Report Series

Non-linear versus Non-gaussian Volatility Models

Christian Schittenkopf
Georg Dorffner
Engelbert J. Dockner

Report No. 39
Oktober 1999

Report Series

Oktober 1999

SFB
'Adaptive Information Systems and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This paper was submitted for publication in:
Journal of Empirical Finance

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

Non-linear versus Non-gaussian Volatility Models

Christian Schittenkopf
Austrian Research Institute for Artificial Intelligence

Georg Dorffner
Austrian Research Institute for Artificial Intelligence and
Dept. of Medical Cybernetics and Artificial Intelligence,
University of Vienna, Austria

Engelbert J. Dockner
Dept. of Business Administration,
University of Vienna, Austria

Abstract

One of the most challenging topics in financial time series analysis is the modeling of conditional variances of asset returns. Although conditional variances are not directly observable there are numerous approaches in the literature to overcome this problem and to predict volatilities on the basis of historical asset returns. The most prominent approach is the class of GARCH models where conditional variances are governed by a linear autoregressive process of past squared returns and variances. Recent research in this field, however, has focused on modeling asymmetries of conditional variances by means of non-linear models. While there is evidence that such an approach improves the fit to empirical asset returns, most non-linear specifications assume conditional normal distributions and ignore the importance of alternative models. Concentrating on the distributional assumptions is, however, essential since asset returns are characterized by excess kurtosis and hence fat tails that cannot be explained by models with sufficient heteroskedasticity.

In this paper we take up the issue of returns' distributions and contrast it with the specification of non-linear GARCH models. We use daily returns for the Dow Jones Industrial Average over a large period of time and evaluate the predictive power of different linear and non-linear volatility specifications under alternative distributional assumptions. Our empirical analysis suggests that while non-linearities do play a role in explaining the dynamics of conditional variances, the predictive power of the models does also depend on the distributional assumptions.

Keywords: volatility, neural networks, GARCH, non-linearity, fat tails

1 Introduction

Recent research in financial time series analysis has put a lot of emphasis on modeling and forecasting asset return volatilities. This interest has at least two roots. One stems from the fact that option prices vary with changes in volatilities of the underlying instrument and hence an accurate prediction of future option prices requires a forecast of volatilities. The second one is related to market risk management. Here the concept of value at risk seems to be the industry standard, which requires a forecast of volatilities of the risk factors (like interest rates, exchange rates, market returns) in order to calculate the market risk of a given portfolio of securities.

Many volatility models have been proposed in the finance literature but all of them can be grouped into two categories. There is the class of models that build on historical asset returns and predict volatilities on the basis of different time series analysis techniques. And there is the concept of implied volatility. This is closely related to option prices and requires an option pricing model in order to calculate the market driven volatilities[1]. While the choice of an option valuation model is arbitrary and hence the concept of implied volatility is model dependent, there is on the contrary an advantage over historical volatilities: implied volatilities do not require any concept of conditional expectation and hence are independent of time series properties of asset returns.

In this paper we do not take up the issue of modeling and deriving implied volatilities, instead we focus on predicting historical ones. The most prominent volatility model that estimates conditional variances of asset returns on the basis of historical observations is the GARCH model (Bollerslev, 1986; Bollerslev et al.,1992; Engle, 1982). It is able to capture several important stylized facts of asset returns, namely heteroskedasticity, volatility clustering and excess kurtosis. In the meantime a large body of literature exists that evaluates the predictive power of GARCH models. These studies have found that there exist additional empirical regularities besides volatility clustering and excess kurtosis. These regularities are the leverage effect, the co-movement of volatilities and the reflection of information that accumulates when financial markets are closed.

Based on these stylized facts many variations of parametric GARCH models have been developed over the past decades to account for them. While ARCH and GARCH models capture fat tailed returns and volatility clustering they are not well suited to capture the leverage effect. To account for the leverage effect it is necessary that the conditional variance at time t depends on both the level and the sign of the lagged residual of asset returns. The Exponential GARCH (EGARCH) model of Nelson (1991) accounts for these effects as does the sign-GARCH model by Glosten, Jagannathan and Runkle (1993). Moreover, many papers found that the autoregressive process of conditional variances is integrated and hence the integrated GARCH model was developed (Engle and Bollerslev, 1986).

While in principle the list of stylized facts narrows the field of GARCH models that can be applied to fit asset returns, the number of possible formulations is still enormous. Bera and Higgins (1993) introduce the class of Non-linear ARCH (NARCH) models that nest

[1]Implied volatilities are derived as follows. On the basis of a current option price, as well as the current price of the underlying security a volatility is calculated such that the theoretical option price based on a given option price model coincides with the observed (actual) price.

both the traditional GARCH as well as the Quadratic ARCH (QARCH) model (see Sentana, 1991). While these specifications are important generalizations, they do not capture the leverage effect. Hence, a large stream of research has focused on this variance asymmetry, i.e. future volatility is asymmetrically related to past return innovations, with negative unexpected returns affecting future volatility more than positive unexpected returns. Engle (1990) introduced a simple version of an Asymmetric ARCH (AARCH) model that accounts for both a level and a sign effect. This model was later generalized through the concept of threshold ARCH (TARCH) models by Zakoian (1990). Modeling asymmetric responses of conditional variances to positive and negative news through threshold specifications allows only for two possible regimes: low and high volatility (i.e. the threshold level is 0). While this approach gives a reasonable fit to empirical data, it was questioned by González-Rivera (1998) who proposes a general smooth transition mechanism within a GARCH model that allows for intermediate regimes and hence goes beyond a single threshold. The introduction of a smooth transition meachnism requires a non-linear model. Donaldson and Kamstra (1997) take up the issue of non-linear modeling and propose a semi-nonparametric approach based on artificial neural networks. They find that a proper modeling of non-linearities captures volatility effects that are overlooked by traditional models like GARCH, EGARCH, and sign-GARCH.

In this paper we take up the issue of non-linear modeling based on neural networks but add a detailed analysis of the distributional assumptions underlying these models. In particular, we evaluate the forecasting performance of linear versus non-linear models with gaussian and non-gaussian conditional distributions. As non-gaussian distributions we employ mixtures of gaussians and the Student's-t distribution. Using daily returns on the Dow Jones Industrial Average over the sample period of 1934 to 1997 we find that the improvement in the forecast of non-linear GARCH models over linear ones is negligable but that the specification of the distribution does matter. This result is economically plausible since returns are characterized by substantial fat tails. While time-dependent volatilities can account for some of the excess kurtosis, there is still some left that is only captured by appropriate distributional assumptions.

Our paper is organized as follows. In the next two section we present the classical and the neural network based models employed. Section 4 discusses the data base that is used in the empirical analysis. Section 5 presents the evaluation of the extensive empirical experiments. Finally, Section 6 concludes the paper.

2 Models for return series

The standard setup for modeling return series is to split the return into a deterministic (predictable) component and a random component:

$$r_t = E(r_t|I_{t-1}) + e_t \tag{1}$$

where E denotes the expectation operator, I_{t-1} the conditioning information set (the information available up to time $t-1$) and the e_t are realizations of independent random variables with

$$E(e_t|I_{t-1}) = 0, E(e_t^2|I_{t-1}) = \sigma_t^2. \tag{2}$$

The conditional expectation $E(r_t|I_{t-1})$ is usually formulated as a function of the previous returns:

$$E(r_t|I_{t-1}) = f_1(r_{t-1}, r_{t-2}, \ldots). \tag{3}$$

If σ_t^2 is known to be time-dependent, the return series is called heteroskedastic. The conditional standard deviation σ_t is refered to as volatility. In general, one assumes a deterministic model for σ_t^2 depending on the previous random shocks, i.e.

$$\sigma_t^2 = f_2(e_{t-1}, e_{t-2}, \ldots), \tag{4}$$

or depending on the squared shocks:

$$\sigma_t^2 = f_2(e_{t-1}^2, e_{t-2}^2, \ldots). \tag{5}$$

At this point it is the expertise and intuition of the model builder which drives the selection of reasonable functions for f_1 and f_2 and of an appropriate probability density function (pdf) for e_t. We can choose linear or non-linear functions and gaussian or non-gaussian pdfs.

In the literature the most prominent class of models are the GARCH(p,q) models. In the original paper of Bollerslev (1986) it is assumed that the random shocks are normally distributed with mean 0 and variance σ_t^2 which is denoted $e_t|I_{t-1} \sim \mathcal{N}(0, \sigma_t^2)$. The conditional variance is modeled by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i e_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \tag{6}$$

where the conditions

$$\alpha_0 > 0, \alpha_i \geq 0, \beta_i \geq 0, \tag{7}$$

are imposed on the parameters to ensure $\sigma_t^2 > 0$. This specification implies that the conditional variance follows an autoregressive process for which stationarity is guaranteed, if $\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1$. In many applications it is sufficient to set $p = q = 1$ which we will adopt in the following. We remark that Eq. (6) is of the types specified in Eqs. (4) and (5). For $p = q = 1$, for instance, (and for parameter values which imply stationarity) the recursive Eq. (6) can be rewritten as

$$\sigma_t^2 = \frac{\alpha_0}{1 - \beta_1} + \alpha_1 \sum_{i=1}^{\infty} \beta_1^{i-1} e_{t-i}^2. \tag{8}$$

The conditional variance is thus a *linear* function of all previous squared random components $e_{t-i}^2, i \geq 1$. Due to the significant autocorrelations found in many return series, the conditional expected value is usually assumed to be constant or to be a linear function of the last return:

$$f_1(r_{t-1}, r_{t-2}, \ldots) = a r_{t-1} + b. \tag{9}$$

In other words, the mean of the conditional distribution of the next return is usually specified by the linear function (9). Within the GARCH framework, the variance of this conditional distribution is also given a linear function and because of the assumption of a

normal distribution, the skewness and kurtosis of the conditional distribution are 0 and 3, respectively. In contrast to the mean and the variance, they are not time-dependent.

In order to allow for leptokurtic conditional distributions Bollerslev later proposed to substitute the normal distribution by the standardized Student's-t distribution with ν degrees of freedom (Bollerslev, 1987), i.e. $e_t|I_{t-1} \sim t_\nu(0, \sigma_t^2)$, $\nu > 2$, with pdf

$$\rho_\nu(e_t|I_{t-1}) = \Gamma\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\pi(\nu-2)\sigma_t^2\right)^{-\frac{1}{2}}\left(1+\frac{e_t^2}{(\nu-2)\sigma_t^2}\right)^{-\frac{\nu+1}{2}}. \qquad (10)$$

For $\nu \to \infty$, $t_\nu(0, \sigma_t^2) \to \mathcal{N}(0, \sigma_t^2)$. Therefore GARCH-$t$ models are more general than GARCH models. The degrees of freedom ν are an additional parameter which determines the kurtosis of the conditional distribution. More precisely, for $\nu > 4$ the kurtosis is given by $3(\nu-2)/(\nu-4)$. Since the conditional pdf (10) is symmetric, the skewness is 0. We remark that the skewness and kurtosis of the conditional distribution are (again) not time-dependent.

3 Neural network models

Over the last decade the modeling paradigm of (artificial) neural networks has gained widespread interest. Many different types of neural networks have been developed (Bishop, 1995, Hertz et al., 1991, Rojas, 1996) and applied in a variety of fields including econometrics and finance (Refenes, 1995). Most frequently, neural networks have been used in the context of non-linear regression. From a statistical point of view, neural networks are a technique for modeling non-linear relationships *semi-nonparametrically*. Depending on the assumed statistical relationship, the following neural network models are of particular interest.

3.1 Multi-layer perceptrons

The most popular neural network for non-linear regression tasks is the multi-layer perceptron (MLP). The statistical framework for the application of an MLP is given by the model[2]

$$y_t = f(x_t) + e_t' \qquad (11)$$

where y_t denotes the dependent variable, x_t denotes the independent variable, and e_t' is a zero mean gaussian random variable of variance σ^2. This model is homoskedastic. For particular sets of realizations $\{x_t : t = 1, \ldots, N\}$ and $\{y_t : t = 1, \ldots, N\}$, the non-linear regression task consists of modeling the unknown map f given the N pairs (x_t, y_t). In the simplest case, an MLP with one input unit, one layer of hidden units and one output unit is defined by

$$\mathrm{MLP}(x_t) = g\left(\sum_{j=1}^{H} v_j h(w_j x_t + c_j) + b\right) \qquad (12)$$

[2]For simplicity, we restrict our attention to one-dimensional models in this section.

where H denotes the number of hidden units, w_j and v_j the weights of the first and second layer, and c_j and b the biases of the first and second layer (see also Fig. 1). The activation function h of the hidden units is usually chosen to be bounded, non-linear, and increasing[3], whereas the output unit is often allowed to produce outputs of arbitrary size (e.g. $g(x) = x$).

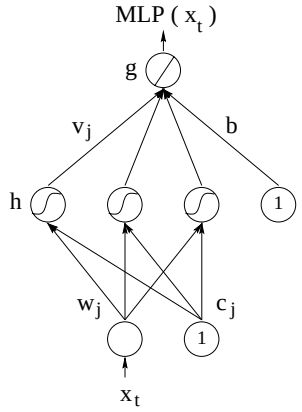

Figure 1: A multi-layer perceptron with one input unit, three non-linear hidden units and one linear output unit.

In contrast to the number of input and output units, which is fixed by the particular regression problem, the number of hidden units H must be chosen ad hoc. In general, MLPs can approximate any smooth, non-linear function to arbitrary accuracy as the number of hidden units tends to infinity (Hornik et al., 1989). However, one always tries to keep H as small as possible to have networks of moderate size. The choice of H is thus a compromise between necessary approximation power with respect to the data set and the number of free parameters (weights). For the network specified in Eq. (12), the number of weights equals $3H + 1$. In order to train MLPs on regression tasks, i.e. to fit the set of input vectors $\{x_t : t = 1, \ldots, N\}$ to the set of output vectors (targets) $\{y_t : t = 1, \ldots, N\}$, one typically minimizes the *mean squared error*

$$\text{MSE} = \frac{1}{N} \sum_{t=1}^{N} (y_t - \text{MLP}(x_t))^2. \tag{13}$$

As stated above, the assumption underlying this model is that the variance of the target (conditioned on the input) is constant, or more precisely, that the conditional pdf of the target is a single gaussian of constant variance.

Besides non-linear regression, MLPs have also been applied to time series processing tasks. The MLP depicted in Fig. 1 for instance, can be interpreted as a non-linear autoregressive model of first order. A similar (multivariate) setup has been applied recently to predict intraday volatilities for the Spanish stock market (González Miranda and Burgess, 1997). The non-linear neural networks are reported to produce forecasts which dominate forecasts from traditional linear methods (out-of-sample). In the context

[3]We took the hyperbolic tangens: $h(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

of heteroskedastic time series modeling, MLPs (with other activation functions) have been used to capture volatility effects in stock returns overlooked by traditional asymmetric volatility models (Donaldson and Kamstra, 1997). The conditional pdf of the return series however, is still assumed to be gaussian in contrast to the conditional pdfs estimated by the neural networks described below.

3.2 Mixture density networks

A much more powerful framework for modeling non-linear statistical dependences is provided by the concept of mixture density networks (MDNs, Bishop, 1994, 1995; Neuneier et al., 1994; Ormoneit, 1998). In the past MDNs have also been proposed to allow for heteroskedasticity in return series models in a semi-nonparametric way (Ormoneit and Neuneier, 1996; Schittenkopf et al., 1998). MDNs are able to learn conditional (target) pdfs of non-constant variance or, more generally, they are able to approximate arbitrary non-gaussian, even multimodal pdfs. Thereby the main idea is to use MLPs to predict the parameters of the conditional pdf of the variable y_t in dependence of the variable x_t[4]. These parameters are the priors, the centres, and the widths of a weighted sum of gaussian pdfs (mixture of gaussians). The conditional pdf of y_t is thus approximated by

$$\rho(y_t|x_t) = \sum_{i=1}^{n} \alpha_{i,t} k(\mu_{i,t}, \sigma_{i,t}^2), \tag{14}$$

$$k(\mu_{i,t}, \sigma_{i,t}^2) = \frac{1}{\sqrt{2\pi\sigma_{i,t}^2}} \exp\left(-\frac{(y_t - \mu_{i,t})^2}{2\sigma_{i,t}^2}\right) \tag{15}$$

where the parameters $\alpha_{i,t}$, $\mu_{i,t}$, and $\sigma_{i,t}^2$ of the n gaussian components are estimated by

$$\begin{aligned}
\alpha_{i,t} &= s(\tilde{\alpha}_{i,t}) = \frac{\exp(\tilde{\alpha}_{i,t})}{\sum_{j=1}^{n} \exp(\tilde{\alpha}_{j,t})}, & (16)\\
\tilde{\alpha}_{i,t} &= \mathrm{MLP1}_i(x_t), & (17)\\
\mu_{i,t} &= \mathrm{MLP2}_i(x_t), & (18)\\
\sigma_{i,t}^2 &= (\mathrm{MLP3}_i(x_t))^2. & (19)
\end{aligned}$$

The softmax function $s(\tilde{\alpha}_{i,t})$ ensures that the priors $\alpha_{i,t}$ are positive and that they sum up to one, which makes the right-hand side of Eq. (14) a pdf. The quadratic output function in Eq. (19) guarantees non-negative variances. As a result, each MLP receives the same (one-dimensional) input x_t and produces a different n-dimensional output[5] where n equals the number of gaussian components[6]. The representation (14) is completely general since gaussian mixture models can approximate any pdf to, in principle, arbitrary accuracy (McLachlan and Basford, 1988). An MDN with one input unit, three non-linear hidden

[4] Again, we only discuss the one-dimensional case.

[5] Choosing again the activation function $g(x) = x$, the ith output of the corresponding MLP is given by $\mathrm{MLP}_i(x_t) = \sum_{j=1}^{H} v_{ij} h(w_j x_t + c_j) + b_i$. In the case of several output units the weight vector v_j of the second layer is thus replaced by a weight matrix v_{ij}, and the bias b is replaced by a bias vector b_i.

[6] This is an extension of the standard MDN (Bishop, 1994) in that it uses three separate MLPs to estimate the parameters.

units and two output units is depicted in Fig. 2. Connections which are not changed during the training, are indicated by weights equal to 1. The total number of (trained) weights[7] is $3(n+2)H+3n$.

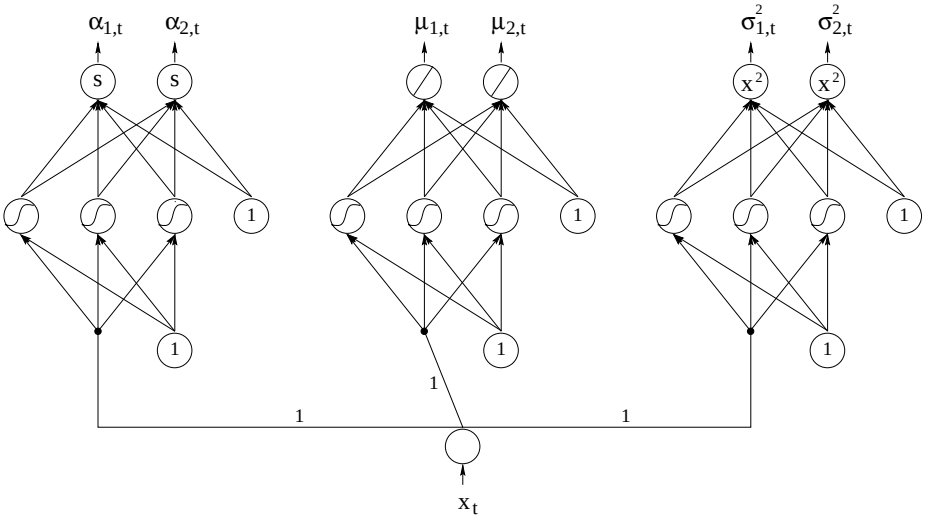

Figure 2: A mixture density network with one input unit, three non-linear hidden units (for each MLP) and two output units (for each MLP).

We emphasize that the assumed statistical model is heteroskedastic:

$$y_t|x_t \sim \mathcal{D}(\mu_t, \sigma_t^2) \tag{20}$$

where $\mathcal{D}(\mu_t, \sigma_t^2)$ denotes any probability distribution, and the conditional mean μ_t and the conditional variance σ_t^2 are non-linear functions of the independent variable x_t (Bishop, 1994):

$$\mu_t = \sum_{i=1}^{n} \alpha_{i,t}\mu_{i,t}, \tag{21}$$

$$\sigma_t^2 = \sum_{i=1}^{n} \alpha_{i,t}\left(\sigma_{i,t}^2 + (\mu_{i,t}-\mu_t)^2\right). \tag{22}$$

We briefly remark here that an MDN with one gaussian component ($n=1$) can be interpreted as a non-linear extension of the ARCH(1) model[8] (Engle, 1982)

$$r_t = \mu_{A,t} + e_t, \mu_{A,t} = b \tag{23}$$

where e_t denotes a zero mean gaussian random variable of variance

$$\sigma_{A,t}^2 = \alpha_0 + \alpha_1 e_{t-1}^2, e_{t-1} = r_{t-1} - b. \tag{24}$$

[7]This is, however, not the number of free parameters of the model. For $n=1$, the prior $\alpha_{1,t}$ equals 1 for all inputs, and therefore MLP1 may be ignored.

[8]In the context of time series, x_t and y_t are subsequent values r_{t-1} and r_t of a time series.

In the context of MDNs, the rather simple models for $\mu_{A,t}$ (constant) and for $\sigma^2_{A,t}$ (quadratic in r_{t-1}) are thus replaced by arbitrary smooth, non-linear functions of r_{t-1} which are represented by MLP2 and MLP3, respectively.

In order to train MDNs or other heteroskedastic models, one must use a more general error function than the MSE (see Eq. (13)). If the samples (x_t, y_t), $t = 1, \ldots, N$, are supposed to be independent, the joint pdf $\rho(x_1, y_1; \ldots; x_N, y_N)$ can be rewritten as

$$\rho(x_1, y_1; \ldots; x_N, y_N) = \prod_{t=1}^{N} \rho(y_t|x_t) \prod_{t=1}^{N} \rho(x_t) \tag{25}$$

where the second product does not depend on the parameters of the MDN. Therefore, one must optimize the weights of the network with respect to the first product only, which is the joint likelihood function $\mathcal{L}$. Now maximizing $\mathcal{L}$ is equivalent to minimizing the average negative loglikelihood function

$$\begin{aligned} \bar{\mathcal{L}} &= -\frac{1}{N} \log \mathcal{L} \\ &= -\frac{1}{N} \sum_{t=1}^{N} \log \rho(y_t|x_t) \end{aligned} \tag{26}$$

where $\rho(y_t|x_t)$ is defined by Eq. (14). The general error function $\bar{\mathcal{L}}$ for training an MDN is thus given by Eq. (26). If we assume for a moment that the MDN has only one component ($n = 1$), this function reduces to

$$\bar{\mathcal{L}}' = \frac{1}{N} \sum_{t=1}^{N} \left(\frac{1}{2} \log(2\pi\sigma_t^2) + \frac{(y_t - \mu_t)^2}{2\sigma_t^2} \right). \tag{27}$$

One can easily see that the standard MLP error function, i.e. the MSE, is a special case of Eq. (27) (σ_t constant)[9]. In summary, MDNs are more general neural network models for non-linear regression than MLPs.

3.3 Recurrent networks

We now come back to our goal of modeling financial time series. Notationally, y_t and x_t of the previous section are replaced by the next return r_t and the previous return r_{t-1} (and the conditioning information set I_{t-1}). Apart from some initial condition, the average negative loglikelihood function, which has to be minimized, is given by

$$\bar{\mathcal{L}} = -\frac{1}{N} \sum_{t=1}^{N} \log \rho(r_t|I_{t-1}). \tag{28}$$

As an extension of the ARCH(1) model, the GARCH(1,1) model gains most of its power to estimate the conditional variance σ_t^2 from considering not only the previous squared error e_{t-1}^2 but also the previous conditional variance σ_{t-1}^2. In this paper we extend the

[9] The factors and terms which do not depend on the weights of the network, may be neglected during minimization.

MDN architecture in a *recurrent* way to allow for "GARCH effects". Our new models are thus non-linear and non-gaussian generalizations of the class of GARCH models. With respect to the definition of an MDN we only change Eq. (19) appropriately. A recurrent MDN with n gaussian pdfs (RMDN(n)) is defined by Eqs. (14) - (18) and by

$$\sigma_{i,t}^2 = \left| \sum_{j=1}^{H} v_{ij} h \left(w_{j0} e_{t-1}^2 + \sum_{k=1}^{n} w_{jk} \sigma_{k,t-1}^2 + c_j \right) + b_i \right| . \tag{29}$$

An RMDN with two gaussian components ($n = 2$) and three hidden units (for each MLP) is depicted in Fig. 3. In fact, only the subnetwork estimating the variances of the gaussian components is recurrent. Its input is $(n+1)$-dimensional (plus bias) and consists of the squared error[10] $e_{t-1}^2 = (r_{t-1} - \mu_{t-1})^2$ and the n previous conditional variances $\sigma_{k,t-1}^2$, $k = 1, \ldots, n$. The activation function of the n output units is chosen as $g(x) = |x|$ to ensure positive network outputs, i.e. conditional variances. The subnetworks estimating the priors and the centres remain standard MLPs. The total number of weights of an RMDN(n) network equals $2(2n+3)H + 3n$.

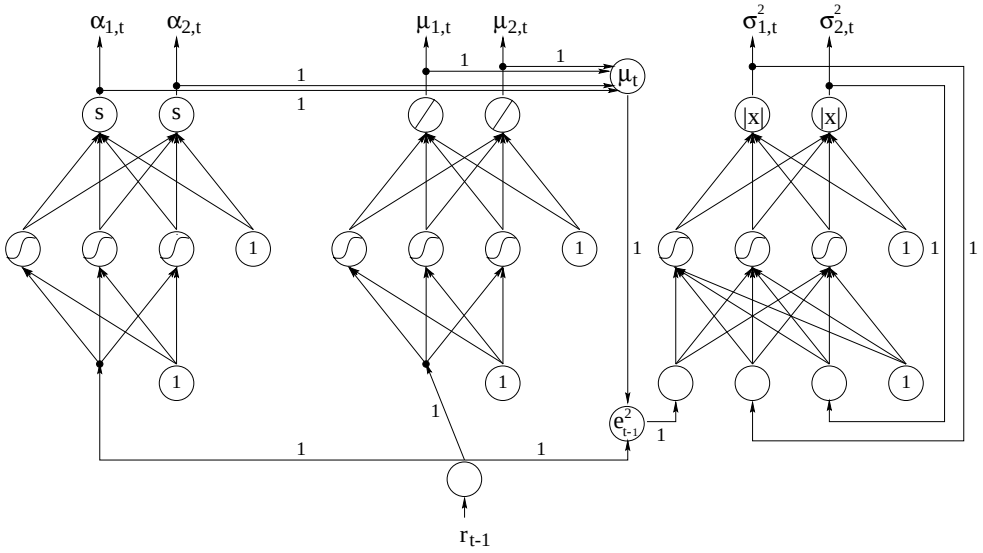

Figure 3: A recurrent mixture density network with two gaussian components.

We briefly remark here that an RMDN with one gaussian component ($n = 1$) can be interpreted as a non-linear extension of a GARCH(1,1) model. The MLP estimating the prior, which is equal to 1, is meaningless and MLP2 and MLP3 perform a non-linear estimation of the mean μ_t and the variance σ_t^2 of the conditional distribution of the next return r_t , respectively.

In general ($n \geq 1$), the mean and the variance of the conditional distribution are given by Eqs. (21) and (22), i.e. by non-linear, time-dependent functions. For the skewness s_t

[10] In Fig. 3, the calculation of μ_t, which is fed back as μ_{t-1} in the next time step, and of e_{t-1}^2 is indicated by extra units.

and the kurtosis k_t of the conditional distribution one gets

$$s_t = \frac{1}{\sigma_t^3} \sum_{i=1}^{n} \alpha_{i,t} \left(3\sigma_{i,t}^2 (\mu_{i,t} - \mu_t) + (\mu_{i,t} - \mu_t)^3 \right), \tag{30}$$

$$k_t = \frac{1}{\sigma_t^4} \sum_{i=1}^{n} \alpha_{i,t} \left(3\sigma_{i,t}^4 + 6\sigma_{i,t}^2 (\mu_{i,t} - \mu_t)^2 + (\mu_{i,t} - \mu_t)^4 \right). \tag{31}$$

Therefore and in contrast to the GARCH and GARCH-t models, the conditional skewness and the conditional kurtosis are also non-linear, time-dependent functions.

There are two other models which should be discussed in order to evaluate the influence of (non-)linear functions and (non-)gaussian pdfs on the performance of the described models in detail. First, it is necessary to study non-linear GARCH-t models. This can be done in the framework of RMDNs ($n = 1$) by replacing the mixture of gaussians in Eq. (14) by the density of the t-distribution given in Eq. (10). We will refer to these models as RMDN(1)-t models. Secondly, it is interesting to study the performance of mixture models for the case that only linear functions are allowed. More precisely, the three MLPs estimating the parameters of the mixture model are replaced by linear functions[11]. For instance, the conditional variances $\sigma_{i,t}^2$, $1 \leq i \leq n$, are now modeled by

$$\sigma_{i,t}^2 = \left| v_{i0} e_{t-1}^2 + \sum_{k=1}^{n} v_{ik} \sigma_{k,t-1}^2 + b_i \right| \tag{32}$$

where taking the absolute value of the outputs is necessary to ensure that all conditional variances are positive. These linear mixture models are refered to as LRMDN(n) models.

4 Empirical Analysis

4.1 Data Set: Dow Jones Industrial Average

The data set we used in our numerical experiments are the daily closing values s_t of the Dow Jones Industrial Average (DJIA) between November 3, 1934 and December 31, 1997. The data[12] were transformed into continuously compounded returns r_t (in percent) by applying the transformation

$$r_t = 100 \log \frac{s_t}{s_{t-1}}. \tag{33}$$

The resulting set of 16630 returns, which is displayed in Fig. 4, was divided into 10 training sets of length 1100 with subsequent test sets of length 563. In Fig. 4 the training and test sets are indicated by dashed and dotted lines, respectively. The exact periods of time[13] covered by the sets together with some basic statistics are available from Table 1 and 2. In particular, the mean, the standard deviation, the skewness, and the kurtosis are reported for all training and test sets as well as the minimum and the maximum return.

[11] This can be easily implemented by changing the non-linear transfer function $h(x) = \tanh(x)$ in Eq. (12) to $h(x) = x$. The resulting MLPs are overparametrized but they in fact perform a linear mapping.

[12] The main part of the data set was obtained from ftp://wueconb.wustl.edu/econ-wp/data/papers/9603/9603001.tar.gz. Several obvious errors in this data set were corrected.

[13] Until June 1952 the NYSE operated on Saturdays. Therefore the first six sets cover periods several months shorter than the remaining sets.

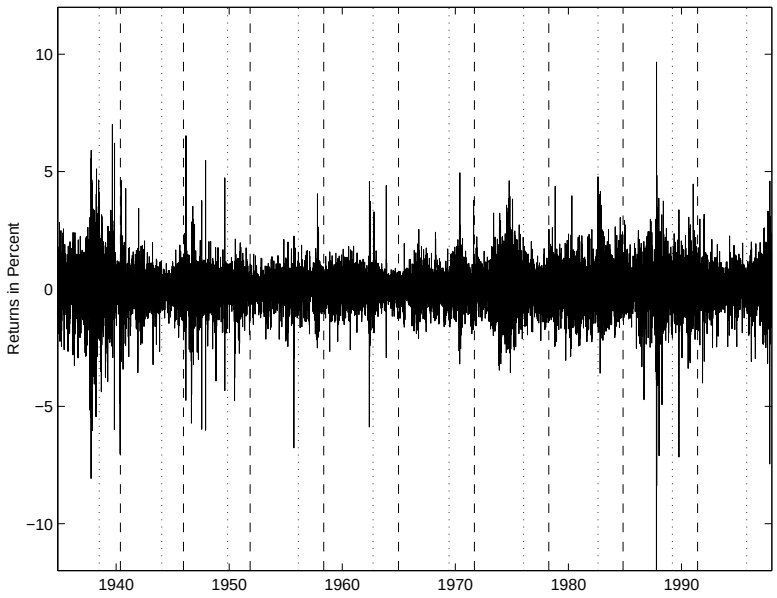

Figure 4: The time series of daily returns of the Dow Jones Industrial Average over a period of more than 60 years.

If volatility is measured by the standard deviation, a very volatile period of time are the thirties (set 1). The 50s and 60s are a rather calm period (sets 4, 5, and 6). In the 70s volatility increases again during the oil crisis (set 7). The large standard deviation in set 9 is due to the crash in 1987 which is also the source of the increased kurtosis. For a more detailed inspection of daily DJIA data between 1928 and 1989, the reader is refered to (Turner and Weigel, 1992).

4.2 Error Measures

The parameters of all models were optimized with respect to the loss function $\bar{\mathcal{L}}$ (see Eq. (28)), which is the average negative loglikelihood of the sample apart from some initial condition. In the following, we use the term *loss function* rather than likelihood because $\bar{\mathcal{L}}$ can also be evaluated on a test set. The optimization routine was a scaled conjugate gradient algorithm. For the models with t-distributions, the degrees-of-freedom parameter was additionally optimized by a one-dimensional search routine. The training of the models was stopped when no further decrease of the loss function could be achieved[14]. Depending on the model and the particular data set, the number of training cycles varied considerably.

In addition to the loss function, we also calculated alternative error measures. Since the actual purpose of a volatility model is to predict future volatility, the performance of all models was compared to the performance of a naive volatility predictor. The squared returns r_t^2 were considered the "true" volatility at time t. Furthermore, the ability of the

[14]except for the GARCH-t models on the sets 2 and 3 (see below)

Set	Period	Mean	Std.	Skew.	Kurt.	Min.	Max.
1	03/11/1934–02/07/1938	0.033	1.265	-0.497	7.897	-8.072	5.900
2	18/05/1940–12/01/1944	0.010	0.768	-0.729	13.442	-7.020	4.625
3	17/12/1945–09/11/1949	0.000	0.902	-0.647	14.647	-6.012	6.528
4	10/11/1951–15/02/1956	0.053	0.610	-1.690	18.404	-6.766	2.251
5	14/05/1958–24/09/1962	0.023	0.712	-0.419	9.614	-5.882	4.579
6	22/12/1964–12/06/1969	0.003	0.609	-0.020	3.930	-2.123	2.543
7	07/09/1971–13/01/1976	0.001	1.072	0.239	4.012	-3.567	4.603
8	10/04/1978–12/08/1982	0.002	0.872	0.085	4.033	-3.038	4.363
9	02/11/1984–10/03/1989	0.058	1.371	-6.054	118.649	-25.632	9.666
10	05/06/1991–06/10/1995	0.041	0.663	-0.215	6.300	-4.006	3.187

Table 1: Statistics for the 10 training sets.

Set	Period	Mean	Std.	Skew.	Kurt.	Min.	Max.
1	05/07/1938–17/05/1940	-0.017	1.224	-0.378	9.722	-7.043	7.012
2	13/01/1944–15/12/1945	0.061	0.508	-0.261	4.620	-2.066	1.867
3	10/11/1949–09/11/1951	0.055	0.711	-1.215	8.778	-4.765	2.128
4	16/02/1956–13/05/1958	-0.004	0.716	0.179	5.270	-2.514	4.048
5	25/09/1962–21/12/1964	0.069	0.559	0.766	12.268	-2.931	4.403
6	13/06/1969–03/09/1971	0.004	0.833	0.550	6.391	-3.193	4.952
7	14/01/1976–07/04/1978	-0.034	0.708	0.010	2.769	-2.199	2.019
8	13/08/1982–01/11/1984	0.077	1.033	0.702	4.582	-3.586	4.781
9	13/03/1989–04/06/1991	0.048	0.996	-0.602	8.681	-7.155	4.466
10	09/10/1995–31/12/1997	0.092	0.962	-0.865	10.616	-7.455	4.601

Table 2: Statistics for the 10 test sets.

models to predict increases and decreases of volatility was investigated. In particular, the following error measures were calculated where $\hat{\sigma}_t^2$ denotes the estimated conditional variance[15]:

$$\text{NMSE} = \sqrt{\frac{\sum_{t=1}^{N}(r_t^2 - \hat{\sigma}_t^2)^2}{\sum_{t=1}^{N}(r_t^2 - r_{t-1}^2)^2}} \tag{34}$$

$$\text{NMAE} = \frac{\sum_{t=1}^{N}|r_t^2 - \hat{\sigma}_t^2|}{\sum_{t=1}^{N}|r_t^2 - r_{t-1}^2|} \tag{35}$$

$$\text{HR} = \frac{1}{N}\sum_{t=1}^{N}\theta_t, \tag{36}$$

$$\theta_t = \begin{cases} 1 & : \quad (\hat{\sigma}_t^2 - r_{t-1}^2)(r_t^2 - r_{t-1}^2) \geq 0 \\ 0 & : \quad \text{else} \end{cases} \tag{37}$$

[15] For the mixture models, the accumulated conditional variance is inserted (see Eq. (22)).

$$\text{WHR} = \frac{\sum_{t=1}^{N} \text{sgn}((\hat{\sigma}_t^2 - r_{t-1}^2)(r_t^2 - r_{t-1}^2))|r_t^2 - r_{t-1}^2|}{\sum_{t=1}^{N} |r_t^2 - r_{t-1}^2|} \tag{38}$$

The *normalized mean squared error NMSE* relates the MSE $1/N \sum_{t=1}^{N}(r_t^2 - \hat{\sigma}_t^2)^2$ of the modeled volatility $\hat{\sigma}_t^2$ to the mean squared error of the naive model $\hat{\sigma}_t^2 = r_{t-1}^2$. The naive model thus serves as a benchmark model which, of course, should be beaten. In this case the NMSE is smaller than 1. The minimum value is 0. The *normalized mean absolut error NMAE* also compares the actual model to the naive model. However, it is more robust against outliers and generally, it gives values closer to 1. The *hit rate HR* is the relative frequency of correctly indicated increases and decreases of volatility, i.e. it measures how often the model gives the correct direction of change of volatility. The HR lies between 0 and 1. A value of 0.5 indicates that the model is not better than a random predictor generating a random sequence of ups and downs (provided that ups and downs are equally likely). The *weighted hit rate WHR* additionally takes the real changes $r_t^2 - r_{t-1}^2$ into account meaning that large changes are considered more important than small changes. The WHR lies between -1 (worst case) and 1 (best case).

4.3 Results

We fitted GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) models to each of the 10 training sets. Consequently, the total number of fitted models is 60. In the rest of the paper the various models are sometimes also refered to as model 1 to model 6. The estimated parameter values for the GARCH(1,1) and the GARCH(1,1)-t models are given in Table 3 and 4, respectively. For the network based models, which are semi-nonparametric, the weights of the MLPs can be hardly interpreted, and they are therefore not reported. The number of hidden units in the MLPs was chosen as $H = 3$. For the six classes of models, the number of free parameters of a particular model is thus given by 5, 33, 6, 34, 48, and 48, respectively (see also Section 3).

Set	a	b	α_0	α_1	β_1
1	*0.044	0.045	0.017	0.065	0.925
2	0.190	*0.011	0.026	0.064	0.883
3	0.148	*0.012	0.144	0.244	0.594
4	0.284	0.024	0.112	0.132	0.531
5	0.211	0.020	0.048	0.146	0.753
6	0.248	*0.010	0.028	0.116	0.806
7	0.242	*0.003	*0.005	0.060	0.937
8	0.095	*0.004	0.022	0.033	0.937
9	*0.040	0.084	0.080	0.147	0.797
10	*0.024	0.039	0.035	0.045	0.875

Table 3: Parameter values for the GARCH(1,1) models. Parameters with an asterisk (*) are not significant (at the 5% level).

The GARCH(1,1) models summarized in Table 3 are all stationary since the condition $\alpha_1 + \beta_1 < 1$ derived in (Bollerslev, 1986) holds[16]. If these conditions are met, the first and second moment of a GARCH(1,1) process exist and they are given by $E(r_t) = b/(1-a)$ and $E((r_t - E^2(r_t))^2) = \alpha_0/((1-a^2)(1-\alpha_1-\beta_1))$. Interestingly, the correlation of returns is strong enough to produce significant parameters a on the sets 2–8.

Set	a	b	α_0	α_1	β_1	ν
1	*0.034	0.076	*0.014	0.062	0.930	5.782
2	0.141	*0.037	0.037	0.091	0.875	2.945
3	0.126	*0.027	0.313	0.310	0.605	2.440
4	0.201	0.055	0.045	0.104	0.756	5.838
5	0.188	0.035	0.056	0.149	0.734	8.634
6	0.241	*0.011	*0.023	0.107	0.833	8.714
7	0.246	*0.002	*0.005	0.059	0.937	37.129
8	0.095	*0.003	0.022	0.029	0.942	8.963
9	*0.008	0.087	*0.026	0.036	0.937	4.461
10	*-0.005	0.050	0.016	0.033	0.927	5.970

Table 4: Parameter values for the GARCH(1,1)-t models. Parameters with an asterisk (*) are not significant.

The condition $\alpha_1 + \beta_1 < 1$ is also required for the GARCH(1,1)-t models[17] which are reported in Table 4. Again, all models are stationary. At this point we remark that the training procedure was modified for the GARCH(1,1)-t models on the sets 2 and 3. On these sets we observed serious overfitting in the conditional variances: After a few steps of the scaled gradient descent algorithm the parameters of the models were already close to optimum in parameter space (in the likelihood sense). These nearly optimal solutions were stationary, and they are listed in Table 4. Since the loss function was still slightly decreasing, the training of the models was continued in the first experiments. In the following steps the values of the parameters changed only slightly but suddenly α_0 and α_1 started to grow rapidly as a consequence of (over)fitting some large returns at the beginning of the corresponding training sets. Finally, the loss function remained constant at some minimum value. The obtained solutions were far from being stationary with degrees-of-freedom ν close to 2. The other error measures were much worse for these solutions than for the other models fitted to the sets 2 and 3. Therefore we decided to stop the training at an intermediate stage where the loss function was nearly minimal but the parameters were such that the model was still stationary.

It is emphasized that these problems did not arise for the RMDN(1)-t models where the conditional variance is modeled by a non-linear MLP. In our models an MLP has a bounded activation function in the hidden units ($h(x) = \tanh(x)$) and an unbounded activation function in the output units ($g(x) = x$, see Eq. (12)). Although an MLP is thus able to produce outputs of arbitrary size, the RMDN(1)-t models and also the

[16] For stationarity we demand, of course, also $|a| < 1$ (see Eq. (9)).

[17] The existence of the second moment requires $\nu > 2$ (see definition (10)).

other non-linear models tend to predict, on average, smaller conditional variances than the linear models. In other words, the boundedness of h seems to reduce the sensitivity of the non-linear models to large returns which may result in more stable training sessions as for the RMDN(1)-t models above. In general, overfitting of the neural network models was observed on some data sets but only to a negligible extent. This robustness of MDNs against large returns, which may be interpreted as outliers, is an appealing property (Ormoneit, 1998), also in the recurrent framework. The larger number of free parameters (with respect to GARCH type models) of density estimating neural networks represents thus no problem in fitting these models to an actual data set.

The parameters of the GARCH(1,1)-t models reported in Table 4 are quite similar to the corresponding parameters of the GARCH(1,1) models. The new degrees-of-freedom parameter ν, however, expresses the fact that the *conditional* distribution of returns is in general leptokurtic and not gaussian. The kurtosis of the conditional distribution is given by $3(\nu-2)/(\nu-4)$, $\nu > 4$. Therefore the kurtosis is far from 3 for most of the models, and it does not even exist for the models on the sets 2 and 3. Only the conditional distribution of the model on set 7 is approximately gaussian (large ν).

To illustrate the differences between the models, the conditional pdfs of the next return estimated by the models on two specific days are plotted in Fig. 5. Both days are from the period covered by the tenth training set: Day 1 is November 15, 1991 after a rather large 4% decrease of the DJIA, and day 2 is October 17, 1993 during a calm period with returns around 0. The conditional pdfs of day 1 are depicted on the left-hand side with the pdfs of the linear models in the upper figure and the pdfs of the non-linear models in the lower figure. The same graphic representation is used for day 2 on the right-hand side of Fig. 5.

Due to the large return on day 1, the conditional variances, i.e. the estimated (squared) volatilities of the next day, are much larger for the pdfs depicted on the left-hand side than for the ones on the right-hand side. Additionally, the conditional variances of the non-linear models are slightly smaller than the ones of the linear models which is barely seen from Fig. 5. The conditional pdfs of the non-gaussian models differ remarkably from the conditional pdfs of the gaussian models in that they are leptokurtic. In particular, the two gaussians of the LRMDN(2) and the RMDN(2) models are weighted and combined such that the resulting conditional pdfs have fat tails. The conditional distributions of these two models are negatively skewed on day 1 in contrast to the other models, which have, by their definition, a skewness of 0.

4.3.1 In-sample diagnostics

Table 5 and 6 summarize the performance of all models on the 10 training sets. We report the five error measures 'Loss', NMSE, NMAE, HR, and WHR as well as some statistics for the standardized residuals $\hat{e}_t/\hat{\sigma}_t$. For each set the six lines give the values for the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively. We remark that some models include other models: The GARCH(1,1) model, for instance, is a GARCH(1,1)-t model with an infinite number of degrees of freedom ($\nu \to \infty$). The LRMDN(2) model is a non-gaussian extension of the GARCH(1,1) model, and the RMDN(1) model is trivially included in the RMDN(2) model. Finally,

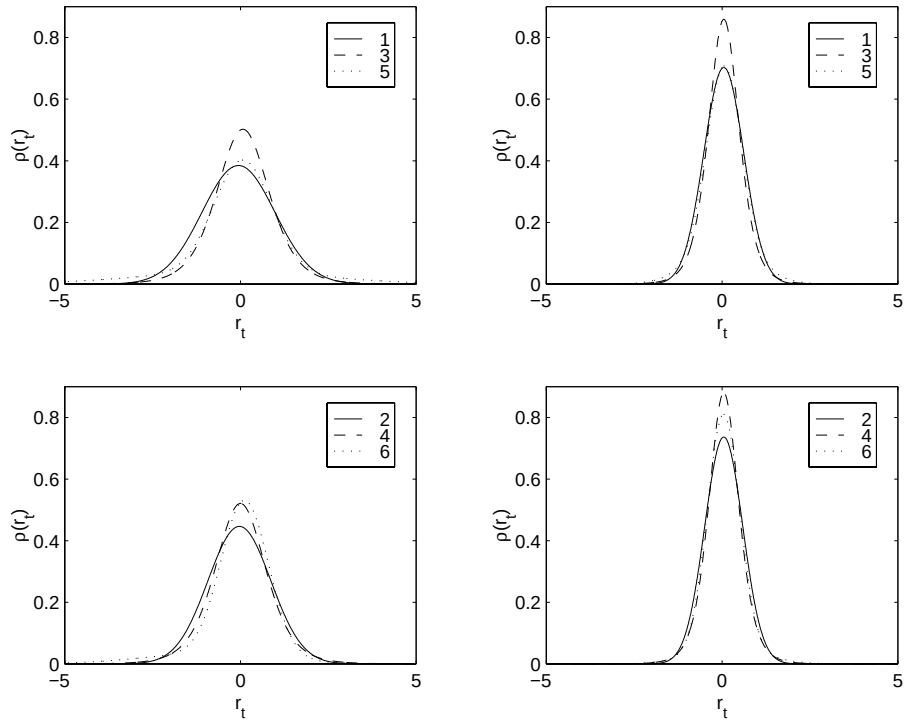

Figure 5: The conditional pdfs predicted by the models on November 15, 1991 and on October 17, 1993 are depicted on the left- and on the right-hand side, respectively. The conditional pdfs of the linear models 1, 3, and 5 are plotted in the upper two figures (solid line: GARCH(1,1), dashed line: GARCH(1,1)-t, dotted line: LRMDN(2)), and the conditional pdfs of the non-linear models 2, 4, and 6 are plotted in the lower two figures (solid line: RMDN(1), dashed line: RMDN(1)-t, dotted line: RMDN(2)).

the RMDN(1)-t model includes the RMDN(1) model (again for $\nu \to \infty$). Therefore the more general models naturally achieve smaller (or at least the same) values for the loss function than the included models on all ten training sets.

Obviously, the non-gaussian models achieve smaller values of the loss function than the gaussian models GARCH(1,1) and RMDN(1) on most sets. Among the non-gaussian models, a non-linear RMDN(1)-t or RMDN(2) model has the best performance on 6 of the 10 sets. Thus, besides non-gaussian conditional distributions, non-linearity may be an issue. The relevance of non-linearity is more emphasized by the other four error measures, especially by the NMAE for which the non-linear models are better than their linear counterparts on each set. On the third set, the GARCH(1,1)-t model (and also the RMDN(1)-t model) is worse than the naive prediction since the NMAE is larger than 1. In summary, a first inspection of the in-sample results reported in Table 5 and 6 indicates that non-gaussian models tend to achieve lower training errors (values of the loss function) than gaussian models and that non-linear components might be useful in modeling and predicting volatility.

Besides reporting the standardized residuals $\hat{e}_t/\hat{\sigma}_t$, it is convenient to run the regression

Set	Loss	NMSE	NMAE	HR	WHR	Standardized Residuals			
						Mean	Std.	Skew.	Kurt.
1	1.493	0.696	0.753	0.692	0.788	-0.007	0.997	-0.434	4.657
	1.487	0.697	0.750	0.701	0.780	-0.006	1.000	-0.450	4.582
	1.462	0.697	0.754	0.691	0.784	-0.036	0.997	-0.436	4.634
	1.460	0.699	0.726	0.697	0.791	-0.038	1.021	-0.448	4.544
	1.469	0.701	0.772	0.677	0.809	-0.011	0.970	-0.463	5.021
	1.486	0.723	0.712	0.689	0.707	-0.007	1.060	-0.493	5.452
2	1.082	0.765	0.796	0.682	0.769	-0.005	1.038	-1.006	12.316
	1.110	0.749	0.801	0.680	0.751	-0.015	1.032	-1.214	14.065
	0.986	0.763	0.916	0.662	0.747	-0.038	0.913	-1.315	16.032
	0.980	0.747	0.861	0.672	0.764	-0.038	0.942	-1.241	14.645
	1.001	0.761	0.817	0.677	0.754	0.013	0.999	-0.936	10.421
	0.980	0.743	0.773	0.689	0.778	-0.004	1.019	-0.705	8.558
3	1.206	0.905	0.937	0.661	0.686	-0.010	1.000	-0.026	11.457
	1.213	0.833	0.859	0.653	0.739	-0.023	1.018	-0.035	11.014
	1.077	0.927	1.289	0.583	0.615	-0.021	0.735	-0.112	11.358
	1.072	0.858	1.021	0.620	0.685	-0.028	0.862	0.259	14.185
	1.083	0.885	0.897	0.652	0.670	-0.004	0.987	-0.196	11.400
	1.084	0.868	0.835	0.654	0.688	-0.000	1.033	-0.294	11.424
4	0.847	0.856	0.819	0.672	0.767	0.015	1.050	-2.177	27.853
	0.896	0.753	0.815	0.684	0.764	0.020	1.022	-1.717	18.544
	0.797	0.814	0.796	0.677	0.780	-0.032	1.071	-2.491	33.387
	0.804	0.770	0.763	0.685	0.806	-0.045	1.067	-2.154	26.287
	0.811	0.805	0.808	0.677	0.780	0.001	1.006	-2.073	24.848
	0.806	0.755	0.776	0.687	0.808	0.004	1.014	-2.116	26.010
5	0.986	0.988	0.868	0.692	0.646	-0.004	1.000	-0.260	4.225
	0.992	0.920	0.836	0.697	0.659	-0.035	1.003	-0.184	4.360
	0.974	0.964	0.861	0.699	0.658	-0.026	0.992	-0.267	4.223
	0.986	0.930	0.839	0.679	0.639	-0.020	1.006	-0.095	5.599
	0.983	0.975	0.878	0.686	0.647	-0.020	0.998	-0.125	5.061
	0.980	0.903	0.811	0.696	0.651	-0.024	1.018	-0.182	4.535

Table 5: In-sample statistics (for set 1–5): The first to sixth line gives the performance of the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively.

Set	Loss	NMSE	NMAE	HR	WHR	Standardized Residuals			
						Mean	Std.	Skew.	Kurt.
6	0.861	0.729	0.797	0.681	0.720	-0.018	1.000	-0.035	3.550
	0.859	0.717	0.783	0.682	0.724	-0.009	1.003	-0.036	3.524
	0.855	0.728	0.803	0.683	0.725	-0.019	0.990	-0.039	3.540
	0.853	0.716	0.795	0.679	0.722	-0.026	0.989	-0.031	3.596
	0.853	0.729	0.800	0.683	0.718	-0.006	0.997	-0.038	3.559
	0.850	0.711	0.776	0.683	0.730	-0.017	1.003	-0.029	3.513
7	1.359	0.701	0.739	0.728	0.747	-0.001	0.996	0.137	3.013
	1.361	0.704	0.728	0.730	0.743	-0.003	1.004	0.156	3.024
	1.359	0.701	0.740	0.727	0.747	0.001	0.994	0.138	3.016
	1.360	0.704	0.730	0.729	0.739	0.000	1.001	0.155	3.025
	1.357	0.702	0.741	0.722	0.743	0.002	0.996	0.140	3.014
	1.361	0.710	0.725	0.723	0.730	-0.001	1.014	0.166	3.082
8	1.268	0.734	0.742	0.704	0.736	-0.001	1.003	0.071	3.745
	1.266	0.725	0.734	0.708	0.743	-0.001	1.006	0.063	3.726
	1.261	0.734	0.746	0.704	0.735	-0.000	0.994	0.074	3.778
	1.260	0.726	0.740	0.705	0.742	-0.003	0.996	0.065	3.780
	1.267	0.736	0.746	0.702	0.739	0.000	1.000	0.069	3.910
	1.262	0.727	0.745	0.700	0.733	-0.001	0.989	0.079	3.821
9	1.447	0.742	0.744	0.682	0.450	-0.014	1.014	-1.083	13.181
	1.446	0.737	0.676	0.679	0.447	-0.008	1.081	-1.765	23.888
	1.365	0.737	0.705	0.686	0.430	-0.019	1.081	-2.925	43.109
	1.373	0.738	0.663	0.686	0.449	-0.018	1.094	-2.269	31.390
	1.377	0.751	0.804	0.674	0.396	0.013	0.999	-0.989	12.403
	1.407	0.738	0.666	0.691	0.456	-0.011	1.143	-2.506	37.157
10	0.994	0.743	0.779	0.702	0.708	-0.000	1.001	-0.358	6.406
	0.993	0.742	0.777	0.705	0.734	-0.000	1.003	-0.363	6.182
	0.953	0.741	0.768	0.707	0.705	-0.014	1.018	-0.382	6.355
	0.952	0.740	0.767	0.705	0.703	-0.012	1.018	-0.366	6.241
	0.957	0.743	0.784	0.697	0.698	-0.013	0.999	-0.177	6.076
	0.950	0.741	0.773	0.705	0.713	-0.002	1.009	-0.339	6.593

Table 6: In-sample statistics (for set 6–10): The first to sixth line gives the performance of the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively.

Set	Par.	Model					
		1	2	3	4	5	6
1	α	0.071	-0.036	0.084	-0.148	0.076	[-1.526]
	β	0.965	1.035	0.950	1.195	0.936	[2.552]
2	α	0.077	-0.703	0.121	-0.093	0.070	-0.361
	β	0.930	2.418	0.640	1.043	0.906	1.816
3	α	0.233	-0.700	0.005	[-1.027]	-0.010	[-2.141]
	β	0.701	2.111	[0.539]	1.849	1.063	[4.483]
4	α	-0.075	[-7.959]	0.060	-0.051	-0.178	-0.280
	β	1.308	[23.449]	0.904	1.284	1.518	1.853
5	α	0.010	-0.343	-0.015	-0.366	-0.016	-0.523
	β	1.003	1.798	1.051	1.848	1.043	2.312
6	α	0.033	-0.065	0.032	0.014	0.043	-0.067
	β	0.902	1.208	0.884	0.932	0.862	1.219
7	α	0.088	-0.175	0.085	-0.154	0.111	[-0.260]
	β	0.901	1.206	0.901	1.175	0.874	[1.323]
8	α	0.085	-0.108	0.017	-0.086	-0.223	-0.331
	β	0.891	1.159	0.968	1.109	1.296	1.415
9	α	0.927	-4.655	0.602	-3.528	[1.074]	-5.518
	β	[0.539]	5.721	0.865	4.911	[0.399]	7.079
10	α	-0.014	-0.045	0.016	-0.050	-0.146	-0.053
	β	1.034	1.111	0.996	1.155	1.340	1.140

Table 7: Estimated parameters α and β from Eq. (39) for the 10 training sets for models 1–6: GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2).

proposed by Pagan and Schwert (1990) in order to detect model biases. In particular, the squared forecast error $\hat{e}_t^2 = (r_t - \hat{r}_t)^2$ is regressed on the forecasted conditional variance $\hat{\sigma}_t^2$:

$$\hat{e}_t^2 = \alpha + \beta \hat{\sigma}_t^2 + \epsilon_t \tag{39}$$

where ϵ_t denotes the residual. The forecasts are supposed to be unbiased, if the obtained parameters are $\alpha = 0$ and $\beta = 1$. In Table 7 we summarize the parameters α and β obtained for all training sets and for all models. The estimated parameters are put into brackets, if they are not within the 95% confidence intervals around their hypothesized values[18]. The non-linear models, i.e. the models 2, 4, and 6, are found to be biased more often than the linear models. In particular, they have negative αs and βs larger than 1 on many sets. An extreme example is the RMDN(1) model on set 4. An inspection of the network weights reveals that this model is nearly homoskedastic, i.e. the conditional variances do not represent the real dynamics of the squared returns. Furthermore, it appears from Table 7 that the distortions of α and β are particularly large for all models on set 9, which includes the historical crash.

In order to compare the in-sample results obtained for the six models, we tested

[18]Heteroskedasticity-consistent standard errors are used (White, 1980).

the hypothesis of higher/lower errors by performing parametric and nonparametric tests. More precisely, we performed a paired t-test and a matched pairs signed rank Wilcoxon test (paired Wilcoxon test) for the five error measures 'Loss', NMSE, NMAE, HR, and WHR for all possible pairs of models (five times 15 paired t-tests/Wilcoxon tests). In our context, the application of *paired* tests[19] is appropriate for the following reason: The error measures of each model vary considerably with the actual segment of the underlying return series but the differences between the error measures of different models are rather small. Therefore the differences can only be detected if a paired test which takes into account the correlations between the error measures, is applied. To illustrate this point the loss function $\bar{\mathcal{L}}$ is plotted for the six models for each of the 10 training sets in Fig. 6.

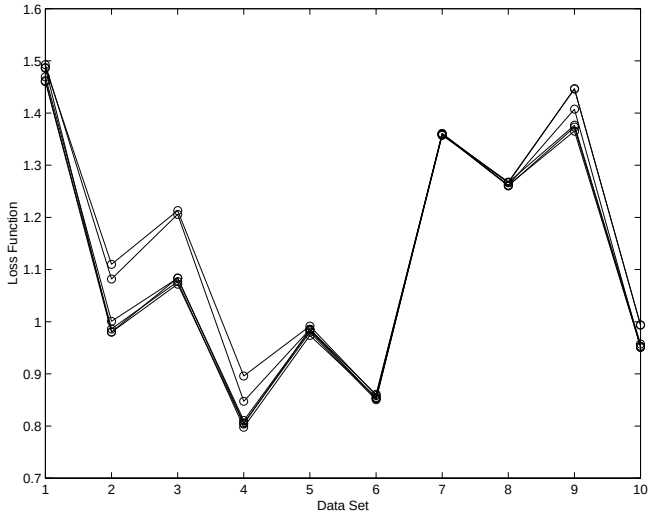

Figure 6: The loss function $\bar{\mathcal{L}}$ of the six models for the 10 training sets.

In Table 8 we report the p-values of the paired t-tests and the R-values[20] of the paired Wilcoxon tests concerning the loss function. The first column gives the model and the second column the mean value of the loss function over the 10 training sets for the particular model. Columns 3 to 8 summarize the p-values and the R-values where the values above the diagonal were obtained from the paired t-tests and the values below the diagonal from the paired Wilcoxon tests. For instance, the GARCH(1,1) model is significantly worse than the LRMDN(2) model for the paired t-test ($p = 0.016$) and the paired Wilcoxon test ($R = 0$).

The results of the parametric and the nonparametric tests are the same in the sense that the models with a non-gaussian conditional distribution, i.e. the models 3–6, are

[19] The common assumption underlying the paired t-test and the paired Wilcoxon test is that the differences of the (paired) error measures are independent. Since the training sets are several years apart (see Table 1), the error measures (and their differences) on different sets can indeed be regarded independent. Additionally, for the paired t-test it is assumed that the differences are normally distributed whereas for the paired Wilcoxon test it is only assumed that the distribution of the differences is symmetric.

[20] For the paired Wilcoxon test, the (integer) R-value has to be compared to a critical (integer) R_α-value to test at the confidence level α. For 10 paired error measures/differences, the critical value for $\alpha = 0.05$ is $R_\alpha = 8$. The null hypothesis is rejected if $R \leq R_\alpha$.

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	1.154	–	0.164	0.010	0.014	0.016	0.022
2: RMDN(1)	1.162	17	–	0.010	0.013	0.014	0.021
3: GARCH(1,1)-t	1.109	1	0	–	0.553	0.003	0.135
4: RMDN(1)-t	1.110	3	0	25	–	0.030	0.157
5: LRMDN(2)	1.116	0	1	3	6	–	0.862
6: RMDN(2)	1.117	1	1	12	18	22	–

Table 8: In-sample statistics (loss function): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

significantly better than the models assuming a gaussian conditional distribution. Among the former, the models with a t-distribution achieve significantly lower errors than the LRMDN(2) model and lower errors than the RMDN(2) model. It is recalled, however, that on three sets the RMDN(2) model achieves the lowest value of the loss function (see Table 5 and 6). Interestingly, the linear models (models 1, 3, and 5) are on average slightly better than their non-linear counterparts.

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.786	–	0.045	0.345	0.039	0.234	0.095
2: RMDN(1)	0.758	5	–	0.056	0.123	0.018	0.402
3: GARCH(1,1)-t	0.780	14	10	–	0.044	0.727	0.093
4: RMDN(1)-t	0.763	5	17	9	–	0.009	0.864
5: LRMDN(2)	0.779	22	2	20	3	–	0.082
6: RMDN(2)	0.762	11	22	13	27	10	–

Table 9: In-sample statistics (NMSE): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

The best model with respect to the NMSE is the RMDN(1) model (see Table 9). It is significantly better than the GARCH(1,1) model and the LRMDN(2) model, and it tends to be better than the GARCH(1,1)-t model. In general, the non-linear models (models 2, 4, and 6) achieve lower errors than their corresponding linear models.

Table 10 summarizes the results of the tests for the NMAE. The best model is the RMDN(2) model which is significantly better than the GARCH(1,1) model, the RMDN(1) model, and the LRMDN(2) model. Furthermore, it tends to be better than the GARCH(1,1)-t model (significance is only obtained for the paired Wilcoxon test). For this error measure, the non-linear models are significantly better than their corresponding linear models (with the exception of the paired t-test for the comparison of the models 3 and 4).

The best models with respect to the hit rate HR are the RMDN(1) and the RMDN(2) models (see Table 11). Note that a higher hit rate corresponds to a better performance

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.797	–	0.045	0.304	0.680	0.370	0.005
2: RMDN(1)	0.776	4	–	0.180	0.454	0.038	0.011
3: GARCH(1,1)-t	0.838	26	7	–	0.094	0.460	0.106
4: RMDN(1)-t	0.791	18	25	0	–	0.522	0.141
5: LRMDN(2)	0.805	15	2	21	15	–	0.005
6: RMDN(2)	0.759	1	5	2	12	0	–

Table 10: In-sample statistics (NMAE): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.689	–	0.190	0.388	0.464	0.034	0.320
2: RMDN(1)	0.692	14	–	0.192	0.128	0.009	0.939
3: GARCH(1,1)-t	0.682	20	12	–	0.428	0.727	0.222
4: RMDN(1)-t	0.686	27	7	15	–	0.847	0.185
5: LRMDN(2)	0.685	8	2	15	20	–	0.006
6: RMDN(2)	0.692	20	22	16	15	2	–

Table 11: In-sample statistics (HR): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

(on average). Both models are, together with the GARCH(1,1) model, significantly better than the LRMDN(2) model. Table 12 summarizes the results for the weighted hit rate WHR which are not significant. The RMDN(1) model is again the best model. For both error measures HR and WHR, the linear models perform slightly worse than their non-linear counterparts.

The in-sample results of the different volatility models reported in Table 5 - 12 can be summarized in the following way: The non-gaussian models achieve significantly smaller values of the loss function than the gaussian models. In other words, conditional distributions which are leptokurtic, provide a more detailed description of return series in the likelihood framework. In this context non-linear specifications can be useful on specific data sets[21] but do not seem to have more explanatory power than linear models

[21]One may argue that the non-linear models benefit from the larger number of parameters. This is, however, only partly true. On the last training set, for instance, where the RMDN(2) model shows the best performance, a simple pruning algorithm was applied to the RMDN(2) model in order to check how many of the 48 parameters were actually needed. The final network (after the pruning) consisted of only 15 weights, and it could be further simplified to a network with just 12 parameters. Although the network size was reduced by 75%, the loss function *decreased* from 0.950 to 0.948. In other words, the smaller network turned out to provide a better fit to the data than the original network. Following the Akaike information criterion (AIC; Akaike, 1973) however, the GARCH(1,1)-t model (AIC=2106) is still preferable to the RMDN(2) model (AIC=2108).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.702	–	0.328	0.272	0.617	0.374	0.833
2: RMDN(1)	0.708	20	–	0.243	0.615	0.218	0.443
3: GARCH(1,1)-t	0.692	18	16	–	0.176	0.691	0.583
4: RMDN(1)-t	0.704	24	22	16	–	0.214	0.645
5: LRMDN(2)	0.696	19	16	27	16	–	0.773
6: RMDN(2)	0.700	20	19	19	21	16	–

Table 12: In-sample statistics (WHR): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

on average. However, non-linearity plays an important role with respect to other error measures which relate the volatility predicted by a model to the squared returns, which are assumed to represent the true volatility of the return series. For these error measures, each non-linear model performs better than its corresponding linear model on average. In particular, significant differences are obtained for the NMAE. Furthermore, distributional assumptions are not crucial to the predictive performance of the models.

At this point it may be also interesting to analyze the performance of the models with respect to the *unconditional* distribution of returns. In particular, it is possible to generate artificial time series using the trained models and to compare some statistics of the original data (training set) to the same statistics of the generated data. In principle, it is possible to generate an arbitrary number of time series of arbitrary length. We performed the following simulations: For each training set the sample mean, standard deviation, skewness, and kurtosis as well as the first five autocorrelation coefficients ρ_i, $1 \leq i \leq 5$, of the *squared* time series were calculated as nine characteristics of the training set. By iteration of each model trained on this set, 100 artificial time series of the same length as the training set ($N = 1100$) were generated, and these nine characteristics were calculated for each of the 100 time series. The characteristics of the original data and the mean values over the 100 generated data sets are reported in Table 13 and 14.

We remark that the parameter values of some models are such that some statistics are meaningless. For a GARCH(1,1) model, for instance, the fourth moment and thus the kurtosis exist only if the condition $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$ is satisfied (Bollerslev, 1986). Plugging in the parameter values of the GARCH(1,1) models we find that this condition is slightly violated on set 7. For the GARCH(1,1)-t models it can be shown[22] that the fourth moment exists if the conditions $\nu > 4$ and $3(\nu-2)/(\nu-4)\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$ hold. The first condition is violated for the models on the second and third set (see Table 4). The second condition is not met for the models on the sets 1 and 7. For the (L)RMDN(n) models it seems to be very hard to prove conditions under which stationarity of the generated sequences is guaranteed. Even in the case of an RMDN(1) model, where the conditional mean and the conditional variance are modeled by simple MLPs, no analytical

[22] The proof in (Bollerslev, 1986) has to be adapted at one point.

results were obtained[23]. However, from the large standard deviations of some statistics it may concluded that the corresponding moments do not exist: the third and fourth moment for the RMDN(1)-t models on the data sets 2 and 3[24]; the fourth moment for the same model on set 1; the third moment for the LRMDN(2) model on set 9; the fourth moment for the same model on the sets 1, 3, 7, and 9. We remark that if the fourth moment does not exist, $E((r_t^2)^2)$ does not exist either, and therefore the expected value of the autocorrelation function of the squared returns does not exist.

In Table 13 and 14 all statistics which are assumed to be meaningless because some moment does not exist, are put into parentheses. For the GARCH(1,1) models, statistics like the unconditional mean, for instance, can be calculated analytically: $E(r_t) = b/(1-a)$. However, we estimated these statistics also by applying the procedure described above in order to arrive at estimations for the original data and for the different models within the same framework[25]. Statistics which are significantly different[26] from the statistics estimated from the original data are put into brackets.

We get the following results: First, the linear models are more sensitive to the issue of non-existing moments than the non-linear models. This property stems from the fact that in the linear models the conditional variance can in principle grow to infinity (if the parameters estimated from the underlying data set are such that e.g. $\alpha_1 + \beta_1 > 1$ holds for a GARCH(1,1) or a GARCH(1,1)-t model). On the other hand, the special type of non-linearity in the hidden units, i.e. the boundedness of the activation function, results in conditional variances on a compact set for the non-linear models. This may be the reason why the statistics of the non-linear models tend to be more stable than the statistics of the linear models. Secondly, it is more difficult for all models to reproduce statistics depending on higher moments such as skewness and kurtosis than statistics such as mean and standard deviation. This seems to be particularly true for the models with a normal conditional distribution, i.e. the GARCH(1,1) and the RMDN(1) models. On one hand, this is not surprising since the skewness of a GARCH(1,1) model is always 0 if the fourth moment exists. On the other hand, there is no restriction for the kurtosis of a GARCH(1,1) model in the sense that it can become arbitrarily large. It might be that this is a consequence of estimating the statistics from rather short generated data sets. Finally, the power of the models to reproduce the autocorrelation function of the squared returns strongly depends on the underlying return series. Sometimes the models generate similar structures (for the sets 6, 8, and 10) and sometimes several statistics are rejected (for the sets 1 and 5).

[23] For an analytical treatment of the homoskedastic case, see however (Leisch et al., 1999).

[24] This is supported by the rather small values for ν: 2.200 and 2.232, respectively.

[25] However, we remark that moments estimated from the generated data sets can be far from their true values if the persistence $\alpha_1 + \beta_1$ is close to 1. For instance, the kurtosis of the random variable represented by the GARCH(1,1) model on the first set ($\alpha_1 = 0.065$, $\beta_1 = 0.925$) is (analytically) given by 5.190. In fact our simulations (for data sets of length $N = 1100$) yield 3.898 (0.893) (see Table 13) where the standard deviation is put into parentheses. For larger data sets ($N = 10^4$), the obtained estimates are better: 4.548 (1.451).

[26] at the 5% level (for a t-test)

Set	Mean	Std.	Skew.	Kurt.	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
1	0.033	1.265	-0.497	7.897	0.100	0.219	0.224	0.162	0.201
	0.052	1.239	[0.009]	[3.898]	0.135	0.137	0.133	0.139	0.134
	0.053	1.111	[-0.010]	[3.410]	0.095	[0.081]	[0.079]	0.075	[0.063]
	0.073	1.169	-0.039	(7.600)	(0.104)	(0.116)	(0.109)	(0.100)	(0.107)
	0.078	1.395	0.001	(6.584)	(0.070)	(0.062)	(0.066)	(0.058)	(0.049)
	0.038	1.321	-0.343	(9.429)	(0.157)	(0.166)	(0.166)	(0.148)	(0.138)
	0.040	1.134	-0.231	[5.029]	0.078	[0.064]	[0.057]	[0.052]	[0.041]
2	0.010	0.768	-0.729	13.442	0.121	0.084	0.031	0.038	0.069
	0.015	0.717	[0.004]	[3.270]	0.125	0.090	0.083	0.076	0.073
	0.015	[0.736]	[-0.007]	[3.084]	0.117	0.027	0.003	-0.001	[0.003]
	0.048	1.000	(-0.233)	(32.770)	(0.086)	(0.082)	(0.067)	(0.071)	(0.073)
	0.040	0.803	(-0.015)	(35.172)	(0.049)	(0.031)	(0.035)	(0.031)	(0.019)
	-0.004	0.766	-0.374	[7.560]	0.190	0.040	0.028	0.007	[0.002]
	0.013	0.738	[-0.254]	[6.152]	0.104	0.065	0.073	0.047	0.053
3	0.000	0.902	-0.647	14.647	0.317	0.065	0.013	0.111	0.026
	0.012	0.955	[0.002]	[4.479]	0.303	0.241	[0.190]	0.163	0.128
	0.020	0.948	[-0.002]	[3.587]	[0.204]	[0.157]	[0.127]	0.103	0.082
	0.028	1.159	(-0.636)	(63.375)	(0.150)	(0.095)	(0.061)	(0.056)	(0.036)
	0.035	0.960	(-0.045)	(45.216)	(0.031)	(0.022)	(0.024)	(0.019)	(0.019)
	0.004	0.888	-0.507	(13.945)	(0.165)	(0.107)	(0.076)	(0.043)	(0.035)
	0.002	[0.819]	-0.558	[7.873]	[0.077]	0.049	0.029	[0.030]	0.020
4	0.053	0.610	-1.690	18.404	0.113	0.039	0.013	0.009	0.117
	0.030	0.601	[-0.014]	[3.270]	0.204	0.101	0.062	0.035	[0.026]
	0.044	0.595	[-0.003]	[2.978]	[0.007]	-0.003	-0.001	-0.000	[-0.001]
	0.069	0.578	[-0.033]	[6.888]	0.141	0.099	0.080	0.060	0.049
	0.074	0.582	[-0.023]	[6.470]	0.147	0.054	0.018	0.010	[0.001]
	0.051	0.611	[-0.703]	[7.192]	0.147	0.057	0.023	0.007	[0.006]
	0.049	0.604	[-0.685]	[6.841]	0.111	0.046	0.025	0.012	[0.008]
5	0.023	0.712	-0.419	9.614	0.387	0.104	0.122	0.216	0.100
	0.023	0.704	[0.002]	[3.756]	[0.224]	0.166	0.145	0.124	0.120
	0.053	0.727	[0.008]	[3.669]	[0.255]	0.180	0.145	[0.119]	0.092
	0.042	0.707	-0.001	6.089	[0.204]	0.156	0.125	0.109	0.100
	0.037	0.698	0.000	[5.616]	[0.206]	0.073	[0.026]	[0.011]	[0.003]
	0.037	0.717	-0.235	[5.172]	0.306	0.170	0.108	[0.072]	0.045
	0.041	0.680	-0.256	[4.258]	[0.172]	0.120	0.094	[0.076]	0.062

Table 13: Additional statistics for set 1–5: for each set the first line gives the values estimated directly from the training set whereas the second to seventh line give the mean values of the statistics calculated from 100 sets of the same length generated by the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively.

Set	Mean	Std.	Skew.	Kurt.	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
6	0.003	0.609	-0.020	3.930	0.111	0.136	0.090	0.114	0.144
	0.013	0.617	0.023	3.627	0.211	0.157	0.134	0.118	0.117
	0.005	0.613	-0.007	[3.304]	0.142	0.091	0.087	0.074	0.068
	0.019	0.647	0.040	5.905	0.202	0.144	0.146	0.119	0.114
	0.022	0.619	0.036	4.883	0.162	0.108	0.100	0.087	0.074
	0.003	0.629	-0.135	4.578	0.215	0.159	0.143	0.129	0.113
	0.016	0.602	-0.081	3.760	0.147	0.099	0.092	0.082	0.078
7	0.001	1.072	0.239	4.012	0.153	0.204	0.180	0.224	0.226
	0.004	1.071	0.005	(4.157)	(0.200)	(0.146)	(0.144)	(0.146)	(0.143)
	0.010	1.023	[-0.023]	3.702	0.161	0.121	0.122	[0.112]	[0.114]
	0.001	1.066	0.019	(4.394)	(0.192)	(0.150)	(0.136)	(0.145)	(0.136)
	0.006	1.045	-0.016	3.969	0.161	0.125	0.114	[0.111]	0.115
	0.001	1.111	0.082	(4.675)	(0.201)	(0.161)	(0.157)	(0.144)	(0.150)
	-0.002	1.038	0.140	3.728	0.143	[0.104]	0.103	[0.092]	[0.094]
8	0.002	0.872	0.085	4.033	0.058	0.011	-0.003	0.071	0.055
	0.002	0.859	-0.004	[3.083]	0.051	0.045	0.045	0.044	0.039
	0.006	0.871	0.010	[3.088]	0.044	0.039	0.041	0.040	0.036
	0.007	0.879	0.029	4.311	0.043	0.038	0.032	0.032	0.028
	0.005	0.852	-0.018	4.314	0.040	0.033	0.041	0.024	0.033
	0.002	0.869	-0.003	3.872	0.052	0.022	0.008	[0.000]	[-0.006]
	0.006	0.882	-0.045	4.089	0.022	0.027	0.031	0.029	0.031
9	0.058	1.371	-6.054	118.649	0.086	0.145	0.063	0.011	0.103
	0.087	1.211	[-0.002]	[4.272]	0.219	0.208	0.196	[0.186]	0.172
	0.085	1.264	[0.007]	[3.518]	0.124	0.117	0.105	0.101	0.092
	0.095	[0.948]	[0.047]	[8.243]	0.040	[0.040]	0.039	0.047	0.043
	0.086	[1.006]	[0.072]	[8.424]	0.048	0.060	0.065	0.055	0.048
	0.063	1.476	(0.131)	(20.733)	(0.128)	(0.164)	(0.137)	(0.127)	(0.121)
	0.086	[1.032]	[-0.138]	[6.843]	0.066	[0.057]	0.055	0.053	0.056
10	0.041	0.663	-0.215	6.300	0.097	0.057	0.002	0.062	0.062
	0.037	0.659	[-0.007]	[3.059]	0.056	0.052	0.048	0.036	0.040
	0.040	0.664	[-0.003]	[3.076]	0.036	0.036	0.036	0.034	0.033
	0.051	0.639	-0.005	6.084	0.039	0.040	0.031	0.034	0.037
	0.052	0.634	0.056	5.827	0.030	0.032	0.036	0.034	0.031
	0.046	0.659	-0.079	5.247	0.049	0.031	0.015	0.003	[-0.003]
	0.041	0.656	-0.176	5.377	0.045	0.050	0.046	0.028	0.033

Table 14: Additional statistics for set 6–10: for each set the first line gives the values estimated directly from the training set whereas the second to seventh line give the mean values of the statistics calculated from 100 sets of the same length generated by the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively.

4.3.2 Out-of-sample performance

Thus far, the various volatility models have been only compared on data sets from which the model parameters had been estimated. These in-sample results, which have been reported in detail above, provide some information on the impact of non-gaussian conditional distributions and non-linearity on the modeling power of different specifications. However, the true test for a volatility model is to predict volatilities out-of-sample, i.e. for a set of returns disjoint of the training set. More precisely, the test data are typically chosen as returns in the future (relative to the training set) such that it is guaranteed that the returns in the training set do not contain information about the returns in the test set. In the rest of the paper we report on the performance of the various volatility models with respect to one-step-ahead predictions. The ten test sets were chosen as a period of roughly two years and three months subsequent to the corresponding training set. Although an overfitting of training data was observed only to negligible extent[27] as mentioned earlier, only the out-of-sample performance[28] provides the basis for a comparison of the various models where issues such as possible overparametrizations may be neglected.

Table 15 and 16 summarize the performance of the models with respect to the error measures described earlier as well as the usual statistics of the standardized residuals. The GARCH(1,1)-t models achieve the lowest value of the loss function on six of the ten test sets whereas the RMDN(1)-t models and the RMDN(2) models perform best on three sets. Therefore the non-gaussian models seem to dominate the gaussian models also out-of-sample (with respect to the loss function). For the other error measures, the GARCH(1,1)-t models and the RMDN(2) models are among the best models on many sets. We remark that on the test sets 2, 3, and 5 the naive predictor is not beaten by some models (with respect to the NMAE). For completeness, the parameters of the regression proposed in (Pagan and Schwert, 1990) are reported in Table 17.

As in the in-sample case, paired t-tests and paired Wilcoxon tests were applied to test wether the differences in performance between the models were significant or not. Table 18 summarizes the results for the loss function: The mean values of the loss function (over the ten test sets) are given in the second column and the p-values and the R-values above and below the diagonal, respectively. The models with non-gaussian conditional distributions are better than the gaussian models as in the in-sample analysis. The best model is the GARCH(1,1)-t model which is significantly better than all the other models except the RMDN(2) model for which it tends to be better.

The RMDN(2) model has, on average, the best performance with respect to the NMSE and NMAE measure. The differences, which are reported in Table 19 and 20, are statistically not significant for the former measure but they are for the latter. In fact, the RMDN(2) models achieve significantly lower errors than the RMDN(1)-t models and the LRMDN(2) models. With respect to the RMDN(1) models and the GARCH(1,1)-t models, the R-values indicate significance. The RMDN(2) model also tends to be better than the GARCH(1,1) model. The dominance of the RMDN(2) model concerning the NMAE is thus confirmed on the test sets.

The results for the HR and the WHR are summarized in Table 21 and 22. Most

[27] However, the training procedure of the GARCH(1,1)-t models on the sets 2 and 3 was modified.

[28] besides the application of information criteria

Set	Loss	NMSE	NMAE	HR	WHR	Standardized Residuals			
						Mean	Std.	Skew.	Kurt.
1	1.543	0.814	0.874	0.641	0.591	-0.085	1.049	-0.982	9.418
	1.587	0.825	0.939	0.639	0.589	-0.086	1.061	-0.879	9.281
	1.443	0.815	0.876	0.648	0.591	-0.115	1.048	-0.993	9.516
	1.495	0.828	0.963	0.639	0.571	-0.115	1.038	-0.872	9.261
	1.431	0.808	0.886	0.641	0.582	-0.073	0.998	-0.640	7.753
	1.428	0.808	0.812	0.648	0.600	-0.075	1.056	-0.795	8.738
2	0.731	0.737	0.926	0.616	0.693	0.049	0.827	-0.470	4.576
	0.822	0.843	1.238	0.584	0.663	0.050	0.714	-0.334	4.671
	0.687	0.809	1.155	0.590	0.678	0.007	0.721	-0.482	4.606
	0.685	0.784	1.058	0.604	0.690	0.005	0.767	-0.485	4.467
	0.716	0.819	1.191	0.584	0.662	0.071	0.710	-0.310	4.680
	0.686	0.746	0.958	0.625	0.694	0.045	0.809	-0.390	4.480
3	1.052	0.742	0.903	0.600	0.695	0.033	0.893	-1.210	8.742
	1.054	0.728	0.886	0.615	0.716	0.031	0.892	-1.218	8.151
	1.000	0.835	1.461	0.536	0.559	0.013	0.650	-1.199	8.666
	0.988	0.745	1.060	0.586	0.662	0.012	0.762	-1.207	8.550
	0.990	0.736	0.914	0.600	0.686	0.042	0.870	-1.213	8.725
	0.979	0.725	0.863	0.606	0.617	0.052	0.883	-1.206	8.474
4	1.072	0.785	0.748	0.710	0.732	-0.063	1.152	0.104	4.117
	1.118	0.736	0.725	0.702	0.649	-0.085	1.194	0.168	5.264
	1.040	0.750	0.743	0.718	0.777	-0.110	1.107	0.026	3.571
	1.071	0.751	0.722	0.719	0.657	-0.133	1.188	0.083	4.598
	1.076	0.757	0.738	0.700	0.650	-0.084	1.138	0.100	4.574
	1.068	0.740	0.724	0.709	0.743	-0.090	1.129	0.082	4.264
5	0.736	0.978	0.968	0.648	0.396	0.038	0.853	0.043	4.536
	0.726	0.917	0.913	0.659	0.434	0.002	0.846	0.024	3.968
	0.716	0.960	0.969	0.655	0.406	0.014	0.838	0.013	4.449
	0.739	0.920	1.005	0.616	0.357	0.032	0.799	0.148	4.486
	0.730	0.964	1.007	0.632	0.359	0.024	0.804	-0.025	3.734
	0.719	0.904	0.907	0.663	0.441	0.021	0.850	0.064	4.275

Table 15: Out-of-sample statistics (for set 1–5): The first to sixth line gives the performance of the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively.

Set	Loss	NMSE	NMAE	HR	WHR	Standardized Residuals			
						Mean	Std.	Skew.	Kurt.
6	1.130	0.736	0.676	0.744	0.709	-0.017	1.116	0.188	4.074
	1.153	0.773	0.669	0.734	0.701	-0.009	1.156	0.315	4.609
	1.110	0.735	0.685	0.735	0.716	-0.018	1.095	0.187	4.069
	1.123	0.765	0.677	0.734	0.699	-0.027	1.133	0.234	4.338
	1.116	0.737	0.683	0.742	0.705	-0.007	1.106	0.177	4.055
	1.151	0.774	0.670	0.737	0.701	-0.020	1.167	0.298	4.558
7	1.073	0.697	0.754	0.739	0.766	-0.041	0.966	-0.110	2.895
	1.077	0.699	0.751	0.735	0.753	-0.051	0.978	-0.121	2.910
	1.074	0.697	0.755	0.741	0.766	-0.039	0.964	-0.109	2.892
	1.077	0.697	0.748	0.741	0.752	-0.046	0.979	-0.118	2.891
	1.079	0.701	0.765	0.732	0.739	-0.038	0.961	-0.111	2.897
	1.085	0.706	0.766	0.735	0.745	-0.049	0.979	-0.137	2.918
8	1.407	0.701	0.704	0.725	0.802	0.058	1.055	0.557	3.828
	1.434	0.717	0.704	0.716	0.760	0.074	1.106	0.585	3.898
	1.397	0.701	0.705	0.723	0.797	0.060	1.050	0.559	3.830
	1.412	0.715	0.703	0.714	0.785	0.070	1.093	0.569	3.906
	1.456	0.728	0.703	0.718	0.743	0.074	1.173	0.676	4.588
	1.427	0.720	0.702	0.719	0.764	0.075	1.118	0.640	4.167
9	1.423	0.780	0.831	0.696	0.760	-0.041	1.021	-1.237	15.750
	1.420	0.766	0.813	0.691	0.741	-0.040	1.052	-1.870	24.038
	1.345	0.764	0.773	0.709	0.785	-0.044	1.044	-1.175	14.532
	1.346	0.765	0.797	0.700	0.765	-0.046	1.038	-1.375	16.986
	1.390	0.785	0.863	0.686	0.726	-0.015	1.025	-1.412	18.026
	1.402	0.764	0.786	0.702	0.760	-0.039	1.059	-1.648	20.332
10	1.368	0.831	0.698	0.732	0.667	0.056	1.212	-0.864	6.736
	1.372	0.831	0.693	0.734	0.672	0.058	1.204	-0.890	7.524
	1.314	0.819	0.700	0.730	0.671	0.046	1.196	-0.944	7.341
	1.324	0.822	0.685	0.734	0.676	0.051	1.233	-0.877	7.349
	1.409	0.821	0.668	0.721	0.631	0.049	1.367	-0.774	7.069
	1.340	0.833	0.679	0.734	0.670	0.065	1.270	-0.656	6.043

Table 16: Out-of-sample statistics (for set 6–10): The first to sixth line gives the performance of the GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2) model, respectively.

Set	Par.	Model					
		1	2	3	4	5	6
1	α	0.439	[0.789]	0.453	[0.870]	0.399	-0.859
	β	0.746	[0.453]	0.735	[0.389]	0.737	1.927
2	α	-0.057	-0.342	-0.042	-0.017	0.005	-0.066
	β	0.856	1.207	[0.616]	[0.635]	[0.499]	0.845
3	α	0.178	0.054	0.056	-0.138	0.100	-0.446
	β	0.495	0.703	[0.376]	0.753	0.605	1.526
4	α	0.138	-3.871	0.099	0.111	0.096	-0.254
	β	0.995	12.290	1.026	1.112	1.075	2.010
5	α	-0.010	-0.553	-0.038	-1.177	-0.060	-0.838
	β	0.868	2.399	0.918	3.683	0.931	3.261
6	α	0.175	-0.675	0.160	-0.338	0.182	-0.623
	β	0.872	[2.963]	0.869	2.057	0.835	[2.891]
7	α	[0.284]	[0.304]	[0.283]	[0.294]	[0.326]	[0.356]
	β	[0.384]	[0.358]	[0.386]	[0.379]	[0.303]	[0.258]
8	α	-0.271	[-0.978]	-0.383	[-1.165]	0.022	[-1.408]
	β	1.447	[2.441]	1.565	[2.621]	1.348	[3.010]
9	α	0.596	0.538	0.481	0.604	[0.698]	0.475
	β	[0.355]	[0.415]	0.529	[0.375]	[0.247]	0.510
10	α	0.036	-1.052	0.133	-1.153	-0.316	-1.648
	β	1.443	3.286	1.228	3.628	[2.561]	4.856

Table 17: Estimated parameters α and β from Eq. (39) for the ten test sets for models 1–6: GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2), and RMDN(2).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	1.153	–	0.045	0.003	0.021	0.360	0.101
2: RMDN(1)	1.176	7	–	0.002	0.006	0.081	0.029
3: GARCH(1,1)-t	1.113	1	0	–	0.047	0.033	0.074
4: RMDN(1)-t	1.126	9	2	7	–	0.327	0.812
5: LRMDN(2)	1.139	18	11	7	18	–	0.258
6: RMDN(2)	1.128	13	4	10	23	17	–

Table 18: Out-sample statistics (loss function): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.780	–	0.811	0.530	0.944	0.582	0.447
2: RMDN(1)	0.784	26	–	0.725	0.537	0.812	0.261
3: GARCH(1,1)-t	0.789	25	24	–	0.422	0.797	0.280
4: RMDN(1)-t	0.779	27	26	26	–	0.413	0.185
5: LRMDN(2)	0.786	24	25	16	20	–	0.218
6: RMDN(2)	0.772	23	20	21	14	16	–

Table 19: Out-sample statistics (NMSE): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

p- and R-values for these measures are such that the differences between the models are not significant. On average, the RMDN(2) model performs best with respect to the HR whereas the GARCH(1,1) model achieves the highest value of the WHR.

Summing up, it may be said that the out-of-sample performance of the models is similar to the in-sample performance: In the context of likelihood, the non-gaussian conditional distributions model the return series better than the gaussian conditional distributions as in the in-sample studies. Among the former, the GARCH(1,1)-t model is still the best model. This holds, of course, only on average, or in other words, the two non-linear, non-gaussian models may also achieve the lowest errors over specific periods of time. Concerning the NMAE measure, the RMDN(2) models are a class of its own in-sample as well as out-of-sample since they are or tend to be significantly better than the other models. For the remaining error measures, however, the situation is not that clear out-of-sample. For instance, the RMDN(2) models as well as the GARCH(1,1) models are among the best performing models on many sets. Non-linear specifications can thus not be considered superior to linear ones as in-sample. Distributional aspects do not seem to be really important either. Consequently, the choice of a particular model for predicting volatility out-of-sample is closely related to the question of how to measure the prediction performance of a model.

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.808	–	0.471	0.239	0.155	0.236	0.057
2: RMDN(1)	0.833	20	–	0.433	0.770	0.556	0.141
3: GARCH(1,1)-t	0.882	14	23	–	0.366	0.496	0.140
4: RMDN(1)-t	0.842	20	27	21	–	0.997	0.047
5: LRMDN(2)	0.842	11	22	22	24	–	0.041
6: RMDN(2)	0.787	9	7	5	8	4	–

Table 20: Out-sample statistics (NMAE): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.685	–	0.299	0.376	0.115	0.011	0.281
2: RMDN(1)	0.681	18	–	0.797	0.736	0.144	0.129
3: GARCH(1,1)-t	0.679	25	15	–	0.980	0.729	0.265
4: RMDN(1)-t	0.679	13	14	20	–	0.500	0.114
5: LRMDN(2)	0.676	0	10	14	20	–	0.022
6: RMDN(2)	0.688	16	6	20	10	3	–

Table 21: Out-sample statistics (HR): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

5 Conclusion

In this paper we study a general class of asset return models that nests several existing models as special cases. In particular we specify a nonlinear mixture density network and analyze the impact of nonlinearity and non-gaussian behaviour on the predictive power of conditional variances. We use return series generated from the DJIA over the sample period November 1934 to Dezember 1997 to evaluate the in sample and out of sample forecasting abilities of six different specifications of our models: (i) a simple GARCH(1,1) process with conditional normal distributions, (ii) a GARCH(1,1)-t model, (iii) a recurrent network with a conditional normal distribution, (iv) a linear recurrent mixture density network with two gaussian distributions, (v) a recurrent network with a t-distribution and (vi) a recurrent mixture density network with two normals. Hence, we are able to empirically evaluate the impact of nonlinearity (models (iii), (v) and (vi)) and of non-gaussian distributions (models (ii), (iv), (v) and (vi)) vis a vis the classical GARCH(1,1) model. We use different summary statistics to rank the models according to their forecasting performance. We get the following general results. Different summary statistics generate different rakings of the models and hence there is no consistent ordering to determine the "best" model. This result triggers two additional questions. Firstly, are we using the correct statistics to evaluate the models and secondly, what are the reasons for the

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	0.681	–	0.238	0.689	0.033	0.002	0.488
2: RMDN(1)	0.668	16	–	0.761	0.578	0.031	0.703
3: GARCH(1,1)-t	0.675	23	17	–	0.464	0.223	0.907
4: RMDN(1)-t	0.661	5	27	12	–	0.121	0.394
5: LRMDN(2)	0.648	0	4	9	13	–	0.111
6: RMDN(2)	0.674	26	17	25	26	9	–

Table 22: Out-sample statistics (WHR): Mean values (second column), p-values for the paired t-tests (above the diagonal) and R-values for the paired Wilcoxon tests (below the diagonal).

inconsistent rankings? Looking at the problem of model selection from a practical point of view one could replace the proper statistical theory with a measure of profitability that makes use of the in sample or out of sample forecasting power of the individual variance models together with an appropriate option trading strategy. This approach was followed in Dockner and Strobl (1999), for example, and leads to interesting results. If we follow this route of model selection, however, it must be understood that we are testing a joint hypothesis: the predictive power of the volatility model together with the informational efficiency of the corresponding market.

As a second conclusion we get that non-gaussian models tend to outperform gaussian ones. This is not a surprising result since we know from financial time series analysis that asset returns are characterized by fat tails and that a non-normal distribution is capable of capturing this fact.

Based on our findings in this paper we propose the following directions for future research. Intensify work on a proper theory for model selection so that classes of nested models like the ones presented here can be evaluated and ranked according to their statistical properties. This theory can include both statistical as well as financial models like the one that uses trading profits as a performance measure. Evalute variance models not only on the basis of their predictive power but also on the basis of how well they are able to predict option prices. This, however, requires a very different approach, one in which emphasis is put more on volatility modelling as part of option pricing theory rather than a topic of itself. Both routes will be part of our future work.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) within the research project "Adaptive Information Systems and Modelling in Economics and Management Science" (SFB 010). The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Science and Transport. The authors want thank F. Leisch, P. Tiňo, A. Trapletti and A. Weingessel for valuable discussions. The models were implemented by extending the NETLAB neural network software which can be obtained from (`http://neural-server.aston.ac.uk/`).

References

Akaike, H., 1973, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov and F. Csáki, eds., 2nd International Symposium on Information Theory (Akadémia Kiadó, Budapest) 267-281.

Bera, A.K. and M.L. Higgins, 1993, ARCH models: properties, estimation and testing, Journal of Economic Surveys 7, 307-366.

Bishop, C.M., 1994, Mixture density networks, Neural Computing Research Group Report: NCRG/94/004 (Aston University, Birmingham).

Bishop, C.M., 1995, Neural networks for pattern recognition (Clarendon Press, Oxford).

Bollerslev, T., 1986, A generalized autoregressive conditional heteroskedasticity, Journal of Econometrics 31, 307-327.

Bollerslev, T., 1987, A conditionally heteroskedastic time series model for speculative prices and rates of return, Review of Economics and Statistics 69, 542-547.

Bollerslev, T., Chou, R.Y. and K.F. Kroner, 1992, ARCH modelling in finance: A review of the theory and empirical evidence, Journal of Econometrics 52, 5-59.

Dockner, E.J. and G. Strobl, 1999, Volatility forecasts and the enhancement of risk/return profiles through automated trading strategies, SFB Working paper 44.

Donaldson, R.G. and M. Kamstra, 1997, An artificial neural network-GARCH model for international stock return volatility, Journal of Empirical Finance 4, 17-46.

Engle, R.F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, Econometrica 50, 987-1008.

Engle, R.F. and T. Bollerslev, 1986, Modelling the persistence of conditional variances, Econometric Reviews 5, 1-50.

Engle, R.F., 1990, Discussion: stock market volatility and the crash of 87, Review of Financial Studies 3, 103-106.

Glosten, L.R., Jagannathan, R. and R. Runkle, 1993, Relationship between the expected value and the volatility of the normal excess return on stocks, Journal of Finance 48, 1779-1801.

González Miranda, F. and N. Burgess, 1997, Modelling market volatilities: the neural network perspective, The European Journal of Finance 3, 137-157.

González-Rivera, G., 1998, Smooth-transition GARCH models, Studies in Nonlinear Dynamics and Econometrics 3, 61-78.

Hertz, J., Krogh, A. and R.G. Palmer, 1991, Introduction to the theory of neural computation (Addison-Wesley, Redwood City).

Hornik, K., Stinchcombe, M. and H. White, 1989, Multilayer feedforward networks are universal approximators, Neural Networks 2, 359-366.

Leisch, F., Trapletti, A. and K. Hornik, 1999, Stationarity and stability of autoregressive neural network processes, in: M.J. Kearns, S.A. Solla and D.A. Cohn, eds., Advances in Neural Information Processing Systems 11 (MIT Press) to appear.

McLachlan, G.J. and K.E. Basford, 1988, Mixture models: inference and applications to clustering (Marcel Dekker, New York).

Neuneier, R., Finnoff, W., Hergert, F. and D. Ormoneit, 1994, Estimation of conditional densities: a comparison of neural network approaches, in: M. Marinaro and P.G. Morasso,

eds., ICANN 94 - Proceedings of the International Conference on Artificial Neural Networks (Springer, Berlin) 689-692.
Ormoneit, D. and R. Neuneier, 1996, Experiments in predicting the German stock index DAX with density estimating neural networks, in: Proceedings of the 1996 Conference on Computational Intelligence in Financial Engineering (CIFEr 96), New York, USA.
Ormoneit, D., 1998, Probability estimating neural networks (Shaker, Aachen).
Pagan, A.R. and G.W. Schwert, 1990, Alternative models for conditional stock volatility, Journal of Econometrics 45, 267-290.
Refenes, A.P., 1995, Neural networks in the capital markets (Wiley, New York).
Rojas, R., 1996, Neural networks: a systematic introduction (Springer, Berlin).
Schittenkopf, C., Dorffner G. and E.J. Dockner, 1998, Volatility prediction with mixture density networks, in: L. Niklasson, M. Bodén and T. Ziemke, eds., ICANN 98 - Proceedings of the 8th International Conference on Artificial Neural Networks (Springer, Berlin) 929-934.
Sentana, E., 1991, Quadratic ARCH models: a potential re-interpretation of ARCH models, unpublished manuscript, London School of Economics.
Turner, A.L. and E.J. Weigel, 1992, Daily stock market volatility: 1928-1989, Management Science 38, 1586-1609.
White, H., 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817-838.
Zakoian, J.M., 1990, Threshold heteroskedastic models, unpublished manuscript, CREST, INSEE.