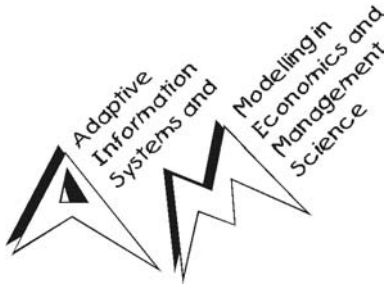


Report Series

Non-linear versus Non-gaussian Volatility Models in Application to Different Financial Markets

Tatiana Miazhynskaia

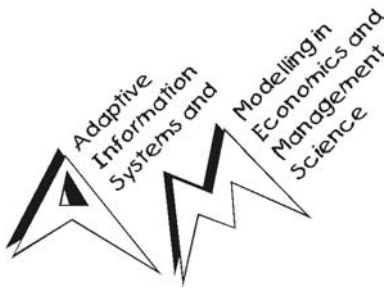
Georg Dorffner

Engelbert J. Dockner

Report No. 84

November, 2003

Report Series



November, 2003

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

Papers published in this report series
are preliminary versions of journal articles
and not for quotations.

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 in Application to Different Financial Markets

Tatiana Miazhynskaia*
Austrian Research Institute for
Artificial Intelligence,
Freyung 6/6, A-1010 Vienna, Austria,
phone:+431.5336112-25,
fax: +431.5336112-77,
tatiana@oefai.at

Georg Dorffner
Austrian Research Institute for
Artificial Intelligence and
Department of Medical Cybernetics
and Artificial Intelligence,
University of Vienna, Freyung 6/2,
A-1010 Vienna, Austria,
georg@ai.univie.ac.at

Engelbert J. Dockner
Department of Business Studies, University of Vienna
Brünner Strasse 72, A - 1210 Vienna, Austria,
Engelbert.Dockner@univie.ac.at

Abstract

We used neural-network based modelling to generalize the linear econometric return models and compare their out-of-sample predictive ability in terms of different performance measures under three density specifications. As error measures we used the likelihood values on the test sets as well as standard volatility measures. The empirical analysis was based on return series of stock indices from different financial markets. The results indicate that for all markets there was found no improvement in the forecast by non-linear models over linear ones, while non-gaussian models significantly dominate the gaussian models with respect to most performance measures. The likelihood performance measure mostly favours the linear model with Student-t distribution, but the significance of its superiority differs between the markets.

Keywords: forecasting, neural networks, time series models, volatility, GARCH

1 Introduction

It is widely agreed that although daily and monthly financial asset returns are unpredictable, return volatility is highly predictable, a phenomen with important implications for financial economics (e.g., Bollerslev *et al.* (1992)). Of course, volatility is inherently unobservable, and most of what we know about volatility has been learned, e.g., by fitting parametric econometric models. Typically a volatility model is used to forecast the absolute magnitude of returns, but it may also be used to predict quantiles or, in fact, the entire density. Such forecasts are used in risk management,

*Corresponding author

derivative pricing and hedging, market making, market timing, portfolio selection and many other financial activities. In each, it is the predictability of volatility that is required. A risk manager must know today the likelihood that his portfolio will decline in the future. An option trader will want to know the volatility that can be expected over the future life of the contract. A portfolio manager may want to sell a stock or a portfolio before it becomes too volatile. A market maker may want to set the bid-ask spread wider when the future is believed to be more volatile. That is why modeling volatility of financial time series have been a very popular research topic for the last several years.

The most famous model widely used in practice is the GARCH (Bollerslev (1986)) where conditional variances are governed by a linear autoregressive process of past squared returns and variances. This model captures several ”stylized facts” of asset return series, namely heteroskedasticity (time-dependent conditional variance), volatility clustering and excess kurtosis. But later studies (e.g., Nelson (1991), Glosten *et al.* (1993), Alles and Kling (1994), Hansen (1994)) have found that there exist additional empirical regularities that can not be described by classical GARCH model, such as leverage effect, negative skewness, fat tails of conditional distribution. Moreover, Alles and Kling (1994) showed for different financial series that the third moments are also time-varied. Further, Harvey and Siddique (2000) presented an asset pricing model which incorporated conditional skewness explicitly and showed its good performance.

We consider the generalization of the classical GARCH model in two directions: the first is to allow for non-linear dependencies in the conditional mean and in the conditional variance and the second concerns specification of the conditional density. As a tool for non-linear regression we used a neural network-based (NN) modeling, so called recurrent mixture density networks, describing the conditional mean and variance by multi-layer perceptrons (the same approach was applied by Schittenkopf *et al.* (1999), Schittenkopf *et al.* (2000) and Bartlmae and Rauscher (2000)). For NN modelling, these conditional moments can be approximated with an arbitrary accuracy if the size of the neural network models is not restricted.

We want to note that NN modeling has become rather popular methodology in the last research literature on financial modeling. We mention only recent ones: Bartlmae and Rauscher (2000), Boero and Cavallil (1997), Dunis and Jalilov (2002), González and Burgess (1997), H.-L. Poh (1998), J. T. Yao (2000), where the NN approach was found to be advantageous. This is the case because NN modeling is a semi-parametric and non-linear modeling technique in which data series themselves identify relationships among variables. And as a semi-parametric model, NN has the following important advantages over the more traditional parametric models. Since it does not rely

on restrictive parametric assumptions such as normality, stationarity, or sample-path continuity, it is robust to specification errors plaguing parametric models. Moreover, a NN model is sufficiently flexible and can easily encompass a wide range of securities and fundamental asset price dynamics. Indeed, NN has considerable flexibility to uncover hidden non-linear relationships among several classes of individual forecasts and realizations (Donaldson and Kamstra (1997)).

Concerning distributions, we compare three different density specifications: 1) the standard GARCH gaussian model and its non-linear generalization with *normal distribution*; 2) the GARCH model and its non-linear neural network generalization with a *Student's t-distribution*; 3) linear and non-linear recurrent mixture density models, which approximate the conditional distributions by a *mixture of gaussians* (two components). All these distributions model heteroskedastic data. The models with t-distribution produce also the conditional leptokurtosis. But only the linear and non-linear mixture models allow the higher moments to be time varying in general that hopefully give more flexibility to this class of models. We should note that the modelling of the dependence of higher-order moments on the past is rather rare in the financial literature. We can mention Gallant *et al.* (1991) with semi non-parametric approach to density estimation, based on a series expansion about the Gaussian density, and Hansen (1994) who applied fixed parameterized forms.

The empirical analysis was based on return series of stock indices from different financial markets. We used return series of the Dow Jones Industrial Average (USA), FTSE 100 (Great Britain) and NIKKEI 225 (Japan) over a period of more than 12 years in order to evaluate in detail the out-of-sample predictive performance of our models. The models were evaluated with respect to likelihood as well as with respect to their volatility forecasting performance. The original return series due to their length could be split into several parts and each of the models was estimated separately on every part. Thus, we could not only compare the models with respect to performance results but also apply statistical tests to find out whether the differences in performance were significant. We want to note that we continue the work of Schittenkopf *et al.* (1999) and Schittenkopf *et al.* (2000), comparing non-linear versus non-gaussian volatility models for data from different financial markets.

The paper is organized as follows. In the next section we present the models we are working with. Section 3 discusses the data that is used in the empirical analysis. Sections 4 present the performance measures together with the estimation procedure. In the section 5 we discuss the results of the extensive empirical experiments. Finally, Section 6 concludes the paper.

2 Description of Models

The usual approach for modeling return series is to split the returns into a predictable deterministic component μ_t (mean) and a stochastic error process e_t with independent realizations and with $\mathbf{E}(e_t|I_{t-1}) = 0$, $\mathbf{E}(e_t^2|I_{t-1}) = \sigma_t^2$, where I_{t-1} denotes series history up to time $t-1$. σ_t is an estimate of the volatility of the return series at time t.

The most prominent model of time-varying volatility is GARCH(p,q) introduced in Bollerslev (1986), where conditional variances are governed by a linear autoregressive process of past squared errors and variances, i.e.

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

with the restrictions $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$ to ensure positive variances. Stationarity in variance imposes the condition $\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1$.

In this paper we consider only the GARCH(1,1) model which is (1) with $p = q = 1$. The GARCH(1,1) specification has proven attractive for models of returns. It typically dominates other GARCH models using Akaike or Schwarz information criteria (see Bollerslev *et al.* (1992)).

Due to the significant autocorrelation found in many return series, we chose the autoregressive process of the order 1 for the mean equation, i.e. $\mu_t = a_1 r_{t-1} + a_0$.

One possible extension of the GARCH model is to substitute the conditional normal distribution by a Student's-t distribution with ν degrees of freedom in order to allow for excess kurtosis in the conditional distribution (see Bollerslev (1987) for details). The conditional variance is again given by the specification (1) and the parameter restrictions for stationarity of the model are the same as those for the GARCH model. Since the conditional density of t-distribution is symmetric, the conditional skewness is again 0. The new parameter degrees of freedom ν determines, among other characteristics, the kurtosis of the conditional distribution. For $\nu > 4$, the conditional kurtosis is given by $3(\nu-2)/(\nu-4)$ which is always larger than 3. Therefore, GARCH-t models exhibit fat tails in the unconditional and in the conditional distribution. As for GARCH models, the higher-order moments of the distribution are not time-dependent.

The second direction of the extension of the classical GARCH model is to allow for non-linear dependencies in the conditional mean and in the conditional variance. As a tool for non-linear regression we used neural network-based modeling, so called recurrent mixture density networks, describing conditional mean and variance by multi-layer perceptrons (MLP) (the approach applied by Schittenkopf *et al.* (2000)).

In the simplest case an MLP with one input unit, one layer of hidden units and one output unit realizes the mapping

$$\tilde{f}(x_t) = g\left(\sum_{j=1}^{H} v_j h(w_j x_t + c_j) + s x_t + b\right), \tag{2}$$

where H denotes the number of hidden units, w_j and v_j the weights of the first and second layer, s the shortcut weight, and c_j and b the bias weights of the first and second layer. In general, the activation function h of the hidden units is chosen to be bounded, non-linear, and increasing as, e.g., the hyperbolic tangent. The activation function of the output unit may be unrestricted, e.g. $g(x) = x$. Hornik *et al.* (1989) showed that MLP can approximate any smooth, non-linear function with arbitrary accuracy as the number of hidden units tends to infinity. In such a way, MLP can be interpreted as a non-linear autoregressive model of first order and can be applied to predict the parameters of conditional density of the return series.

Recurrent mixture density network models RMDN(n) approximate the conditional distributions of returns by a mixture of n Gaussians:

$$\rho(r_t|I_{t-1}) = \sum_{i=1}^{n} \pi_{i,t} k(\mu_{i,t}, \sigma_{i,t}^2), \tag{3}$$

where $k(\mu_{i,t}, \sigma_{i,t}^2)$ is the gaussian density and the parameters $\pi_{i,t}$, $\mu_{i,t}$, and $\sigma_{i,t}^2$ of the n gaussian components are estimated by three MLPs:

$$\pi_{i,t} = s(\tilde{\pi}_{i,t}) = \frac{\exp(\tilde{\pi}_{i,t})}{\sum_{j=1}^{n} \exp(\tilde{\pi}_{j,t})} \tag{4}$$

$$\tilde{\pi}_{i,t} = \tilde{f}_{1,i}(r_{t-1}) \tag{5}$$

$$\mu_{i,t} = \tilde{f}_{2,i}(r_{t-1}) \tag{6}$$

$$\sigma_{i,t}^2 = \tilde{f}_{3,i}(\sigma_{1,t-1}^2, \sigma_{2,t-1}^2, \ldots, \sigma_{n,t-1}^2, e_{t-1}^2), \tag{7}$$

where $\tilde{f}_{m,i}$ denotes the ith component of the output of the mth MLP. The softmax function $s(\tilde{\pi}_{i,t})$ in (4) ensures that the priors $\pi_{i,t}$ are positive and that they sum up to 1 which makes the right-hand side of (3) a density. The MLPs $\tilde{f}_{1,i}$ and $\tilde{f}_{2,i}$ estimating the priors and the centers are standard MLPs (2). The MLP $\tilde{f}_{3,i}$ estimating the variances of the normal densities is recurrent and has the form

$$\sigma_{i,t}^2 = g\left(\sum_{j=1}^{H} v_{ij} h\left(w_{j0} e_{t-1}^2 + \sum_{k=1}^{n} w_{jk} \sigma_{k,t-1}^2 + c_j\right) + s_{i0} e_{t-1}^2 + \sum_{k=1}^{n} s_{ik} \sigma_{k,t-1}^2 + b_i\right).$$

Its input is $(n+1)$-dimensional (plus bias) and consists of the squared error $e_{t-1}^2 = (r_{t-1} - \mu_{t-1})^2$ 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 non-negative network outputs, i.e., conditional variances.

The total number of weights of an RMDN(n) model equals $2(2n+3)H + n^2 + 6n$.

In terms of the parameters of the mixture distribution, the conditional mean and the conditional variance of future return are given by

$$\mu_t = \sum_{i=1}^{n} \pi_{i,t}\mu_{i,t}, \tag{8}$$

$$\sigma_t^2 = \sum_{i=1}^{n} \pi_{i,t}\left(\sigma_{i,t}^2 + (\mu_{i,t} - \mu_t)^2\right). \tag{9}$$

The skewness s_t and the kurtosis k_t of the conditional distribution can also be calculated analytically:

$$s_t = \frac{1}{\sigma_t^3}\sum_{i=1}^{n} \pi_{i,t}\left(3\sigma_{i,t}^2(\mu_{i,t} - \mu_t) + (\mu_{i,t} - \mu_t)^3\right) \tag{10}$$

$$k_t = \frac{1}{\sigma_t^4}\sum_{i=1}^{n} \pi_{i,t}\left(3\sigma_{i,t}^4 + 6\sigma_{i,t}^2(\mu_{i,t} - \mu_t)^2 + (\mu_{i,t} - \mu_t)^4\right) \tag{11}$$

This time-dependence of the higher-order moments is an appealing feature of RMDN models and it is in contrast to the properties of GARCH and GARCH-t models.

We note that an RMDN model with one Gaussian component ($n = 1$) can be interpreted as a non-linear extension of a GARCH model.

There are two other models that must be introduced in order to analyze the influence of linear and non-linear functions and density specification on the performance of return series models in detail. First, we consider non-linear GARCH-t models in the framework of RMDN models by replacing the weighted sum of normal densities in (3) by the density of the t-distribution. These models will be called RMDN(1)-t models in the following. Secondly, one must study the performance of mixture models for the case that only linear functions are allowed. More precisely, in all three MLPs estimating the parameters of the mixture model the activation function h of the hidden units are supposed to be linear. These linear mixture models are referred to as LRMDN(n) models in the following. LRMDN(1) is again the classical GARCH model. We limited ourselves to the cases $n = 1$ and $n = 2$, mainly focusing on the non-linearity aspects.

In such a way, we concentrate further on the comparison of the performance of the following six models according to two dimensions: linearity issue and distributional aspect:

type of distribution	Linear	Non-linear
gaussian	**GARCH(1,1)**	**RMDN(1)**
t-distribution	**GARCH(1,1)-t**	**RMDN(1)-t**
mixture of gaussians	**LRMDN(2)**	**RMDN(2)**

3 Data Sets

In our numerical experiments we used three data sets related to different financial markets:

1. daily closing values of the American stock index Dow Jones Industrial Average (DJIA);
2. daily closing values of the FTSE 100 traded at the London Stock Exchange;
3. daily closing values of the Japan index NIKKEI 225.

The index series were taken from public sources. The time interval for all data sets was 13 years from 1985 to 1997. All data were transformed into continuously compounded returns η_t (in percent) in the standard way by the natural logarithm of the ratio of consecutive daily closing levels. The time series of returns together with their unconditional kernel density approximations are depicted in Fig.1. We want to note that we did not perform the data cleaning to delete the outliers from the further analysis. Thus, the extremal negative market returns in October, 1987 are left and used in the estimation procedure.

In order to take care of stationarity issues and increase the reliability of the empirical analysis, all time series were divided into overlapping segments of a fixed length of 700 trading days, where the first 500 returns of each segment form a training set, the next 100 points form a validation set and the remaining 100 returns are used for testing. The detailed segment structure is presented in see Fig.2. The first segment starts on trading day 1 and ends on day 700, the second segment begins on day 101 and ends on trading day 800 and so on. The training sets are used to optimize the parameters of each model. The validation sets are used for an "early stopping" strategy to avoid overfitting for the neural networks models and independent test sets are reserved for out-of-sample model performance evaluation. The test sets are not overlapping.

In such a way, according to the available data, we got around 25 segments for the discussed return series. The summary of the descriptive statistics of the data are plotted in Fig.3. It can be seen that all logarithmic series exhibit time-dependent significant skewness and excess kurtosis indicating non-normality of the unconditional distributions. Enormous skewness and kurtosis on segments 1-7 can be explained by the influence of the October 1987 default.

4 Error Measures and Estimation of Models

We fitted GARCH(1,1), RMDN(1), GARCH(1,1)-t, RMDN(1)-t, LRMDN(2) and RMDN(2) models to each of the training sets separately. The number of optimized parameters of a particular model

is 5 for the GARCH(1,1) model, 26 for RMDN(1), 6 for GARCH(1,1)-t, 27 for RMDN(1)-t, 16 for LRMDN(2) and 54 for RMDN(2). The number of hidden units of the MLPs in the RMDN-models was chosen to be $H = 3$. The parameters of all models were optimized with respect to the average negative log likelihood of the sample

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

where N denotes the sample size and $\rho(r_t|I_{t-1})$ is the conditional probability density function of the corresponding distribution. We refer to $\mathcal{L}$ as the *loss function* of a data set, since we will make use of values of $\mathcal{L}$ calculated for data sets which were not used to estimate the model parameters.

The optimization routine was a scaled conjugate gradient algorithm. We performed optimization of RMDN models with several parameter initializations in an attempt to approach a global optimum. For the models with t-distribution, the degrees-of-freedom parameter was additionally optimized by a one-dimensional search routine.

Since the main goal of this work is out-of-sample diagnostic, i.e., comparison of model performance on a future data set (test set), we are interested in obtaining models with optimal generalization performance. However, all standard neural network architectures such as the fully connected multi-layer perceptron are prone to overfitting (see, e.g., Geman *et al.* (1992), Reed (1993)): while the network seems to become better and better , i.e., the error (in our case - the value of the loss function) on the training set decreases, beginning with some point during training the error on an unseen sample increases. In order to prevent the RMDN models from overfitting the training data, the generalization error is estimated by the performance of the model on a validation set and an "early stopping" strategy (Prechelt (1998)) is applied . More precisely, the model parameters are optimized with respect to the loss function on the training set and after each iteration the loss function on the validation set is calculated. Finally, the RMDN model on the optimization iteration t^* is selected, where

$$t^* = \arg \min_{t_0<t<T} \mathcal{L}_{\text{validation}}(t),$$

where t is an iteration number;

T is the number of all iterations performed;

t_0 - minimal iteration number chosen to avoid artefact behaviour of the loss function on the validation set in such a way that the parallel value of the loss function on the training set of simpler (less parametrized) model is beaten.

In addition to the loss function, some other error measures common in literature are applied to

analyze the performance of the models. Since the actual purpose of a volatility model is to predict future volatility, we need some "true" measure of volatility. Because the volatility process is not observed, researchers have used a variety of empirical measures of daily return variability, often called realized volatility. The most common method for computing a daily realized volatility which we also apply here is to square the daily period returns. Andersen *et al.* (2001) are critical about the squared daily returns to measure the realized volatility and propose to use high-frequency intraday returns for this purpose. But our main point is that even "bad" volatility forecasts according to evaluation criteria do not necessarily imply that these volatility forecasts are not useful for comparing the model performances. In such a way, denoting the estimated conditional variance as $\hat{\sigma}_t^2$ (for time step t), we calculate

-the normalized mean squared error

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

relating mean square error of the modeled volatility $\hat{\sigma}_t^2$ to the mean square error of the naive model $\hat{\sigma}_t^2 = r_{t-1}^2$.

-the normalized mean absolute error

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

which is more robust against outliers in comparison with NMSE.

Furthermore, the ability of the model to predict increases and decreases of volatility is investigated with the help of the following measures:

- the hit rate

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

with

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

as a measure of how often the model gives the correct direction of change of volatility. HR$\in (0,1)$, where a value of 0.5 indicates that the model is not better than a random predictor generating a random sequence of up and down moves with equal probability.

-the weighted hit rate

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

WHR takes also the real changes $r_t^2 - r_{t-1}^2$ into account meaning that large changes are considered more important. WHR $\in$ (-1,1) with a value of 1 in the perfect case.

In such a way, we compare out-of-sample model performances based on different volatility measures.

We want to note that the volatility measures above consider only the second conditional moment, not taking into account the flexible behaviour of the mixture density models with respect to the higher order moments. Moreover, the models adjustment was based on the optimization of the likelihood function (without any relation to the volatility measures). That is, we think about the likelihood values on the test sets as being more appropriate error measure and base our conclusions mostly on likelihood results.

5 Results

We investigated out-of-sample performance of the models, i.e. error values on data sets disjoint from the training data. The parameters of the models were estimated by the procedure described above using training and validation sets and then, keeping the parameters fixed, we computed the error measures on the test sets of the corresponding segments. The favour to the out-of sample criterion for a comparison of the models was given because in this case possible overparametrizations may be neglected.

In spite of the numerous different starting values in the optimization procedure for the neural network models, we obtained on some test sets for single models values of the loss function that were three-five times larger than the average level. Based on the smooth behaviour of the loss function for other models on the same test set, we considered such models to be "non-indicated" overfitting cases and deleted these test sets parallel for all data sets from analysis. After eliminating, we got 24 test sets for model evaluation.

We checked the GARCH models for the stationatiry (the condition $\alpha_1 + \beta_1 < 1$). All the models were found stationary except for NIKKEI 225 series for the sets 10 and 11 (years 1987-1988).

To illustrate the flexible structure of the mixture models, we check the prediction behaviour of the higher order moments for DJIA returns for two test sets (in 1990 year) in Fig.4. While the models with t-distribution keep the conditional skewness to be 0 and the kurtosis to be 4.33 and 4.89 for the first and the second part of the presented test sample, the mixture models predict skewness and kurtosis plots far from being stationary.

The performance of the models on each of the test sets for DJIA data with respect to the loss

measure is summarized in Fig.5. For convenience of the analysis, all the results are presented with respect to the functional form of conditional variance equation (linearity versus non-linearity) and a type of conditional distribution. We compare the performance of the gaussian model versus the model with t-distribution versus the mixture of gaussians. Thus, three lines in the upper panel of the figure give the values of the relevant statistic for the linear models GARCH(1,1), GARCH(1,1)-t and LRMDN(2). The bottom panels present non-linear models RMDN(1), RMDN(1)-t and RMDN(2). Based on Fig.5, we can make the following preliminary conclusion: in general, the differences between the models over the most test sets are negligible. On single sets (test set 5,6 and 10-12) the linear and non-linear gaussian models show the worse results, while the models with Student-t distribution exhibit the smallest likelihood values. If we compare the upper and the lower plots in Fig.5, we can note that the linear models and their non-linear neural network generalizations reach mostly equal likelihoods. Single cases, like test set 11, where non-linear RMDN(1) model show the loss value close to 1.6 against 1.2 for the linear GARCH(1,1), and the sets 23-24, where the non-linear mixture density RMDN(2) behaves significantly worse than its linear version, can be explained by the problems with the maximum likelihood estimation of the non-linear models.

In order to be statistically consistent in the model selection process, we tested 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. The application of the paired tests 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. Additionally, for the paired t-test it is assumed that the differences are normally distributed what is not always the case and whereas for the paired Wilcoxon test it is only assumed that the distribution of the differences is symmetric. Because of this fact, our conclusions are mostly based on the results of the paired Wilcoxon test. For our most reliable evaluation criterion (loss function) we present the results (Tables 1, 4, 7) of both tests, while for the volatility measures - only the results of the paired Wilcoxon test (Tables 2-3, 5-6, 8-9) (to save the space).

The results of the paired test for DJIA return series are summarized in Tables 1-3. The column "mean" gives the mean value of the corresponding statistic over all test sets. The minimal mean value 1.184 of the loss function is reached by the GARCH(1,1)-t model. The p-values of both paired tests

between this model and all other models are less than 0.025, indicating that these differences are significant, i.e. GARCH(1,1)-t significantly outperforms all other models. Its non-linear generalization RMDN(1)-t is either among the best with the mean loss function value 1.201, but the p-value of the Wilcoxon test for the differences between this model and the linear mixture density model LRMDN(2) is 0.103 (or even 0.776 by t-test). At that, RMDN(1)-t performance statistically does not differ much from the LRMDN(2) performance over all test sets in consideration. The third group of the models consists of both gaussian models and the non-linear mixture model. The performance of this group with respect to the loss values is the worst. It also seems that on average linearity plays some positive role since linear models reached in a whole smaller values of loss function compared to their non-linear analogs but this differences are mostly not significant (the p-values between the linear models and their non-linear versions are 0.710, 0.015 and 0.319 for the gaussian, Student-t and the mixture of gaussians conditional distributions, respectively.

The average results with respect to the volatility measures (Tables 2-3) show some consistency with the conclusions above. All the measures favour the models with Student-t conditional distribution, with the non-linear RMDN(1)-t to be slightly better. But the performance of the non-linear gaussian model RMDN(1) appeared to differ not significantly from them (with p-values more than 0.212). On the contrary to the results above, the linear mixture model together with its non-linear generalization demonstrates the bad performance with respect to all volatility error measures.

Table 4 together with Fig.6 show the results for FTSE 100 returns with respect to the loss function. The graphical plots give no preferences to any model or any class of models. With respect to the average statistics, the linear models GARCH(1,1)-t and the mixture LRMDN(2) outperform all other models, but their advantage appeared to be not significant comparing with all other models (the p-values of the Wilcoxon test between LRMDN(2) and all other models are more than 0.107). Moreover, according to the paired tests, there is no statistical difference between the linear and non-linear mixture models at all (the corresponding p-value of the Wilcoxon test is 1.00). The non-linear model with Student-t conditional distribution appears in this case to be among the worst with respect to the mean value of the loss function over all test sets.

For the alternative error measures we included in the paper again only the results of the Wilcoxon paired test (Tables 5-6). We found no consistency with respect to the model ordering for these measures. In addition, almost all differences in the models' performance are declared by the Wilcoxon test to be not significant. The numerical square errors slightly prefer both models with t-distribution and place the mixture models on the last position. On the contrary, the hit rates' measures put the

mixture models before the GARCH(1,1)-t and RMDN(1)-t. Non-linearity seems to play again no significant role in model performance. So, in general, the obtained results clearly favour the models with t-distribution and mixtures of gaussians over the gaussian models, but are uncertain with some strict model ordering within non-gaussian models.

Out-of-sample diagnostic of NIKKEI 225 series is given in Fig.7 and Tables 7-9. As in the case with FTSE 100 data, the graphical plot of the likelihood values over all test sets in Fig.7 gives no clear preferences to any model. But the results of the paired statistical tests with respect to the loss function and the volatility error measures show the advantage of the linear and non-linear models with t-distribution. The mixture models take the second place, but their performance lose to the models with t-distribution not significantly (the corresponding p-values are more than 0.109). But both gaussian models again show the worst efficiency with respect to all error measures.

6 Discussion and Conclusions

We analyzed the impact of non-linearity and of non-gaussian distributions versus the classical GARCH model. The empirical analysis was based on return series of stock indices from different financial markets. We divided data into a number of segments in order to take into account stationarity issues and to perform reliable model selection in the maximum likelihood framework. The parameters of all models were first estimated on the training part of every segment by the usual maximum likelihood methodology and then we performed out-of-sample forecasts and forecast evaluations on the test sets. The models were evaluated with respect to the likelihood as well as standard volatility performance measures.

In analyzing the obtained results, we emphasize the likelihood characteristic of the test sets because we consider it to be a more reliable error measure compared to the alternative volatility measures. This is explained by, first, our uncertainty concerning the squared daily returns as the estimates for the realized volatility. Second, the volatility measures above consider only the second conditional moment, not taking into account the flexible behaviour of the mixture density models with respect to higher order moments, while the likelihood measure reflects the conditional distribution as a whole. And the last point is that we fit the models by optimizing the likelihood function and, therefore, it is more adequate to compare models performances.

Summing up, we derived the following conclusions:

- All statistical tests clearly confirmed the expected conclusion that non-gaussian models significantly dominated the gaussian ones with respect to the most performance measures for all

stock indices considered.

- At the same time, within non-gaussian models themselves there was some difference across the markets, namely, for DJIA series almost all tests gave the significant superiority to the models with t-distributions, while for FTSE 100 data the mixture models were among the best and for the japanese index NIKKEI the models with t-distribution and mixture density networks were almost statistically equal in their performance. But, while both the models with t-distribution and mixture density networks are capable to capture fat tail elements in the conditional distribution, only mixture density networks allow for time-varying skewness and kurtosis which are found to be common in financial markets.

- The likelihood performance measure mostly favours the linear GARCH(1,1)-t model over all data sets considered, but the significance of its superiority differs between the markets. The results with respect to the alternative volatility measures do not show any consistent preferences to one definite model, but give slight advantage to the class of the models with t-distribution.

- For all markets we did not find any improvement in the forecast by non-linear models over linear ones for all the error measures applied.

To extend the paper, the alternative approach (to the discussed statistical measures) to compare the model performance can be based on some practical model applications, e.g., in risk management, to evaluate and compare Value-at-risk predictions, or to define the trading strategies.

Acknowledgements

This work was funded by the Austrian Science Fund (FWF) under grant SFB#010: "Adaptive Information Systems and Modeling in Economics and Management Science". The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.

The models were estimated using the NETLAB neural network software (can be downloaded from http://neural-server.aston.ac.uk) adapted by Christian Schittenkopf to the model specifications considered in the paper.

References

Alles, L., Kling, J., 1994. Regularities in the variation of skewness in asset returns, *Journal of Financial Research*, 17 427–438.

Andersen, T., Bollerslev, T., Diebold, F., Ebens, H., 2001. The distribution of realized stock return volatility, *Journal of Financial Economics*, 61 43–76.

Bartlmae, K., Rauscher, F. A., 2000. Measuring dax market risk: A neural network volatility mixture approach, presentation at the FFM2000 Conference, London, 31 May-2 June, can be downloaded from http://www.gloriamundi.org/picsresources/kbrauscher.pdf.

Boero, G., Cavallil, E., 1997. Exchange rate forecasting: Neural networks versus linear econometric models, *Neural Network World*, 1 29–42.

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., Kroner, K., 1992. ARCH modelling in finance: A review of the theory and empirical evidence, *Journal of Econometrics*, 52 5–59.

Donaldson, R., Kamstra, M., 1997. An artificial neural network-GARCH model for international stock return volatility, *Journal of Empirical Finance*, 4(1) 17–46.

Dunis, C. L., Jalilov, J., 2002. Neural network regression and alternative forecasting techniques for predicting financial variables, *Neural Network World*, 12 113–139.

Gallant, R., Hsieh, D., Tauchen, G., 1991. On fitting a recalcitrant series:the pound/dollar exchange rate 1974-83, in: Barnett, W., Powell, J., Tauchen, G. (Eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, Cambridge University Press.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma, *Neural Computation*, 4 1–58.

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

González, M., Burgess, N., 1997. Modelling market volatilities: the neural network perspective, *The European Journal of Finance*, 3 137–157.

H.-L. Poh, T. J., J. T. Yao, 1998. Neural networks for the analysis and forecasting of advertising and promotion impact, *International Journal of Intelligent Systems in Accounting, Finance and Management*, 7(4) 253–268.

Hansen, B., 1994. Autoregressive conditional density estimation, *International Economic Review*, 35 705–730.

Harvey, C., Siddique, A., 2000. Conditional skewness in asset pricing tests, *Journal of Finance*, 55 1263–1295.

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

J. T. Yao, Y. L. L., C. L. Tan, 2000. Option prices forecasting using neural networks, *International Journal of Management Science*, 28(4) 455–466.

Nelson, D., 1991. Conditional heteroskedasticity in asset returns: a new approach, *Econometrica*, 59 347–370.

Prechelt, L., 1998. Early stopping - but when?, in: *Neural Networks : Tricks of the Trade*, G.B. Orr, Univirsität Kalsruhe, Berlin, pp. 55–69.

Reed, R., 1993. Prunung algorithm - a survey, *IEEE Transactions on Neural Networks*, 4(5) 740–746.

Schittenkopf, C., Dorffner, G., Dockner, E. J., 1999. Non-linear versus non-gaussian volatility models, Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science, Vienna.

Schittenkopf, C., Dorffner, G., Dockner, E. J., 2000. Forecasting time-dependent conditional densities: a seminonparametric neural network approach, *Journal of Forecasting*, 19 355–374.

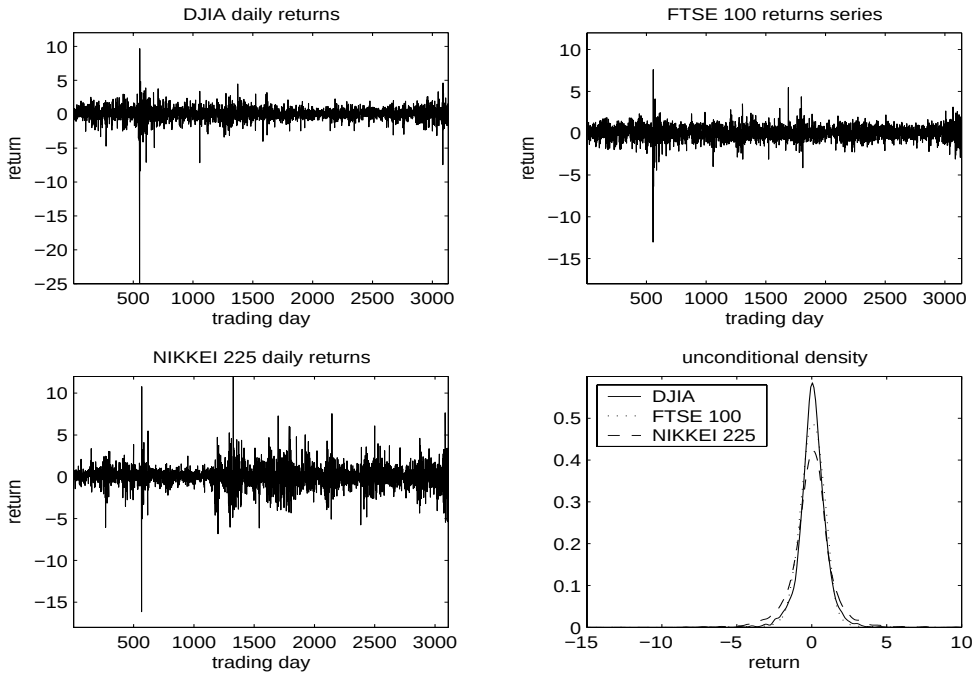


Figure 1: The plots and kernel density approximation of the unconditional densities.

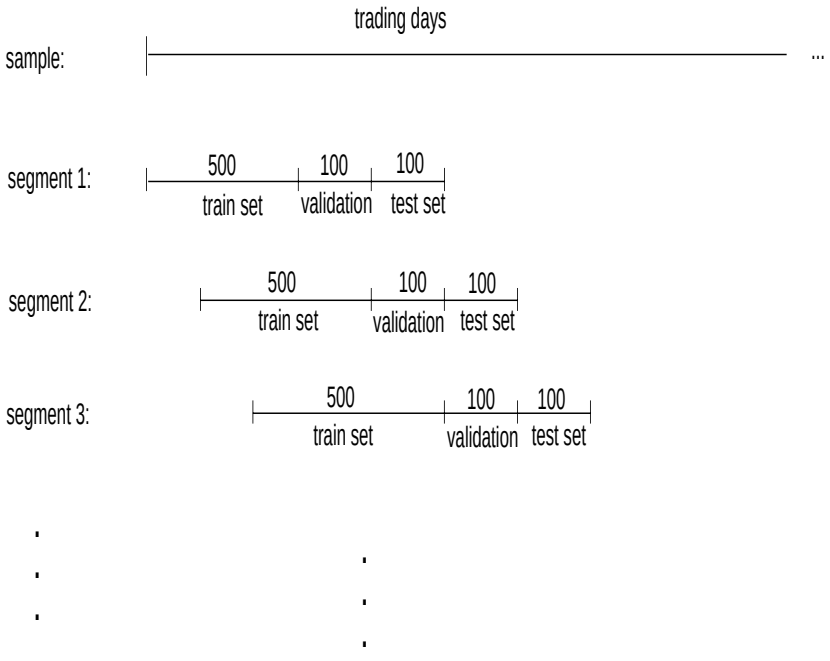


Figure 2: Segment structure of the data sets

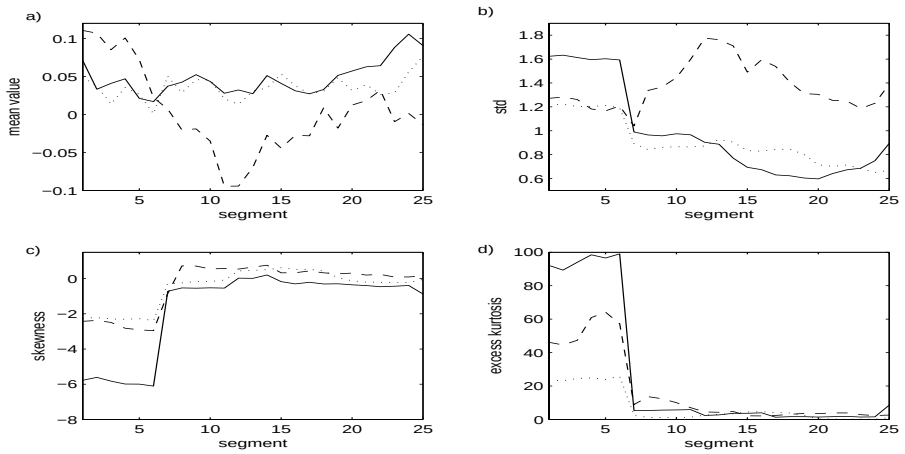


Figure 3: The basic statistics of the return series for all segments (a) - mean values; (b) - standard deviation; (c) - skewness; (d) - kurtosis. Solid line corresponds DJIA return series, dotted line denotes the results for FTSE 100 data and dashed line is used for NIKKEI 225 data.

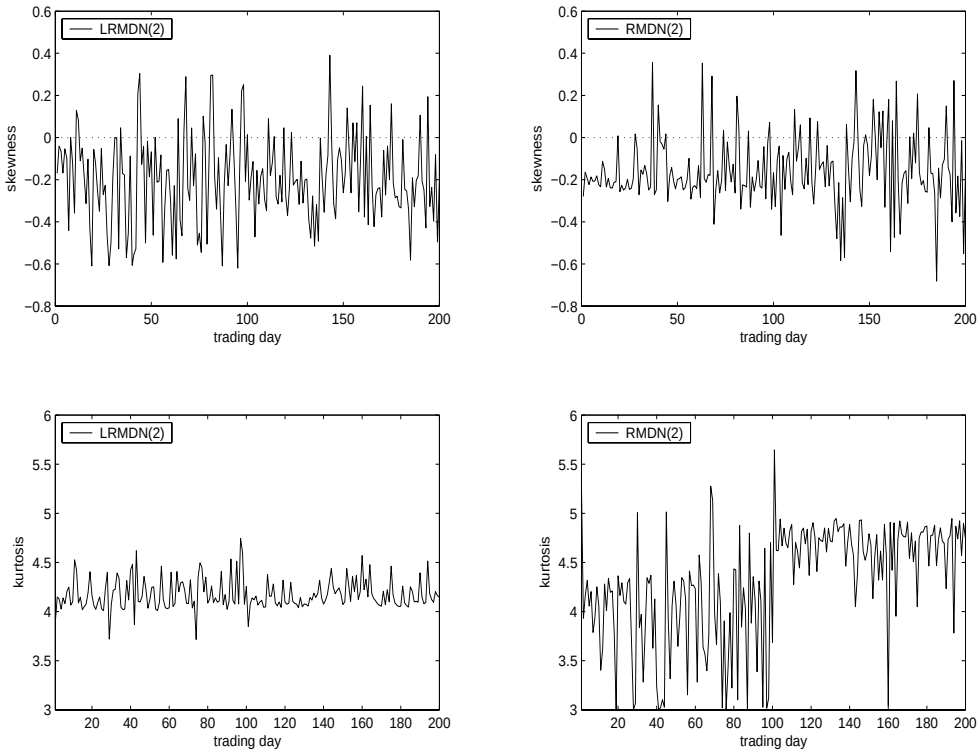


Figure 4: The predicted skewness and kurtosis for DJIA returns (two test sets) by the mixture models.

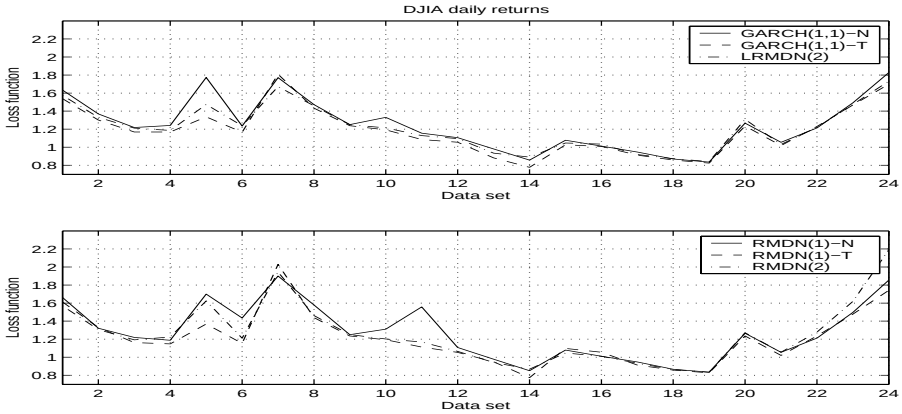


Figure 5: DJIA: the loss function values of linear (in the upper figures) and non-linear (in the lower figures) models compared with respect to conditional distributions.

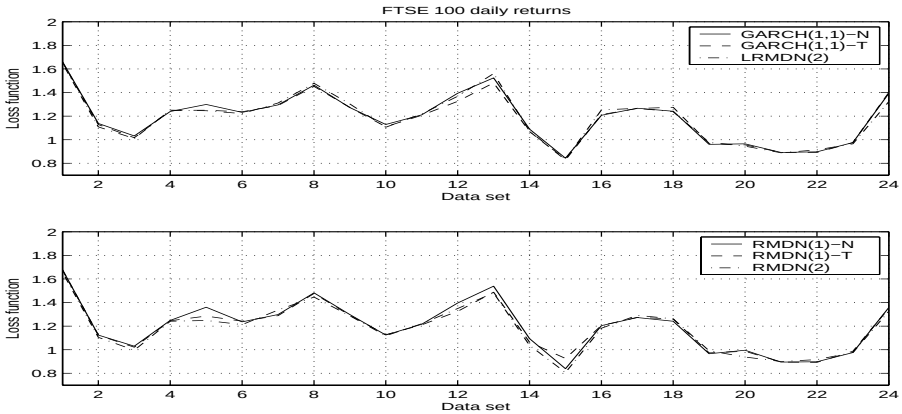


Figure 6: FTSE 100: the loss function values of linear (in the upper figures) and non-linear (in the lower figures) models compared with respect to conditional distributions.

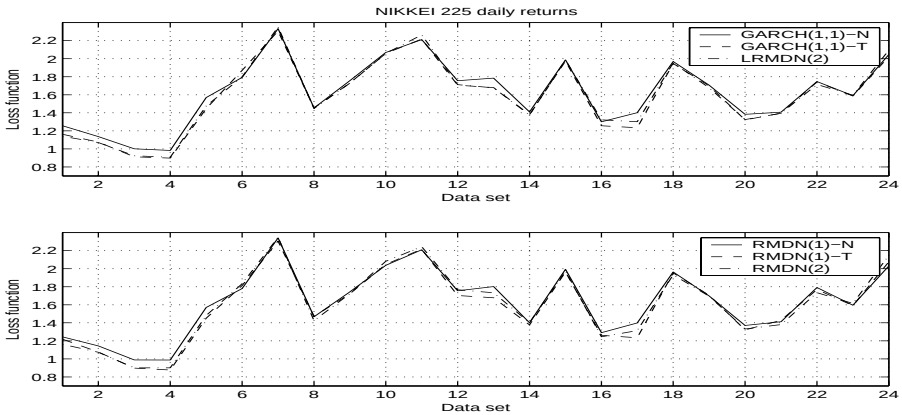


Figure 7: NIKKEI 225: the loss function values of linear (in the upper figures) and non-linear (in the lower figures) models compared with respect to conditional distributions.

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	1.249	-	0.165	0.002	0.073	0.019	0.778
2: RMDN(1)	1.278	0.710	-	0.001	0.012	0.009	0.385
3: GARCH(1,1)-t	1.184	0.000	0.000	-	0.025	0.007	0.004
4: RMDN(1)-t	1.209	0.004	0.004	0.015	-	0.776	0.050
5: LRMDN(2)	1.214	0.008	0.010	0.000	0.103	-	0.090
6: RMDN(2)	1.255	0.265	0.230	0.000	0.032	0.391	-

Table 1: DJIA daily returns: Loss function statistics. Mean values (second column), p-values for the paired t-tests (above the diagonal) and p-values for the paired Wilcoxon signed rank tests (below the diagonal).

Model	NMSE mean	1	2	3	4	5	6	NMAE mean
1: GARCH(1,1)	0.758	-	0.230	0.219	0.013	0.710	0.290	0.872
2: RMDN(1)	0.747	0.407	-	0.886	0.230	0.086	0.954	0.841
3: GARCH(1,1)-t	0.739	0.008	0.304	-	0.021	0.097	0.797	0.833
4: RMDN(1)-t	0.730	0.049	0.932	0.732	-	0.011	0.241	0.785
5: LRMDN(2)	0.765	0.253	0.024	0.003	0.012	-	0.067	0.885
6: RMDN(2)	0.764	0.424	0.024	0.032	0.015	0.886	-	0.860

Table 2: DJIA daily returns: NMSE and NMAE statistics. Mean values of NMSE (second column) together with p-values for the paired Wilcoxon signed rank tests (below the diagonal). Mean values of NMAE (the last column) together with p-values for the paired Wilcoxon signed rank tests (above the diagonal).

Model	HR mean	1	2	3	4	5	6	WHR mean
1: GARCH(1,1)	0.693	-	0.420	0.035	0.021	0.131	0.058	0.729
2: RMDN(1)	0.700	0.421	-	0.212	0.248	0.117	0.036	0.736
3: GARCH(1,1)-t	0.698	0.394	0.943	-	0.286	0.039	0.012	0.750
4: RMDN(1)-t	0.719	0.013	0.019	0.001	-	0.033	0.006	0.756
5: LRMDN(2)	0.687	0.404	0.124	0.360	0.002	-	0.548	0.714
6: RMDN(2)	0.693	0.968	0.240	0.520	0.005	0.363	-	0.714

Table 3: DJIA daily returns: HR and WHR statistics. Mean values of HR (second column) together with p-values for the paired Wilcoxon signed rank tests (below the diagonal). Mean values of WHR (the last column) together with p-values for the paired Wilcoxon signed rank tests (above the diagonal).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	1.189	-	0.227	0.028	0.389	0.457	0.612
2: RMDN(1)	1.217	0.153	-	0.125	0.830	0.190	0.214
3: GARCH(1,1)-t	1.179	0.012	0.007	-	0.250	0.111	0.083
4: RMDN(1)-t	1.215	0.831	0.059	0.048	-	0.325	0.364
5: LRMDN(2)	1.184	0.648	0.107	0.236	0.394	-	0.635
6: RMDN(2)	1.187	0.855	0.094	0.107	0.927	1.000	-

Table 4: FTSE 100 daily returns: Loss function statistics. Mean values (second column), p-values for the paired t-tests (above the diagonal) and p-values for the paired Wilcoxon signed rank tests (below the diagonal).

Model	NMSE mean	1	2	3	4	5	6	NMAE mean
1: GARCH(1,1)	0.733	-	0.301	0.001	0.016	0.260	0.429	0.807
2: RMDN(1)	0.732	0.715	-	0.927	0.006	0.761	0.784	0.802
3: GARCH(1,1)-t	0.729	0.004	0.412	-	0.121	0.274	0.648	0.798
4: RMDN(1)-t	0.729	0.162	0.135	0.605	-	0.144	0.107	0.789
5: LRMDN(2)	0.733	0.670	0.429	0.101	0.274	-	0.301	0.803
6: RMDN(2)	0.735	0.484	0.089	0.023	0.073	0.808	-	0.800

Table 5: FTSE 100 daily returns: NMSE and NMAE statistics. Mean values of NMSE (second column) together with p-values for the paired Wilcoxon signed rank tests (below the diagonal). Mean values of NMAE (the last column) together with p-values for the paired Wilcoxon signed rank tests (above the diagonal).

Model	HR mean	1	2	3	4	5	6	WHR mean
1: GARCH(1,1)	0.708	-	0.877	0.059	0.616	0.109	0.398	0.749
2: RMDN(1)	0.711	0.379	-	0.227	0.638	0.149	0.122	0.747
3: GARCH(1,1)-t	0.714	0.013	0.345	-	0.879	0.528	0.936	0.755
4: RMDN(1)-t	0.714	0.240	0.382	0.925	-	0.527	0.469	0.746
5: LRMDN(2)	0.717	0.006	0.088	0.093	0.486	-	0.940	0.757
6: RMDN(2)	0.718	0.072	0.112	0.298	0.435	0.730	-	0.756

Table 6: FTSE 100 daily returns: HR and WHR statistics. Mean values of HR (second column) together with p-values for the paired Wilcoxon signed rank tests (below the diagonal). Mean values of WHR (the last column) together with p-values for the paired Wilcoxon signed rank tests (above the diagonal).

Model	Mean	1	2	3	4	5	6
1: GARCH(1,1)	1.598	-	0.643	0.000	0.001	0.002	0.011
2: RMDN(1)	1.597	0.367	-	0.001	0.002	0.004	0.022
3: GARCH(1,1)-t	1.557	0.000	0.001	-	0.531	0.207	0.058
4: RMDN(1)-t	1.559	0.002	0.003	0.840	-	0.421	0.124
5: LRMDN(2)	1.565	0.004	0.007	0.253	0.253	-	0.300
6: RMDN(2)	1.571	0.016	0.021	0.174	0.109	0.242	-

Table 7: NIKKEI 225 daily returns: Loss function statistics. Mean values (second column), p-values for the paired t-tests (above the diagonal) and p-values for the paired Wilcoxon signed rank tests (below the diagonal).

Model	NMSE mean	1	2	3	4	5	6	NMAE mean
1: GARCH(1,1)	0.829	-	0.397	0.002	0.023	0.065	0.778	0.947
2: RMDN(1)	0.855	0.840	-	0.074	0.069	0.211	0.353	0.942
3: GARCH(1,1)-t	0.793	0.048	0.264	-	0.382	0.861	0.009	0.891
4: RMDN(1)-t	0.789	0.061	0.412	0.397	-	0.192	0.017	0.886
5: LRMDN(2)	0.797	0.061	0.221	0.493	0.174	-	0.061	0.894
6: RMDN(2)	0.810	0.115	0.353	0.098	0.051	0.192	-	0.936

Table 8: NIKKEI 225 daily returns: NMSE and NMAE statistics. Mean values of NMSE (second column) together with p-values for the paired Wilcoxon signed rank tests (below the diagonal). Mean values of NMAE (the last column) together with p-values for the paired Wilcoxon signed rank tests (above the diagonal).

Model	HR mean	1	2	3	4	5	6	WHR mean
1: GARCH(1,1)	0.661	-	0.150	0.123	0.394	0.153	0.376	0.621
2: RMDN(1)	0.665	0.597	-	0.019	0.081	0.029	0.109	0.596
3: GARCH(1,1)-t	0.680	0.014	0.033	-	0.861	0.445	0.201	0.671
4: RMDN(1)-t	0.684	0.020	0.030	0.352	-	0.677	0.076	0.660
5: LRMDN(2)	0.680	0.040	0.030	0.896	0.408	-	0.476	0.658
6: RMDN(2)	0.670	0.306	0.681	0.126	0.091	0.171	-	0.638

Table 9: NIKKEI 225 daily returns: HR and WHR statistics. Mean values of HR (second column) together with p-values for the paired Wilcoxon signed rank tests (below the diagonal). Mean values of WHR (the last column) together with p-values for the paired Wilcoxon signed rank tests (above the diagonal).