Modelling Volatility with Mixture Density Networks

Fahed Mostafa
DEBII, Curtin University
fahed.mostafa@postgrad.curtin.edu.au

Tharam Dillon
DEBII, Curtin University
Tharam.Dillon@cbs.curtin.edu.au

Abstract

Volatility is an important variable in financial forecasting. Forecasting volatility requires a development of a suitable model for it. In this paper, we examine different time series models for volatility modelling. Specifically, we will study the use of recurrent mixture density networks, GARCH and EGARCH models to model volatility. In addition, we demonstrate the impact of different factors on the accuracy and completeness of each of these models.

1. Introduction

Volatility modelling and forecasting are of vital interest in the field of risk management and option pricing. Investors are interested in understanding and predicting the movements of the underlying instruments. The probability of such movement is depicted as a volatility variable. Predicting volatility becomes very important in forecasting option prices, and it is an essential variable in calculating the market risk of any portfolio. Since the introduction of the Arch [13] model by Engle and the GARCH model [7], there has been a great deal of interest in this field. This is due to the models ability to forecast market volatility. The ability of the GARCH models to capture the stylised facts of asset returns, such as heteroskedasticity, excess kurtosis, and volatility clustering, has contributed to the enhancements of such models. Some researchers have showed many more empirical irregularities existed in the return series that the GARCH models fail to capture, such as leverage effect and co-movement of volatilities. Many models have evolved to capture these stylised effects, such as the EGARCH Model [21].

Neural networks has been successfully applied to time series prediction. Bishop [5], introduced a new class of neural networks called Mixture Density Networks, in which he combines a neural network with a mixture density model. The MDN was found to do much better than the conventional neural network in modelling the underlying process. The MDN was extended to be recurrent hence mimicking the GARCH models. The recurrent MDN was found to be a very useful tool for forecasting volatility [25], [26]. In this paper an emphasis is placed on the issues surrounding wrong model selection and the length of the time series used in the model training (in-sample data set). To demonstrate these issues, data from four different time series were studied. The performances of all models are compared by using 1000 and 750 returns. The structure of the MDN was tested by varying the number of hidden units and Gaussians. The results obtained demonstrate the effect of the in-sample length on the model performance. That is the same model preformed differently when trained on the same time series with different in-sample lengths. This was also applied to GARCH models.

2. The GARCH Model

Let's assume that the return process r_t is generated by equation (1).

$$r_t = X_t'\xi + \varepsilon_t, t = 1,.., T \qquad (1)$$
$$\varepsilon_t \mid \psi_{t-1} \sim N(0, \delta_t)$$

where X_t is a $k \times 1$ vector of exogenous variables ,ξ is a $k \times 1$ vector of regression parameters. Bollerslev [7] formulated the GARCH model by generalised the ARCH [13]. The GARCH model suggests that the conditional variance can be specified as

$$\delta_t^2 = \alpha_0 + \alpha(B)\varepsilon_t^2 + \beta(B)\delta_{t-1}^2 \text{ ,where } \begin{matrix}\alpha_0 > 0 \\ \alpha_i \geq 0 \\ \beta_i \geq 0\end{matrix} \qquad (2)$$

The inequalities are imposed to ensure the conditional variance is positive. A GARCH process with order p and q is denoted by GARCH(p,q). The GARCH model is considered as a generalisation of an ARCH(∞) process, since the conditional variance depends linearly on all previous squared residuals.

3. The EGARCH

The GARCH model does well in capturing the thick tailed returns and volatility clustering. Although the GARCH model is very successful at this, it is not

well suited to capture leverage effect, since the variance equation is a function of the magnitudes of the lagged residuals and not their signs. Nelson [21] first proposed Exponential GARCH (EGARCH). The EGARCH model was formulated with the variance equation that depends on the sign and size of the lagged residual. Hence, capturing the leverage asymmetric effects. The presence of leverage effects can be tested by the hypothesis that $\gamma > 0$ and the impact is asymmetric if $\gamma \neq 0$.

$$\ln(\delta_t^2) = \alpha_0 + \sum_{i=1}^{p} \beta_i \ln(\delta_{t-1}^2) + \sum_{j=1}^{q} \left(\alpha_j \left| \frac{\varepsilon_{t-j}}{\delta_{t-j}} - \sqrt{\frac{2}{\pi}} \right| + \gamma_j \frac{\varepsilon_{t-j}}{\delta_{t-j}} \right) \qquad (3)$$

4. Recurrent Mixture Density Network

Given the recurrent nature of the GARCH model, the mixture density network Bishop [5] is extended to include the lagged variance as input parameter. As noted in [26] if the input to the MDN is extended to include lagged values of the variance δ^2_{t-1}, the networks become a generalisation of a GARCH model. With the generalisation capability of the MDN to approximating the distribution of the underlying data, the MDN should yield better forecasts as observed by [22],[25] and [26]. The recurrent mixture density network (rMDN) can approximate the distribution of the underlying data, hence overcoming the assumption of normality in the GARCH Model.

To demonstrate the importance of selecting the right model for a given time series, different rMDN structures is investigated. The rMDN structures were varied by changing the hidden units and the number of Gaussians. As shown in figure 1, the conditional variance of the previous period δ^2_{t-1} is fed back as an input at time t which is in-line with the GARCH model. The calculation on the conditional mean and conditional variance are as follows;

$$\mu_{t+1} = \sum_{i=1}^{n} \alpha_{i,t+1} \mu_{i,t+1} \qquad (4)$$

$$\delta^2_{t+1} = \sum_{i=1}^{n} \alpha_{i,t+1} (\delta^2_{i,t+1} + (\mu_{i,t+1} - \mu_{t+1})^2) \qquad (5)$$

$$L = \prod_{t=1}^{T} \sum_{j=1}^{M} p_j(x_t) \frac{1}{\sqrt{2\pi\delta_j^2(x_t)}} \exp\left(\frac{-(y_t - \mu_j(x_t))^2}{2\delta_j^2(x_t)} \right) \qquad (6)$$

$$\ell = -\frac{1}{N} \log L \qquad (7)$$

Maximising the log likelihood function L is equivalent to minimising the average negative log likelihood function ℓ in equation (7), where ℓ is the error function used for training the MDN.

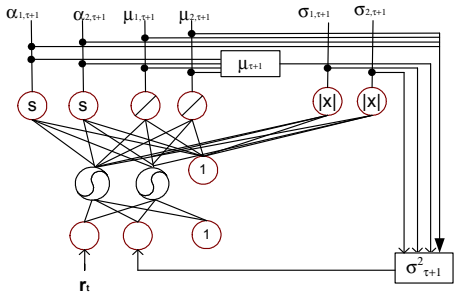

Figure 1 : rMDN with 2 Gaussian and 2 hidden units

5. Forecast Evaluation

The performance of the volatility models are measured by the loss function ℓ, normalised mean absolute error (NMAE) and Hit rate(HR). The NMAE measures the mean absolute error of the volatility model, compared to the true volatility r^2_t, which should be smaller than the naive model $\delta^2_t = r^2_t$. This measure is the least reliable out of the three measuring criteria used. Since it compares the forecasted volatility with the returns squared, this has some major drawbacks as stated previously. The hit rate (HR) is the measure of frequency of the correctly predicted increase (or decrease) in volatility. A HR value of 0.5 indicates a forecast that is no better then a random predictor of increase (or decrease) in volatility. This measure is a better indicator than the NMAE since it does not rely on the magnitude of the squared return; only that the predicted direction is taken into account.

$$NMAE = \frac{\sum_{t=1}^{N} | r^2_{t+1} - \hat{\delta}^2_{t+1} |}{\sum_{t=1}^{N} | r^2_{t+1} - r^2_t |} \qquad (8)$$

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

$$\theta_t = \begin{cases} 1 : (\delta^2_{t+1} - r^2_t)(r^2_{t+1} - r^2_t) \geq 0 \\ 0 : else \end{cases} \qquad (9)$$

The loss function ℓ is used as an error measurement for forecast performance. This measure is the most reliable, since it does not rely on the return squared. The model performance would be ranked according to the smallest value for the loss function with the highest HR and smallest NMAE respectively.

6. Data Analysis

The time series used in this research are the daily close prices for an aviation company QANTAS (QAN), a resource company BHP Billiton Limited (BHP), and two companies from the banking sector

National Australia Bank (NAB) and ANZ Banking Group Ltd (ANZ). They are all listed on the Australian Stock Exchange. The in-sample data consists of two sets, 1000 and 750 returns. The daily close prices were transformed to daily returns through equation (10)

$$r_t = \ln\left(\frac{close\_price_{t+1}}{close\_price_t}\right) \quad (10)$$

7. Experiment set-up

In this study we carry out a study of the volatility forecast performance of the rMDN, GARCH and EGARCH models. This is achieved by forecasting the volatility 1 and 10 days ahead. Each model is compared based on the error measures as explained in section 5. To optimise the performance of the rMDN, the hidden units and number of Gaussian were varied. For the one day forecast, we performed 100 one day forecasts. This is achieved by training the model using returns from period 1 to T, then forecasting volatility at time T+1. The model is retrained using returns from period 2 to T+1 and then forecasting the volatility at period T+2. Long term forecasts with the rMDN are not as easily obtained. The returns r_{t+1}, are not directly observed, hence, we need to simulate the future returns using Monte Carol simulation. For all volatility models the accumulated conditional variance for the day after tomorrow is the average conditional variance.

8. Results

The parameters for the GARCH and EGARCH models for 1000 returns are shown in tables 1-2. The parameters for 750 returns are omitted due to restrictions on the size of this paper. All GARCH models are stationary with high persistence ($\omega+\beta < 1$) for both 1000 and 750 except for BHP. The parameters for the EGARCH display asymmetric and leverage effect ($\gamma \neq 0$ and $\gamma < 0$) for all companies except for BHP, which only exhibits asymmetric effects ($\gamma \neq 0$).

Table1: average GARCH parameters for 1000 returns

	ANZ	NAB	BHP	QAN
μ	0.00129	0.0005	-7E-05	0.0006
α	0.00004	0.00002	0.0001	0.00001
β	0.197	0.053	0.261	0.082
ω	0.648	0.825	0.085	0.896

Table2: average EGARCH parameters for 1000 returns

	ANZ	NAB	BHP	QAN
μ	0.00092	0.00054	-0.0002	3E-05
ω	-1.865	-1.56775	-2.593	-0.98
α	0.303	0.117	0.092	0.229
γ	-0.116	-0.01	0.017	-0.026
β	0.84	0.86	0.764	0.919

This research places emphasis on the right model selection for a given time series. The models are compared based on 1 and 10 steps ahead forecast. The results for the best performing models for the 1 day forecast are shown in the tables 3-6. The naming convention used in the tables for the rMDN is number of Gaussian, number of hidden units then the length of the time series used.

For the 1000 and 750 returns the BHP and NAB, the EGARCH is best suited whereas the GARCH is better suited for the QAN and ANZ. The rMDN displays a different behaviour where different network structures are needed to obtain optimal performance for the same time series. This is also the case when analysing the performance with respect to different in-sample lengths. The overall performance of the rMDN is slightly better to that of the GARCH models. This highlights the importance of optimising the model with respect to the in-sample length.

Table 3: (E)GARCH 1 days results for 1000 returns

	Model	NMAE	HR	ℓ
BHP	EGARCH	0.667	0.74	2.19
ANZ	GARCH	0.796	0.73	2.578
NAB	EGARCH	0.745	0.74	2.754
QAN	GARCH	0.914	0.68	2.379

The 10 day forecast results are displayed in table 7-10. The same conclusion can be drawn with regards to the optimising the rMDN relative to network structure and in-sample length. Also rMDN outperforms the GARCH models on all time series where it has a significantly lower value for the loss function ℓ.

Table 4: (E)GARCH 1 day results for 750 returns

	Model	NMAE	HR	ℓ
BHP	EGARCH	0.847	0.65	3.007
ANZ	GARCH	0.705	0.74	2.836
NAB	EGARCH	0.705	0.74	2.48
QAN	GARCH	0.746	0.72	2.52

Table 5: rMDN 1 day results for 750 returns

	rMDN	NMAE	HR	ℓ
BHP	3G5H750	0.742	0.72	3.906
ANZ	3G5H750	0.857	0.65	1.369
NAB	3G2H750	0.722	0.73	1.34
QAN	3G5H750	0.738	0.76	1.25

Table 6: rMDN 1 day results for 1000 returns

	rMDN	NMAE	HR	ℓ
BHP	3G2H1k	0.662	0.74	2.261
ANZ	3G4H1k	0.753	0.75	1.282
NAB	3G2H1k	0.72	0.75	1.34
QAN	3G4H1k	0.991	0.64	1.342

Table 7: (E)GARCH 10 day results using 750 returns

	Model	NMAE	HR	ℓ
BHP	EGARCH	0.755	0.6	2.889
ANZ	GARCH	0.724	0.6	3.027
NAB	EGARCH	0.777	0.8	2.549
QAN	EGARCH	0.783	0.7	2.549

Table 8: (E)GARCH 10 day results using 1000 returns

	Model	NMAE	HR	ℓ
BHP	GARCH	0.96	0.6	0.727
ANZ	EGARCH	0.524	0.9	2.103
NAB	GARCH	0.606	0.8	2.529
QAN	GARCH	0.702	0.6	2.726

Table 9: rMDN 10 day results using 1000 returns

	rMDN	NMAE	HR	ℓ
BHP	2G2H1k	0.919	0.7	1.297
ANZ	2G8H1k	0.535	0.9	2.095
NAB	3G2H1k	0.571	0.8	2.051
QAN	3G3H1k	0.681	0.8	3.018

As indicated by the results, choosing the optimal training data set is crucial to the model forecasting performance. Also the structure of the rMDN including number of hidden units should be careful selected relative to time series

Table 10: rMDN 10 day results using 750 returns

	rMDN	NMAE	HR	ℓ
BHP	2G5H750	0.719	0.6	3.341
ANZ	3G3H750	0.797	0.6	3.371
NAB	2G8H750	0.774	0.8	3.633
QAN	2G5H750	0.585	0.8	3.605

9. Discussion

In time series modelling the length of the in-sample data series is often neglected and it is assumed to be of a certain length. Over fitting and under fitting of the model parameters is normally caused by such an oversight. In particular when using too few in-samples data points, the parameters of the model would not be fully optimised after the model training is completed. The instability of the model would lead to a poor forecast performance. Also when using too many in-sample data points the model becomes over trained and would lead to memorisation and over fitting issues. The degree to which over fitting and under fitting is possible is related to the number of training patterns and the number of parameters in the model. For optimal performance it is the right balance between in-sample length and number of parameters in the model. As seen in the results, NAB and QAN provide much better forecasts using 750 days on returns, whereas, ANZ and BHP provide better forecasts with 1000 days of returns. This may be due to a discrete change in the environment. This leads to the conclusion that the in-sample length should contain enough information to explain future behaviour.

Several alternative models are often proposed to explain the same data, and objective criteria are needed to choose among models. While adding extra parameters to a model is often desirable, the increased complexity comes with a cost. In general, the more parameters contained in a model, the less reliable the parameter estimates are. The criteria to select among models must weigh the trade-off between increased information and decreased reliability. This has been formulated in the famous Akaike Information Criterion (AIC). There have been many theories to determine the optimal neural network size, such as the NIC(network information criteria) [1] which is a generalisation of the AIC. In this paper the optimal model was selected based on the error measures explained in section 5. It was interesting to notice the optimal model differs among the time series. The optimal model also differed for by the forecast horizons. The rMDN could be further tuned by using pruning algorithms such as in [24] Pruning a neural network reduces the effect of spurious data, which will improve the accuracy of the forecast. The generalisation power of the rMDN and its ability to capture the underlying dynamics of the returns series such as high volatility persistence makes it a better time series model than the traditional GARCH models. As demonstrated by [10] and [18] high volatility persistence in the GARCH model could be due to structural changes in the variance process. The rMDN seem to have the capability to capture such dynamic behaviour.

10. Conclusion

This study has demonstrated the ability of the rMDN to model complex time series data. The superior performance of the rMDN relative to traditional GARCH models is attributed to the flexibility and the dynamic nature of the rMDN. However, key modelling issues have been addressed and demonstrated in this paper.

Given the success of the rMDN, there is still room for improvements, such as pruning the neural network and inclusion of other variables. The emphasis was mainly on the right model selection. The structure of the model and the length of the time series are very important issues and need to be examined extensively, to provide a reliable and stable result. This was demonstrated by varying the number of Gaussian, hidden units, and by using different in-sample lengths.

The performance of the models varied dramatically with these factors, which explains the mixed results shown in research when comparing rMDN's to other traditional models. The statistical evaluation criterion used has no economic meaning. Also they do not incorporate uncertainty, due to parameter estimations. It is recommended that the model forecast should be evaluated according to economic loss rather than statistical loss as suggested in [19].

11. References

[1] Amari, S., Learning and statistical inference, *in* M. A. Arbib, ed., 'The Handbook of Brain Theory and Neural Networks', MIT Press, Cambridge, assachusetts, (1995)

[2] T. Andersen, and T. Bollerslev, "Answering the skeptics: yes, standard volatility models do provide accurate forecasts", International Economic Review, 39, 4, 1998,pp. 885-905

[3] TG Andersen, T Bollerslev, N Meddahi, "Analytical Evaluation of Volatility Forecasts," *Working Paper*, Northwestern, Duke, and University of Montreal, (2002b)

[4] K. Bartlmae, and F. A. Rauscher, "Measuring DAX Market Risk: A Neural Network Volatility Mixture Approach", *Presentation at the FFM2000 Conference*, London, 2000.

[5]C. M, Bishop, "Mixture density networks." Neural Computing Research Croup Report NCRG/4288. Aston University, United Kingdom.1996

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

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

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

[9] 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.

[10] F. X. Diebold and P. Pauly, "Structural change and the combination of forecasts." Journal of Forecasting 6, 1987,pp 21-40.

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

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

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

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

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

[16] D.B Fogel, **"**An information criterion for optimal Neural Nnetwork selection", IEEE Transactions on Neural Networks Volume: 2, Issue: 5,1991,pp 490-497

[17] I. T. Nabney and H. W. Cheng, "Estimating conditional volatility with neural networks". 4th International Conference on Forecasting Financial Markets,1997.

[18] C.G. Lamoureux. and W.D. Lastrapes, "Persistence in variance, structural change and the GARCH model", Journal of Business and Economic Statistics, 8, 2, 1990, pp 225-234.

[19] J.A. Lopez, "Evaluating the predictive accuracy of volatility models", Journal of Forecasting, 20, 2, 2001, pp 87-109.

[20] Neuneier, R., F. Hergert, et al. "Estimation of conditional densities: A comparison of neural network approaches", Springer, 1994, pp 689-692.

[21] D. B. Nelson, "Conditional Heteroskedasticity in Asset Returns: A New Approach." Econometrica **59**(2), 1991,pp 347-370.

[22] D. Ormoneit and R. Neuneier, "Experiments in predicting the German stock index DAX with density estimating neural networks", Conference on Computational Intelligence in Financial Engineering (CIFEr 96), 1996.

[23] S. Poon and C. Granger, "Forecasting Volatility in Financial Markets: A Review ", Journal of Economic Literature Vol. 41, No. 2, 2003

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

[25] C. Schittenkopf, G. Dorner. "Volatility prediction with mixture density networks." Proceedings of the International Conference on Artificial Neural Networks, 1998.

[26] C. Schittenkopf, G. Dorffner & E. J. Dockner ``Forecasting Time-dependent Conditional Densities: A Seminonparametric Neural Network Approach", *Journal of Forecasting* 19, 2000,pp 355-374