Financial statistical modelling with a new nature-inspired technique

Nikos S. Thomaidis*[1], George D. Dounias[1], and Nick Kondakis[1,2]

[1]Dept. of Financial Engineering & Management, University of the Aegean,, 31 Fostini Str., GR-821 00, Chios, GREECE, tel: +30-2271-0-35454, fax: +30-2271-0-35499, email: {nthomaid, kondakis, g.dounias}@fme.aegean.gr, URL: http://decision.fme.aegean.gr

[2]Kapa International, Ltd, 140 Broadway, 46th Floor, New York, NY 10005, e-mail: nick@kondakis.com

Abstract

This paper introduces a nature-inspired intelligent model suitable for high-frequency financial time series. It combines a neural network parametrization for the mean with a linear (GARCH) parametrization for the variance. We propose a complete model-building cycle for the family of NN-GARCH specifications that includes all three stages of econometric modelling (specification, estimation and evaluation). Based on the maximum likelihood theory, we device procedures for statistical inference in the framework of NN-GARCH models and thus offer the modeler the opportunity to test hypotheses of interest concerning both the mean and variance structure of the data-generating process. We demonstrate the model-building cycle by constructing a NN-GARCH dynamic model for the returns on the DAX Stock Index.

Keywords: Neural Networks, Volatility Forecasting, GARCH Models, Maximum Likelihood Theory

1 Introduction

Nature - inspired computational intelligence (NICI) schemes, like neural networks, genetic algorithms, etc, are nowadays widely appreciated in time-series analysis and forecasting, especially in economics and finance (see [4, 23] for relatively recent surveys). The use of NICI algorithms in empirical work is generally motivated by several advantages that these modelling methodologies offer. First of all, they provide the analyst with semi-parametric specifications, which in contrast to other parametric methodologies (like linear regression), do not make explicit assumptions on the functional relationship between target and explanatory variables. In addition, they possess a *universal approximation property*, meaning that under mild regularity conditions they are capable of approximating a quite general family of functions to any given degree of accuracy [11, 12].

In the last twenty years, the modelling and forecasting of volatility has gained exceptional attention in the financial literature. The accurate representation of the volatility dynamics has become a key

*We are grateful to Dr Marcelo Medeiros, Department of Economics Pontifical Catholic University of Rio de Janeiro, and Prof Timo Teräsvirta, Stockholm School of Economics, for providing us with the Matlab code for their Neural Network model building methodology. This research is funded by the Public Benefit Foundation "Alexander S. Onassis" (www.onassis.gr), under the 2003-2004 Scholarships Programme, and by a grant from "Empeirikion" Foundation (www.empirikion.gr).

issue for both researchers in financial economics and analysts in financial markets. Banks and other financial institutions make volatility assessments as a part of monitoring their risk exposure. The research interest on models of volatility was initiated in the middle 80's by R. Engle [8] who proposed a new concept for describing the volatility dynamics of a time series: the *Autoregressive Conditional Heteroskedasticity*[1] *(ARCH)*. In the ARCH family of models, the conditional variance of the innovation term of a mean model is assumed to depend on past shocks. Hence, ARCH-type models impose a certain type of *conditional* heteroskedasticity characterised by successive periods of high and low volatility in the history of a time series. They also assume a steady long-run volatility level (i.e. unconditional homoskedasticity) to which conditional variance has a tendency to return. This type of heteroskedasticity is different from that imposed by other heteroskedastic models which assume that the overall volatility level of a time series changes over time (i.e. *unconditional* heteroskedasticity). Since their inception ARCH models have been extended in many directions, most important being the class of *Generalized Autoregressive Conditional Heteroskedasticity (GARCH)* models proposed by T. Bollerslev [1] (other extensions of ARCH models are discussed in the paper). Those offer more parsimonious representations of the volatility structure and are therefore more popular in practice.

Given the success of both NICI and GARCH models, a natural idea is to combine both "technologies" in a single model. By that we hope to gain from the high ability of NICI algorithms in approximating complicated functional relations and the simple but accurate representations of the volatility dynamics that GARCH models offer. In this paper, we introduce a combined artificial neural network (NN) linear-GARCH model and explore its statistical properties. We propose a complete model-building cycle for NN-GARCH models (specification, estimation and evaluation) and device procedures for statistical inference. In hypothesis testing, we adopt robustified versions of classical statistics (e.g. the Lagrange Multiplier (LM)) which preserve their validity even in the presence of various structural or distributional misspecifications. Our approach is similar to [6], who consider a special linear in-mean AR(1) and nonlinear-in-volatility NN-based model. However, our methodological framework incorporates a greater variety of parametrizations. We also provide a coherent modelling strategy for specification and evaluation of the final model as well as procedures for statistical inference.

The rest of the paper is organised as follows: In section 2 we discuss basic issues in statistical modelling. Our starting point is that in an inherently stochastic environment, as is financial markets, the focus of interest is on finding a *probabilistic* rather than an exact functional relationship between the target and a set of explanatory variables. This is summarised by the conditional distribution of the target variable given the set of explanatory ones. The success of a neural network is ultimately evaluated at the extend to which it can uncover interesting features of the distribution from which observations were sampled. Section 3 reviews some basic results in maximum likelihood theory, pertaining to estimation of parameters, construction of confidence intervals and parameter restrictions testing. This section provides the theoretical background for various procedures proposed in the paper for specification, estimation and evaluation of NN-GARCH models. The specification of NN regression models under the ML principle is considered in more detail in section 4, where the recent approach of Medeiros, Teräsvirta and Rech [15] is presented. This follows the spirit of other ML approaches to the specification of nonlinear regression models and forms the basis for our proposed NN-GARCH model. Section 5.1 reviews current trends in the modelling of volatility with GARCH-type processes and section 5.2 introduces the NN-GARCH model. In section 5.3 we propose a simple-to-complicate model building procedure for the family of NN-GARCH models and in section 5.4 we present a general framework for in-sample misspecification testing on the estimated model. In order to demonstrate the model building procedure presented in section 5.3, we apply a NN-GARCH model to the task of forecasting returns on the DAX Stock Index. Section 7 summarises the main findings of the paper and provides directions for further research.

2 Principles of econometric modelling

In most econometric time series applications, NNs are traditionally treated as approximators to the functional relationship between a set of explanatory variables X and another target variable Y which

[1] A time series is said to be *heteroskedastic* if its variance changes over time.

is of interest to the modeler. However, such approach is justified in application domains where the modeler has complete control over the data-generating process, i.e. in an experimental or computer setting. In any situation where "nature", or the environment in general, has a hand in generating measurements, the application of ANNs has to be treated with care[2]. Nature's role may be partial, as in an experimental setting where observations are "noisy" i.e. their precise value is largely determined by chance, or complete, as in many economic phenomena where the modeler does not have any 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 *probabilistic relationship*.

In a statistical setting, the sample of observations $\{(x_t, y_t),\ t = 1, ..., T\}$ available to the modeler is viewed as a realization of the jointly distributed stochastic processes $\{(X_t, Y_t),\ t = 1, 2, ..., T\}$. The probabilistic relationship between X_t and Y_t is completely summarised by the *conditional probability law* of Y_t given X_t, $P(Y_t \in A | X_t \in B)$, which essentially describes the relative frequency of Y_t taking values in A given that X_t is in B. In this probabilistic context, the focus of interest (and thus the focus of learning for the NN) 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.

In the language of econometrics, the data that there is available to the modeler for explaining the behaviour of Y_t make up an *information structure* $\{\mathcal{I}_t, t = 1, 2, ..., T\}$, whose individual component $\mathcal{I}_t$ represents all the relevant information that is available to the modeler just before the realization of Y_t. Several things can fit in $\mathcal{I}_t$: "environmental" or exogenous variables Z_t, which seem not to be influenced by Y_t (and hence can be treated as fixed in repeated samples), as well as lags of Z_t and Y_t. Assuming that the conditional probability law is well behaved, one can define the *conditional probability density function* $p(y_t|\mathcal{I}_t)$, which is the probability of the random variable Y_t taking a value close to y_t given the information set $\mathcal{I}_t$.

Several features of the conditional density are of particular interest in financial applications:

- The *first moment* or *conditional expectation* $E(Y_t|\mathcal{I}_t)$. It is a measure of the value of Y_t that will be realized on average given $\mathcal{I}_t$. Since the actual realization of Y_t will (almost certainly) differ from $E(Y_t|\mathcal{I}_t)$, there will be an *expectational error* $\epsilon_t \equiv Y_t - E(Y_t|\mathcal{I}_t)$, which is often called the *innovation* or *"surprise"* term. Note that by definition and the properties of conditional expectation $E(\epsilon_t|\mathcal{I}_t) = 0$, i.e. the innovation cannot be predicted by I_t. Note that the conditional expectation is only a function of the information set and hence of X_t, i.e. $E(Y_t|\mathcal{I}_t) = g(X_t)$. By turning the above relationship around we obtain

$$Y_t = g(X_t) + \epsilon_t, \tag{1}$$

  which is very similar to the "function + noise" data-generating process that is typically assumed in most time series applications of NNs. However, the analysis given above assigns a fundamentally different interpretation to (1). The process that generates Y_t can be decomposed into a part which can be explained though X_t (and consequently $\mathcal{I}_t$) and the remainder or unexplained part ϵ_t. Note that in a "noisy" experimental setting, where ϵ_t is an exogenous perturbation to Y_t, the only object of interest is the conditional expectation $E(Y_t|\mathcal{I}_t) = g(X_t)$. Hence, the task of an ANN is to provide accurate approximations to the mapping $g(.)$. In economics, however, where the modeler does not have control over the values that economic variables take, ϵ_t does not simply enter (1) in the form of "exogenous" noise. On the contrary, it conveys important features of the conditional distribution of Y_t which might deserve further investigation.
- The *second moment* or *conditional variance* $\mathrm{Var}(Y_t|\mathcal{I}_t)$. It is a measure of the average uncertainty or "risk" about the realisation of Y_t given the information set I_t. The accurate modelling of the conditional variance is in particular important to the management of financial risk.

The process of specifying an econometric model is schematically depicted in figure 1. An integral part

[2]See [20] for an excellent discussion on the topic.

of the procedure are the so-called misspecification tests, which mainly concerned with the adequacy of the model as well as the significance of the various parts thereof.

The adequacy of the model is related to whether it adequately describes the data-generating process or leaves interesting features unmodelled the residuals of the model. In an adequate statistical model the residuals highly behave as white noise. The existence of "strong" properties in the error term indicates some model *specification bias*. 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, and in particular the explanatory variables. The main trend in econometric modelling is to identify models with the least possible number of independent variables, enough to capture the salient features or "driving forces" of the data-generating process, relegating all minor and random influences to the noise term. This is the well-known principle of *parsimony* or Occam's razor, often found in econometrics textbooks (see for example [3]).

From an econometric point of view, NNs can be seen as nonlinear semi-parametric statistical models for $p(y_t|\mathcal{I}_t)$, the conditional density. From that perspective, various NN specification problems (selection of the number of hidden neurons, the connections from inputs to neurons, etc) 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 network and test hypotheses regarding the obtained network specification (irrelevant inputs/hidden units/connections, etc). These issues are discussed in detail in the next section.

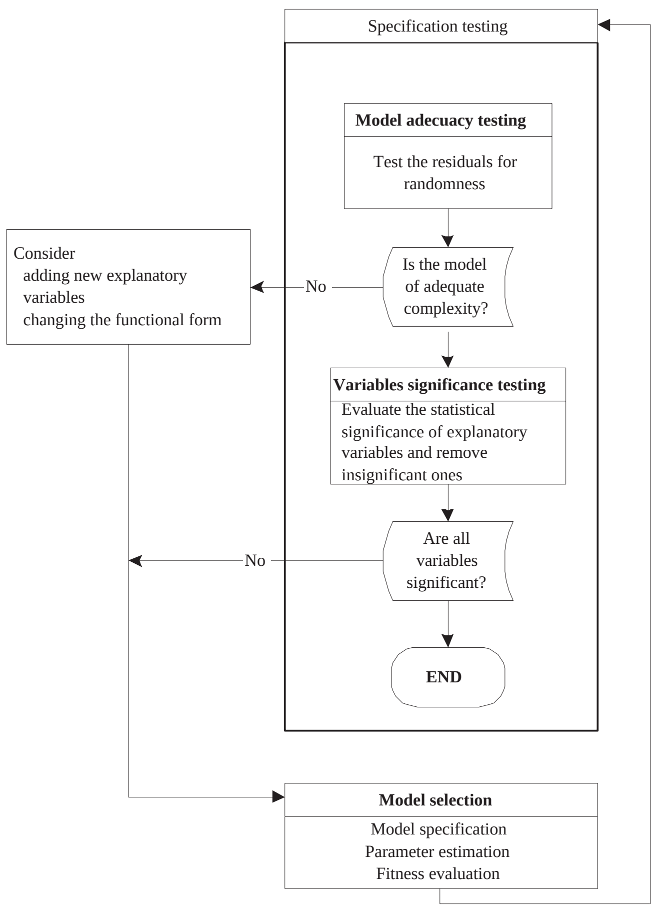

Figure 1: The process of specifying an econometric model (adapted from [22]).

3 The maximum likelihood theory

3.1 The general idea

In the maximum-likelihood approach to estimation, one starts with a parametrised family $\mathcal{D} = \{p(y_t|\mathcal{I}_t;\delta),\ \delta \in \Theta, \Theta \subset \mathbb{R}^n\}$ of candidate specifications for the "true" conditional probability (i.e. the density that generated the observations set $\{y_t, t = 1, 2, ...T\}$). The probability that sample observations were generated by a given specification of $\mathcal{D}$ is

$$L(\delta) \equiv \prod_{t=1}^{T} p(y_t|\mathcal{I}_t;\delta)$$

By maximising the above with respect to δ we ask for the most *likely* member of the family given the sample of observations $\{x_t, y_t,\ t = 1, 2, ..., T\}$. This is equivalent to solving the following maximization problem

$$\max_{\delta\in\Theta} l(\delta) \tag{2}$$

where $l(\delta) \equiv T^{-1}\log L(\delta) = T^{-1}\sum_{t=1}^{T}\log p(y_t|\mathcal{I}_t;\delta)$. $l(\delta)$ is referred to as the *log-likelihood* function generated by the parametric family $\mathcal{D}$ and the value $\hat{\delta}_T$ that solves (2) is the *maximum likelihood estimator*. When the actual conditional density is included in the pre-assumed parametric family then $\hat{\delta}_T$ is truly the ML estimator. In the case where the true conditional density is not correctly specified within $\mathcal{D}$ then the MLE is referred to as the *Quasi Maximum Likelihood Estimator* (QMLE). QMLE has an information theoretic interpretation: it is the parameter vector which provides the best possible approximation to the true probability density within the parametric family of densities. The adequacy of approximation, or else the "distance" between the true and the proxy density, is measured by the Kullback-Leibler Information Criterion (KLIC) (see [21], ch.2).

3.2 Properties of the QMLEs

Under certain regularity and moment conditions, estimators obtained under the ML principle possess some desirable properties, which justify their use in empirical work (see [21], ch. 3, 6 and 7).

As the sample size T increases, the QMLE $\hat{\delta}_T$ tends stochastically to δ_0, a parameter vector which minimizes the average Kullback-Leibler "distance" of the approximating from the true probability density. This is a so-called *consistency* property, which is formally written as

$$\hat{\delta}_T \xrightarrow{P} \delta_0 \quad \text{as} \quad T \to \infty$$

or

$$\hat{\delta}_T \to \delta_0 \quad \text{almost surely as} \quad T \to \infty$$

where

$\xrightarrow{P}$ denotes convergence in probability.

In addition to that, $\hat{\delta}_T$ is asymptotically normally distributed around δ_0 with variance-covariance matrix C ($\overset{\alpha}{\sim}$ is read as "asymptotically distributed"):

$$\sqrt{T}(\delta_T - \delta_0) \overset{\alpha}{\sim} N(0, C)$$

where

$$C \equiv A^{-1} I A^{-1} \tag{3a}$$

$$A = E\left(-\frac{\partial^2 l_0}{\partial\delta\partial\delta'}\right) \tag{3b}$$

$$I = E\left(T\frac{\partial l_0}{\partial\delta}\frac{\partial l_0}{\partial\delta'}\right) \tag{3c}$$

$$\frac{\partial^2 l_0}{\partial\delta\partial\delta'} \equiv \left.\frac{\partial^2 l}{\partial\delta\partial\delta'}\right|_{\delta=\delta_0} \quad \text{and} \quad \frac{\partial l_0}{\partial\delta} \equiv \left.\frac{\partial l}{\partial\delta}\right|_{\delta=\delta_0}$$

The asymptotic variance-covariance matrix is particularly useful in obtaining *confidence intervals* for the estimates of δ_0 and *testing hypothesis of interest* involving δ_0 (see next subsection). One tricky issue is that C cannot be explicitly computed as it involves knowledge of the expectation $E(.)$ and hence the true probability distribution. However, under certain circumstances, we can construct a consistent sample estimator $\hat{C}_T$ of C, given by

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

where

$$\hat{A}_T = -\frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial^2 l_t}{\partial\delta\partial\delta'}\right|_{\delta=\hat{\delta}_T}$$

$$\hat{I}_T = \frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial l_t}{\partial\delta}\frac{\partial l_t}{\partial\delta'}\right|_{\delta=\hat{\delta}_T}$$

l_t is the time t contribution to the log-likelihood function ($l_t(\delta) \equiv log(p(y_t|\mathcal{I}_t;\delta))$)). Conditions on the existence of a sample estimator of the variance-covariance matrix are given in [21], ch. 8. In the case where the actual density is correctly specified within the parametric family $\mathcal{D}$, the variance covariance matrix takes the simplified form $C = A^{-1} = -I^{-1}$.

The QMLE is *asymptotically efficient* relative to the class of consistent and asymptotically normally distributed (CAN) estimators. If the probability model is correctly specified within $\mathcal{D}$ then the QMLE attains also the *minimum asymptotic variance bound*, I^{-1}, which is the inverse of the so-called *Fisher information matrix*. This is the minimum possible asymptotic variance that a CAN estimator can achieve.

In summary, we see that the QMLE preserves its desirable properties (consistency and asymptotical normal distribution) despite the possibility that the probability model assumed by the analyst is a priori misspecified from various points of view (insufficient complexity, distributional misspecification, etc). The form (3) of C which takes into account this possibility is said to be *robust against misspecifications*.

3.3 Statistical Inference

In the framework of ML theory, there are three classical statistics one can use to test hypotheses of interest: (1) The Wald (W), (2) the Lagrange Multiplier (LM) and (3) the Likelihood Ratio (LR). Hypotheses of the "true parameters" δ_0 are usually expressed as a set of restrictions on the parameters of interest, embodied in a system of (possibly nonlinear) equations such as

$$H_0 : r(\delta_0) = 0,$$

where $r : \Theta \to R^p$, $p \leq n$. An alternative hypothesis that is frequently of interest is

$$H_1 : r(\delta_0) \neq 0$$

Hypothesis tests are conducted by using the asymptotic distribution of the above test statistics to decide whether the tests statistic, as estimated from the sample, takes on a plausible value under the null hypothesis H_0. Let $\tilde{\delta}_T$ denote the constrained QMLE obtained by maximising $l(\delta)$ under the constraint $r(\delta_0) = 0$ imposed by the null hypothesis. Then, the afore mentioned test statistics take on the following forms (see [21], §8.1):

Wald

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

Lagrange Multiplier

$$LM_T = T\,\frac{\partial \tilde{l}}{\partial \delta'}\left(\frac{\partial^2 \tilde{l}}{\partial\delta\partial\delta'}\right)^{-1}\frac{\partial \tilde{r}_T}{\partial \delta}\left(\frac{\partial \tilde{r}_T}{\partial \delta'}\tilde{C}_T\frac{\partial \tilde{r}_T}{\partial \delta}\right)^{-1}\frac{\partial \tilde{r}_T}{\partial \delta'}\left(\frac{\partial^2 \tilde{l}}{\partial\delta\partial\delta'}\right)^{-1}\frac{\partial \tilde{l}}{\partial \delta} \tag{4}$$

Likelihood Ratio

$$LR_T = -2T\left[l(\tilde{\delta}) - l(\hat{\delta})\right]$$

where $\hat{r}_T \equiv r(\hat{\delta}_T)$ $\tilde{l} \equiv l(\tilde{\delta}_T)$

In hypothesis testing, many authors replace $\hat{C}_T$ in the formulae above with $\hat{A}_T$. Such substitution however is not theoretically justified unless the parametric family of distributions is correctly specified for the real density.

It can be shown (see [21], theorem 8.10) that under the null hypothesis both W_T and LM_T follow *asymptotically* the χ^2_p distribution, despite any misspecifications of the probability model, as long as one uses the robustified version of C. The same does not however hold for the LR_T statistic unless the QMLE is in fact the MLE, i.e. the conditional distribution of Y_t given $\mathcal{I}_t$ is correctly specified within the parametric family $\mathcal{D}$. The LM or Wald statistics do not require the validity of the information matrix equality, as they can be modified by replacing classical estimators of the asymptotic covariance matrix with specification-robust ones. Therefore, in practice it is safer to base hypothesis testing either on W_T or LM_T.

If we want to construct a test such that the probability of wrongly rejecting the null hypothesis is not higher than $a\%$, we compare the value obtained for the statistic with $\chi^2_{p,1-\alpha}$, the $1-\alpha$ percentile of the χ^2_p distribution. One rejects H_0 at confidence level α if $W_T > \chi^2_{p,1-\alpha}$ or $LM_T > \chi^2_{p,1-\alpha}$.

The choice between these two statistics is solely a matter of computational effort. Note that LM_T is based on the constrained estimator while W_T requires in addition the unconstrained one. Since constrained models are simpler in form, LM_T might be more convenient in some cases.

4 The specification of neural network models under the ML principle

The machinery of likelihood theory can be applied to various problems in neural network specification

- Estimating the parameters of the network
- Determining the appropriate complexity of the network (i.e. number of hidden neurons) as dictated by the available set of observations
- testing the significance of various network connections (connections either from inputs to hidden layer or from hidden layer to output)

In the following section we present a typical ML approach to estimating NN models, which has been recently proposed by Medeiros, Teräsvirta and Rech (henceforth MTR approach) [15]. This builds on early work by [17–19]. The MTR methodology forms the basis for our proposed model and hence we analytically review it in the next section.

4.1 The approach of Medeiros, Teräsvirta and Rech (MTR)

In their setting, MTR assume that observations are generated according to the following statistical model

$$y_t = \phi' \tilde{x}_t + f(x_t; \theta) + \epsilon_t \tag{5a}$$

$$\epsilon_t | \mathcal{I}_t \sim N\left(0, \sigma^2\right) \tag{5b}$$

where $y_t \in \mathbb{R}$, $\tilde{x}_t' = [1, x_t']$, $x_t \in \mathbb{R}^n$ is the vector of regressors (possibly including lags of y_t and/or other exogenous variables of the information set) and $f(x_t; \theta)$ represents a feedforward neural network with a single hidden layer and h neurons, i.e.

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

where $F(z) = \frac{1}{1+e^{-z}}$ is the logistic function, $c_j, \gamma_j, \lambda_j \in \mathbb{R}$ and $w_j \in \mathbb{R}^n$.

The parameters for the linear and nonlinear part of the conditional mean structure are $\phi \in \mathbb{R}^{n+1}$ and $\theta = (\lambda_1, ..., \lambda_h, \gamma_1, ..., \gamma_h, w_1', ..., w_n', c_1, ..., c_h)' \in \mathbb{R}^{(n+2)h}$ respectively. σ is assumed time-independent.

In building a neural network model, MTR adopt a specific-to-general strategy, which is roughly summarised in the following steps:

- Start with specifying a linear model. Select variables according to an information-criterion such as AIC, SBIC, etc.
- Test the hypothesis of a linear against an additive nonlinear NN model with $h = 1$ neurons using the LM test statistic.
- If null cannot be rejected then stop. Otherwise, estimate a NN model (5) with a single neuron.
- Carry on with testing the hypothesis of h against $h+1$ neurons, for $h = 1, 2, ...$, until the first acceptance of the null.

This specific-to-general approach is intended to produce simple parsimonious models that are less likely to overfit the data. Complex and overidentified models involve more parameters and are harder to estimate from an algorithmic point of view. Apart from standard arguments of parsimony, overidentified models are also problematic from a theoretical point of view. The parameters of the model cannot be estimated consistently and most important proper statistical inference is not available, if the model contains nuisance parameters that are not identified under the null hypothesis. Test statistics in the framework of ML theory (Wald, LM, LR) do not follow the standard χ^2 distribution which renders impossible to access the right size of the test (see [10]).

MTR provide a new toolkit for in-sample evaluation of linear neural network models (5), including tests for autocorrelation in the residuals, parameters constancy and additional nonlinearity in the condition mean model.

Simulations results (see p. 30 of [15]) seem to favour the MTR approach as concerns the specification of a NN model. The right set of variables is usually selected, although the method has a tendency to oversimplify models. In small sample sizes, in particular, there is a strong tendency to select a linear model, but this is somewhat expected as nonlinearity becomes more apparent with an increasing sample size. Overfitting is rare and can be effectively controlled if the initial significance level is lowered as the sample of observations increases. The MTR approach also seems superior when compared to other statistics-oriented approaches to NNs specification.

Although the MTR approach follows the principles of maximum likelihood, it is still no more than a nonlinear least squares method to neural network specification. This is because it is only concerned with modelling the expectational part of the conditional distribution of the target variable. In the model proposed in this paper, we extend specification (5) to account for time-varying conditional volatility. This extension helps also to avoid overfitting. As neural networks have high approximating abilities, some affects due to the presence of changing variance in the data-generating process may be wrongly interpreted as "neglected nonlinearity" and hence captured by the parameters of the network. By modelling the volatility part of the distribution, we hope to robustify the neural network model against time-varying volatility.

5 A combined Neural Network-GARCH model

In this section, we introduce a combined neural network (NN)- GARCH model and explore its statistical properties. Our proposed model is an attempt to jointly specify the conditional mean and variance structure of the data-generating process. Such a joint specification is advantageous as it can be used to draw conclusions on the *interaction* between the first two moments of the conditional distribution, for example how mean innovations affect volatility.

As a first approach, we specify the model so as to comprise a nonlinear NN for the conditional mean and a linear GARCH model for the conditional variance. The motivation for such structure is to allow for much flexibility in the modelling of the conditional mean while at the same time provide simple parsimonious representations of the volatility dynamics. The following section briefly reviews GARCH models and its various extensions proposed in the literature. Section 5.2 introduces the combined NN-GARCH model.

5.1 GARCH models

In a typical high-frequency financial time series one often distinguishes succeeding periods of high and low volatility as well as a tendency for the conditional variance to settle down to a long-run level. These findings, documented in a number of empirical studies of financial prices/returns (see), gave inspire to the class of *(Generalised) Autoregressive Conditional Heteroskedasticity* (ARCH/GARCH) models, introduced by [1, 8]. ARCH/GARCH models can be seen as an extension of time series analysis concepts from the mean of a process to the variance of the errors. They introduce some kind of memory to the second moment of a process, allowing the conditional variance of Y_t (i.e. the "uncertainty" of the process conditional on the information set) to change over time as a function of past errors.

GARCH models have been originally introduced in combination with linear regression models, such as AR, MA and ARMA. A typical ARMA(r, s)-GARCH(p, q) specification takes the form:

$$y_t = \sum_{i=1}^{r} \phi_i y_{t-i} + \sum_{j=1}^{s} \theta_j \epsilon_{t-j} + z_t + \epsilon_t$$

where z_t is a vector of exogenous variables and ϵ_t is a conditionally heteroskedastic normal process with variance h_t, i.e.

$$\epsilon_t | \mathcal{I}_t \sim N\left(0, h_t\right)$$
$$h_t = a_0 + \sum_{i=1}^{p} a_i h_{t-i} + \sum_{j=1}^{q} b_j \epsilon_{t-j}^2$$

To ensure positivity and non-explosive behaviour of h_t, it is typically assumed that $\alpha_0 > 0$, $\alpha_i, \beta_j \geq 0$, $i = 1, 2, ..., p$, $j = 1, 2, ..., q$, and $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$. If $b_i = 0$ for all i then the GARCH(p , q) reduces to an ARCH(q) model.

Since their inception, GARCH models (and especially the simple GARCH(1,1)) have seen exceptional popularity in financial time series applications (see [2] for a relatively recent survey). A significant number of publications have been also devoted to extending the basic GARCH model to account for other effects on the volatility structure. Note that the original GARCH model imposes a symmetry in the response of the variance to past shocks or "news" , where the volatility depends only on the size but not on the sign of the shock. Many authors have argued, however, that in many financial time series the direction of "news" has an asymmetric effect on the conditional volatility and therefore proposed several nonlinear parametrizations of the GARCH structure which account for this asymmetry (see [5, 9, 16]).

GARCH-type models for the volatility have also been introduced in combination with nonlinear (in mean) regression models. See [13] and [14] for examples of threshold and smooth transition heteroskedastic models.

5.2 A Neural Network GARCH model

In our combined NN-GARCH model, we specify the condition mean of the model as a joint linear and neural network process (5), i.e.

$$y_t = \phi' \tilde{x}_t + f(x_t; \theta) + \epsilon_t \tag{6a}$$

The error process is assumed to follow a GARCH(p, q) model

$$\epsilon_t | \mathcal{I}_t \sim N(0, h_t) \tag{6b}$$

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i h_{t-i} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}^2 \tag{6c}$$

Let $\alpha' = (\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q)$ and $\delta = (\phi', \theta', \alpha')' \in \mathbb{R}^m$, where $m = (n+1)(h+1) + p + q + h + 1$. The conditional density of the process is

$$p(y_t | \mathcal{I}_t; \delta) = \frac{1}{\sqrt{2\pi h_t}} \exp\left\{-\frac{1}{2}\frac{\epsilon_t^2}{h_t}\right\}$$

and hence the log-likelihood function is $l = \frac{1}{T}\sum_{t=1}^{T} l_t$ where

$$l_t(\delta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log h_t - \frac{\epsilon_t^2}{2h_t}$$

The maximum likelihood estimator is given by the following equation

$$\hat{\delta} = \underset{\delta \in \Theta}{\operatorname{argmax}}\ l(\delta)$$

When the distribution of ϵ_t is not normal the MLE can be treated as a Quasi-MLE (QMLE).

Notice that the conditional variance equation is recursive and, in general, estimation of a GARCH model requires p and q initial presample values for h_t and ϵ_t. For this reason, the objective function shown above is actually a *conditional* log-likelihood function.

The expected hessian and the Fisher information matrix of the log likelihood are

$$A = E\left(-\frac{\partial^2 l_0}{\partial\delta\partial\delta'}\right)$$

and

$$I = E\left(T\frac{\partial l_0}{\partial\delta}\frac{\partial l_0}{\partial\delta'}\right)$$

The derivatives of the likelihood function are given in the appendix. It is also shown that A is *block-diagonal* between the parameters of the condition mean and variance model. Block-diagonality is convenient in many ways. First of all, several numeric algorithms used to optimize the log-likelihood function (e.g. the Gauss-Newton) require the inversion of the hessian matrix (which is equivalent to A in large samples). With block-diagonality we can invert each block separately and thus save computational effort. Block-diagonality has in addition important implications for the estimation of model parameters: if the conditional mean structure is estimated with a consistent estimator (e.g. NLS) the parameter vector α of the variance model can be separately estimated using the residuals of the conditional mean model. This procedure is asymptotically equivalent to ML, and hence preserves asymptotic efficiency (see [21], Theorem 7.9)).

Several restrictions have to imposed on the parameters of the specification (6) to make it functional. In the expectational part, we impose the following set of restrictions adopted from [15] a) $c_1 < ... < c_h$ which precludes permutation of hidden neurons and b) $\gamma_j > 0, j = 1, 2, ..., h$, which remedies any possible misidentification attributed to the property of the logistic function $F(x) = 1 - F(-x)$. Finally, care has to be taken with overidentified NN models, i.e. models that contain neurons which are never fired, as the value of the log-likelihood function remains unchanged for any value of w_j and c_j. This also implies multiple maxima. To circumvent this problem we have to start with simple models and apply statistical tests of an irrelevant or redundant neuron. Such tests can be easily designed in the framework of ML theory (see below).

The final set of restrictions refers to the conditional variance specification and are similar to those imposed in other GARCH-type models. First, $\alpha_0 > 0$ and $\alpha_i, \beta_j \geq 0$, $i = 1, 2, ..., p$, $j = 1, 2, ..., q$ ensure positivity of h_t. Notice that when the equality holds we obtain a homoskedastic process for $Y_t|\mathcal{I}_t$. In addition, $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$ implies stability and second order stationarity of the variance process (i.e. stationarity in mean and variance).

5.3 A model building methodology for NN-GARCH models

Assuming a specific structure for the NN-GARCH model (6), 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 linear or the conditional variance constant over time. These make in turn the addition of an extra neuron or the specification of a GARCH model redundant. Therefore, we need to adopt a bottom-up approach in model building, starting from the simplest possible specification and when necessary complicating the structure. Apart from reasons stated in section 4.1, this procedure is totally justified in the presence of conditional heteroskedasticity. As neural networks have high approximating abilities it is very likely that some affects due to the presence of conditional heteroskedasticity in the data-generating process be wrongly interpreted as neglected nonlinearity and hence "absorbed" by the parameters of an over-parametrised NN model. The methodology used for specifying the mean part must offer some kind of robustification against these adverse affects. Robustified tests are available within ML theory.

The general rule is to first specify the conditional mean equation, using for example a NLS criterion, and then the conditional variance. This seems the proper way to follow, as the parameters of the conditional mean are consistently estimated by NLS even if the conditional variance is not specified (i.e. errors are assumed homoskedastic) (provided of course that the expectational part is correctly specified for the conditional mean) [21].

Our model building procedure is presented below. The estimation of the conditional mean follows from [15]. However, contrary to this approach, we suggest the use of misspecification-robust formulae both for the asymptotic variance-covariance matric and the various statistics in the framework of ML theory (in particular, the LM).

1. Start with a linear model and under the assumption of homoscedastic errors estimate the regression equation using the LS criterion.
2. Choose a significance level (say $\alpha\%$) and test the null hypothesis of linearity against a neural network model (1) with a single hidden neuron. If linearity is not rejected at the given significance then stop. Otherwise estimate the NN model with a single neuron and test it against a NN model with an additional neuron at half the significance level $1/2\alpha\%$. If null is not rejected

then stop. Otherwise, repeat the above procedure for NN models with h neurons $h = 2, ...,$ until first acceptance of the null. At each subsequent test half the significance level.

3. Given the conditional mean structure, test the null hypothesis of no ARCH against ARCH of a given order, using the LM test proposed in [8]. If the null is rejected then
   (a) Start with estimating a GARCH(1,1) model, using the residuals of the conditional mean.
   (b) Jointly estimate the parameters of an NN-GARCH(1,1) model
   (c) Omit insignificant parameters and reestimate the model. Perform in-sample diagnostic tests (see below). If the model passes all tests accept it. Otherwise, further complicate the mean and/or the variance structure of the model whenever tests indicate it.

We suggest simultaneous estimation of the parameters of the whole model despite the block-diagonality between the mean and variance parts. Although this procedure increases the computational effort, it finally leads to more parsimonious models. As noted in other nonlinear regression models [14], the two-step estimation has a tendency to yield over-parametrised models because some effects due to the non-constant conditional variance may at first be captured by the estimated conditional mean. Although not experimentally verified, similar results are expected to hold for our NN-based methodology. Even in the simultaneous estimation of the parameters, we can still take advantage of block-diagonality for obtaining good initial guesses for the ML estimates of θ and α.

After the estimation of the final model, we suggest a *backward elimination procedure*. This amounts to repeatedly eliminating the parameters of the model with least-significant estimates and re-estimating the reduced model. The backward elimination should terminate when all parts of the model are significant.

The issue of in-sample diagnostic tests is discussed in the following section.

5.4 Misspecification tests

Many in-sample misspecification tests exist in the literature, such as the Ljung - Box - Pierce for autocorrelation in the residuals or the Breysch - Pagan for heteroskedasticity. However, these tests have been developed in the framework of linear regression models. Hence they cannot be directly applied to our case as the asymptotic distribution of the test statistics is not known for NN-GARCH processes. Hence, all the misspecification tests have to be developed from scratch. In this section we propose a unifying framework for various misspecification tests on the NN-GARCH model (6) based on the LM statistic. This is typical in many nonlinear regression methodologies (see e.g. [7,14,17,19]).

5.4.1 The general framework

Consider a pre-specified NN-GARCH parametrization. A unifying framework for misspecification tests can be based on the additive extension of (6):

$$y_t = \phi' \tilde{x}_t + f(x_t;\theta) + A(z_t;\pi_c) + \epsilon_t \tag{7a}$$

$$\epsilon_t | \mathcal{I}_t \sim N(0, h_t) \tag{7b}$$

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i h_{t-i} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}^2 + B(z_t;\pi_v) \tag{7c}$$

where $A(z_t;\pi_c)$, $B(z_t;\pi_v)$ are assumed continuous and twice differentiable for all π_c, π_v everywhere in the corresponding sample spaces. Without loss of generality, we assume that $A(z_t;\pi_c) = 0$ and $B(z_t;\pi_v) = 0$ for $\pi_c = 0$ and $\pi_v = 0$, respectively. The null hypothesis of no additional structure in the conditional mean and variance part coincides with $H_0 : \pi_c = \pi_v = 0$.

To derive a formula for the LM test statistic, note that the inverse of the expected Hessian is block-diagonal i.e.

$$A^{-1} = \begin{pmatrix} J_{\phi\phi'} & J_{\phi\theta'} & J_{\phi\pi_c'} & 0 & 0 \\ J_{\theta\phi'} & J_{\theta\theta'} & J_{\theta\pi_c'} & 0 & 0 \\ J_{\pi_c\phi'} & J_{\pi_c\theta'} & J_{\pi_c\pi_c'} & 0 & 0 \\ 0 & 0 & 0 & J_{\alpha\alpha'} & J_{\alpha\pi_v'} \\ 0 & 0 & 0 & J_{\pi_v\alpha'} & J_{\alpha\alpha'} \end{pmatrix}$$

Hence (4) takes the form

$$LM_T = \frac{1}{T}\begin{pmatrix} \frac{\partial\tilde{l}}{\partial\pi_c'}\tilde{J}_{\pi_c\pi_c'} & \frac{\partial\tilde{l}}{\partial\pi_v'}\tilde{J}_{\pi_v\pi_v'} \end{pmatrix}\begin{pmatrix} \tilde{C}_{\pi_c\pi_c'} & \tilde{C}_{\pi_c\pi_v'} \\ \tilde{C}_{\pi_v\pi_c'} & \tilde{C}_{\pi_v\pi_v'} \end{pmatrix}^{-1}\begin{pmatrix} \tilde{J}_{\pi_c\pi_c'}\frac{\partial\tilde{l}}{\partial\pi_c} \\ \tilde{J}_{\pi_v\pi_v'}\frac{\partial\tilde{l}}{\partial\pi_v} \end{pmatrix} \tag{8}$$

where $\tilde{C}_{\pi_.\pi_.'}$ are the corresponding $\pi_.$-blocks of sample variance-covariance matric and '$\sim$' means evaluated under the null $H_0: \pi_c = \pi_v = 0$.

Provided that H_0 is true, LM_T follows *asymptotically* the χ^2 distribution with $\dim(\pi_c) + \dim(\pi_v)$ degrees of freedom. At a given confidence level $\alpha\%$, one would reject the hypothesis of no sufficient structure in the conditional mean and variance if the value of the test statistic is greater that $\chi^2_{1-\alpha}$, the $1-\alpha$ percentile of the χ^2 distribution. By letting either $A(z_t;\pi_c)$ or $B(z_t;\pi_v) = 0$ be trivially equal to 0 also under the alternative amounts to testing the conditional mean and variance specification separately. This is more customary in practice.

Based on the afore-presented general framework we can derive explicit formulas for various diagnostic tests:

5.4.2 A test for serial correlation

To test the null hypothesis of no serial dependence in the conditional mean we state the alternative as remaining autocorrelation of order r_c in the ordinary residual process, i.e. $A(z_t;\pi_c) = \pi_c' z_t$ with $z_t = (\epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-r_c})$. The null hypothesis of no remaining serial dependence is thus equivalent to testing $\pi_c = 0$. Under this null, the LM statistic is asymptotically χ^2-distributed with r_c degrees of freedom.

5.4.3 A test for neglected nonlinearity

In this test, the alternative hypothesis is the extended model (7) with $A(z_t;\pi_c)$ corresponding to an additional neuron in the network. To avoid problems of identification we approximate $A(z_t;\pi_c)$, as in [17], by a 3rd order Taylor expansion around the null hypothesis, i.e. $A(z_t;\pi_c) = \pi_c' z_t + R_3(z_t)$ where

$$z_t = (x_{1,t}^2, x_{1,t}x_{2,t}, ..., x_{i,t}x_{j,t}, ..., x_{1,t}^3, ..., x_{i,t}x_{j,t}x_{k,t}, ..., x_{h,t}^3)$$

and $R_3(z_t)$ is the remainder. Under the null hypothesis, $\pi_c = 0$, $R_3(z_t) = 0$ and the LM statistic follows asymptotically the χ^2 distribution with $n(n+1)/2 + n(n+1)(n+2)/6$ degrees of freedom.

5.4.4 A test for neglected heteroskedasticity

We can test the presence of additional autoregressive structure in the variance model, by stating the alternative as $B(z_t;\pi_v) = \pi_v' z_t$, with $z_t = (e^2_{t-q-1}, e^2_{t-q-2}, ..., e^2_{t-q-r_v})$. The null hypothesis of no remaining heteroskedasticity is equivalent to testing $\pi_v = 0$. Under this null, the LM statistic is asymptotically χ^2-distributed with r_v degrees of freedom.

6 Application

In order to demonstrate the model building procedure presented in section 5.3, we apply the NN-GARCH model to the task of describing the dynamic structure of returns on the DAX Stock Index. Our model provides a means of investigating how past unpredictable returns relate to both future returns and the volatility of returns. Our sample consists of daily log-price differences from 03/05/1993 to 22/03/2005, excluding weekends, public holidays and other non-trading days[3]. This amounts to a total of 3000 observations, two-thirds of which are used for estimation and the last one third for out-of-sample evaluation. The plots of DAX quotes and returns for the whole sample period are given in figures 2 and 3, respectively. Observe that the returns series is highly heteroskedastic, characterised by succeeding periods of high and low volatility. Several sample statistics for the returns series are given in table 1. Figure 4 depicts the empirical distribution along with an optimally fitted normal probability density function. Observe that the sample distribution highly deviates from the normal prototype and has the characteristic leptokurtotic shape (i.e. pointy peak and fat tails) frequently documented in studies of financial time series.

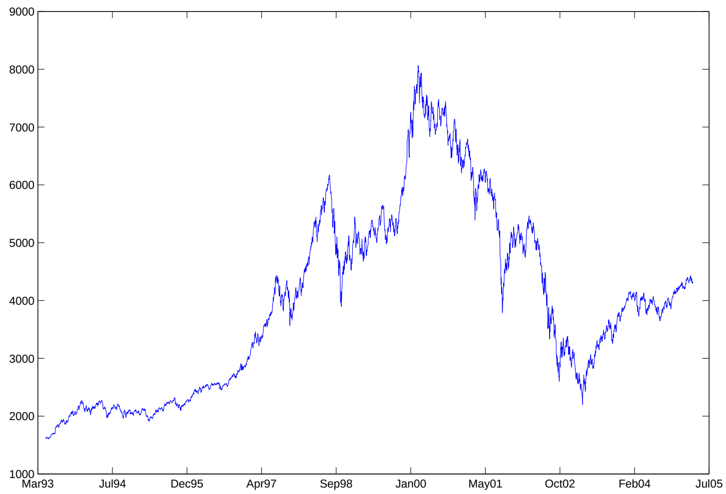

Figure 2: The DAX Stock Index levels for the whole sample period (03/05/1993 - 22/03/2005).

Descriptive statistics	
min	-0.1371
max	0.1200
mean	2.164×10^{-4}
standard deviation	0.012
skewness	-0.2078
kurtosis	10.7262
Jarque-Bera normality test[a]	
statistic	2.843×10^{3}
p-value[b]	0

Table 1: Some sample statistics for the DAX log-returns series.

[a] H_0: The series is sampled from a normal distribution.

[b] The probability that the statistic takes a value as high as the estimated one, given that the null hypothesis is true.

[3] Raw data were taken from Bloomberg.

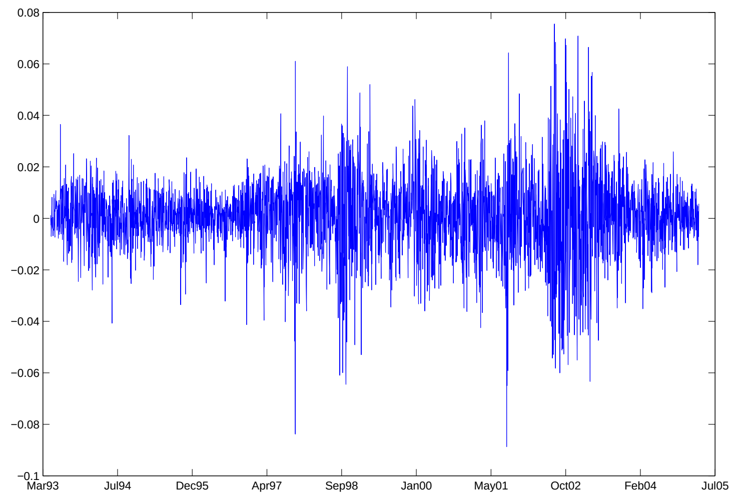

Figure 3: The DAX Stock Index returns for the whole sample period (03/05/1993 - 22/03/2005).

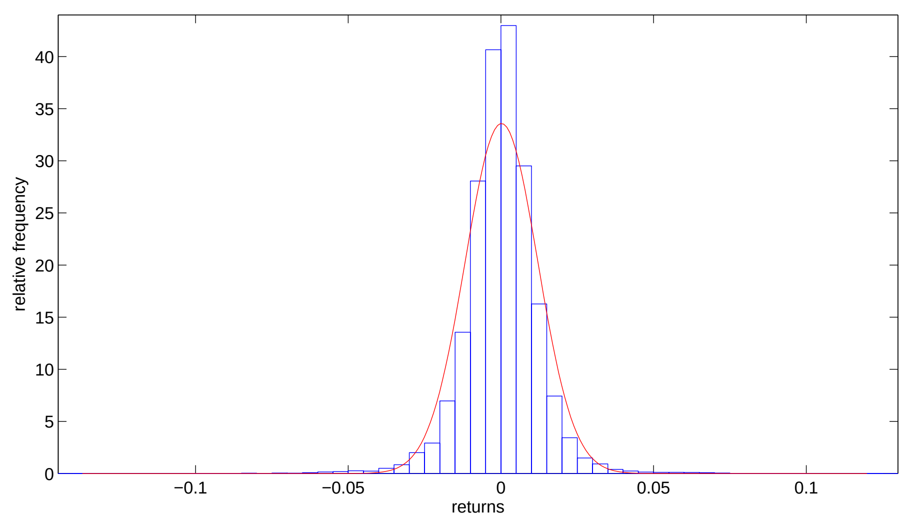

Figure 4: The empirical distribution of DAX returns and the optimal normal pdf fit.

We begin our model building cycle with estimating the linear autoregressive (AR) part of the expectational model. Our explanatory variables set includes lags 1 to 5. The relevant variables are chosen by means of AIC and SBIC. AIC chooses lags 1 to 5, while SBIC surprisingly indicates no lags for the linear part of the model. The residuals of the no-lag model were tested for autocorrelation of order 1 to 15. The hypothesis of no autocorrelation is not rejected at confidence level 5% (see table 2), hence we choose to proceed with SBIC and include no lags in the linear part of the mean model. Next step is to test for nonlinearity. The misspecification-robust test did not indicate the inclusion of any hidden neuron after the linear model was estimated (the significance of the linear model is 18.81%).

Hence the final expectational model includes only a constant term. This is indicative of a no memory process; if the variance of the residuals is also constant over time then DAX returns are pure white noise.

Order l of autocorrelation	LM-statistic	$p-$value
1	0.1051	0.7458
2	1.9399	0.3791
3	2.0166	0.5690
4	2.0453	0.7274
5	2.1732	0.8247
10	7.9840	0.6304
15	22.4585	0.0963

Table 2: The results of the residuals autocorrelation test at the first stage of the model building methodology. The formula used for the test statistic is robust against heteroskedasticity. The null hypothesis is no serial dependence of order l in residuals. Note that at confidence level 5% the null hypothesis cannot be rejected of all orders.

In order to investigate this claim, we test the residuals of the mean model for autoregressive conditional heteroskedasticity using the LM test proposed by Engle [8]. The results of the ARCH test are given in table 3. We see that the hypothesis of no ARCH effects is strongly rejected against an ARCH model of any order, at any reasonable significance level. Hence, the inclusion of a volatility part is meaningful. The next step is to jointly estimate the mean and volatility model, i.e. a constant term with a GARCH(1,1) component. The final model is of the form:

$$y_t = \phi_0 + \epsilon_t \tag{9a}$$

$$\epsilon_t \sim N(0, h_t) \tag{9b}$$

$$h_t = \alpha_0 + \alpha_1 h_{t-1} + \beta_1 \epsilon_{t-1}^2 \tag{9c}$$

where the values of the coefficients are found in table 4. We also report the asymptotic standard error, the z-statistic and the p-value. This is the probability of getting a value of the $z-$statistic greater in magnitude than the observed value, under the hypothesis that the corresponding coefficient is zero. We see that all coefficients are highly significant except for the constant term ϕ_0, which is probably close to zero.

In-sample diagnostics for our model are given in table 5. Panel A reports basic sample moments for the residuals series normalised by the estimated conditional heteroskedasticity ($\hat{e}_t/\hat{h}_t$). We see that normalised residuals have smaller kurtosis than the original time series, thus closely approaching a normal distribution, although their skewness is slightly higher. The value of the Jarque-Bera statistic is also lower. Panels B & C present the results of the misspecification tests for serial correlation and neglected heteroskedasticity (see section 5.4) in the residuals. As seen, the hypothesis of no serial correlation both in mean and the variance of model cannot be rejected for any lag order. Finally, panel D reports the LM test for neglected nonlinearity in the mean of the model. The test rejects the hypothesis of an extra neuron at confidence level 5%. The smallest p-value of the null (0.0904) was obtained for lags 1 to 5.

After the in-sample evaluation, we apply the estimated model to the unseen sample period (13/04/01-22/03/05 - 1000 observations) to obtain one-step ahead forecasts. The sample characteristics of the forecast-error series are given in the first panel of table 6 (under the labels $h = 0$, GARCH(1,1)) and compared to the characteristics of the in-sample residuals. Figure 6, panel B, depicts the out-of-sample empirical distribution of errors. For comparison purposes, we also give the out-of-sample statistics and the empirical distribution of the residuals obtained from a pure expectational model without a GARCH component (see the first panels of table 6 and figure 6). Observe that in the evaluation period the residuals obtained from the estimated GARCH model closely follow a normal distribution. On the contrary, the empirical distribution of the forecast errors obtained from a pure

Order*	ARCHstat	Critical value**	$p-$value
1	125.4137	3.8415	0
2	171.3999	5.9915	0
3	188.2319	7.8147	0
4	205.2957	9.4877	0
5	220.7509	11.0705	0
10	251.7714	18.3070	0
15	266.1981	24.9958	0

* The number of lags of the squared sample residuals included in the ARCH test.
** At 5% confidence level.

Table 3: The results of the ARCH test in the residuals of the expectational model. The null hypothesis is no ARCH effect of order l.

Coefficient	Estimate	Std. Error	z-statistic	p-value
Mean equation				
ϕ_0	0.0129	0.0171	0.7582	0.4483
Variance equation				
a_0	0.0191	0.0008	24.6936	0
a_1	0.8878	0.0013	693.6949	0
b_1	0.0948	0.0037	25.9686	0

Table 4: The estimated linear GARCH(1,1) model.

mean model keeps much of the original leptokurtotic shape. The above claims are also supported by the Jarque-Bera normality tests. Figure 5 shows the mean and volatility forecasts of the model for the evaluation period. The red line represents a 95% confidence "envelope" for the one-step ahead forecasts. This envelope closely follows changes in the fluctuation level of the actual returns series. All the above evidence allows us to conclude that model (9) provides an adequate description of the statistical properties of the returns time series. A pure mean model, on the other hand, cannot capture all the statistical essence of the time series, as it focuses only in mean relations.

Still remains the issue of whether the LM-based statistical tests have correctly identified the appropriate complexity for the model, especially the mean part. As stated above, the inclusion of any neurons was not indicated at any of the usual confidence levels (1, 5 or 10%). However, there is still a possibility of hidden nonlinearity which statistical tests were not able to identify. We can investigate this issue by "forcing" the inclusion of extra neurons in the mean part of the final model. Table 6 shows the out-of-sample statistical performance of four estimated NN-based models, with 1 and 2 neurons and with GARCH or no components. Observe that the complication of the mean structure of the original model does not significantly improve its statistical performance. This is immediately seen by the more or less similar values of the sample and the normality test statistics. Figure 7 provides additional support to this claim by separately showing the contribution of each neuron (degree of activation) for the estimation period. We see that all neurons are intended to capture exceptional spikes of the returns series and therefore do not significantly improve the statistical properties of the returns model.

Sample characteristics (Standardised residuals)					
Sample moments					
mean	-0.0183				
std	0.9981				
skewness	-0.2445				
kurtosis	4.0496				
Jarque-Bera normality test					
statistic	129.8806				
p-value	0				
LM-test for remaining autocorrelation					
mean[1]					
order r	1	2	3	4	5
statistic	0.4251	0.5635	0.5449	0.5970	0.6664
p-values	0.5144	0.7545	0.9089	0.9634	0.9848
variance (neglected heteroskedasticity)[2]					
order r	1	2	3	4	5
statistic	0.2054	0.2608	0.5009	5.9180	7.2206
p-values	0.6504	0.8778	0.9187	0.2054	0.2047
LM-test for neglected nonlinearity [3]					
p-value[4]	0.0904				

[1] H_0: no serial correlation in the residuals.
[2] H_0: no remaining ARCH effects in the residuals.
[3] H_0: sufficient complexity of the mean structure.
[4] The smallest p-value obtained for an extra neuron with input lags 1 to 5 attached to it. The hypothesis of sufficient structure cannot be rejected at confidence 5%.

Table 5: In-sample diagnostics for the estimated linear GARCH(1,1) model.

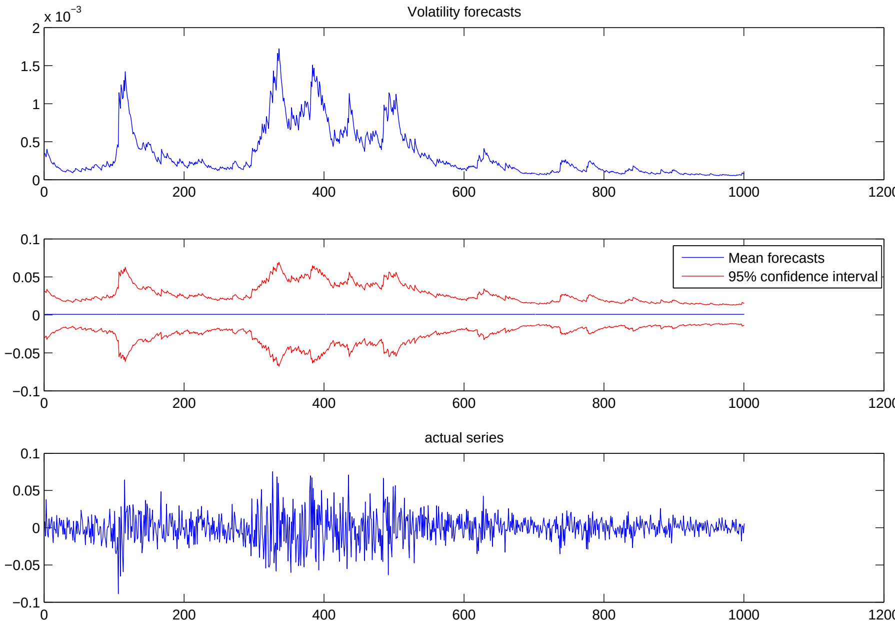

Figure 5: The mean and volatility forecasts provided from the estimated model for the evaluation period.

Number of neurons	Volatility model[b]		In sample	Out of sample
$h = 0$ (only constant term)	no GARCH	*Sample statistics*		
		mean	-3.537×10^{-6}	-9.7571×10^{-4}
		std	0.0132	0.0188
		skewness	-0.4457	-0.0253
		kurtosis	5.7328	5.1778
		Jarque-Bera normality test		
		statistic	686.1980	196.050
		p-value	0	0
	GARCH(1,1)[c]	*Sample statistics*		
		mean	-0.0183	-0.0814
		std	0.9981	1.0168
		skewness	-0.2445	-0.1373
		kurtosis	4.0496	3.1589
		Jarque-Bera normality test		
		statistic	129.8806	4.1079
		p-value	0	0.1282[a]
$h = 1$	no GARCH	*Sample statistics*		
		mean	-6.096×10^{-6}	-0.0010
		std	0.0130	0.0187
		skewness	-0.5228	-0.0749
		kurtosis	5.6941	5.0532
		Jarque-Bera normality test		
		statistic	693.6159	175.0382
		p-value	0	0
	GARCH(1,1)	*Sample statistics*		
		mean	-0.0208	-0.0847
		std	0.9968	1.0140
		skewness	-0.2399	-0.1393
		kurtosis	4.1100	3.1658
		Jarque-Bera normality test		
		statistic	140.4296	4.2895
		p-value	0	0.1171[a]

continued on next page

continued from previous page

Number of neurons	Volatility model		In sample	Out of sample
$h = 2$	no GARCH	*Sample statistics*		
		mean	-2.9303×10^{-6}	-0.0011
		std	0.0129	0.0190
		skewness	-0.5035	-0.0553
		kurtosis	5.7367	5.1941
		Jarque-Bera normality test		
		statistic	706.2277	199.4096
		p-value	0	0
	GARCH(1,1)	*Sample statistics*		
		mean	-0.0184	-0.0837
		std	0.9968	1.0129
		skewness	-0.2461	-0.1376
		kurtosis	4.1540	3.1770
		Jarque-Bera normality test		
		statistic	150.1290	4.3628
		p-value	0	0.1129[a]

[a] The hypothesis that residuals follow a normal distribution is not rejected at confidence 1, 5 or 10%.

[b] For models that include a GARCH component the reported values of statistics refer to the *standardised* residuals.

[c] This is the model indicated by the proposed methodology.

Table 6: The in- and out-of-sample statistical performance of different models, with varying number of neurons and volatility or no component.

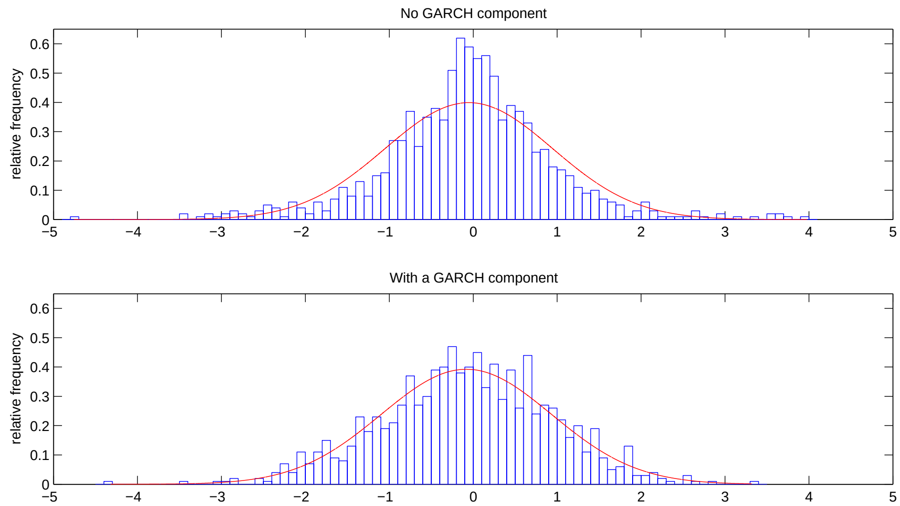

Figure 6: The distribution of the residuals for the evaluation sample period. Panel A refers to the residuals obtained from the pure expectational model (including only a constant term) and panel B those from the combined mean-volatility model.

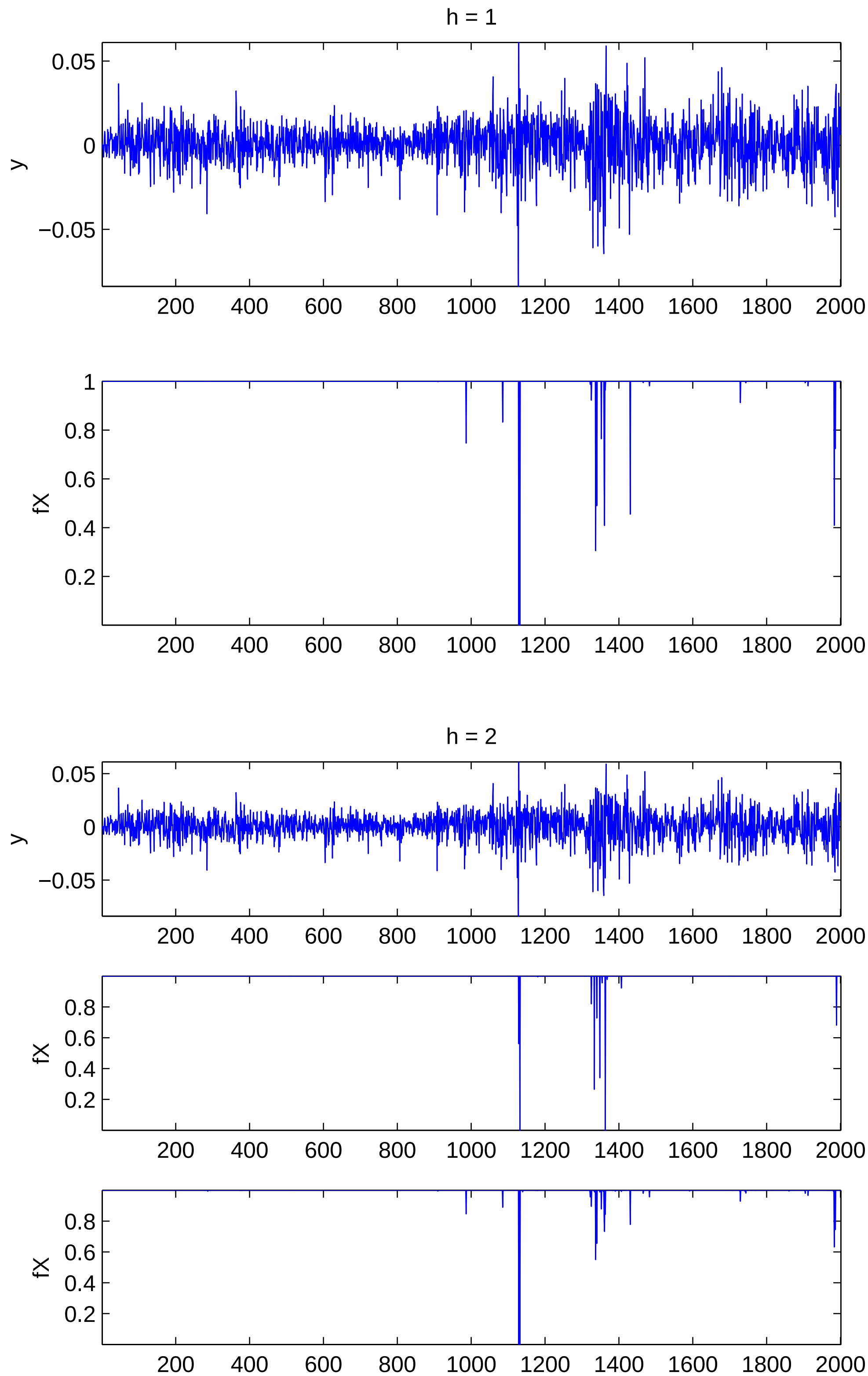

Figure 7: The contribution of extra neurons in each of the nonlinear-in-mean models for the estimation sample period. Top (bottom) panel refers to the NN with one 1 (2) hidden neurons. For each model, we show the raw returns series and the contribution of each hidden neuron (the degree of activation). Note that the output of all neurons is most of time unchanged (i.e. neurons fire at the upper saturation area), and changes on exceptional spikes of the sample series.

7 Conclusions and further research

In this paper, we introduce a new nature-inspired intelligent technique suitable for the modelling of high-frequency financial time series. It combines a neural network parametrization for the mean with a linear GARCH parametrization for the variance of a process. By jointly modelling the conditional mean and volatility of the data-generating process, we extend the scope of NNs from function approximation to statistical modelling 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 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 specification, estimation and evaluation of the final model. We pay special attention to in-sample misspecification tests on the estimated model (residuals autocorrelation, neglected nonlinearity, neglected heteroskedasticity, etc.). Based on the maximum likelihood theory, we device procedures for statistical inference in the framework of NN-GARCH models and thus offer the modeler the opportunity to test hypotheses of interest concerning both the mean and variance structure of the probability distribution. To preserve the validity of the tests in the presence of various forms of misspecification (structure or distributional), we adopt robustified versions of classical ML statistics.

After establishing a complete model building cycle for the family of NN-GARCH specifications, our next goal is to apply this methodology to a number of financial time-series forecasting tasks and see its effectiveness compared to linear or other nonlinear regression models. 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 (Quasi)-Maximum Likelihood Estimator. These are very important as far as the validity of diagnostic tests is concerned.

Although the initial version of the proposed NN-GARCH model is linear-in-variance, it can be easily extended to include nonlinear parametrizations in the variance part and thus account for asymmetric affects in the response of volatility to shocks. The nonlinear-in-mean NN model can be combined with literally any asymmetric model of conditional volatility discussed in section 5.1. One can estimate the joint model by forming the proper log-likelihood function and apply results of section 3 for constructing confidence intervals on the values of the estimates as well as performing statistical inference. One of our future research directions is to extend the neural network parametrization 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. [14] for Smooth Transition).

A Analytical Derivatives of the NN-GARCH model

A.1 Gradient of the log-likelihood

$$\frac{\partial l_t}{\partial \delta'} = \left(\frac{\partial l_t}{\partial \phi'}, \frac{\partial l_t}{\partial \theta'}, \frac{\partial l_t}{\partial \alpha'} \right)$$

$$\frac{\partial l_t}{\partial \phi} = \frac{\epsilon_t}{h_t} \tilde{x}_t + \frac{1}{2h_t} \frac{\partial h_t}{\partial \phi} \left(\frac{\epsilon_t^2}{h_t} - 1 \right) \tag{10}$$

$$\frac{\partial l_t}{\partial \theta} = \frac{\epsilon_t}{h_t} \frac{\partial f_t}{\partial \theta} + \frac{1}{2h_t} \frac{\partial h_t}{\partial \theta} \left(\frac{\epsilon_t^2}{h_t} - 1 \right) \tag{11}$$

$$\frac{\partial l_t}{\partial \alpha} = \frac{1}{2h_t} \left(\frac{\epsilon_t^2}{h_t} - 1 \right) \frac{\partial h_t}{\partial \alpha} \tag{12}$$

with

$$\frac{\partial h_t}{\partial \alpha} = (1, h_{t-1}, ..., h_{t-p}, \epsilon_{t-1}^2, ..., \epsilon_{t-q}^2)' \tag{13}$$

$$\frac{\partial h_t}{\partial \phi} = \sum_{i=1}^{p} \alpha_i \frac{\partial h_{t-i}}{\partial \phi} - 2 \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \tilde{x}_{t-j} \tag{14}$$

$$\frac{\partial h_t}{\partial \theta} = \sum_{i=1}^{p} \alpha_i \frac{\partial h_{t-i}}{\partial \theta} - 2 \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \frac{\partial f_{t-j}}{\partial \theta} \tag{15}$$

for $t = 1, 2, ..., T$ and

$$\frac{\partial f_t}{\partial \theta} = \left(F_1, ..., F_h, \lambda_1 \frac{\partial F_1}{\partial \gamma_1}, ..., \lambda_h \frac{\partial F_h}{\partial \gamma_h}, \lambda_1 \frac{\partial F_1}{\partial w_1'}, \ldots, \lambda_h \frac{\partial F_h}{\partial w_h'}, \lambda_1 \frac{\partial F_1}{\partial c_1}, \ldots, \lambda_h \frac{\partial F_h}{\partial c_h} \right)'$$

where

$$F_j \equiv F(\gamma_j(w_j' x_j - c_j)), \quad \frac{\partial F_j}{\partial \gamma_j} = (w_j' x_j - c_j) F_j', \quad \frac{\partial F_j}{\partial w_j} = \gamma_j x_j F_j', \quad \frac{\partial F_j}{\partial c_j} = -\gamma_j F_j'$$

and

$$F_j' = F'[\gamma_j(w_j' x_j - c_j)] = F(\gamma_j(w_j' x_j - c_j))[1 - F(\gamma_j(w_j' x_j - c_j))] = 2\cosh(\gamma_j(w_j' x_j - c_j))]^{-2}$$

A.2 Hessian of the log-likelihood

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

$$\frac{\partial^2 l_t}{\partial\phi\partial\phi'} = -\frac{1}{h_t}\tilde{x}_t\tilde{x}_t' - 2\frac{\epsilon_t}{h_t^2}\frac{\partial h_t}{\partial\phi}\tilde{x}_t' + \left(\frac{\epsilon_t^2}{h_t} - 1\right)\frac{\partial}{\partial\phi}\left(\frac{1}{2h_t}\frac{\partial h_t}{\partial\phi'}\right) - \frac{1}{2h_t^2}\left(\frac{\epsilon_t^2}{h_t}\right)\frac{\partial h_t}{\partial\phi}\frac{\partial h_t}{\partial\phi'} \quad (16)$$

$$\begin{aligned}\frac{\partial^2 l_t}{\partial\phi\partial\theta'} = &-\frac{1}{h_t}\tilde{x}_t\frac{\partial f_t}{\partial\theta'} - \frac{\epsilon_t^2}{2h_t^3}\frac{\partial h_t}{\partial\phi}\frac{\partial h_t}{\partial\theta'} \\ &- \frac{\epsilon_t}{h_t^2}\left(\frac{\partial h_t}{\partial\phi}\frac{\partial f_t}{\partial\theta'} + \tilde{x}_t\frac{\partial h_t}{\partial\theta'}\right) + \frac{1}{2h_t^2}\left(\frac{\epsilon_t^2}{h_t} - 1\right)\left(\frac{\partial^2 h_t}{\partial\phi\partial\theta'} - \frac{1}{h_t}\frac{\partial h_t}{\partial\phi}\frac{\partial h_t}{\partial\theta'}\right) \quad (17)\end{aligned}$$

$$\frac{\partial^2 l_t}{\partial\phi\partial\alpha'} = \frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t} - 1\right)\left(\frac{\partial^2 h_t}{\partial\phi\partial\alpha'} - \frac{1}{h_t}\frac{\partial h_t}{\partial\phi}\frac{\partial h_t}{\partial\alpha'}\right) - \frac{\epsilon_t\tilde{x}_t}{h_t^2}\frac{\partial h_t}{\partial\alpha'} - \frac{1}{2h_t^2}\left(\frac{\epsilon_t^2}{h_t}\right)\frac{\partial h_t}{\partial\phi}\frac{\partial h_t}{\partial\alpha'} \quad (18)$$

$$\begin{aligned}\frac{\partial^2 l_t}{\partial\theta\partial\theta'} = &-\frac{1}{h_t}\frac{\partial f_t}{\partial\theta}\frac{\partial f_t}{\partial\theta'} + \frac{\epsilon_t}{h_t}\left(\frac{\partial^2 f_t}{\partial\theta\partial\theta'} - \frac{2}{h_t}\frac{\partial h_t}{\partial\theta}\frac{\partial f_t}{\partial\theta'}\right) \\ &+ \left(\frac{\epsilon_t^2}{h_t} - 1\right)\frac{\partial}{\partial\theta}\left(\frac{1}{2h_t}\frac{\partial h_t}{\partial\theta'}\right) - \frac{1}{2h_t^2}\left(\frac{\epsilon_t^2}{h_t}\right)\frac{\partial h_t}{\partial\theta}\frac{\partial h_t}{\partial\theta'} \quad (19)\end{aligned}$$

$$\frac{\partial^2 l_t}{\partial\theta\partial\alpha'} = \frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t} - 1\right)\left(\frac{\partial^2 h_t}{\partial\theta\partial\alpha'} - \frac{1}{h_t}\frac{\partial h_t}{\partial\theta}\frac{\partial h_t}{\partial\alpha'}\right) - \frac{\epsilon_t}{h_t^2}\frac{\partial f_t}{\partial\theta}\frac{\partial h_t}{\partial\alpha'} - \frac{1}{2h_t^2}\left(\frac{\epsilon_t^2}{h_t}\right)\frac{\partial h_t}{\partial\theta}\frac{\partial h_t}{\partial\alpha'} \quad (20)$$

$$\frac{\partial^2 l_t}{\partial\alpha\partial\alpha'} = -\frac{1}{2h_t}\left(\frac{\epsilon_t^2}{h_t} - 1\right)\left(\frac{\partial^2 h_t}{\partial\alpha\partial\alpha'} - \frac{1}{h_t}\frac{\partial h_t}{\partial\alpha}\frac{\partial h_t}{\partial\alpha'}\right) - \frac{1}{2h_t^2}\left(\frac{\epsilon_t^2}{h_t}\right)\frac{\partial h_t}{\partial\alpha}\frac{\partial h_t}{\partial\alpha'} \quad (21)$$

The minus expected hessian of the log-likelihood function is

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

Using the derivatives calculated above and the property $E(.) = E(E(.|\mathcal{I}_t))$ of conditional expectations we have for the diagonal blocks of A

$$A_{\phi\phi'} = \frac{1}{T}E\left\{\sum_{t=1}^{T}\frac{1}{h_t}\tilde{x}_t\tilde{x}_t' + \frac{1}{2h_t^2}\frac{\partial h_t}{\partial\phi}\frac{\partial h_t}{\partial\phi'}\right\}$$

$$A_{\theta\theta'} = \frac{1}{T}E\left\{\sum_{t=1}^{T}\frac{1}{h_t}\frac{\partial f_t}{\partial\theta}\frac{\partial f_t}{\partial\theta'} + \frac{1}{2h_t^2}\frac{\partial h_t}{\partial\theta}\frac{\partial h_t}{\partial\theta'}\right\}$$

$$A_{\alpha\alpha'} = \frac{1}{T}E\left\{\sum_{t=1}^{T}\frac{1}{2h_t^2}\frac{\partial h_t}{\partial\alpha}\frac{\partial h_t}{\partial\alpha'}\right\}$$

and the off-diagonal blocks

$$A_{\phi\theta'} = \frac{1}{T} E \left\{ \sum_{t=1}^{T} \frac{1}{h_t} \tilde{x}_t \frac{\partial f_t}{\partial \theta'} + \frac{1}{2h_t^2} \frac{\partial h_t}{\partial \phi} \frac{\partial h_t}{\partial \theta'} \right\}$$

$$A_{\phi\alpha'} = \frac{1}{T} E \left\{ \sum_{t=1}^{T} \frac{1}{2h_t^2} \frac{\partial h_t}{\partial \phi} \frac{\partial h_t}{\partial \alpha'} \right\}$$

$$A_{\theta\alpha'} = \frac{1}{T} E \left\{ \sum_{t=1}^{T} \frac{1}{2h_t^2} \frac{\partial h_t}{\partial \theta} \frac{\partial h_t}{\partial \alpha'} \right\}$$

It can be proven by arguments similar to [8] 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 [1]), so the combined NN-GARCH model presented above possesses this property.

References

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

[2] T. Bollerslev, R. F Engle, and D.B Nelson. ARCH models. In R. F. Engle and D. McFadden, editors, *The Handbook of Econometrics*, volume 4, pages 2959–3038. North-Holland, 1994.

[3] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. Pentice-Hall International, Inc., 1994.

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

[5] Z. Ding, C.W. Granger, and R.F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1:83–106, 1993.

[6] 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.

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

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

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

[10] B. E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64:413–430, 1996.

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

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

[13] 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.

[14] S. Lundbergh and T. Teräsvirta. Modelling economic high-frequency time series with STAR-STGARCH models. Scandinavian Working Papers, 1998.

[15] M.C. Medeiros, T. Teräsvirta, and G. Rech. Building neural network models for time series: A statistical approach. Scandinavian Working Papers, 2002.

[16] D.B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1991.

[17] G. Rech, T. Teräsvirta, and R. Tschernig. A simple variable selection technique for nonlinear models. *Communications is Statistics, Theory and Methods*, 30:1227–1241, 2001.

[18] 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.

[19] T. Teräsvirta, C.-F. J. Lin, and C.W.J. Granger. Power of the neural network linearity test. *Journal of Time Series Analysis*, 14:309–323, 1993.

[20] H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1:425–464, 1989.

[21] H. White. *Estimation, Inference and Specification Analysis*, volume 22 of *Econometric Society Monographs*. 1996.

[22] 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.

[23] 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.