PERGAMON Journal of the Franklin Institute 336 (1999) 227–236

Journal of The Franklin Institute

Neural network forecast combining with interaction effects

R. Glen Donaldson[a, 1], Mark Kamstra[b, *, 1]

[a] Faculty of Commerce, University of British Columbia, Vancouver, BC, Canada V6T 1Z2
[b] Department of Economics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

Abstract

In this paper we discuss and expand recent innovations in forecast combining with artificial neural networks (ANNs). In particular, we demonstrate that ANNs can outperform traditional forecast combining procedures, such as least-squares weighting, because ANNs can account for traditionally uncaptured interaction effects between time series forecasts. Data employed in this study are price volatility forecasts for the S & P 500 stock index. © 1998 The Franklin Institute. Published by Elsevier Science Ltd.

Keywords: Financial data; S & P 500; Volatility forecasting; ARCH

1. Introduction

Different forecasters typically have access to different information and therefore produce different forecasts. Given this, one would ideally combine the individual forecasters' various *information sets* to produce a single superior information set from which a single superior forecast could be produced. In practice, however, it is often possible to only obtain an individual forecaster's *forecast*. In such cases, a popular alternative is to combine in some fashion the available forecasts to produce a single combined forecast in the hope of reducing the variance of the forecasting error without inducing a bias to the forecast.

Traditional forecast combining methods produce the single superior forecast from a linear combination of the various individual forecasts, where weights on each individual forecast are chosen by a wide variety of methods including ordinary

* Corresponding author. E-mail:kamstra@kamstra.econ.sfu.ca
[1] Supported by Social Sciences and Humanities Research Council of Canada.

PII: S0016-0032(98)00018-0

least-squares, mean absolute deviations, etc.[2] A potential drawback of these traditional combining methods is their restriction to consider only linear combinations of the individual forecasts. There is in fact no guarantee that a linear combination is efficient if the individual forecasts are based on non-linear models or if the true underlying conditional expectation is a non-linear function of the information sets on which the individual forecasts are based.

For example, consider the case of a dependent variable $y = x_1 \cdot x_2 + \varepsilon$, where ε is an innovation and x_1 and x_2 are explanatory variables known to us. Suppose we do not know the true relationship between y and x_1 and x_2. If forecaster 1 has the model $f_1 = \alpha_1 x_1$ (hence ignoring, incorrectly, x_2) and forecaster 2 has the model $f_2 = \alpha_2 x_2$ (hence ignoring, incorrectly, x_1) then any linear combination of the two forecasts will be inferior to the non-linear forecast $\beta(f_1 \cdot f_2)$, with $\beta = (1/\alpha_1 \cdot \alpha_2)$. Complications such as these have become widely appreciated in modelling economic and financial data.[3]

Semiparametric artificial neural networks can be used to select optimal weights on nonlinear functions of the individual forecasts, instead of using the traditional classical parametric techniques to linearly weight individual forecasts. Since ANNs have the ability to approximate arbitrarily well a large class of functions, ANNs should provide considerable flexibility to uncover hidden relationships between a group of individual forecasts and realizations of the variable being forecasted.[4] In particular, the ANN model allows interaction effects between forecasts, which permits an ANN combination to "turn-off" one forecast in favour of the other in what could be interpreted as a state-dependent fashion (e.g. one forecast "on" during recessions, the other "on" during expansions), something which simple linear models cannot do.[5]

Work by Donaldson and Kamstra [8], and Harrald and Kamstra [19] both make use of ANNs to combine forecasts, Harrald–Kamstra being distinguished from Donaldson–Kamstra by the use of evolutionary programming to "evolve" an optimal neural network model. Here we focus on the ability of the ANN models from these papers to capture interaction effects that "turn-off" one forecast in favour of the other in a state-dependent fashion. We explain, largely with the use of figures, the reasons why ANN combining is superior to other methods and what this implies for future research and application of the combining technology. In Sections 2 and 3 we describe the various methods of combining forecasts, we compare them and present our data: forecasts of return volatility for the S & P500 stock market index. In Section 4 we discuss the criteria commonly used to compare combined forecasts and present our results. Section 5 gives our conclusions.

---

[2] See surveys [1, 2] for a review of traditional combining methods. For additional insight into recent developments in the use of Bayesian analysis to form forecast combinations, see [3].

[3] See, for example [4–12].

[4] For technical details on ANN estimation and properties see [13–18].

[5] There are modeling techniques for regime-switching, such as Markov-switching, which formally incorporate state dependencies and are also an interesting avenue of research, but will not be pursued here.

2. Combining methods

As mentioned above, traditional combining methods rest on the implicit assumption that the conditional expectation of the variable being forecasted is a linear combination of the available forecasts. Thus, when combining two individual forecasts, $f_{1,t}$ and $f_{2,t}$, a single combined forecast F_t is produced according to Eq. (1) by appropriate choice of weights β_0, β_1, and β_2.

$$F_t = \beta_0 + \beta_1 f_{1,t} + \beta_2 f_{2,t} \tag{1}$$

Perhaps the most popular combining method takes the cross-sectional average of the individual forecasts, i.e. $\beta_0 = 0$, $\beta_1 = \beta_2 = 0.5$. Another popular approach is to run a multivariate ordinary least-squares (OLS) regression of the variable being forecasted on the individual forecasts in-sample to obtain "optimal" forecast weights β_0, β_1, and β_2 for use in out-of-sample combining.[6] A potential drawback of these and other traditional methods is their restriction to consider only linear combinations of the individual forecasts; a restriction which may lead to inefficient combined forecasts. We therefore now discuss a semiparametric modelling technique, artificial neural networks (ANNs), which allows substantial nonlinearity in the conditional expectation of combined forecasts.[7]

Let $f_{j,t}$ denote the forecast from model j for time t. Let $\bar{y}$ and S_y denote, respectively, the in-sample mean and in-sample standard deviation of the dependent variable being forecasted. We consider ANN models of the form in Eqs (2–4), as in [6]:

$$z_{j,t} = (f_{j,t} - \bar{y})/S_y, \quad j \in \{1, 2\} \tag{2}$$

$$\Psi(z_t \gamma_i) = \left(1 + \exp\left[-\left(\gamma_{0,i} + \sum_{j=1}^{2} \gamma_{1,i,j} z_{j,t}\right)\right]\right)^{-1} \tag{3}$$

$$F_t = \beta_0 + \sum_{j=1}^{k} \beta_j f_{j,t} + \sum_{i=1}^{p} \delta_i \Psi(z_t \gamma_i) \tag{4}$$

$$k \in \{0, 2\}, \quad p \in \{0, 1, 2, 3\}$$

Note that the function Ψ maps into (0, 1). The forecasts from each model are standardized as in Eq. (2). This standardization, together with the appropriate choice of γ's, is employed to ensure that the Ψ function in Eq. (3) typically maps into the region close to $\frac{1}{2}$. Equation (4) displays the form of the final combined forecast.

A computationally simple approach to implement estimation of Eq. (4) which yields a universal approximation result for the neural nets functions with arbitrary choice of

[6] It is shown in [20] that the OLS combination will in general be more efficient than the simple average. Further, in [21, 22] it is demonstrated that weight-selection methods that are robust with respect to undue influence from outliers have the potential to outperform the standard OLS procedure in some cases.

[7] ANNs have been previously used for forecast combining in [8, 19]. For a review of the mechanics of ANN modelling with special attention to application to economics data, the interested reader is directed to [10]. As the mechanics of ANN modelling are now fairly well understood we restrict our attention to the application of ANNs to forecast combination.

γ's is presented in [16]. The strategy of setting the γ's with a uniform random number generator to lie between -1 and $+1$, and then estimate the values of the δ parameters with least squares is adopted in [8]. Little difference is made in the performance of the ANN model with deviations from this convention.

3. Forecast data

Stock market returns (i.e., the rate of change in stock prices) are well known to be serially dependent, non-identically distributed, and quite fat-tailed, i.e., leptokurtic. This has been considered troubling by some economists whose theories have suggested that stock returns should be closer to normally independently identically distributed. The serial dependence of stock returns is largely captured with a simple autoregressive term of order 1: AR(1), i.e., a single lagged return is used to forecast the current return. Depending on the return series and its periodicity (daily, weekly, etc.) this sort of simple model will typically explain between 1 and 15% of the variation of the return.[8] The non-identical distribution of the data of stock returns is revealed by the observation that large price swings tend to occur in clusters so that the second moment of the stock return distribution appears to vary conditionally on past squared return innovations, a feature referred to as autoregressive conditional heteroskedasticity (ARCH) by Engle [23]. ARCH effects imply fat-tails unconditionally, so ARCH models hope to resolve both the non-identical distribution of the data and its leptokurtosis. Stock volatility-forecasting models are therefore designed to produce a normalized return residual series (i.e., the return residual divided by its forecasted standard deviation) which is homoskedastic and less leptokurtic than the raw series.[9]

Individual forecasts for use in the combining exercises in [8, 19], and the current paper are forecasts of the volatility in daily returns on the S & P 500 stock index, for the period January 1969–September 1987, as produced by two popular models of stock returns volatility: the moving average variance model (MAV) and the generalized ARCH model (GARCH). Given the preceding discussion of stock return regularities, these models begin by defining r_t as the daily AR(1) stock return:

$$r_t = \rho_0 + \rho_1 r_{t-1} + \varepsilon_t.$$

The error ε_t has zero mean and conditional variance

$$E(\varepsilon_t^2 | I_t) = \sigma_t^2,$$

where I_t is the available information.

[8] Some of this AR(1) component appears to be generated by the so-called "non-synchronous trading effect" caused by the fact that small less liquid stocks trade less frequently, and therefore incorporate information less quickly, than large heavily traded stocks. The mixing of small and large stocks into a single index, such as the S & P 500, thus causes an AR(1) effect in the broad index as news enters prices of large stocks quickly and small stocks more slowly.

[9] See [24] for a detailed discussion of the ARCH literature.

Our task in volatility forecasting is to form a forecast of σ_t^2: the volatility of stock returns. Although the true volatility is unobserved, it is related to ε_t^2. Let $\hat{\rho}_0$ and $\hat{\rho}_1$ be estimates of the parameters ρ_0 and ρ_1, and let

$$\hat{\varepsilon}_t = r_t - \hat{\rho}_0 - \hat{\rho}_1 r_{t-1}. \quad (5)$$

The market volatility measure is $\hat{\varepsilon}_t^2$, which has expectation σ_t^2. The conditional variance forecast from the MAV model is

$$\hat{\sigma}_t^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_{t-i}^2,$$

with n chosen to minimize the Schwarz criterion and the parameters ρ_0 and ρ_1 estimated with OLS. The conditional variance forecast from the GARCH(1, 1) model is

$$\hat{\sigma}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\sigma}_{t-1}^2 + \hat{\alpha}_2 \hat{\varepsilon}_{t-1}^2,$$

with parameters $\rho_0, \rho_1, \alpha_0, \alpha_1$, and α_2 estimated jointly with maximum likelihood methods and the assumption of conditional normality of ε_t.

The application of MAV, GARCH and other ARCH-type models to the task of forecasting stock market volatility is documented and motivated in [24, 25] and the numerous references cited therein. The most important feature of these forecasts, for our purposes, is that the MAV and GARCH models used to produce them employ partially non-overlapping information sets. Thus, there may be an advantage to use a combined forecast as opposed to either of the individual forecasts. The period 1969:1–1979:12 was used for model selection and specification tests—the in-sample period. The remaining data was used for out-of-sample forecasting and forecast evaluation. For details, see [8, 19].

4. Comparing forecasts

There are a number of statistical criteria commonly used to evaluate combining models. The most basic focus is on the ability of the models to reproduce broad features of the data, including comparisons of summary statistics on the unconditional moments. A satisfactory combining method must at least pass such basic specification tests and do no worse than the forecasts incorporated in the combination. There are also comparisons of the forecasts on the basis of root mean squared forecast error (RMSFE) and mean absolute forecast error (MAFE), both in- and out-of-sample. Finally, there are comparisons of the combining methods on the basis of statistical tests of superior performance, including "encompassing tests" on out-of-sample forecasts. Such tests address the criticism that ranking models by RMSFE and MAFE do not provide a measure of statistically significant difference in performance across forecasting models.

For a comparison of the in- and out-of-sample forecasting performance of the MAV and GARCH individual models discussed above, as well as combined models based on OLS and the ANN model in Eqs. (2–4), see [8, 19]. They find that all the models attain a satisfactory level of performance both in- and out-of-sample, that the RMSFE and MAFE do not give a clear ranking of the models, and that the encompassing tests clearly favour the ANN models. The purpose of the current paper is not to reproduce these statistical results; the interested reader is directed to the original articles for such details. This paper's purpose is to explain graphically the reasons why ANN combining is superior to other methods and what this implies for future research and application of the combining technology. It should be noted that these are statistical criteria, with sampling error properties—questions of theoretical global optimality properties of ANN's are not answered by this work. See footnote 4 for work addressing optimality concerns.

Figure 1 plots the surfaces of the forecasting functions for the ANN and OLS methods which combine the individual MAV and GARCH forecasts (horizontal axes) into a single combined forecast (vertical axis) normalized so that their average value is equal to one. This surface corresponds to the model parameters estimated on the in-sample period, 1969:1–1979:12. The vertical height of each surface therefore represents the value of the combination forecast for each pair of individual forecasts listed on the basis of the figure. The flat surface of circles in Fig. 1 plots the relationship between the MAV and GARCH volatility forecasts and the OLS combined forecast. The curved surface of diamonds reveals the way in which the MAV and GARCH individual forecasts are combined to produce the ANN forecast.

A comparison of the surfaces in Fig. 1 provides visual confirmation that the ANN nonlinear combining method is doing something very different from the OLS linear method. For example, when the GARCH forecast is 6, a fall in MAV from 6 to 5 produces almost no change in the ANN combined forecast, which remains at approximately 3. However, when GARCH is 3 a decline in MAV from 6 to 5 produces a huge fall in ANN-combined volatility from 9 down to 3. Similarly, when MAV is 5, an increase in GARCH from 2 to 6 produces very little movement in the combined ANN forecast, which stays around 3. But, when the MAV forecast is 3, changes in GARCH between values of 2 and 6 produce huge swings in the combined ANN forecast, from a low of close to zero up to a high of 3. In other words, the impact of a change in GARCH on the ANN combined forecast is highly influenced by the interaction with the MAV forecast, and vice versa. In contrast to this highly flexible ANN response function, the response of the OLS linear combination forecast to changes in MAV and GARCH inputs remains constant no matter what values MAV and GARCH assume.

Whether the functional differences between ANN and OLS are of any practical importance depends, of course, on whether the data ever falls into regions of difference between OLS and ANN. Figure 1 therefore also plots the actual data points from the period 1969:4–1979: 12, presented as pillars with their height slightly raised so that they are easily visible. (The flagged pillars—roughly half of all the data points—are data points for which the ANN combination produces a smaller forecast error than does the OLS combination.) From these pillars we see that the majority of data points in Fig. 1 cluster in the areas of intersection between the ANN and OLS forecasting functions where the ANN and OLS functions are approximately equal (e.g., near the front right side of the figure where MAV and GARCH are both close to zero).

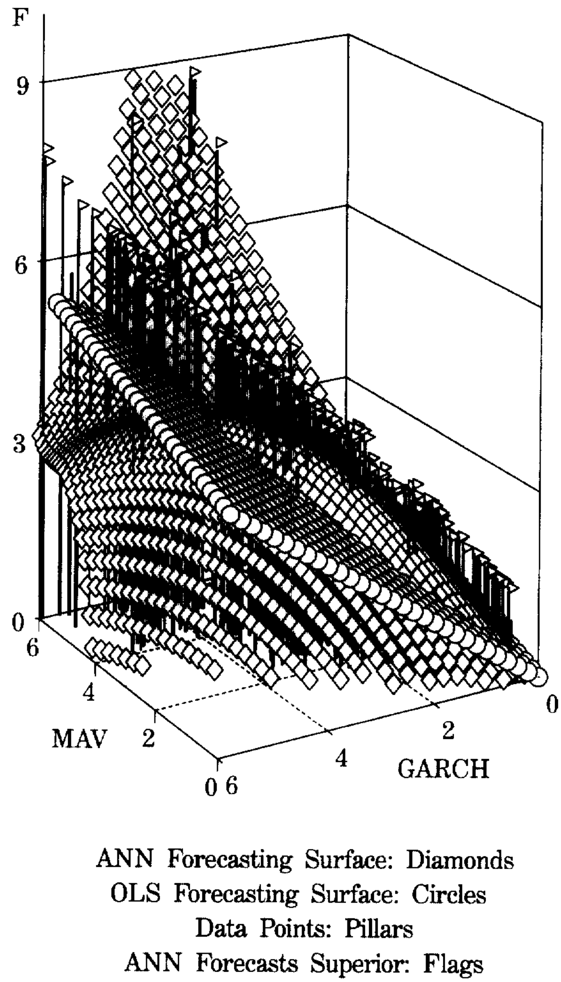

Fig. 1. S & P 500 stock market volatility forecasting surfaces.

However, there are a significant number of data points in areas where the OLS and ANN curves are far apart (e.g., near the back left side of the figure where the MAV and GARCH forecasts are large). In these regions one would therefore expect the interactive effects which the nonlinear ANN captures to be potentially important.[10]

As an example of the importance of the ANN's flexibility to treat different combinations of inputs differently, Fig. 2 plots various volatility forecasts for the period

[10] It should be noted that questions of "superiority" of one method over another cannot, of course, be resolved with inspection of such figures. Tests of forecast encompassing and comparisons of RMSFE and MAFE, mentioned above, are appropriate.

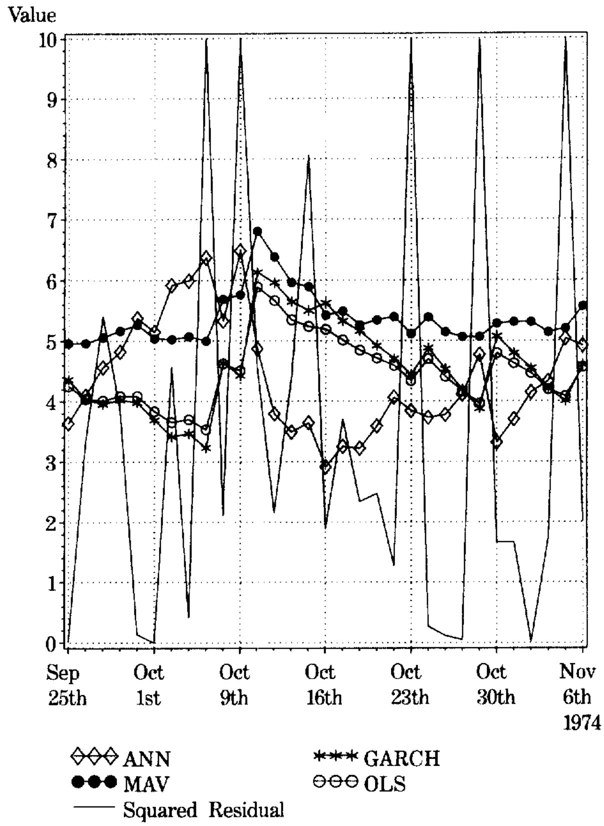

Fig. 2. S & P 500 stock market volatility forecasts.

surrounding the month of October 1974 (i.e., the fourth week in September through the first week in November 1974). The plain line in Fig. 2 plots the squared residual from Eq. (5), i.e., the dependent variable the forecasts are trying to explain. As in Fig. 1, ANN is represented by diamonds and OLS by circles. Solid dots represent MAV and stars GARCH.

Note from Fig. 2 that all of the series capture the magnitude of the peak in "true" volatility (plain line) during the second week of October. However, only ANN captures the sudden decline in "true" volatility immediately after the peak. Both MAV and GARCH start to decline immediately after the mid-October peak—and so does OLS, which is a simple linear combination of MAV and GARCH—but it takes several weeks for the decline to be completed. The rapid decline in ANN arises in spite of the slow decline in MAV and GARCH because ANN is able to condition on the information that both MAV and GARCH assume relatively extreme values during mid-October.

To see the dependence of the ANN combined forecast on the GARCH/MAV interaction, and the link between Figs 1 and 2, we can start tracking the various forecasts beginning 9th October, when MAV is 5.9 and GARCH is 4.4. These values put us in the middle back portion of Fig. 1 where the ANN surface rises to produce an ANN combined forecast of 6.5, as seen in Fig. 2 for 9th October: the peak (note that OLS is only 4.5 on 9th October). In contrast, on 16th October *both* MAV and GARCH are in excess of 5. This puts us at the far left back corner of Fig. 1 such that the ANN model delivers a combined forecast of only 3. Thus, even though GARCH and MAV have each fallen only approximately one point over the week (i.e., from their 10th October peak values to their values on 16th October), the ANN combination has fallen over three points. The ANN "interaction effect" therefore pushes the ANN nonlinearly combined forecast down much more quickly than either of the individual forecasts or the OLS linear combination.

5. Conclusions

In this paper we have argued that the nonlinear combination of forecasts with artificial neural networks can provide substantial improvements over traditional linear combining methods. We have explained this as occurring as a function of the ability of the ANN model to capture interaction effects between inputs to the ANN model. These interaction effects can be interpreted as state dependencies of the functional relationship between the dependent variable and the explanatory variables. The ability of ANNs to capture such interaction suggests the potential of future work in economic forecasting where it is likely important to allow for interplay with the economic state (e.g., recession, expansion, etc.). This feature of ANNs also has obvious implications for engineering applications like multi-sensor tracking systems in robotics and elsewhere.

References

[1] R.T. Clemen, Combining forecasts: a review and annotated bibliography, Int. J. Forecasting (1989) 559–583.
[2] C.W.J. Granger, Invited review: combining forecasts—twenty years later, J. Forecasting 8 (1989) 167–173.
[3] C. Min, A. Zellner, Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates, J. Economet. 56 (1993) 89–118.
[4] W.S. Chan, H. Tong, On tests for non-linearity in time series analysis, J. Forecasting 5 (1986) 217–228.
[5] R.G. Donaldson, M. Kamstra, Forecasting fundamental asset return distributions and tests for excess volatility and bubbles. University of British Columbia Working Paper, 1996.
[6] R.G. Donaldson, M. Kamstra, An artificial neural network GARCH model for international stock volatility. J. Empirical Finance 4 (1) (1997) 17–46.
[7] R.G. Donaldson, M. Kamstra, A new dividend forecasting procedure that rejects bubbles in asset prices, Rev. Financial Stud. 9 (1996) 345–398.
[8] R.G. Donaldson, M. Kamstra, Forecast combining with neural networks, J. Forecasting 15 (1996) 49–61.
[9] D.M. Keenan, A tukey non-additive type test for time series non-linearity, Biometrica 72 (1985) 39–44.
[10] C. Kuan, H. White, Artificial neural networks: an econometric perspective, Econometric Rev. 13 (1) (1994) 1–91.

[11] T. Lee, C.W.J. Granger, H. White, Testing for neglected nonlinearity in time series models, J. Economet. 56 (1993) 269–290.

[12] R.S. Tsay, Non-linearity tests for time series, Biometrica 73 (1986) 461–466.

[13] J. Hertz, A. Krogh, R.G. Palmer, Introduction to the Theory of Neural Computing, Addison-Wesley, Redwood City, CA, 1991.

[14] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Network 4 (1991) 251–257.

[15] K. Hornik, M. Stinchcombe, H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, Neural Network 3 (1990) 551–560.

[16] M. Stinchcombe, H. White, Using feedforward networks to distinguish multivariate populations, Proc. Int. Joint Conf. on Neural Networks, ICANN '94, Sorrento, Italy, May 26–29 1994.

[17] H. White, Some asymptotic results for learning in single hidden layer feedforward network models, J. Am. Statist. Assoc. 84 (1989) 1003–1013.

[18] H. White, Connectionist nonparametric regression multilayer feedforward networks can learn arbitrary mappings, Neural Network 3 (1990) 535–549.

[19] P.G. Harrald, M. Kamstra, Evolving artificial neural networks to combine financial forecasts, IEEE Trans. Evolutionary Comput. 1 (1) (1997) 40–52.

[20] C.W.J. Granger, R. Ramanathan, Improved methods of combining forecasts, J. Forecasting 3 (1984) 197–204.

[21] J. Hallman, M. Kamstra, Combining algorithms based on robust estimation techniques and cointegrating restrictions. J. Forecasting 8 (1989) 189–198.

[22] J. LeSage, M. Magura, A mixture-model approach to combining forecasts, J. Business Economic Statist. 10 (1992) 445–452.

[23] R.F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50 (1982) 987–1007.

[24] T. Bollerslev, R.T. Chou, K.F. Kroner, ARCH modelling in finance: a review of the theory and empirical results, J. Economet. 54 (1991) 1–30.

[25] A. Pagan, W. Schwert, Alternative models for conditional stock volatility, J. Economet. 45 (1990) 267–290.