Fat Tails and Non-linearity in Volatility Models: What is more important?

Christian Schittenkopf
Austrian Research Institute for Artificial Intelligence
chris@ai.univie.ac.at

Georg Dorffner
Dept. of Medical Cybernetics and Artificial Intelligence,
University of Vienna
georg@ai.univie.ac.at

Engelbert J. Dockner
Dept. of Business Administration, University of Vienna
dockner@finance2.bwl.univie.ac.at

1 Overview

Since the seminal works of Engle [7] and Bollerslev [3] about heteroskedastic return series models, many extensions of their (G)ARCH models have been proposed in the literature. In particular, the functional dependence of conditional variances and the shape of the conditional distribution of returns have been varied in several ways (see [1] and [5] for an extensive overview).

These two issues have been addressed by the neural network community using multi-layer perceptrons (MLPs) and mixture density networks (MDNs) (see, e.g., [6, 8, 10]). In this paper we extend the concept of MDNs in a *recurrent* way to allow for "GARCH effects". These recurrent MDNs (RMDNs) offer a consistent framework to analyze the impact of non-linearity and of non-gaussian (leptokurtic) conditional distributions on the explanatory power of volatility models. We present numerical experiments on a very large return data set the size of which allows to perform detailed statistical tests to compare the obtained results.

In summary, conditional non-gaussian distributions (fat tails in the conditional distributions) tend to be more important than non-linear specifications for conditional means and variances *in the likelihood framework*. With respect to other error measures however, the application of non-linear neural networks seems to be promising. We think that the choice of a particular model for predicting volatility

is closely related to the question of how to measure the prediction performance of a model.

2 Models for return series

The standard setup for modeling return series is to split the return into a deterministic (predictable) component μ_t and a random component ϵ_t which is assumed to be white noise of time-dependent variance σ_t^2: $r_t = \mu_t + \epsilon_t$. Time series of this type are 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. $\sigma_t^2 = f(\epsilon_{t-1}, \epsilon_{t-2}, \ldots)$. At this point it is the expertise and intuition of the model builder which drives the selection of a reasonable function f and of an appropriate probability density function (pdf) for ϵ_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[1] [3] where it is assumed that the random shocks are normally distributed with mean 0 and a variance which follows an autoregressive process:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2, \alpha_0 > 0, \alpha_i \geq 0, \beta_i \geq 0. \qquad (1)$$

σ_t^2 is thus a *linear* function of all previous squared random components $\epsilon_{t-i}^2, i \geq 1$. We remark that the skewness (0) and kurtosis (3) of the conditional distribution are not time-dependent. The conditional expectation is usually assumed to be constant or to be a linear function of the last return: $\mu_t = ar_{t-1} + b$. In order to allow for leptokurtic conditional distributions Bollerslev [4] later proposed to substitute the normal distribution by a t-distribution with ν degrees of freedom (GARCH-t models). We remark that the skewness (0) and the kurtosis ($3(\nu-2)/(\nu-4)$) of the conditional distribution are (still) not time-dependent.

Over the last decade the modeling paradigm of artificial neural networks has been applied in a variety of fields including econometrics and finance [9]. In this context MDNs have been proposed to allow for heteroskedasticity [8, 10] in return series models in a semi-non-parametric way. MDNs are able to approximate arbitrary non-gaussian, even multimodal pdfs [2]. Thereby the main idea is to use MLPs[2]) to predict the parameters of the conditional pdf of the next return in dependence of the previous returns. In this paper we extend the MDN architecture

[1] In many applications it is sufficient to set $p = q = 1$ which we will adopt in the following.

[2] The standard functionality of an MLP with H hidden neurons and input $(u_1, \ldots, u_p)$ is given by MLP$(u_1, \ldots, u_p) = \sum_{j=1}^{H} v_j h\left(\sum_{k=1}^{p} w_{jk} u_k + c_j\right) + b$ with $h(x) = \tanh(x)$.

in a *recurrent* way to take into account the previous conditional variances as in the GARCH framework. A recurrent MDN with n gaussian pdfs (RMDN(n)) is defined by

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

where I_{t-1} denotes the conditioning information set (the information available at time $t-1$) and $k(\mu_{i,t}, \sigma_{i,t}^2)$ denotes a gaussian pdf with mean $\mu_{i,t}$ and variance $\sigma_{i,t}^2$. The parameters $\alpha_{i,t}$, $\mu_{i,t}$, and $\sigma_{i,t}^2$ are the priors, centres, and widths of the mixture distribution in Eq. (2). They are estimated by separate MLPs with n outputs:

$$\alpha_{i,t} = s(\tilde{\alpha}_{i,t}) = \frac{\exp(\tilde{\alpha}_{i,t})}{\sum_{j=1}^{n} \exp(\tilde{\alpha}_{j,t})}, \tilde{\alpha}_{i,t} = \text{MLP1}_i(r_{t-1}), \tag{3}$$

$$\mu_{i,t} = \text{MLP2}_i(r_{t-1}), \tag{4}$$

$$\sigma_{i,t}^2 = |\tilde{\sigma}_{i,t}^2|, \tilde{\sigma}_{i,t}^2 = \text{MLP3}_i(\epsilon_{t-1}^2, \sigma_{1,t-1}^2, \ldots, \sigma_{n,t-1}^2). \tag{5}$$

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. (2) a pdf. An RMDN with two gaussian components ($n = 2$) and three hidden neurons (for each MLP) is depicted in Fig. 1. In fact, only the MLP estimating the conditional variances is recurrent (Eq. (5)). We remark that the mean, variance, skewness, and kurtosis of the conditional pdf in Eq. (2) are (in contrast to the GARCH and GARCH-t models) *all* non-linear, time-dependent functions (for $n \geq 2$). An RMDN with one gaussian component ($n = 1$) can be interpreted as a non-linear extension of a GARCH(1,1) model.

There are two other models which are discussed in this paper in order to evaluate the influence of (non-)linear functions and (non-)gaussian pdfs on the performance of the described models in detail. First, we examine non-linear GARCH-t models. This can be done in the framework of RMDNs ($n = 1$) by replacing the gaussian pdf in Eq. (2) by the density of the t-distribution (RMDN(1)-t models). Secondly, it is interesting to study the performance of RMDNs for the case that only linear functions (instead of the MLPs[3]) are allowed (LRMDN(n) models, $n \geq 2$). We remark that some classes of models include other classes: A GARCH(1,1) model, for instance, is a GARCH(1,1)-t model with an infinite number of degrees of freedom ($\nu \to \infty$). An LRMDN(n) model, $n \geq 2$, is a multimodal extension of a GARCH(1,1) model, and the RMDN(n) models are trivially included in the class of RMDN($n+1$) models. Finally, the class of RMDN(1)-t models includes the RMDN(1) models (again for $\nu \to \infty$).

[3]This can be easily implemented by specifying $h(x) = x$.

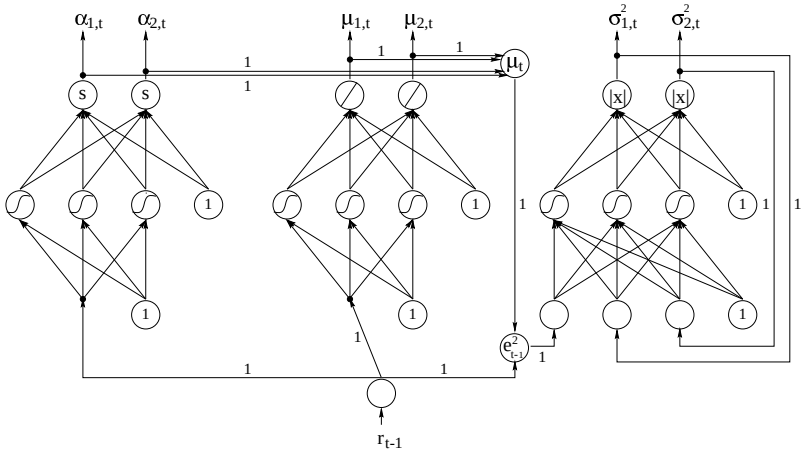


Figure 1: A recursive mixture density network with two gaussian components.

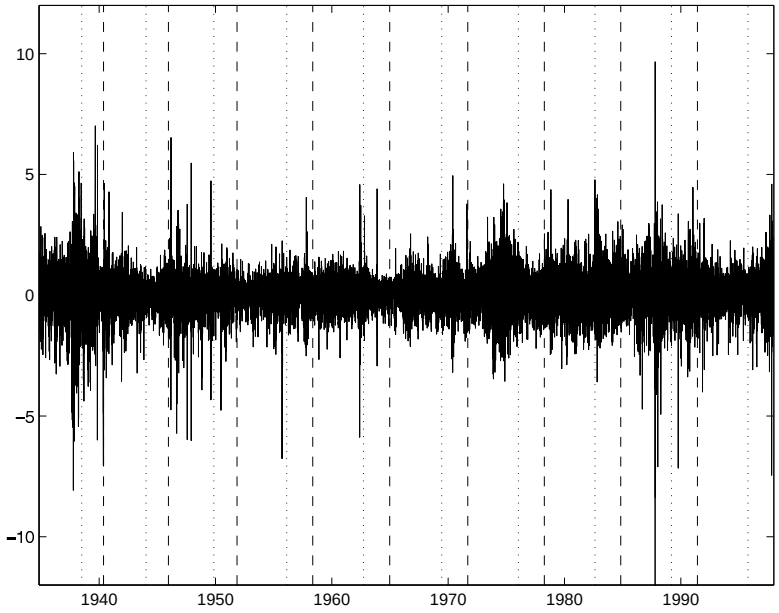
3 Empirical analysis

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 were transformed into returns r_t (in percent) by applying the transformation $r_t = 100 \log(s_t/s_{t-1})$. The resulting set of 16630 returns, which is displayed in Fig. 2, was divided into 10 training sets of length 1100 with subsequent test sets of length 563.

The parameters of all models were optimized with respect to the average negative loglikelihood which is, apart from some initial condition, given by

$$\bar{\mathcal{L}} = -\frac{1}{N} \sum_{t=1}^{N} \log \rho(r_t | I_{t-1}) \tag{6}$$

where $\rho(r_t|I_{t-1})$ denotes the conditional pdf of the corresponding model. In the following, we use the term loss function rather than likelihood because $\bar{\mathcal{L}}$ can also be calculated on a test set. 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. The optimization routine was a scaled conjugate gradient algorithm. Besides the loss function, the models were also evaluated with respect to the normalized mean absolute error NMAE and the hit rate HR. For both error measures, it is assumed



$$= \frac{\sum_{t=1}^{N} |\hat{\sigma}_t^2 - r_t^2|}{\sum_{t=1}^{N} |r_{t-1}^2 - r_t^2|} \tag{7}$$

where $\hat{\sigma}_t^2$ denotes the estimated conditional variance[4]. The NMAE is a robust error measure, and it relates the mean absolute error of the modeled volatility $\hat{\sigma}_t^2$ to the mean absolute 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 ($0 \leq$ NMAE < 1). The HR is the percentage of correctly predicted directions of change (up or down) of volatility, i.e.

$$\text{HR } = \frac{1}{N} \sum_{t=1}^{N} \theta_t, \theta_t = \begin{cases} 1 & : \quad (\hat{\sigma}_t^2 - r_{t-1}^2)(r_t^2 - r_{t-1}^2) \geq 0 \\ 0 & : \quad \text{else} \end{cases} \tag{8}$$

[4]For the (L)RMDN(2) models, the accumulated conditional variance is inserted.

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 1s (ups) and 0s (downs).

In order to compare the in-sample and out-of-sample results obtained for the six models, we tested the hypothesis of higher/lower errors by performing parametric and non-parametric tests. More precisely, we performed a paired t-test and a matched pairs signed rank Wilcoxon test (paired Wilcoxon test) for the three error measures 'Loss', NMAE, and HR for all possible pairs of models (3 times 15 paired t-tests/Wilcoxon tests). In our context, the application of *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.

3.1 In-sample results

In Table 1 we report the p-values of the paired t-tests and the R-values[5] 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) models are significantly worse than the LRMDN(2) models for the paired t-test ($p = 0.016$) and the paired Wilcoxon test ($R = 0$).

The results of the parametric and the non-parametric tests are the same in the sense that the models with a non-gaussian conditional distribution, i.e. the models 3–6, are 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) models and lower errors than the RMDN(2) models. We remark, however, that on three sets the RMDN(2) model achieves the lowest value of the loss function. Interestingly, the linear models (models 1, 3, and 5) are on average slightly better than their non-linear counterparts.

Table 2 summarizes the results of the tests for the NMAE. The best models are the RMDN(2) models which are significantly better than the GARCH(1,1) models, the RMDN(1) models, and the LRMDN(2) models. Furthermore, they tend to be better than the GARCH(1,1)-t models (significance is only obtained for the paired

[5]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 1: Mean values of the loss function and p-values (R-values) for the paired t-tests (Wilcoxon tests) above (below) the diagonal.

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 2: Mean values of the NMAE and p-values (R-values) for the paired t-tests (Wilcoxon tests) above (below) the diagonal.

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 3). Note that a higher hit rate corresponds to a better performance (on average). Both models are, together with the GARCH(1,1) models, significantly better than the LRMDN(2) models.

The in-sample results of the different volatility models 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 but do not seem to have more explanatory power than linear

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 3: Mean values of the HR and p-values (R-values) for the paired t-tests (Wilcoxon tests) above (below) the diagonal.

models 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. In particular, significant differences are obtained for the NMAE.

3.2 Out-of-sample results

Thus far, the various volatility models have been only compared on data sets from which the model parameters had been estimated. The true test for a volatility model however, is to predict volatilities out-of-sample, i.e. for a set of returns disjoint of the training set. Only the out-of-sample performance[6] provides the basis for a comparison of the various models where issues such as possible overparametrizations may be neglected.

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 4 summarizes the results for the loss function. The models with non-gaussian conditional distributions are better than the gaussian models as in the in-sample analysis. The best models are the GARCH(1,1)-t models which are significantly better than all the other models except the RMDN(2) models for which they tend to be better.

The RMDN(2) models have, on average, the best performance with respect to the NMAE measure (see Table 5). They 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

[6] besides the application of information criteria

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 4: Mean values of the loss function and p-values (R-values) for the paired t-tests (Wilcoxon tests) above (below) the diagonal.

RMDN(2) models also tend to be better than the GARCH(1,1) models. The dominance of the RMDN(2) models concerning the NMAE is thus confirmed on the test sets.

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 5: Mean values of the NMAE and p-values (R-values) for the paired t-tests (Wilcoxon tests) above (below) the diagonal.

The results for the HR are summarized in Table 6. Most p-values and R-values for this measure are such that the differences between the models are not significant. On average, the RMDN(2) models show the best performance.

Suming 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. Among the former, the GARCH(1,1)-t models are still the best models. 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

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 6: Mean values of the HR and p-values (R-values) for the paired t-tests (Wilcoxon tests) above (below) the diagonal.

significantly better than the other models. On average, the RMDN(2) models also show the best performance with respect to the HR.

The evaluation of the various volatility models with respect to the hit rate is a first step towards the implementation of real trading strategies using option prices from the market. One possible strategy is to sell/buy a straddle whenever volatility decreases/increases. We will report on these issues in much more detail in a forthcoming paper.

Acknowledgements

The models were implemented by extending the NETLAB neural network software (`http://neural-server.aston.ac.uk/`). 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.

References

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

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] Refenes, A.P., 1995, Neural networks in the capital markets (Wiley, New York).

[10] 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.