Report Series

Volatility Prediction with Mixture Density Networks

Christian Schittenkopf
Georg Dorffner
Engelbert J. Dockner

Report No. 15
May 1998

Report Series

May 1998

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 paper was accepted for publication in:
Proceedings of the International Conference on
Artificial Neural Networks, Sept.2-4 1998, Skövde, Sweden.

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').

Volatility Prediction with Mixture Density Networks

Christian Schittenkopf, Georg Dorffner
Austrian Research Institute for Artificial Intelligence
Dept. of Medical Cybernetics and Artificial Intelligence,
University of Vienna, Austria
chris@ai.univie.ac.at, georg@ai.univie.ac.at

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

Abstract

Despite the lack of a precise definition of volatility in finance, the estimation of volatility and its prediction is an important problem. In this paper we compare the performance of standard volatility models and the performance of a class of neural models, i.e. mixture density networks (MDNs). First experimental results indicate the importance of long-term memory of the models as well as the benefit of using non-gaussian probability densities for practical applications.

1 Introduction

Stock market returns typically exhibit the following time series characteristics. While the returns are uncorrelated, the squared returns show a rich structure that can be approximated by linear and non-linear models. Especially the appearance of volatility clustering renders the assumption of a constant variance (homoscedasticity) doubtful. This assumption is usually made when feedforward networks are trained to fit a given time series by gradient descent on the standard error function (mean squared error).

In this paper we apply the concept of mixture density networks (MDNs) [1] to estimate and predict the volatility of the Austrian stock market index ATX. Consequently, our neural models are *heteroscedastic*. Furthermore, the multidimensional networks (several gaussian distributions in the output) are able to also approximate non-gaussian, typically leptocurtic (fat tailed) distributions. We measure the performance of the MDNs and of some standard models of volatility with respect to the likelihood function evaluated on test sets and with respect to a prediction error of change of volatility. We find that small errors on the test sets do not necessarily imply good predictions and a profitable application of volatility models in terms of trading strategies.

Standard models for volatility estimation are briefly described in Section 2. The architecture and training of MDNs is summarized in Section 3. Section 4 includes our preliminary results on the ATX. We discuss planned and partially implemented extensions of our MDN architecture in Section 5.

2 Classical Models

Basic to standard models and our models of volatility is the notion that the financial time series $\{x_t\}$ under study can be decomposed into a predictable component μ_t and an unpredictable component e_t, which is assumed to be zero mean gaussian noise of variance σ_t^2: $x_t = \mu_t + e_t$. The models are thus characterized by *time-varying* conditional variances σ_t^2 and are thus well suited to explain volatility clusters. The most widely used (standard) models of volatility are ARCH/GARCH models and the GJR approach [2, 3, 4]. Financial time series often exhibit means close to zero and negligibly small correlations. In these cases the corresponding models can be forced to predict a conditional mean $\mu_t = 0$. If there are reasons to believe that the conditional mean is significantly different from zero, an extra component for μ_t should be provided in the model. The classical ARCH(q) model [2] is given by

$$x_t \sim N(\mu_t; \sigma_t^2), \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i e_{t-i}^2, \tag{1}$$

where $N(\mu_t; \sigma_t^2)$ denotes a Gaussian random variable of mean μ_t and of variance σ_t^2. To ensure that the variance σ_t^2 is positive for each t the restrictions $\alpha_0 > 0, \alpha_i \geq 0, i = 1, \ldots, q$, are imposed on the parameters. A GARCH(p, q) model [3] is an extension of an ARCH(q) model because the variance is calculated recursively by

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

Additionally to the constraints of the ARCH model we require that $\beta_i \geq 0$, $i = 1, \ldots, p$. This specification implies that the conditional variance σ_t^2 follows an autoregressive process for which stationarity is guaranteed, if the sum of coefficients $\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1$. In many applications it is sufficient to choose $p = q = 1$. Finally, the GJR model [4], which is an extension of the GARCH(1,1) model, has been successfully applied to financial time series. It incorporates asymmetric effects, and it is defined by

$$\sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \alpha_2 s_{t-1} e_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{3}$$

where $s_{t-1} = 1$ if $e_{t-1} < 0$ and $s_{t-1} = 0$ otherwise. The use of the GJR model has been proposed because stock returns are characterized by a leverage effect, i.e. volatility increases as returns for stocks decrease.

3 Network Architecture

Within the last years MDNs [1, 5] have turned out to be a very useful tool to model conditional probability density functions (pdfs) in different fields such

as non-linear inverse problems [1] and time series analysis [6]. The main idea of MDNs is to use multi-layer perceptrons (MLPs) to predict the parameters of the pdf of the next observation x_t in dependence of the past observations $x_{t-1}, \ldots, x_{t-m}$. A very natural way to approximate the conditional pdf of x_t is to choose a weighted sum of n gaussian densities, i.e.

$$x_t \sim \sum_{i=1}^{n} \alpha_{i,t} N(\mu_{i,t}; \sigma_{i,t}^2), \tag{4}$$

$$\alpha_{i,t} = s(\tilde{\alpha}_{i,t}), \tilde{\alpha}_{i,t} = \mathrm{MLP}_j(x_{t-1}, \ldots, x_{t-m}), 1 \leq j \leq n, \tag{5}$$

$$\mu_{i,t} = \mathrm{MLP}_j(x_{t-1}, \ldots, x_{t-m}), n+1 \leq j \leq 2n, \tag{6}$$

$$\sigma_{i,t}^2 = \exp(\mathrm{MLP}_j(x_{t-1}, \ldots, x_{t-m})), 2n+1 \leq j \leq 3n. \tag{7}$$

The softmax function

$$s(\tilde{\alpha}_{i,t}) = \frac{\exp(\tilde{\alpha}_{i,t})}{\sum_{j=1}^{n} \exp(\tilde{\alpha}_{j,t})} \tag{8}$$

ensures that the priors $\alpha_{i,t}$ are positive and that they sum up to one, which makes the right hand side of Eq. (4) a pdf. The exponential function in Eq. (7) guarantees positive conditional variances. As a result the MDN receives the m-dimensional input $x_{t-1}, \ldots, x_{t-m}$ and produces a $3n$-dimensional output. The first n components MLP_j, $1 \leq j \leq n$, are used to calculate the priors, the outputs MLP_j, $n+1 \leq j \leq 2n$, are the conditional means, and the components MLP_j, $2n+1 \leq j \leq 3n$, are squared to estimate the conditional variances. The parameters of the MDN (the MLP) and of the models of Section 2 are updated according to scaled gradient descent on the negative logarithm of the likelihood function [1]. To test the performance of the models on independent test sets, the same function applied to the test data can be used as a loss or generalized error function.

4 Experimental Results

The time series $\{x_t\}$ of the Austrian stock market index ATX from 7 January 1986 until 14 June 1996 (2575 measurements) was preprocessed using the transformation $r_t = 100(\log x_{t+1} - \log x_t)$. The resulting time series of returns r_t and the autocorrelation functions of r_t and r_t^2 are depicted in Fig. 1. There is an obvious change in structure in the time series at time $t \approx 950$ when the trading conditions at the stock exchange in Vienna were changed. Several volatility clusters (accumulations of large positive and negative returns) are clearly visible. The two horizontal lines on the right hand side of Fig. 1 indicate the 95% confidence interval for an identically and independently distributed (i.i.d.) process (white noise). Consequently, only the first autocorrelation of r_t should be assumed to be statistically significant. The squared returns r_t^2, however, show a very regular structure which is significant for all lags k ($1 \leq k \leq 25$). The quasiperiodicity of period five might indicate that the volatilities of identical days of the week are particularly correlated.

First, we fitted an ARCH(1), a GARCH(1,1) and a GJR model to the time series of returns r_t. Due to the correlation analysis the mean component μ_t was modelled by an autoregressive process of first order, i.e. $\mu_t = a x_{t-1}$. In order to

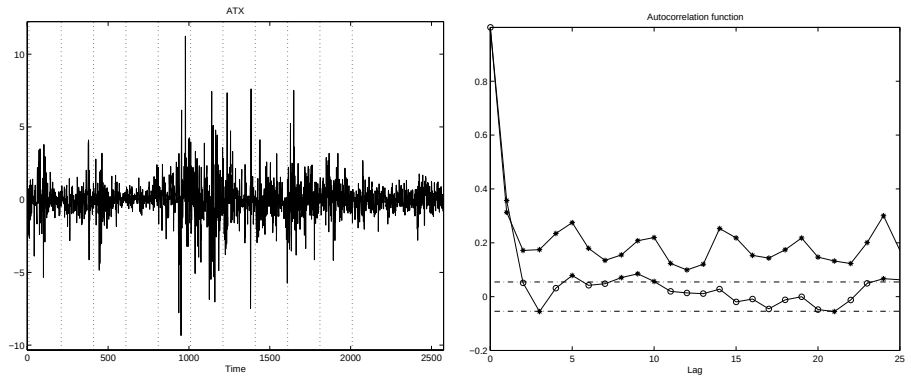

Figure 1: (Left) The returns r_t of the ATX from 7 January 1986 until 14 June 1996 and the division into training and test sets (dotted lines). (Right) The autocorrelation function of r_t (lower curve) and r_t^2 (upper curve) and the 95% confidence intervals for white noise.

evaluate the performance of the models we used the concept of cross validation. More precisely, the time series was divided into ten subsequent intervals of equal size: $I_1 = (r_{11}, \ldots, r_{210}), \ldots, I_{10} = (r_{1811}, \ldots, r_{2010})$. The rest of the data $T = (r_{2011}, \ldots, r_{2575})$ was used as an independent test set (see Fig. 1). Then each model was trained on nine of these ten intervals and the error E_j on the missing interval I_j was calculated ($1 \leq j \leq 10$). Additionally, each model was trained on the whole training data set $I = (r_{11}, \ldots, r_{2010})$ and evaluated on the test set T. The first set I_1 starts with r_{11} since we wanted to present the same training sets to the models (of different order m).

The mean value and the standard deviation of the errors E_j are summarized in Table 1. We see that the GARCH(1,1) model has the best performance of the standard models. Table 1 also gives the results for the trained MDNs. A network with two inputs (x_{t-1}, x_{t-2}), three hidden neurons and two gaussian distributions is denoted MDN(2-3-2), for instance. The best network is MDN(2-3-2) which is also better than the best standard model GARCH(1,1). The performance of the largest network MDN (5-4-3) is slightly worse which could be the result of insufficient training. We emphasize that the standard deviation is very large for all models (in comparison to the mean) because of the change in structure at $t \approx 950$. In fact, there are subintervals I_j which can be easily modelled ($j = 4$, for instance), whereas some periods are characterized by large returns which are hard to predict ($j = 5$, for instance). If the models are trained on the whole training data set I and tested on the independent set T, the GARCH(1,1) model performs best.

Another test for the quality of volatility forecasts is the analysis of the profitability of trading strategies based on the predicted volatilities. More precisely, the volatility forecast based on historical returns gives us the information if the volatility is going to increase or decrease in the next period. This information can be interpreted as a buying or selling signal for a straddle [7]. If the predicted volatility is lower than the current one (volatility decreases) we go short, and if the volatility increases we take a long position. Therefore the quality of a volatil-

Model	Parameters	mean	std.	T	correct
ARCH(1)	3	1.617	0.415	1.138	53.9%
GARCH(1,1)	4	1.505	0.411	0.925	67.6%
GJR	5	1.514	0.404	0.934	65.4%
MDN(1-3-1)	18	1.601	0.375	1.157	51.6%
MDN(2-3-2)	33	1.448	0.368	0.994	52.0%
MDN(5-4-3)	69	1.458	0.414	1.002	56.6%

Table 1: Overview of models fitted to the Austrian stock market index ATX.

ity model can be measured by the percentage of correctly predicted directions of change of volatility from this period to the next (increase or decrease).

The first part of the ATX data set, i.e. I, was used to train the classical and neural models which were evaluated on the independent test set T afterwards. The performance of the models concerning the correctly predicted directions of change of volatility is summarized in the last column of Table 1 (the concrete implementation of trading strategies is planned for the future). Strictly speaking, the squared returns r_t^2 are considered the "true" volatility and compared to the forecasted volatility σ_t^2. A prediction is thus classified as correct if and only if $(\sigma_t^2 - r_{t-1}^2)(r_t^2 - r_{t-1}^2) > 0$. For the MDNs with several gaussian distributions the "accumulated" variance of the distribution [1] is used. The best model is again the GARCH(1,1) model with impressive 67.6%. From Table 1 we also learn that the predictive quality of the MDNs increases with the number of inputs (past values).

5 Discussion and Conclusion

These results indicate the importance of long-term memory of the models if they are implemented in trading strategies. For the GARCH(1,1) model the conditional variance σ_t^2 is heavily influenced by the previous conditional variances $\sigma_{t-1}^2, \ldots$ owing to the parameter $\beta_1 \approx 0.918$ (for the training set I). A promising idea is thus to include *recurrent structures* into the MDNs. Our new architecture, which is currently investigated, consists of three MLPs which estimate the priors, the means and the variances separately. Following the GARCH(1,1) specification the MLP estimating the variance σ_t^2 receives a two-dimensional input: the squared error e_{t-1}^2 and the previous variance σ_{t-1}^2. We think that a comparison of the performance of standard and neural models is only fair if this extended MDN architecture is considered.

Furthermore, the implementation and evaluation of different trading strategies might provide further valuable insights into the behavior and predictive power of standard and neural volatility models.

Acknowledgements

The MDNs were implemented using the NETLAB neural network software written by I. Nabney and C. Bishop (`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 A. Weingessel and F. Leisch for valuable discussions.

References

[1] Bishop CM. Mixture density networks, Neural Computing Research Group Report: NCRG/94/004, Aston University, Birmingham, 1994

[2] Engle RF. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica 1982; 50:987-1008

[3] Bollerslev T. A generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 1986; 31:307-327

[4] Glosten LR, Jagannathan R., Runkle DE. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 1993; 48:1779-1801

[5] Neuneier R, Finnoff W, Hergert F, Ormoneit D. Estimation of conditional densities: a comparison of neural network approaches. In: Marinaro M, Morasso PG (ed) ICANN 94 - Proceedings of the International Conference on Artificial Neural Networks. Springer-Verlag, Berlin, 1994, pp 689-692

[6] Schittenkopf C, Deco G. Testing nonlinear Markovian hypotheses in dynamical systems. Physica D 1997; 104:61-74

[7] Noh J, Engle RF, Kane A. Forecasting volatility and option prices of the S & P 500 index. Journal of Derivatives 1994; 17-30