from the long-run relation. A typical *equilibrium correction* model is specified as:

$$\Delta y_t = a_0^1 + \gamma_1 z_{t-1} + \sum_i a_i^1 \Delta y_{t-i} + \sum_j b_j^1 \Delta x_{t-j} + e_t^1 \tag{8.0.1a}$$

$$\Delta x_t = a_0^2 + \gamma_2 z_{t-1} + \sum_i a_i^2 \Delta y_{t-i} + \sum_j b_j^2 \Delta x_{t-j} + e_t^2 \tag{8.0.1b}$$

where a's b's and γ's are free parameters, Δ denotes the first difference operator ($\Delta x_t = x_t - x_{t-1}$) and e_t^1, e_t^2 are two (not necessarily independent) innovation processes. According to the above specification, market Y is not caused by X if the coefficients of Δx_{t-j} in the first regression, including γ_1, are *jointly* statistically insignificant. In other words, a causality test is equivalent to testing the hypothesis

H_0: b_j^1's and γ_1 are altogether 0

against the alternative

H_1: at least one b_j^1 or γ_1 is different from 0

Such tests can be easily designing using an F or Wald statistic.

Note that causality is immediately implied if there exists an equilibrium between the two time-series. Since series are tied up to a long-run relation, at least one of them must "take care" of preserving this relationship by reacting in such a way that corrects temporal deviations.

The majority of causality tests applied in studies of international equity markets are based on linear regression models such as (8.0.1), further equipped with an error correction term if cointegration exists. However, such models cannot capture nonlinear influences between equity markets, which might be especially profound in periods of instability and crises. Hence, (8.0.1) is in fact a device for testing *linear* short-term influences and rejection of the null does *not* imply lack of causality but rather lack of *linear* causality.

In order to draw finer conclusions on the nature of short-term dynamics between the two markets, one can use a wider class of regression models:

$$\Delta y_t = g^1(z_{t-1}, \text{lagged}(\Delta y_t, \Delta x_t)) + e_t^1 \tag{8.0.2a}$$

$$\Delta x_t = g^2(z_{t-1}, \text{lagged}(\Delta y_t, \Delta x_t)) + e_t^2 \tag{8.0.2b}$$

where g^1 and g^2 are two possibly nonlinear mappings. In Thomaidis et al. (2005b) we employed neural network regression models of the form (4.2.1a) to approximate g^1 and g^2 in a study of the cross-dynamics between the German and French equity market. We investigated causalities and long-run equilibria in daily values of DAX and CAC indexes from July 1987 to March 2005, by slitting the whole sampling period into four subperiods characterised by important economic and political events. By means of Wald significance tests on proper sets of parameters in the linear and the neural network part of the models[1], we investigated a variety of hypotheses concerning the nature of inter-market dynamics:

- Do lagged returns on one index provide any predictive information about returns on the other?
- Is this information mainly the result of linear or nonlinear reactions?
- Which market is responsible for preserving the long-run equilibrium (whenever existed)?
- How does each market respond to temporal deviations from the equilibrium?

The results of this study showed that interdependencies between indexes are varying in nature and highly dependent on the historic period under study. Frequently, linear with nonlinear effects are combined to create a complex dynamic behaviour.

We believe that this study reveals a broader utility of artificial neural networks, and in fact any intelligent learning technique, which goes far beyond the estimation of data relationships. From a statistical perspective, NNs can be considered as semi-parametric devices for testing interesting hypotheses regarding the nature of financial/economic phenomena. As NNs have the ability to extract complex nonlinear interactive effects between the variables of interest, they can be utilised in testing hypotheses that are possibly beyond the reach of traditional linear models[2]. The separation between linear and nonlinear causalities, discussed in the above study, is a good an example.

Multivariate density models

It is important to note that mean causalities are not the only type of information exchanged between markets. Interdependencies are also observed among the second moment of the distribution of returns, i.e. the volatility. Empirical observation confirms that a sudden price increase or drop in one market can have an impact on the short-term volatility of the domestic *as well as* "neighbouring" markets[3]. This fact adds additional levels or dimensions into the analysis of the cross-dynamics between international markets and this is where a NN-GARCH modelling framework can be exploited. In order to model the transmission of news and volatility between markets, one could incorporate additional GARCH or ARCH terms, corresponding to exogenous influences, in the volatility equation of each market.

One of our future research objectives is to extend the two-equation model, presented above, so that it can cover a broader range of markets. Apart from equities, our financial network could also include bond and foreign exchange markets, which seem to be an important determinant of the course of stocks. Much of the methodology presented above can be almost directly applied to more than two time-series,

[1] See e.g. Kuan and White (1994) for the application of the Wald test in a neural network regression model.

[2] See White and Racine (2001) for a discussion on the use of NNs in statistical inference.

[3] This is phenomenon is often termed in the literature as *volatility spill-overs* (see e.g. Engle (1987)).

although in this case more sophisticated tools are needed for detecting long-run equilibria between more than two variables[4].

Another important research direction is concerned with a *behavioural* analysis of inter-market dynamics. Knowing how markets behave over time, especially in extreme events, gives us important insight into their functioning and helps to predict their reaction to similar, low-probability, events that may occur in the future. A behavioural analysis of financial markets attempts to address many interesting questions, such as how much time it takes on average until a shock is absorbed, how shocks of different magnitude or sign ("good/bad news") are propagated through the financial network, how long-run equilibriums are restored, etc. The ultimate goal of this research direction is to create statistical models that can capture the main features of cross-market dependencies, as those are imprinted in the *joint* empirical distribution of returns. These models can in turn be used as Monte Carlo *simulators* of realistic market scenarios upon which other financial engineering tasks can be based, such as the estimation of financial risk (relevant to trading purposes) and the pricing of more advanced financial products (derivatives).

[4] See e.g Enders (1995) and the references therein.

Appendix A

Analytical derivatives of the NN-GARCH model

1.1 Introduction

In appendix A we provide analytical expressions for the gradient and the hessian of the log-likelihood function of a NN-GARCH model, under the assumption of a normal distribution of errors. These are very useful in optimisation routines or in the computation of test statistics. Section 1.2 computes the gradient and section 1.3 the hessian of the log-likelihood function. For purposes of brevity, we adopt the following notation in the subsequent formulae: for a scalar function $f(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, where $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ and $y = (y_1, \ldots, y_m)' \in \mathbb{R}^m$ are two vector variables, we denote by $\nabla_x f$ the gradient of $f(.)$ across the direction of x and by $\nabla^2_{xy'} f$ the xy block of the hessian matrix of $f(.)$.

1.2 Gradient of the log-likelihood

$$\nabla_\delta l = (1/T) \sum_{t=1}^{T} \nabla_\delta l_t$$

$$\nabla_\delta l = \left(\nabla'_\phi l_t\, , \, \nabla'_\theta l_t, \, \nabla'_\alpha l_t\right)$$

$$\begin{aligned}
\nabla_\phi l_t &= \frac{\epsilon_t}{h_t} \bar{x}_t + \frac{1}{2h_t} \left(\frac{\epsilon_t^2}{h_t} - 1\right) \nabla_\phi h_t \\
\nabla_\theta l_t &= \frac{\epsilon_t}{h_t} \nabla_\theta f_t + \frac{1}{2h_t} \left(\frac{\epsilon_t^2}{h_t} - 1\right) \nabla_\theta h_t \\
\nabla_\alpha l_t &= \frac{1}{2h_t} \left(\frac{\epsilon_t^2}{h_t} - 1\right) \nabla_\alpha h_t
\end{aligned}$$

The gradient of the neural network part of the mean model is

$$\nabla_\theta f_t = \left(F_1, \ldots, F_h, \lambda_1 \nabla'_{w_1} F_1, \ldots, \lambda_h \nabla'_{w_h} F_h,\ \lambda_1 \nabla_{c_1} F_1, \ldots, \lambda_h \nabla_{c_h} F_h\right)'$$

where

$$F_j \equiv F(w'_j x_t - c_j), \quad \nabla_{w_j} F_j = x_t F'_j, \quad \nabla_{c_j} F_j = -F'_j$$

and

$$F'_j = F'[(w'_j x_t - c_j)] = F((w'_j x_t - c_j))[1 - F((w'_j x_t - c_j))] = 2\cosh((w'_j x_t - c_j))]^{-2}$$

The gradient of the volatility part for a GARCH model is

$$\begin{aligned}
\nabla_\alpha h_t &= (1, h^2_{t-1}, \ldots, h^2_{t-p}, \epsilon^2_{t-1}, \ldots, \epsilon^2_{t-q})' + \sum_{i=1}^{p} a_i \nabla_\alpha h^2_{t-i} \\
\nabla_\phi h_t &= \sum_{i=1}^{p} a_i \nabla_\phi h^2_{t-i} - 2 \sum_{j=1}^{q} b_j \epsilon_{t-j} \bar{x}_{t-j} \\
\nabla_\theta h_t &= \sum_{i=1}^{p} a_i \nabla_\theta h^2_{t-i} - 2 \sum_{j=1}^{q} b_j \epsilon_{t-j} \nabla_\theta f_{t-j}
\end{aligned}$$

1.3 Hessian of the log-likelihood

The hessian of the log-likelihood is

$$\nabla^2_{\delta\delta'} l(\delta) = (1/T) \sum_{t=1}^{T} \nabla^2_{\delta\delta'} l_t(\delta)$$

$$\nabla^2_{\delta\delta'} l_t = \left[\begin{array}{cc|c} \nabla^2_{\phi\phi'} l_t & \nabla^2_{\phi\theta'} l_t & \nabla^2_{\phi\alpha'} l_t \\ \nabla^2_{\theta\phi'} l_t & \nabla^2_{\theta\theta'} l_t & \nabla^2_{\theta\alpha'} l_t \\ \hline \nabla^2_{\alpha\phi'} l_t & \nabla^2_{\alpha\theta'} l_t & \nabla^2_{\alpha\alpha'} l_t \end{array}\right]$$

The above matrix is block-symmetric hence we only provide formulae for the upper diagonal blocks

$$\begin{aligned}
\nabla^2_{\phi\phi'} l_t = &-\frac{1}{h_t} \bar{x}_t \bar{x}'_t - \frac{1}{2(h_t)^2} \left(\frac{\epsilon^2_t}{h_t}\right) \nabla_\phi h_t \nabla'_\phi h_t \\
&- \frac{\epsilon_t}{(h_t)^2} \left(\nabla_\phi h_t \bar{x}'_t + \bar{x}_t \nabla'_\phi h_t\right) + \frac{1}{2h_t} \left(\frac{\epsilon^2_t}{h_t} - 1\right) \left(\nabla^2_{\phi\phi'} h_t - \frac{1}{h_t} \nabla_\phi h_t \nabla'_\phi h_t\right)
\end{aligned}$$

$$\begin{aligned}
\nabla^2_{\phi\theta'} l_t = &-\frac{1}{h_t} \bar{x}_t \nabla'_\theta f_t - \frac{1}{2(h_t)^2} \left(\frac{\epsilon^2_t}{h_t}\right) \nabla_\phi h_t \nabla'_\theta h_t \\
&- \frac{\epsilon_t}{(h_t)^2} \left(\nabla_\phi h_t \nabla'_\theta f_t + \bar{x}_t \nabla'_\theta h_t\right) + \frac{1}{2h_t} \left(\frac{\epsilon^2_t}{h_t} - 1\right) \left(\nabla^2_{\phi\theta'} h_t - \frac{1}{h_t} \nabla_\phi h_t \nabla'_\theta h_t\right)
\end{aligned}$$

$$\nabla^2_{\phi\alpha'} l_t = -\frac{\epsilon_t}{(h_t)^2}\bar{x}_t \nabla'_\alpha h_t - \frac{1}{2(h_t)^2}\left(\frac{\epsilon_t^2}{h_t}\right)\nabla_\phi h_t \nabla'_\alpha h_t + \frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t}-1\right)\left(\nabla^2_{\phi\alpha'} h_t - \frac{1}{h_t}\nabla_\phi h_t \nabla'_\alpha h_t\right)$$

$$\nabla^2_{\theta\theta'} l_t = -\frac{1}{h_t}\nabla_\theta f_t \nabla'_\theta f_t - \frac{1}{2(h_t)^2}\left(\frac{\epsilon_t^2}{h_t}\right)\nabla_\theta h_t \nabla'_\theta h_t + \frac{\epsilon_t}{h_t}\nabla^2_{\theta\theta'} f_t - \frac{\epsilon_t}{(h_t)^2}\left(\nabla_\theta h_t \nabla'_\theta f_t + \nabla_\theta f_t \nabla'_\theta h_t\right) + \frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t}-1\right)\left(\nabla^2_{\theta\theta'} h_t - \frac{1}{h_t}\nabla_\theta h_t \nabla'_\theta h_t\right)$$

$$\nabla^2_{\theta\alpha'} l_t = -\frac{\epsilon_t}{(h_t)^2}\nabla_\theta f_t \nabla'_\alpha h_t - \frac{1}{2(h_t)^2}\left(\frac{\epsilon_t^2}{h_t}\right)\nabla_\theta h_t \nabla'_\alpha h_t + \frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t}-1\right)\left(\nabla^2_{\theta\alpha'} h_t - \frac{1}{h_t}\nabla_\theta h_t \nabla'_\alpha h_t\right)$$

$$\nabla^2_{\alpha\alpha'} l_t = -\frac{1}{2(h_t)^2}\left(\frac{\epsilon_t^2}{h_t}\right)\nabla_\alpha h_t \nabla'_\alpha h_t + \frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t}-1\right)\left(\nabla^2_{\alpha\alpha'} h_t - \frac{1}{h_t}\nabla_\alpha h_t \nabla'_\alpha h_t\right)$$

Let A denote the minus expected hessian of the log-likelihood:

$$A = T^{-1} E\left(-\sum_{t=1}^{T} H_t\right)$$

Assuming that conditions 4.3.6 are fulfilled, i.e. the NN-GARCH model is structurally correct, and by the property $E(.) = E(E(.|x_t))$ of conditional expectations we have for the diagonal blocks of A

$$A_{\phi\phi'} = T^{-1}\sum_{t=1}^{T}\left(\frac{1}{h_t}\bar{x}_t\bar{x}'_t + \frac{1}{2(h_t)^2}\nabla_\phi h_t \nabla'_\phi h_t\right)$$

$$A_{\theta\theta'} = T^{-1}\sum_{t=1}^{T}\left(\frac{1}{h_t}\nabla_\theta f_t \nabla'_\theta f_t + \frac{1}{2(h_t)^2}\nabla_\theta h_t \nabla'_\theta h_t\right)$$

$$A_{\alpha\alpha'} = T^{-1}\sum_{t=1}^{T}\frac{1}{2(h_t)^2}\nabla_\alpha h_t \nabla'_\alpha h_t$$

and the off-diagonal blocks

$$A_{\phi\theta'} = T^{-1}\sum_{t=1}^{T}\left(\frac{1}{h_t}\bar{x}_t \nabla'_\theta f_t + \frac{1}{2(h_t)^2}\nabla_\phi h_t \nabla'_\theta h_t\right)$$

$$A_{\phi\alpha'} = T^{-1}\sum_{t=1}^{T}\frac{1}{2(h_t)^2}\nabla_{\phi}h_t\nabla'_{\alpha}h_t$$

$$A_{\theta\alpha'} = T^{-1}\sum_{t=1}^{T}\frac{1}{2(h_t)^2}\nabla_{\theta}h_t\nabla'_{\alpha}h_t$$

It can be proven along the lines of Engle (1982), theorem 4, that the off-diagonal blocks $A_{\phi\alpha'}$ and $A_{\theta\alpha'}$ are zero provided that the conditional variance model is symmetric (in the sense that the model responds similarly to positive and negative inputs of the same size) and satisfies certain regularity conditions. The classical GARCH specification is symmetric and satisfies the regularity conditions (see Bollershev (1986)) and so the combined NN-GARCH model presented above possesses this property.

Bibliography

Y. Abu-Mostafa, A. Atiya, M. Magdon-Ismail, and H. White. *IEEE Transactions On Neural Networks*, 12(4):653–949, 2001. Special Issue on Financial Engineering.

U. Anders and O. Korn. Model selection in neural networks. *Neural Networks*, 12: 309–323, 1999.

E. M. Azoff. *Neural Network time-series Forecasting of Financial Markets*. John Wiley & Sons, 1994.

G. Bakshi, C. Cao, and Z. Chen. Empirical performance of alternative options pricing models. *Journal of Finance*, 52:2003–2049, 1997.

N. Barberis and R. Thaler. A survey of behavioural finance. In G.M. Constantinides, M. Harris, and R. Stulz, editors, *Handbook of the Economics of Finance*. Elsevier, 2001.

K. Bartlmae and F.A. Rauscher. Measuring DAX market risk: A neural network volatility mixture approach. URL `http://www.smartquant.com/references/VaR/var15.pdf[November2004]`. Working Paper, 2000.

E. Barucci and L. Landi. A neural network model for short term interest rate forecasting: The 12-month bot italian auction rate. *Neural Network World*, 3:625–656, 1993.

Basle. Amendment to the capital accord to incorporate market risks. Technical report, Basle Committee On Banking Supervision, 1996. URL `http://www.bis.org[November2007]`.

R. Becker and A.S Hurn. Testing for nonlinearity in mean in the presence of heteroskedasticity. Discussion Papers, School of Economics and Finance, Queensland University of Technology, 2006. URL `http://ideas.repec.org/p/qut/sthurn/2006-02.html[February2007]`.

J. Bennell and C. Sutcliffe. Black-scholes versus artificial neural networks in pricing FTSE 100 options. *Intelligent Systems in Accounting, Management and Finance*, 12:243–260, 2004.

A. K. Bera and M. L. Higgins. ARCH models: Properties, estimating and testing. *Journal of Economic Surveys*, 7:305–362, 1993.

R. P. Berben and W. J. Jansen. Co-movement in international equity markets: a sectoral view. *Journal of International Money and Finance*, 24(5):832–857, 2005.

K. Bergerson and D.C. Wunsch. A commodity trading model based on a neural networkexpert system hybrid. In *Proceedings of the IEEE International Conference on Neural Networks, Seattle, WA*, volume 1, pages 1289–1293, 1991.

C.M. Bishop. Mixture density networks. Neural Computing Research Group Report (NCRG/94/04), Dept. of Computer Science and Applied Mathematics, Aston University, Birmingham, UK, 1994.

C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:673–654, 1973.

P. J. Bolland, J. T. Connor, and A.-P. N. Refenes. Application of neural networks to forecast high frequency data: Foreign exchange. In C. Dunis and B. Zhou, editors, *Nonlinear Modelling of High Frequency Financial time-series.* John Wiley & Sons, 1998.

T. Bollershev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 72:307–327, 1986.

T. Bollerslev. A conditionally heteroskedastic time-series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69(3):542–547, 1987.

T. Bollerslev, R. Y. Chou, and K. F. Kroner. ARCH modeling in finance : A review of the theory and empirical evidence. *Journal of Econometrics*, 52(1-2):5–59, 1992.

T. Bollerslev and J. M. Wooldridge. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric reviews*, 11 (2):143–172, 1992.

P. Bougerol and N. Picard. Stationarity of GARCH processes and of some nonnegative time-series. *Journal of Econometrics*, 52:115–127, 1992.

G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *time-series Analysis: Forecasting and Control.* Pentice-Hall International, Inc., 1994.

N. Burgess. Statistical arbitrage models of the FTSE 100. In Y. Abu-Mostafa, B. LeBaron, A. W Lo, and A. S. Weigend, editors, *Computational Finance 1999*, pages 297–312. The MIT Press, 2000.

N. Burgess. Cointegration. In J. Shadbolt and J. G. Taylor, editors, *Neural Networks and the Financial Markets: predicting, combining and portfolio optimisation*, pages 181–191. Springer, 2002.

N. Burgess and A.N Refenes. Modelling nonlinear cointegration in international equity index futures. In A. Refenes, Y. Abu-Mostafa, J. Moody, and A. Weigend, editors, *Neural Networks in Financial Engineering*, pages 50–63. World Scientific, 1996.

S. Chen, C. Yeh, and W. Lee. Option pricing with genetic programming. In *Genetic Programming 1998: Proceedings of the Third Annual Conference, San Francisco, CA*, pages 32–37. Morgan Kaufmann, 1998.

Sh.-H. Chen. *Genetic Algorithms and Genetic Programming in Computational Finance*. Kluwer Academic Pub, 2002.

Sh.-H. Chen, W.-Ch. Lee, and Ch.-H. Yeh. Hedging derivative securities with genetic programming. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 8(4):237–251, 1999.

Sh.-H. Chen and Ch.-H. Yeh. Toward a computable approach to the efficient market hypothesis: An application of genetic programming. *Journal of Economic Dynamics and Control*, 21:1043–1063, 1997.

Z. Chen. *Computational Intelligence for Decision Support*. The CRC Press International Series on Computational Intelligence. CRC Press, U.S.A., 2000.

V. Cherkassky, D. Gehring, and F. Mulier. Comparison of adaptive methods for function estimation from samples. *IEEE Transactions on Neural Networks*, 7(4): 969 – 984, 1996.

N. K. Chidambaran, C. W. J. Lee, and J. R. Trigueros. Adapting Black-Scholes to a non-Black-Scholes environment via genetic programming. In *Proceedings of the 1998 Conference on Computational Intelligence for Financial Engineering (CIFEr), 29-31 Mar 1998, New York, USA*, pages 197–211, 1998.

N. K. Chidambaran, C. W. J. Lee, and J. R. Trigueros. An adaptive evolutionary approach to option pricing via genetic programming. In *Proceedings of the 6th International Conference on Computational Finance*, 1999.

K. L. Chuah. *A Nonlinear Approach to Return Predictability in the Securities Markets Using Feedforward Neural Networks*. PhD thesis, Washington State University, 1992.

R. Cont and J. Forseca. Dynamics of implied volatility surfaces. *Quantative Finance*, 2:45–60, 1997.

B. Cornell and K. R. French. The pricing of stock index futures. *The Journal of Futures Markets*, 3:1–14, 1983a.

B. Cornell and K. R. French. Taxes and the pricing of stock index futures. *Journal of Finance*, 38:675–694, 1983b.

P. Cortez, M. Rocha, and J. Neves. Genetic and evolutionary algorithms for time-series forecasting. In *Proceedings of the 14th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems: Engineering of Intelligent Systems*, Lecture Notes In Computer Science, pages 393–402. Springer-Verlag, 2001.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman & Hall, London, 1974.

J. C. Cox, S. A. Ross, and A. Stephen. The valuation of options for alternative stochastic processes. *Journal of Financial Economics*, 3:145–166, 1976.

R. B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64:247254, 1977.

G. J. Deboeck and M. Cader. Trading U.S. treasury notes with a portfolio of neural net models. In G. J. Deboeck, editor, *Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial Markets.* John Wiley & Sons, NY, 1994.

F. X. Diebold. *Empirical Modeling of Exchange Rate Dynamics.* Springer-Verlag, New York, 1988.

R. G. Donaldson and M. Kamstra. An Artificial Neural Network-GARCH Model for International Stock Return Volatility. *Journal of Empirical Finance*, 4(1):17–46, 1997.

K. Dowd. *Beyond Value at Risk: The New Science of Risk Management.* John Wiley & Sons, 1998.

C. L. Dunis and X. Huang. Forecasting and trading currency volatility: an application of recurrent neural regression and model combination. *Journal of Forecasting*, 21 (5):317–354, 2001.

O. Eitrheim and T. Teräsvirta. Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics*, 74:59–75, 1996.

E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann. *Modern Portfolio Theory and Investment Analysis.* John Wiley & Sons, Inc., 6th edition, 2003.

W. Enders. *Applied Econometric Analysis.* John Wiley & Sons, 1995.

A. P. Engelbrecht. *Computational Intelligence: An Introduction.* John Wiley & Sons, 2003.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation. *Econometrica*, 50:987–1008, 1982.

R. F. Engle. Multivariate GARCH with factor structures-cointegration in variance. unpublished paper, Department of Economics, University of California, San Diego, 1987.

R. F. Engle and S.J. Brown. Model selection for forecasting. *Neural Computation*, 20:313–327, 1986.

R. F. Engle and V. K. Ng. Measuring and testing the impact of news on volatility. *Journal of Finance*, 48:1749–1777, 1993.

R. F. Engle and H. White, editors. *Cointegration, Causality, and Forecasting. A festschrift in Honour of Clive W. J. Granger*. Oxford University Press, 1999.

E. Fama. Efficient capital markets: A review of theory and empirical evidence. *The Journal of Finance*, 25:338–417, 1970.

E. Fama. Efficient capital markets: II. *The Journal of Finance*, 46(5):1575–1617, 1991.

E. Fama. Market efficiency, long-term returns, and behavioral finance. *The Journal of Financial Economics*, 49:283–306, 1998.

A. M. Farley and S. Jones. Using a genetic algorithm to determine an index of leading economic indicators. *Computational Economics*, 7:163–173, 1996.

J. D. Farmer and A. W. Lo. Frontiers of finance: Evolution and efficient markets. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 96, pages 9991–9992, 1999.

S. Figlewski. Explaining the early discounts on stock index futures: The case for disequilibrium. *Financial Analysts Journal*, 40:43–47, 1984.

T. L. Fine. *Feedforward Neural Network Methodology*. Springer, 1999.

A. Fiordaliso. A nonlinear forecasts combination method based on Takagi-Sugeno fuzzy systems. *International Journal of Forecasting*, 14:367–379, 1998.

K. R French, G. W. Schwert, and R. F. Stambaugh. Expected stock returns and volatility. *Journal of Financial Economics*, 19:3–29, 1987.

M. Furno. LM tests in the presence of non-normal error distributions. *Econometric Theory*, 16:249–261, 2000.

L. Galitz. *Financial Engineering: Tools and Techniques to Manage Financial Risk*. Irwin Professional Publishers, 1995.

A. R. Gallant, P. E. Rossi, and G. Tauchen. Stock prices and volume. *Review of Financial Studies*, 5(2):199–242, 1992.

R. Garcia and R. Gencay. Pricing and hedging derivative securities with neural networks and a homogeneity hint. *Journal of Econometrics*, 94:93–115, 2000.

I. Garrett and N. Taylor. Intraday and interday basis dynamics: Evidence from the FTSE 100 index futures market. *Studies in Nonlinear Dynamics nd Econometrics*, 5(2):133–152(20), 2001.

R. Gencay and M. Qi. Pricing and hedging derivative securities with neural networks: bayesian regularization, early stopping and bagging. *IEEE Transations on Neural Networks*, 12(4):726–734, 2001.

R. Gencay, F. Selcuk, and B. Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics.* Academic Press, 2001.

S. Ghoshray. Application of fuzzy regression models to predict exchange rates for composite currencies. In *Proceedings of the 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr), 24-26 Mar 1996, New York, USA*, pages 264–270, 1996a.

S. Ghoshray. Foreign exchange rate prediction by fuzzy inferencing on deterministic chaos. In *Proceedings of the 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr), 24-26 Mar 1996, New York, USA*, pages 96–102, 1996b.

L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48:1779–1801, 1993.

C. W. J. Granger and T. Teräsvirta. *Modelling Economic Nonlinear Relationships.* Oxford University Press, Oxford, 1998.

Ch. M. Hafner and H. Herwartz. Testing for linear autoregressive dynamics under heteroskedasticity. *Econometrics Journal*, 3(2):177–197, 2000.

Y. Hamao, R. W. Masulis, and V. Ng. Correlations in price changes and volatility across international stock markets. *Review of Financial Studies*, 3:280–307, 1990.

W. Hardle, H. Lütkepohl, and R. Chen. A review of non-parametric time series analysis. *International Statistical Review*, 65(1):49–72, 1997.

T. H. Harm and E. Steurer. Much ado about nothing? exchange rate forecasting: Neural networks vs. linear models using monthly and weekly data. *Neurocomputing*, 10:323–339, 1996.

P. Harrald and M. Kamstra. Evolving artificial neural networks to combine financial forecasts. *IEEE Transactions on Evolutionary Computation*, 1(1):40–52, 1997.

J. M. Harrison and D. M. Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20:381–408, 1979.

J. M. Harrison and S.R. Pliska. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*, 11:215–260, 1981.

B. Hassibi and D. Stork. Second order derivatives for network pruning: optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pages 164–171. 1993.

S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 1999.

D. F. Hendry. *Dynamic Econometrics.* Oxford University Press, 1995.

Y. Hong and H. White. Asymptotic distribution theory for non-parametric entropy measures of serial dependence. *Econometrica*, 73(3):837–901, 2005.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251257, 1991.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359366, 1989.

D. A. Hsieh. Testing for nonlinearity in daily foreign exchange rate changes. *Journal of Business*, 62:339–368, 1989.

D. A. Hsieh. Implications of nonlinear dynamics for financial risk management. *Journal of Financial and Quantitative Analysis*, 28:41–64, 1993.

J. C. Hull. *Options, Futures and Other Derivatives.* Pearson Education, 5th edition, 1998.

J. C. Hull and A. White. The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42:281–300, 1987.

J. M. Hutchinson, A. W. Lo, and T. Poggio. A non-parametric approach to pricing dericative securities via learning networks. *Journal of Finance*, 59:851–889, 1994.

J. T. G. Hwang and A. A. Ding. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92:109–125, 1997.

C. Jackwerth and M. Rubinstein. Recovering probabilities from option prices. *Journal of Finance*, 51:1611–1631, 1996.

R. Jang, Ch.-T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence.* Pearson Education, 1996.

F. Jawadi. A nonlinear measurement of adjustment of stock price towards equilibrium: Estimation of an ESTECM model. In *Proceedings of the Third International Finance Conference (IFC3), 3-5 March 2005, Tunish*, 2005.

Ph. Jorion. *Value at risk: the new benchmark for controlling market risk.* McGraw-Hill, 1997.

J. F. Kaashoek and H. K. van Dijk. A simple strategy to prune neural networks with an application to economic time-series. Econometric Institute Report 9854/A, Econometric Institute, Erasmus University, 1998.

R. P. Kanungo. Genetic algorithms: Genesis of stock evaluation. Technical report, Economics Working Paper Archive at WUSTL, 2004. URL `http://ideas.repec.org/p/wpa/wuwpex/0404007.html[November2004]`.

D. Kim and Ch. Kim. Forecasting time-series with genetic fuzzy predictor ensemble. *IEEE Transactions on Fuzzy Systems*, 5(4):523–535, 1997.

T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka. A stock market prediction system with modular neural networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks, San Diego, California*, volume 2, pages 11–16, 1990.

J. Kingdon, J. G. Taylor, and C. L. Mannion. *Intelligent Systems and Financial Forecasting*. Springer-Verlag, New York, 1997.

P. D. Koch and T. W. Koch. Evolution in dynamic linkages across daily national stock indexes. *Journal of International Money and Finance*, 10:231–251, 1991.

K. Kohara, T. Ishikawa, Y. Fukuhara, and Y. Nakamura. Stock price prediction using prior knowledge and neural networks. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 6(1):11–22, 1997.

A. Konar. *Computational Intelligence: Principles, Techniques and Applications*. Springer, 2005.

H. Konno, D. B. Luenberger, and J. M. Mulvey. *Annals of Operations Research*, 45 (1):1–450, 1993. Special Issue on Financial Engineering.

J. R. Koza. A genetic approach to econometric modeling. In Paul Bourgine and Bernard Walliser, editors, *Economics and Cognitive Science*, pages 57–75. Pergamon Press, Oxford, UK, 1991.

C.-M. Kuan and T. Liu. Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Economics*, 10(5):347–364, 1995.

C.-M. Kuan and H. White. Artificial neural networks: An econometric perspective. *Econometric Reviews*, 13:1–91, 1994.

P. Lajbcygier, C. Boek, M. Palaniswami, and A. Flitman. Neural network pricing of All Ordinaries SPI Options on Futures. In *Proceedings of the 3rd Internation Conference on Neural Networks in the Capital Markets*, pages 64–77, London, 1996. World Scientific.

Y. Le Cun, J. S. Denker, and S.A. Solla. Optimal brain damage. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 598–605. University of South California Press, 1990.

T.-H. Lee, H. White, and C. W. J. Granger. Testing for neglected nonlinearity in time-series models : A comparison of neural network methods and alternative. *Journal of Econometrics*, 56(3):269–290, 1993.

C. W. Li and W. K. Li. On a double-threshold autoregressive heteroscedastic time-series model. *Journal of Applied Econometrics*, 11(3):253–74, 1996.

T. Li, L. Fang, D. Guo, and S. Klasa. Predicting exchange rates using a fuzzy learning system. In *Proceedings of the 1995 Conference on Computational Intelligence for Financial Engineering (CIFEr), 9-11 Apr 1995, New York, USA*, pages 103–107, 1995.

J. Litner. Security prices, risk and maximal gains from diversification. *Journal of Finance*, 20:305–362, 1965.

I. Lobato, J.C Nankervis, and E. Savin. Testing for autocorrelation using a modified Box-Pierce Q test. *International Economic Review*, 42(1):187–205, 2001.

S. Lundbergh and T. Teräsvirta. Modelling economic high-frequency time-series with STAR-STGARCH models. Scandinavian Working Papers, 1998.

S. Lundbergh and T. Teräsvirta. Evaluating GARCH models. *Journal of Econometrics*, 110(2):417–435, 2002.

R. Luukkonen, P. Saikkonen, and T. Teräsvirta. Testing linearity in univariate time-series models. *Scandinavian Journal of Statistics*, 15:161175, 1988.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

H. M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

A. M. M. Masih and R. Masih. Dynamic linkages and the propagation mechanism driving major international stock markets: An analysis of the pre- and post-crash eras. *The Quarterly Review of Economics and Finance*, 37(4):859–885, 1997.

P. C. McCluskey. *Feedforward and Recurrent Neural Networks and Genetic Programs for Stock and time-series Forecasting.* PhD thesis, 1993. Department of Computer Science, Brown University.

D. G. McMillan. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time-series: A re-examination. *Review of Financial Economics*, 14:81–91, 2005.

M. C. Medeiros, T. Teräsvirta, and G. Rech. Building neural network models for time-series: A statistical approach. *Journal of Forecasting*, 25:49–75, 2006.

M. C. Medeiros and A. Veiga. Diagnostic checking in a flexible nonlinear time-series model. *Journal of time-series Analysis*, 24:461–482, 2003.

R. C. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4:141–183, 1973.

R. C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 4:125–144, 1976.

M. Michalopoulos, N. S. Thomaidis, G. Dounias, and C. Zopounidis. Using a fuzzy sets approach to select a portfolio of greek government bonds. *Fuzzy Economic Review*, 9(2):27–48, 2004.

M. Mohammadian and M. Kingham. An adaptive hierarchical fuzzy logic system for modelling of financial systems. *Intelligent Systems in Accounting, Finance & Management*, 12(1):61–82, 2004.

J. Mossin. Equilibrium in a capital asset market. *Econometrica*, 34:768–783, 1966.

J. M. Mulvey, D. P. Rosenbaum, and B. Shetty. Strategic financial risk management and operations research: A review. *European Journal of Operational Research*, 97 (1):1–16, 1997.

A. K. Nag and A. Mitra. Forecasting daily foreign exchange rates using genetically optimized neural networks. *Journal of Forecasting*, 21(7):501–11, 2002.

Ch. J. Neely and P. A.Weller. Using a genetic program to predict exchange rate volatility. In Sh.-H. Chen, editor, *Genetic Algorithms and Genetic Programming in Computational Finance*. Kluwer Academic Publishers, 2002.

D. Nelson. Conditional heteroskedasticty in asset returns: A new approach. *Econometrica*, 59:347–370, 1990.

N. Nikolaev and H. Iba. Genetic programming of polynomial harmonic models using the discrete fourier transform. In *Proceedings of the 2001 Congress on Evolutionary Computation, CEC2001, IEEE Press,Piscataway, NJ*, pages 267–274, 2001.

D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2006.

P. K. H. Phua, Z. Xiaotian, and Chung H. K. Forecasting stock index increments using neural networks with trust region methods. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 260–265, 2003. 9/8/2003-9/10/2003, Benalmadena, Spain.

Risk Metrics. RiskMetrics Group, J.P. Morgan, 4th edition, 1996. URL `http://www.jpmorgan.com/RiskManagement/RiskMetrics[November2004]`.

S. A. Ross. The arbitrage pricing theory of capital asset pricing. *Journal of Economic Theory*, 13:341–360, 1976.

S. A. Ross. Return, risk and arbitrage. In I. Friend and J. L. Bicksler, editors, *Risk and Return in Finance*, pages 189–218. Ballinger, Cambridge, MA, 1977.

RSAS. Time-series econometrics: Cointegration and autoregressive conditional heteroskadasticity. Technical report, The Royal Swedish Academy of Sciences, 2003. URL `http://www.kwa.se[February2007]`.

J. M. Samuels, F. M. Wilkes, and R. E. Brayshaw. *Financial Management and Decision Making*. International Thomson Business Press, 1998.

Ch. Schittenkopf, G. Dorffner, and E. J. Dockner. Forecasting time-dependent conditional densities: a semi non-parametric neural network approach. *Journal of Forecasting*, 19(4):355–374, 2000.

E. Schoneburg. Stock price prediction using neural networks: A project report. *Neurocomputing*, 2:17–27, 1990.

M. G. Schuster. A multiobjective genetic programming approach for pricing and hedging derivative securities. In *Proceedings of the 2003 International Conference on Computational Intelligence for Financial Engineering (CIFEr), New York*, pages 77–84, 2003.

G. W. Schwert. Stock volatility and the crash of '87. *The Review of Financial Studies*, 3(1):77–102, 1990.

M. Setnes and O.J.H. van Drempt. Fuzzy modelling in stock-market analysis. In *Proceedings of the 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr), 28-30 Mar 1999, New York, USA*, pages 250–258, 1999.

W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Portfolio Selection*, 19(3):425–442, 1964.

A. Shleifer. *Inefficient Markets: An Inroduction to Behavioural Finance*. Clarendon Lectures in Economics. Oxford University Press, 2000.

S. Singh and J. Fieldsend. Financial time-series forecasts using fuzzy and long memory pattern recognition systems. In *Proceedings of the 2000 Conference on Computational Intelligence for Financial Engineering (CIFEr), 26-28 Mar 2000, New York, USA*, pages 166–169, 2000.

T. Sinha and F. Chamu. Comparing different methods of calculating value at risk. URL `http://ssrn.com/abstract=706582[February2007]`. 2000.

E. Steurer and T.H Hann. Exchange rate forecasting comparison: neural networks, machine learning and linear models. In A. Refenes, Y. Abu-Mostafa, J. Moody, and A. Weigend, editors, *Neural Networks in Financial Engineering*, pages 113–121. World Scientific, 1996.

N. R. Swanson and H. White. A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics*, 13:265–275, 1995.

N. R. Swanson and H. White. Forecasting economic time-series using flexible vs fixed specification and linear vs nonlinear econometric models. *International Journal of Forecasting*, 13:439–461, 1997a.

N. R. Swanson and H. White. A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economic and Statistics*, 79:540–550, 1997b.

G. G. Szpiro. Forecasting chaotic time-series with genetic algorithms. *Physical Review E*, 55:2557–2568, 1997.

T. Takagi and M. Sugeno. Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15 (1):116–132, 1985.

A. S. Tay and K. F. Wallis. Density forecasting: A survey. In M. P. Clements and D. F. Hendry, editors, *A Companion to Economic Forecasting*, pages 45–68. Blackwell, 2002.

P. Tenti. Forecasting foreign exchange rates using recurrent neural networks. *Applied Artificial Intelligence*, 10:567–581, 1996.

T. Teräsvirta. Specification, estimation and evaluation of smooth transition autoregression models. *Journal of Americal Statistician*, 89:208–218, 1994.

T. Teräsvirta, C.-F. Lin, and C. W. J. Granger. Power of the neural network linearity test. *Journal of Time Series Analysis*, 14(2):209–220, 1993.

T. Teräsvirta and C.-F. J. Lin. Determining the number of hidden units in a single hidden-layer neural network model. Technical report, 1993. Bank of Norwey.

T. Teräsvirta, D. van Dijk, and M. Medeiros. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time-series: A re-examination. *International Journal of Forecasting*, 21:755–774, 2005.

H. Theil. *Principle of Econometrics.* John Wiley & Sons, 1971.

N. S. Thomaidis. The implications of behavioural finance to the modelling of securities prices. In D. Satish and P. Krishna Kishore, editors, *Behavioral Finance*, Finance Series. The ICFAI University Press, 2006.

N. S. Thomaidis and D. Dounias. Cointegration and error-correction models: towards a reconcilation between behavioural finance and econometrics. *The ICFAI Journal of Behavioral Finance*, 3(3):51–73, 2006a.

N. S. Thomaidis and G. Dounias. Equilibrium correction models in the framework of computational intelligence. URL `http://ssrn.com[February2007]`. Working Paper, 2005.

N. S. Thomaidis and G. Dounias. A general class of combined neural network GARCH models for financial time-series analysis. URL `http://ssrn.com[February2007]`. Working Paper, 2006b.

N. S. Thomaidis, G. Dounias, and N. Kondakis. Financial statistical modelling with a new nature-inspired technique. In *Proceedings of the 1st European Symposium on Nature-Inspired Smart Information Systems (NISIS), Albufeira, Portugal*, 2005a.

N. S. Thomaidis, N. Kondakis, and G. Dounias. The cross-dynamics of international financial markets: a suitable domain for applying nature inspired intelligent techniques. In *Proceedings of the 1st European Symposium on Nature-Inspired Smart Information Systems (NISIS), Albufeira, Portugal*, 2005b.

N. S. Thomaidis, N. Kondakis, and G. Dounias. An intelligent statistical arbitrage trading system. *Lecture Notes in Artificial Intelligence*, 3955:596–599, 2006.

N. S. Thomaidis, V. Tzastoudis, and G. Dounias. A comparison of neural network model selection strategies for the pricing of S&P 500 stock index options. *International Journal of Artificial Intelligence Tools*, 2007. forthcoming.

D. Tikk, L. T. Kóczy, and T. D. Gedeon. A survey on the universal approximation and its limits in soft computing techniques. *International Journal of Approximate Reasoning*, 33(2):185–202, 2003.

N. Towers. Joint optimisation in statistical arbitrage trading. In J. Shadbolt and J. G. Taylor, editors, *Neural Networks and the Financial Markets: predicting, combining and portfolio optimisation*, pages 193–201. Springer, 2002.

A. Trapletti, F. Leisch, and K. Hornik. Stationary and integrated autoregressive neural network processes. *Neural Computation*, 12:2427–2450, 2000.

R. R. Trippi and E. Turban. *Neural Networks in Finance and Investment: Using Artificial Intelligence to Improve Real-World Performance.* IRWIN, 1996.

A. Tsakonas and G. Dounias. Hybrid computational intelligence schemes in complex domains: An extended review. *Lecture Notes in Artificial Intelligence*, 2308:494–511, 2002.

V. Tzastoudis, N. S. Thomaidis, and G. Dounias. Improving neural network based option price forecasting. *Lecture Notes in Artificial Intelligence*, 3955:378–388, 2006.

H. Vladimirou. *Annals of Operations Research*, 151(1):1–336, 2007a. Special Issue on Financial Modeling.

H. Vladimirou. *Annals of Operations Research*, 152(1):1–420, 2007b. Special Issue on Financial Optimization.

S. Watanabe. Algebraic geormetrical methods for hierarchical learning machines. *Neural Networks*, 14:1049–1060, 2001.

A. S. Weigend, B. A. Huberman, and D. E. Rumelhart. Predicting sunspots and exchange rates with connectionist networks. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, pages 395–432. Addison-Wesley, 1992.

A. S. Weigend, D. E. Rumelhart, and B. A. Huberman. Generalization by weight-elimination with application to forecasting. In J. Moody R. P. Lippmann and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems.* Morgan Kaufmann, San Mateo, CA, 1991.

Y.-J. Whang. A test of autocorrelation in the presence of heteroskedasticity of unknown form. *Econometric Theory*, 14:87–122, 1998.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50: 1–26, 1982.

H. White. Economic prediction using neural networks: The case of IBM daily stock returns. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 2, pages 451–458, 1988a.

H. White. Economic prediction using neural networks: the case of IBM dailystock returns. In *IEEE International Conference on Neural Networks*, volume 2, pages 451–458, 1988b. 9/8/2003-9/10/2003, Benalmadena, Spain.

H. White. An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks. In *Proceedings Of The International Joint Conference On Neural Networks*, pages 451–455. Washington, DC, 1989a.

H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1:425–464, 1989b.

H. White. *Estimation, Inference and Specification Analysis*, volume 22 of *Econometric Society Monographs*. 1996.

H. White and J. Racine. Statistical inference, the bootstrap and neural network modelling with applications to foreign exchange rates. *IEEE Transacations on Neural Networks*, 12(4):657–673, 2001.

J. M. Wooldridge. A unified approach to robust, regression-based specification tests. *Econometric Theory*, 6:17–43, 1990.

J. M. Wooldridge. On the application of robust, regressionbased diagnostics to models of conditional means and conditional variances. *Journal of Econometrics*, 47:5–46, 1991.

S.-I. Wu and H. Zheng. Stock index forecasting using recurrent neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Applications*, 2003. 9/8/2003-9/10/2003, Benalmadena, Spain.

J. Yao, Y. Li, and Ch. T. Tan. Option price forecasting using neural networks. *International Journal of Management Science*, 28:455–466, 2000.

A. Zapranis and A.-P. Refenes. *Principles of Neural Model Identification, Selection and Adequacy: with Applications to Financial Econometrics.* Perspectives in Neural Computing. Springer, 1999.

S. A. Zenios. *Financial Optimization.* Cambridge University Press, 2002.

G. Zhang, B. E. Patuwo, and M. Y. Hu. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14:35–62, 1998.