Evolving Artificial Neural Networks to Combine Financial Forecasts

Paul G. Harrald and Mark Kamstra

***Abstract*—We conduct evolutionary programming experiments to evolve artificial neural networks for forecast combination. Using stock price volatility forecast data we find evolved networks compare favorably with a naïve average combination, a least squares method, and a Kernel method on out-of-sample forecasting ability—the best evolved network showed strong superiority in statistical tests of encompassing. Further, we find that the result is not sensitive to the nature of the randomness inherent in the evolutionary optimization process.**

***Index Terms*— Evolutionary programming, financial forecasting, forecast combination, neural networks, self-adaptive evolutionary programming.**

I. Introduction

POLICY and decision makers must typically adopt a position based on a host of conflicting opinions about the future course of events. Processing multiple and conflicting data is an inherently complex and in many cases subjective procedure. We focus here on this process of consensus-making when the information available to the decision maker is quantitative. For instance, the governor of a central bank may have several forecasts of exchange rate movements available that need to be considered in setting the prime interest rate. These forecasts may come from forecasters of differing ability and reputation and from forecasters with different information at hand. Forming some sort of consensus of the exchange rate movement forecasts is required before the prime rate can be set. A simple average of the forecasts could be taken, but this ignores the ability and reputation of the forecasters, since equal weight would be given to each forecast. It is this basic observation that has led to what is now a voluminous literature on forecast *combining*.

The "optimal" weighting scheme of [1] is based on the covariances of the individual forecasts with the actual realized values of the variable being forecasted. Subsequent work [6] extended this notion of using past forecasts to form a superior combination. Examples from this body of work include the use of unconstrained least squares regression of the actual values on the forecasts to form optimal weights, the correction of serial correlation in the forecast errors, and the use of Bayesian updating techniques to formally allow the weights on forecasts to evolve with new information [19], [37], [20].

Only linear combinations of the individual forecasts have thus far been considered. This is a substantial, most likely inappropriate, restriction and one with serious implications for the efficiency or even the consistency of the combined forecast. For example, consider the case of a dependent variable $y = x_1 \cdot x_2 + \epsilon$, where ϵ is an innovation and x_1 and x_2 are known explanatory variables. If forecaster 1 has the model $f_1 = \alpha_1 x_1$ and forecaster 2 has the model $f_2 = \alpha_2 x_2$, then any linear combination of the two forecasts will be inferior to the nonlinear forecast $\beta(f_1 \cdot f_2)$, with $\beta = (1/(\alpha_1 \cdot \alpha_2))$. There exists some empirical evidence for the gains from incorporating nonlinear combination of forecasts, specifically in the context of combining financial forecasts of stock market volatility [7]. Such complications with nonlinearity are becoming more widely appreciated in modeling economic data in particular, as evidenced in recent work [32] and the wealth of new tests for nonlinearity [4], [30], [33], [49], [50].

Artificial neural networks (ANN's) have the ability to approximate arbitrarily well a large class of functions [23], [25]–[27], [32], [46], [52]–[54]. ANN's, therefore, have at least the potential to capture complex nonlinear relationships between a group of individual forecasts and the variable being forecasted, which simple linear models are unable to capture. Readers should be aware that ANN modeling has been criticized as a "black box" (e.g., [11]). If care is not taken, all intuition for the relationship between the forecasts and what is being forecast may be lost. It is also important to recognize that the very act of combining forecasts is an admission of some sort of failure of the models from which the forecasts are produced. If we are given all the information used in generating the individual forecasts being combined, it is always better to construct a single "super model" that encompasses this full information set and not combine the individual forecasts at all. What we are attempting to demonstrate here is the utility of the ANN model if we are faced with only forecasts to combine.[1]

Manuscript received August 4, 1996; revised February 15, 1997 and March 3, 1997. This work was supported by the Social Sciences and Humanities Research Council of Canada and the Sir Peter Allen Travelling Scholarship.

P. G. Harrald is with the Manchester School of Management, University of Manchester Institute of Science and Technology, Manchester, M60 1QD U.K.

M. Kamstra is with the Department of Economics, Simon Fraser University, Burnaby, B.C. V5A 1S6 Canada.

Publisher Item Identifier S 1089-778X(97)03850-2.

[1] In an unrelated literature [55], it is noted that competing nonlinear "generalizers" making use of identical information sets can be eliminated with a winner-takes-all strategy or they can be combined, termed "stacked generalization," in a possibly nonlinear fashion. This is similar to what we hope to do here. Our problem, however, is to construct the best forecast as a function of forecasts each based on at least partially unique information, without access to anything but the final forecasts. We are not facing the problem of how best to use all the available information, but rather how best to combine individual forecasts themselves based on information unknown to us.

A typical training method for the ANN modeling is some manner of supervised learning on a training sample, of which the familiar backpropagation is an example. It is almost certain, however, that in many optimization problems for which ANN's are considered to be appropriate architectures for search over mappings, the state space will exhibit many local optima [32], [2], rendering gradient-descent methods such as backpropagation unreliable.

In response to this problem of local optima, techniques of evolutionary optimization such as genetic algorithms and evolutionary programming (EP) have been applied to the training of ANN's (e.g., [31], [36], [38], [41]).

This paper describes the ANN method of forecast combination of [7] and reports on EP experiments to evolve suitable parameterizations of a given ANN architecture. In Section II we present combining methods. In Sections III, IV, and V the evolutionary programming approach to estimation of the ANN model is presented in detail. Section VI is a discussion of the data to be used to illustrate these techniques in an application. In Section VII formal comparisons of the different techniques of forecast combining are presented. Section VIII offers a discussion of the results with some intuition. Section IX concludes.

II. Combining Methods

First we present some traditional combining methods to be used in comparison with our ANN method, second an alternative method which is nonparametric, namely the Kernel method, and third, our ANN combining technique.

A. Traditional Combining Methods

There are now many commonly employed methods for combining forecasts [6], [20]. The assumption that the conditional expectation of the variable being forecasted is a linear combination of the available forecasts is consistent across all combining methods. Thus when combining two individual forecasts $f_{1,t}$ and $f_{2,t}$, a single combined forecast F_t is produced according to (1) by appropriate choice of weights β_0, β_1, and β_2

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

The cross-sectional average of the individual forecasts (denoted "Average"), perhaps the most widely used combining method, sets $\beta_0 = 0$ and $\beta_1 = \beta_2 = 0.5$.

It can be shown that a multivariate ordinary least squares (OLS) regression of the variable being forecasted on the individual forecasts in-sample can be used to obtain "optimal" forecast weights β_0, β_1, and β_2 for use in out-of-sample combining [19]. This combination will in general be more efficient than the simple average. We have also considered Bayesian combination methods, but these have been found to have little advantage over classical methods at the forecasting horizon we investigate in this paper [37] and are not pursued further here as a point of comparison.

B. Kernel Estimation

Kernel estimation is a nonparametric smoothing technique widely used in modeling of economic and other data; see [22]. The technique essentially forms multidimensional histograms of the data to discover associations between the dependent and the independent variables. The nature of these associations can be almost arbitrarily nonlinear, but uncovering these associations has one very troublesome technicality—determining the width of the histogram bars, called the *window-width*. The choice of window-width can be data driven. One popular data-driven method is cross validation, in which the window-width is chosen by minimizing or maximizing some objective function on the cross-validated data over a grid of possible window-widths.[2]

A second technicality, but less troublesome as a matter of practice, is the choice of the form of the histogram itself. Simple on–off bars is one option, although a more popular choice has a probability distribution (such as the normal distribution) centered on the middle of the bar.

C. Artificial Neural Network Combining

The mechanics of ANN modeling are now fairly well understood (e.g., [32]), and a review of ANN's in general is not undertaken here: we restrict attention to the application of ANN's to forecast combination.

Consider the task of combining two forecasts. Let $f_{j,t}$ denote the forecast from model j for time t. Let $\overline{d}$ and S_d denote, respectively, the in-sample mean and in-sample standard deviation of the variable being forecasted out-of-sample. We consider ANN models of the form in (2)–(4), from [7]

$$z_{j,t} = (f_{j,t} - \overline{d})/S_d \;\; ; \;\; j \in \{1,2\} \tag{2}$$

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

$$F_{t,p} = \beta_0 + \sum_{j=1}^{2} \beta_j f_{j,t} + \sum_{i=1}^{3} \delta_i \Psi(z_t, \gamma_{i,p}) \tag{4}$$

where p is an index whose use will be described below, $\gamma_{i,p,r}, \beta_j$, and δ_i are parameters to be estimated in a manner described below, $\gamma_{i,p}$ is a vector of the $\gamma_{i,p,r}$ parameters, and $F_{t,p}$ is the forecast of the dependent variable produced by the ANN method. From (3) we see that the input layer accepts as activation the forecasts to be combined. The input nodes are linked to a hidden layer of three nodes, and also directly to an output node, as is a bias. The hidden nodes and output node use a nonlinear sigmoidal filter Ψ which

[2] To implement cross validation, the Kernel weights are estimated on a subset of the data, and then the estimated weights are used to forecast the remaining portion of the data. The "out-of-sample" forecasts are then collected and the process repeated, leaving out a different subset of the data each time, until "out-of-sample" forecasts for the entire data set are produced. (A cross-validation estimation which has 1/N of the data omitted at a time is called an N-fold cross-validation.) The model specification that produces cross-validated forecasts with, say, the lowest mean squared error, is then selected as "best" [24], [29], [48]. Further studies include the derivation of optimality results for a pseudo-likelihood function [40] and many empirical studies (e.g., [47] and [9]).

maps into (0,1). The standardization of forecasts from each model is given in (2). This standardization is employed, together with the appropriate choice of the $\gamma_{i,p,r}$, to ensure that the Ψ function in (3) typically maps into the region close to 1/2 and not typically close to zero or one. Equation (4) makes explicit the manner in which the outputs from the Ψ functions are to be used to form the final combined forecast. As the estimation of (4) will be based on evolutionary programming and "self-adaptive" evolutionary programming, we will refer to these models generically as the EP-NN method and the SEP-NN method, respectively. The EP-NN method is described in Section IV and SEP-NN in Section V.

III. Evolutionary Programming

Evolutionary programming was developed in the early 1960's [17] as a means of solving complex optimization problems by a stochastic numerical process that has features in common with natural evolution. Consider the problem of minimizing a function $F(\gamma)$ where γ is a vector of real values. A simple implementation of EP would proceed according to the following pseudocode.

1) Generate n random vectors $\gamma_1, \cdots, \gamma_n$.
2) Until finished
   2.1) Sort $\gamma_1, \cdots, \gamma_n$ by $F(\gamma_i)$, smallest to largest
   2.2) Delete bottom half of γ_i.
   2.3) Replace bottom half by $\gamma_i + \eta_i$, $i = 1, \cdots, n/2$.

Typically, the random mutation η_i is made by sampling from a multidimensional normal distribution with small variance (or covariance). There are many variations of this classic EP (see [14]). The utility of EP has been demonstrated in a variety of contexts (e.g., [12] and [13]). The EP has also been applied to evolving neural networks [16], [34], [41], with the essential idea that the vector γ represents the parameters of the ANN and that $F(\gamma)$ is a measure of the ANN's performance, chosen in our case to be the mean square error (MSE) of the forecast of the ANN on in-sample data.

IV. ANN Forecast Combination EP

The EP-NN procedure used here is as follows.

0) Define:
$\hat{\epsilon}_t^2$ as the dependent variable, and our forecast of $\hat{\epsilon}_t^2$ as $F_{t,p}$, with

$$F_{t,p} = \beta_{1,p} + \beta_{2,p} f_{1,t} + \beta_{3,p} f_{2,t} + \delta_{1,p}\Psi(z_t, \gamma_{1,p}) + \delta_{2,p}\Psi(z_t, \gamma_{2,p}) + \delta_{3,p}\Psi(z_t, \gamma_{3,p}).$$

1) For each parent $p \in \{1, 2, \cdots, n\}$
   1.1) For each parameter vector $\gamma_{i,p}$, $i = 1, 2, 3$
      1.1.1) Generate $\gamma_{i,p}$ uniformly distributed on [−1, 1], independent of all other trials p.
   1.2) For each observation t
      1.2.1)

$$\Psi(z_t, \gamma_{1,p}) = (1 + \exp(-(\gamma_{1,p,0} + \gamma_{1,p,1} z_{1,t} + \gamma_{1,p,2} z_{2,t})))^{-1}$$

$$\Psi(z_t, \gamma_{2,p}) = (1 + \exp(-(\gamma_{2,p,0} + \gamma_{2,p,1} z_{1,t} + \gamma_{2,p,2} z_{2,t})))^{-1}$$

$$\Psi(z_t, \gamma_{3,p}) = (1 + \exp(-(\gamma_{3,p,0} + \gamma_{3,p,1} z_{1,t} + \gamma_{3,p,2} z_{2,t})))^{-1}.$$

   1.3) For each parent $p \in \{1, 2, \cdots, n\}$
      1.3.1) Estimate $\beta_{1,p}, \beta_{2,p}, \beta_{3,p}, \delta_{1,p}, \delta_{2,p}, \delta_{3,p}$ by OLS, population regression $E[\hat{\epsilon}_t^2] = F_{t,p}$.
   1.4) Sort the array γ_p over p by ascending MSE from regression of $\hat{\epsilon}_t^2$ on a constant, $f_{1,t}$, $f_{2,t}$, $\Psi(z_t, \gamma_{1,p})$, $\Psi(z_t, \gamma_{2,p})$, and $\Psi(z_t, \gamma_{3,p})$.
   1.5) For each $p > n/2$
      1.5.1) $\gamma_p = \gamma_p + \eta_p$, $\eta_p \sim$ N(0,σ), η_p a vector of the same dimension as γ_p.

This represents a single generation of the EP-NN algorithm, for a single ANN model. The set of parents had $n = 20$, a single ANN model estimation consisted of a 1000-generation run, and σ was set to 0.05. A total of 29 independent ANN models were estimated in the above fashion, each with Step 1.1) and 1.5.1) repeated independently of the other 28 model estimations.

The ANN models were ranked by MSE and the median in-sample MSE trial was selected as our ANN model used to forecast out-of-sample, denoted as EP-NN(M) in the tables of results. We also include for comparison and contrast the best MSE ANN model and the worst MSE ANN model, denoted EP-NN(B) and EP-NN(W), respectively. We anticipated that the best MSE ANN model would overfit the data and perform badly on the out-of-sample forecasting period.[3]

V. Self-Adaptive Evolutionary Programming

The choice of $\sigma = 0.05$ was, in fact, our own first *ad hoc* choice, but it remained quite robust to other challengers. We also considered, however, the possibility of self-adaptive mutation, an algorithm we term self-adaptive EP (SEP-NN). The SEP-NN algorithm proceeds much in the same way as the EP-NN, except that each trial solution carries with it a vector of terms describing the mutation variances to be used in the production of offspring (see [43]). We denote the evolvable weights and biases of an ANN p as γ_p, then in the SEP scheme a trial solution is appended with a vector σ_p of the same dimension. When mutation of the jth component of γ_p is undertaken, a normal deviate of mean zero and standard deviation equal to the jth component of σ_p is added. The extant vectors σ_p are updated at the beginning of each generation. If we denote by $\sigma_{p,j}$ the jth component of σ_p, then the update takes the form of

$$\sigma_{p,j} = \sigma_{p,j} \exp[\tau' N(0,1) + \tau N_j(0,1)]$$

where $N_j(0,1)$ indicates a standard normal random variable drawn independently of the other components of the vector σ_p.

The parameters τ and τ' adopted the conventional values of $\left(\sqrt{2\sqrt{k}}\right)^{-1}$ and $(\sqrt{2k})^{-1}$, respectively, where k is the total

[3] Ranking by cross validation, such as described for Kernel estimation, or ranking by an information criterion, such as the Schwarz Criterion, are possible mechanisms to avoid overfitting.

number of weights and biases to be evolved. This scheme was first proposed in [45] and has proved useful in similar and earlier work to our own (e.g., [15] and [42]).[4]

Scope for further experimentation is unlimited: ANN's can achieve arbitrary mappings and be the output quantitative or qualitative, and the EP is similarly a very flexible procedure, since mutation is simply a probability distribution mapping a space into itself.

VI. Data on Individual Forecasts

We require a data set with well-known properties and a large number of observations so that we have both the power to discriminate between the various combining methods and intuition for why some combining methods outperform others. We conduct our exercise on forecasts of daily stock market volatility—conditional stock return variance—from two well-known models over the period 1969–1987, yielding thousands of observations.[5] The fact that the models we combine have well-understood properties aids in interpreting test results and also suggests specification checks of the combined models. The exercise itself—forecasting volatility—is interesting in its own right. Forecasting the volatility (the riskiness) of stocks facilitates tracking the risk/return characteristics of a portfolio which includes stocks and adjusting the portfolio composition when the risk/return characteristics become undesirable. This application certainly does not exhaust the pool of applications in finance, let alone other fields with similar "problems" of a wealth of competing forecasts. Ongoing work by the authors and others include credit-rating problems, mean-return forecasting, and value-at-risk estimation (the monetary risk of holding certain portfolios, often over short horizons of a day to a month). Work in other fields, primarily engineering though stretching to biology, robotics, and beyond, refers to combining problems as data "fusion" exercises. Many of these exercises are directed to "fusing" data from a multitude of sensors, for targeting or tracking.

Stock returns, often assumed to be independent and identically normally distributed (i.i.d.), are in fact dependent, nonidentically distributed, and quite fat-tailed—nonnormal. The dependence of returns is largely captured with a simple autoregressive term of order one [AR(1)]—a single lagged return is used to forecast the current return. This sort of simple model will explain between 1% and 15% of the variation of the return, depending on the return series and its periodicity—daily, weekly, and so on. Market closings and sluggish flows of information leading to nonsynchronous trading patterns are believed by some (e.g., [10] and [44]) to lead to this AR(1) structure in returns. The nonidentical distribution of the data is revealed by the clustering of highly volatile periods. In examining data with similar properties, [8] modeled the second moment of inflation rates as varying conditionally on past squared modeling errors to capture such clustering, termed autoregressive conditional heteroskedasticity[6] (ARCH). In the context of stock returns, the residual from the model of the return is modeled as having time-varying variance. ARCH effects imply fat-tails unconditionally, so modeling ARCH effects promised to resolve both the nonidentical distribution of the data and its leptokurtosis.[7] Typical in the literature investigating time-varying variances of stock returns, then, is the assumption of conditional normality of the returns, with the first moment of the returns and the second moment of the return residuals modeled as autoregressive processes, though a number of other approaches have been adopted [3].

Individual forecasts for use in the combining exercise are forecasts of the volatility in daily returns on the S&P 500 stock index, for the period January 1969 to September 1987, as produced by two popular models of stock returns volatility: 1) the moving average variance model (MAV) and 2) the generalized autoregressive conditional heteroskedasticity (GARCH) model. These two methods for forecasting volatility are widely used by professional portfolio managers and academics and perform admirably, as will be demonstrated below. Improving on them should not be regarded as a matter of course—they are not "straw men."

Following the stock returns volatility literature, define r_t as the daily stock return, then generically

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

The error ϵ_t has zero mean and has conditional variance

$$E(\epsilon_t^2 | I_t) = \sigma_t^2$$

where I_t is any available information the expectation may be conditioned upon. That is, volatility is unobserved but related to ϵ_t^2. Our task is to form a forecast of σ_t^2: the volatility of stock returns. Let $\hat{\rho}_0$ and $\hat{\rho}_1$ be estimates of the parameters ρ_0 and ρ_1, and let

$$\hat{\epsilon}_t = r_t - \hat{\rho}_0 - \hat{\rho}_1 r_{t-1}. \tag{5}$$

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

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

with n chosen to minimize the Schwarz Criterion[8] and the parameters ρ_0 and ρ_1 estimated with OLS.[9] The conditional variance forecast of the GARCH(1,1) model has

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

[4] Since we update mutation variances before offspring are created, we adopt what is known as a "sigma-first" strategy; see [18].

[5] Our data extend only to September 1987 to exclude the infamous stock market crash of October 1987. The context of stable conditional data generating processes is the only context in which model comparisons such as ours make sense, and the crash of 1987 produced such large and abrupt changes in stock volatility that the assumption of stability may not be valid over this period.

[6] A heteroskedastic random variable does not exhibit a constant second moment.

[7] A leptokurtotic random variable is random variable with a fourth moment which is larger than that of a normally distributed random variable.

[8] The Schwarz Criterion is a likelihood-based information criterion which assigns a penalty for use of modeling degrees of freedom. It equals the log of the likelihood function minus $\frac{1}{2} K log(T)$, where K is the number of parameters in the model and T is the number of observations available.

[9] The value chosen for n over the range 1 to 40 was 28.

with parameters $\rho_0, \rho_1, \alpha_0, \alpha_1$, and α_2 estimated jointly with maximum likelihood methods under the assumption of conditional normality of ϵ_t. The wide application of such models to the task of forecasting stock market volatility is well documented and motivated [3], [39].

We use one-step-ahead out-of-sample GARCH and MAV forecasts beginning the first day of trading in January 1980. The MAV and GARCH model parameters are estimated with data from the first day of April 1969 to the last day of 1979 and then used to produce the one-step-ahead out-of-sample MAV and GARCH forecasts for the first trading day of 1980. Our data set is then updated by adding the first trading day of 1980 and dropping the first from 1969, the models then re-estimated to produce a one-step-ahead out-of-sample forecast for the second day in 1980, and so on.[10] This procedure of updating and one-step-ahead out-of-sample forecasting is repeated until one-step-ahead out-of-sample MAV and GARCH forecasts of daily returns volatility are produced for each trading day from January 1, 1980 to September 30, 1987, constituting the individual out-of-sample forecasts used in the combining exercise. The most important feature of these forecasts, for our purposes, is that the MAV and GARCH models used to produce them employ partially nonoverlapping information sets. Thus there may be an advantage to using a combined forecast as opposed to either of the individual forecasts.[11]

The in-sample observations 1969–1979 are used for two purposes. First, the training and selection of the EP-NN and SEP-NN models and choice of the window-width of the Kernel estimator[12] is performed exclusively with the 1969–1979 data using the in-sample forecasts from MAV and GARCH. Second, the first out-of-sample forecast from all of the models is made with only this in-sample data. The MAV and GARCH in-sample 1969–1979 forecasts are used to form the combining parameter weights for the forecast of the volatility of the first trading day of 1980 from the OLS, Kernel, and EP-NN and SEP-NN models.[13] The information set is then updated one day at a time, in a rolling window fashion just as with the formation of the MAV and GARCH out-of-sample forecasts, and one-step-ahead *out-of-sample* combined forecasts of all the methods are obtained, January 1, 1980 to September 30, 1987. These are the forecasts used to evaluate the performance of the combining models. We must stress that the training and selection of the EP-NN and SEP-NN models is carried out once and once only, on the in-sample data 1969-1979. Although it is sensible to retrain and re-evaluate the EP-NN and SEP-NN as we update the data set, this is computationally too onerous to attempt.

It may be helpful at this point to plot some of the data and compare the two forecasts which will be used in the combining exercise, the MAV and GARCH forecasts. We also plot the SEP-NN forecasts, but these will not be discussed until Section VIII.

Figs. 1 and 2 plot subsets of the in-sample data, covering periods 1970 and 1974, respectively. The data plotted are the squared residuals from (5) and volatility estimates from the MAV, GARCH, and self-adaptive evolutionary model, normalized so that their average value is equal to one. Of course, since the data are strictly positive and quite skewed (the residuals are approximately distributed as χ_1^2), the points below one cluster more closely to one. The points (dots) are the squared residuals, the line of circles is the MAV forecast, the boxed line is the GARCH forecast, and the line with diamonds is the SEP-NN forecast.

Fig. 1 plots a year which presented a volatile period May through July and two less volatile periods on either side. In such quiet periods the MAV and GARCH forecasts are virtually overlaid while in volatile periods the two are often quite different; GARCH reacts much more swiftly than MAV to changes in volatility. We see much the same pattern in Fig. 2, though the year 1974 was much more volatile due in part to the OPEC oil price shock. We see remarkable divergences of forecasts from MAV and GARCH during volatile periods, and the very slow adjustment of the MAV forecast causes it to forecast substantially higher volatility than GARCH when brief periods of calm present themselves, such as in August, October, and December 1974. While in some cases it appears the quick adjustment provided by GARCH was appropriate, as in June of 1970, the slower adjustment of MAV provides a better guide through the last half of 1974. Hence there is hope that combining these two forecasts may provide us with a better volatility estimate than the use of one or the other alone.

VII. Comparing Forecasts

First, we evaluate the models' abilities to reproduce broad features of the data, characterized by summary statistics on the unconditional moments of the data and a test for residual autoregressive conditional heteroskedasticity (ARCH) effects and normality. If a combining forecast method is to be relied on, then as a minimum it must pass basic specification tests and do no worse than the forecasts incorporated in the combination. These summary statistics and specification tests give some insight on this minimum level of performance.

Next, we compare forecasts on the basis of root mean squared forecast error (RMSFE) and mean absolute forecast error (MAFE), in- and out-of-sample. A traditional measure of the best forecasting method is a simple comparison of RMSFE and MAFE across competing methods, but this provides no measure of statistically significant difference in performance.

Finally, we compare combining methods on the basis of statistical tests of superior performance—encompassing tests

[10] "Rolling window" model estimation is common in the combining literature (e.g., [21]).

[11] In practice, having access to the forecasters' information sets means any combination of the forecasts is an inefficient use of the available information. The application to conditional variance forecasts in our paper should therefore be viewed as an exercise to compare the combining methods.

[12] We made use of a normal distribution Kernel and picked the window-width to minimize the cross-validated Gaussian log-likelihood among all Kernel estimators that removed evidence of ARCH at the 10% significance level on the 1969–1979 period. The test performed was the ARCH Lagrange Multiplier (LM) test at lags 5 and 20. For a description of this test see footnote 14.

[13] The average combined forecast has weights $\beta_0 = 0, \beta_1 = \beta_2 = 0.5$ and hence does not need to be estimated. Once the combining parameter weights have been estimated—with only *in-sample* MAV and GARCH forecasts—the *out-of-sample* MAV and GARCH forecasts are used to produce the *out-of-sample* forecast of the volatility of the first trading day of 1980.

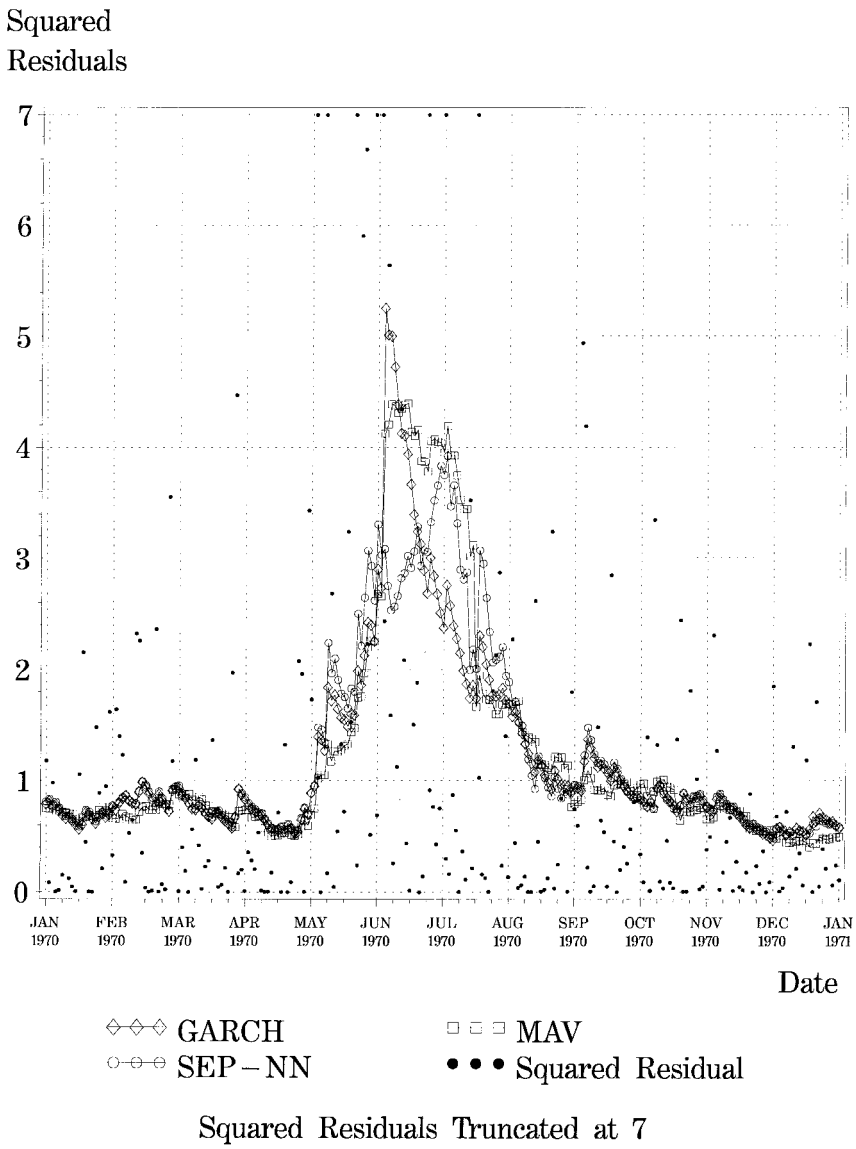

Fig. 1. This plots a subset of the in-sample data, covering the year 1970. The data plotted are the squared residuals and the volatility estimates from the MAV, GARCH, and evolutionary models, normalized so that their average value is equal to one. The MAV model volatility estimate is an average of past squared residuals, and the GARCH(1, 1) volatility estimate is a function of the squared residual and volatility estimate of the last period. The SEP model volatility estimate is a nonlinear function of both the MAV and GARCH volatility estimates.

on out-of-sample data. This addresses the criticism that ranking models by RMSFE and MAFE does not provide a measure of statistically significant difference in performance across forecasting models.

The importance of these different criteria are inversely related to their order of presentation. The summary statistics presented below indicate that all the methods studied attain a satisfactory minimum of performance. The RMSFE and MAFE are traditional measures of performance but we will argue that they provide little information. The encompassing tests provide a sound statistical basis for model comparison, and our argument for the superior utility of the EP methods rests largely on the evidence from the encompassing tests.

A. Summary Statistics

Table I contains summary statistics on the volatility forecasts and on the implied standardized return residuals from our models for stock volatility for the in-sample period April 1, 1969 to December 31, 1979 and the out-of-sample period, January 1, 1980 to September 30, 1987.

The first column in Table I contains the forecast method name. Columns 2 and 3 contain the mean and standard deviation of actual stock returns volatility for the S&P 500 index as well as the mean and standard deviation of the stock volatility forecasts produced by each method. We expect the mean of the various volatility forecasts to be similar to the actual mean and the standard deviation of the forecasts to be smaller than that of the actual data. The mean should be identical in-sample for the OLS, EP-NN, and SEP-NN forecasts, which it is.

Columns 4–6 of Table I (the first three columns of the standardized returns statistics) contain summary statistics on the standardized return residuals from the S&P 500 index; i.e., $\hat{\epsilon}_t / \sqrt{\hat{\sigma}_t^2}$. When divided by its forecasted standard deviation,

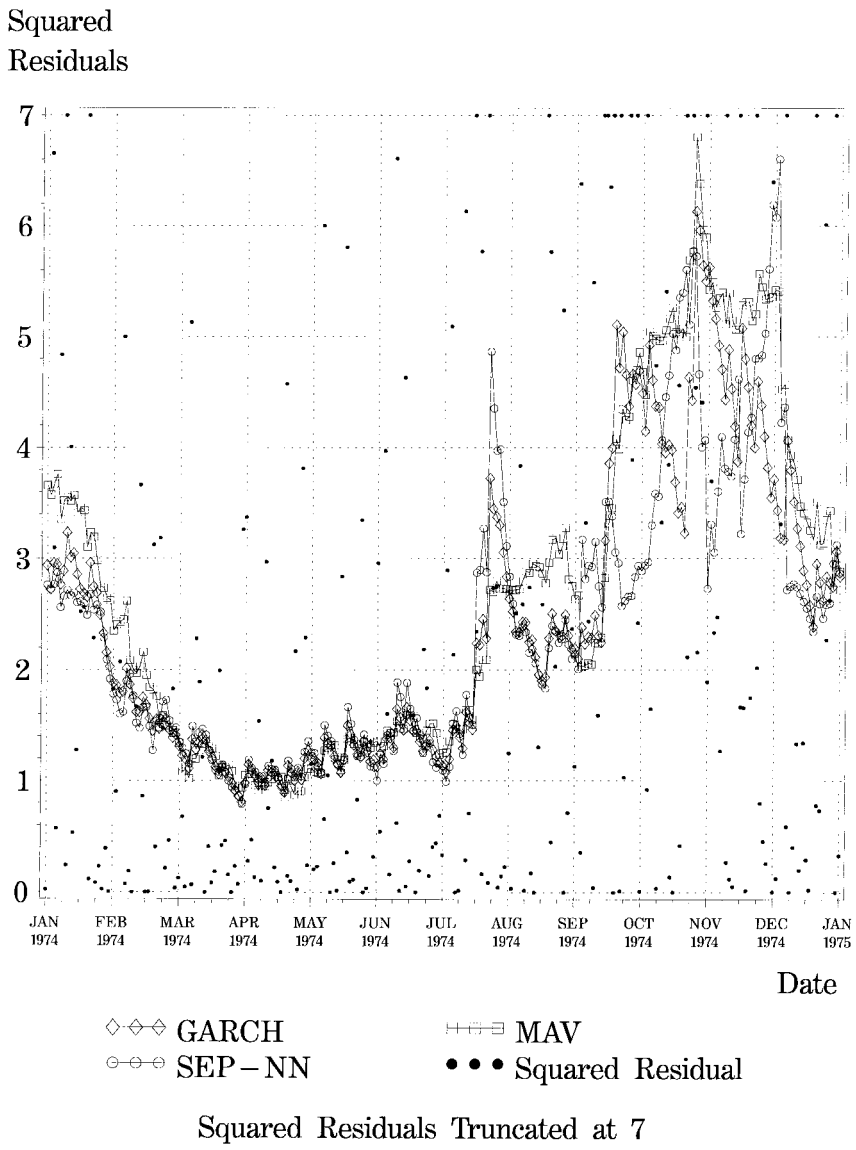

Fig. 2. Plots a subset of the in-sample data, covering the year 1974. The data plotted are the squared residuals and the volatility estimates from the MAV, GARCH, and evolutionary model, normalized so that their average value is equal to one. The MAV model variance forecast is an average of past squared residuals, and the GARCH(1, 1) model variance forecast is function of the squared residual and variance forecast of the last period. The SEP model variance forecast is a nonlinear function of both the MAV and GARCH variance forecasts.

the return residuals should have a standard deviation of one (as is the case for the actual or raw return residuals divided by their sample standard deviation, shown in the first row). Stock returns distributions are well known to be leptokurtotic compared to the normal distribution, and we see this in our raw data, with a kurtosis of 5.749 in-sample instead of three. As stock volatility-forecasting models are designed to produce a standardized return residual series which is less leptokurtotic than the raw series, a reliable method for combining volatility forecasts should have this property. Table I shows that all the forecasting and combining methods produce standardized return residuals which are less skewed and leptokurtotic than the raw series. The skewness is close to zero and the kurtosis is close to three, both on the in-sample data, shown in Table I(a), and on the out-of-sample data, shown in Table I(b).

Column 7 (the fourth column of the standardized returns statistics) contains the p-value from an LM test of the null hypothesis that the standardized return residuals do not display ARCH.[14] As expected, the raw return residuals display strong evidence of ARCH with a p-value of zero to three decimal places for both the in-sample and the out-of-sample data. Stock volatility-forecasting models are designed to remove ARCH from return residuals and all the forecasting and combining methods we investigate here succeed in removing significant

[14]The ARCH LM test is a regression-based test, which takes the R^2 from the regression of squared errors on lagged squared errors, multiplies this R^2 by the sample size, and compares this to the critical value from a χ^2 distribution, degrees of freedom equal to the number of lagged squared errors in the regression. This test is appropriate under the null of no heteroskedasticity in the errors, and is performed with 12 lags. The results are qualitatively identical with 24 lags. For a more detailed discussion of this test for ARCH; see [3].

TABLE I

(a) REPORTS SUMMARY STATISTICS ON THE IN-SAMPLE FITTED CONDITIONAL VARIANCES, $\hat{\sigma}_t^2$, AND STANDARDIZED RETURNS, $\hat{\epsilon}_t/\hat{\sigma}_t^2$, FOR NEW YORK'S S&P 500 INDEX ON DAILY DATA 1969–1979. (b) REPORTS SUMMARY STATISTICS ON THE ONE-STEP-AHEAD OUT-OF-SAMPLE FORECASTED CONDITIONAL VARIANCES, $\hat{\sigma}_t^2$, AND STANDARDIZED RETURNS, $\hat{\epsilon}_t/\hat{\sigma}_t^2$, FOR NEW YORK'S S&P 500 INDEX ON DAILY DATA 1980:1–1987:9. THE MEAN AND STANDARD DEVIATION (DENOTED "STD" IN THE TABLE) OF THE VARIOUS VARIANCE FORECASTS ARE PRESENTED. FOR THE VARIOUS STANDARDIZED RETURNS, THE STANDARD DEVIATION, SKEWNESS (DENOTED "SKEW" IN THE TABLE), KURTOSIS, ARCH LM TEST PROBABILITY VALUE, AND BERA-JARQUE NORMALITY TEST PROBABILITY VALUE ARE SHOWN

Method	Variance Forecasts		Standardized Returns				
	Mean $\times 10^{-5}$	Std $\times 10^{-5}$	Std	Skew	Kurtosis	ARCH Test p-Value	Normality Test p-Value
Raw data	6.702	14.597	1.000	0.350	5.749	0.000	0.000
MAV	6.700	6.286	1.049	0.067	3.588	0.306	0.000
GARCH	6.681	5.345	1.000	0.083	3.419	0.652	0.000
Average	6.691	5.768	1.019	0.077	3.481	0.653	0.000
Kernel	6.564	5.523	0.995	0.097	3.403	0.275	0.000
OLS	6.702	5.120	0.991	0.100	3.475	0.489	0.000
EP NN(B)	6.702	5.537	1.001	0.079	3.346	0.636	0.000
EP-NN(W)	6.702	5.381	1.012	0.054	3.386	0.254	0.000
EP-NN(M)	6.702	5.482	1.029	0.009	3.433	0.659	0.000
SEP-NN	6.702	5.437	1.003	0.061	3.339	0.656	0.001

(a)

Method	Variance Forecasts		Standardized Returns				
	Mean $\times 10^{-5}$	Std $\times 10^{-5}$	Std	Skew	Kurtosis	ARCH Test p-Value	Normality Test p-Value
Raw data	8.331	15.514	1.000	0.062	4.468	0.000	0.000
MAV	8.394	4.773	1.053	0.051	4.329	0.768	0.000
GARCH	8.247	3.887	1.028	0.013	4.223	0.370	0.000
Average	8.321	4.272	1.036	0.029	4.257	0.637	0.000
Kernel	8.187	4.086	1.033	0.011	4.342	0.403	0.000
OLS	8.217	3.654	1.023	0.015	4.263	0.569	0.000
EP-NN(B)	8.314	4.008	1.025	0.028	4.400	0.138	0.000
EP-NN(W)	8.423	3.801	1.019	-.017	4.347	0.388	0.000
EP NN(M)	8.484	3.775	1.018	0.047	4.440	0.602	0.000
SEP-NN	8.408	3.632	1.018	0.022	4.318	0.578	0.000

(b)

evidence of ARCH effects. Further, assuming conditional normality, as is often done in empirical studies of return volatility, implies that the standardized return residuals should be normally distributed. The last column of Table I contains the p-value of the Bera-Jarque normality test on the standardized return residuals.[15] These tests for normality show a statistically significant deviation from the normal distribution. This is a widely documented feature of stock return data, not peculiar to a particular stock index or time period [3].

These summary statistics and tests indicate that all of the forecasting and combining methods perform reasonably, though the assumption of conditional normality of the data is clearly violated. The lack of conditional normality impinges on the efficiency of the GARCH method, but does not lead to inconsistent estimation or biased forecasting, and so this failure is of second-order significance.

[15]The Bera-Jarque test is an LM test for normality, based on standardized third and fourth moments [28].

TABLE II

THE ROOT MEAN SQUARED FORECAST ERROR AND MEAN ABSOLUTE FORECAST ERROR FOR ALL THE MODELS, FOR NEW YORK'S S&P 500 INDEX ON DAILY DATA, IN-SAMPLE 1969:4-1979:12 AND OUT-OF-SAMPLE 1980:1-1987:9

Method	In-Sample, 1969:4-1979:12		Out-of-Sample, 1980:1-1987:9	
	RMSFE $\times 10^{-4}$	MAFE $\times 10^{-3}$	RMSFE $\times 10^{-4}$	MAFE $\times 10^{-3}$
MAV	1.375	8.210	1.545	9.365
GARCH	1.368	8.175	1.530	9.292
Average	1.369	8.183	1.535	9.323
Kernel	1.287	8.024	1.538	9.283
OLS	1.367	8.177	1.531	9.279
EP-NN(B)	1.350	8.166	1.535	9.321
EP-NN(W)	1.357	8.180	1.532	9.345
EP-NN(M)	1.353	8.190	1.535	9.351
SEP-NN	1.354	8.172	1.532	9.323

B. Summary Measures of Performance

The traditional summary measures of the forecasting methods is a comparison of RMSFE and MAFE across competing methods. Although all of our methods were picked with a MSE criterion method, comparison by mean absolute deviation is typically considered interesting because researchers often have little confidence in their choice of criteria to minimize. If we knew beyond a shadow of a doubt that the data were normally distributed, there would be no interest in MAFE, but as we typically do not, alternative measures of goodness of fit to MSE are provided. Table II contains the root mean squared forecast error and mean absolute forecast error for each of the individual models and each of the combining methods for the in-sample period April 1, 1969 to December 31, 1979, and the out-of-sample period, January 1, 1980 to September 30, 1987.

Since the parametric modeling approaches provide the greatest degrees of freedom, the EP methods naturally perform best in-sample among the parametric models. The nonparametric Kernel method performs best overall in-sample. It is the out-of-sample behavior that is informative, for if a method "over-fits" the data, the out-of-sample performance typically falls far short of the in-sample performance. What we find in Table II is no evidence of this problem with the EP-NN forecasts, from the best, worst, or median in-sample MSE ranked models, and similarly no such problem with the SEP-NN. On the out-of-sample data these methods are in the middle of the pack, with lower RMSFE's than the MAV method, the worst performer, and not much higher RMSFE's than the best performer GARCH. Similar rankings come from MAFE measures. We also see no problem of overfitting with the Kernel method. It is in the middle of the pack by measure

TABLE III
THE PROBABILITY VALUES OF THE TESTS OF ENCOMPASSING ON THE ONE-STEP-AHEAD FORECASTED CONDITIONAL VARIANCES, $\hat{\sigma}_t^2$, FOR NEW YORK'S S&P 500 INDEX ON DAILY DATA 1980–1987. THE TESTS ARE ON θ_1 PARAMETER IN THE REGRESSION $\hat{\xi}_{j,t} = \theta_0 + \theta_1 \hat{\sigma}_{k,t}^2 + \eta_t$ WHERE $\hat{\xi}_{j,t} = \hat{\epsilon}_t^2 - \hat{\sigma}_{j,t}^2$ IS MODEL j'S OUT-OF-SAMPLE FORECAST ERROR AND $\hat{\sigma}_{k,t}^2$ IS MODEL k'S OUT-OF-SAMPLE FORECAST. COLUMNS 2–10 (COLUMNS 1–9 OF THE DATA ENTRIES) CONTAIN p-VALUES ASSOCIATED WITH THE t-STATISTICS ON θ_1 FOR ALL POSSIBLE $j-k$ COMPARISONS

Error	Forecast								
	MAV	GARCH	Average	Kernel	OLS	EP-NN(B)	EP-NN(W)	EP-NN(M)	SEP-NN
MAV	—	.001	.000	.001	.000	.000	.000	.000	.000
GARCH	.021	—	.024	.023	.025	.022	.025	.010	.013
Average	.000	.005		.004	.002	.003	.004	.001	.001
Kernel	.038	.052	.041	—	.035	.023	.034	.012	.017
OLS	.060	.141	.085	.091	—	.083	.090	.044	.049
EP-NN(B)	.020	.029	.022	.012	.017		.009	.004	.007
EP-NN(W)	.073	.103	.081	.068	.063	.035	—	.019	.025
EP-NN(M)	.071	.100	.079	.053	.062	.039	.042		.020
SEP-NN	.094	.183	.123	.113	.105	.102	.086	.037	

of RMSFE, and it has the best MAFE measure on the out-of-sample data. Comparison by RMSFE and by MAFE provides us with no indication of whether any one model is performing significantly better than the other models, however. We, therefore, investigate in the next section an additional means of comparison between forecasting models that allows for tests of significance: comparison by forecast encompassing.

C. Tests of Forecast Encompassing

Encompassing-in-forecast tests [5], [21] revolve around the intuition that a Model j should be preferred to a Model k if Model j can explain what Model k cannot explain, without Model k being able to explain what Model j cannot explain. Encompassing-in-forecast tests are designed to provide a statistically significant test of this characteristic. As such, the test provides an obvious method for ranking forecasts.

A set of OLS regressions of the out-of-sample forecast error from one model on the out-of-sample forecast from the other provide the formal test for encompassing-in-forecast. Let $\hat{\xi}_{j,t} = \hat{\epsilon}_t^2 - \hat{\sigma}_{j,t}^2$ be Model j's out-of-sample forecast error and $\hat{\sigma}_{k,t}^2$ be Model k's out-of-sample forecast. The tests for encompassing involve testing for significance of the θ_1 parameter in the regression in (6)

$$\hat{\xi}_{j,t} = \theta_0 + \theta_1 \hat{\sigma}_{k,t}^2 + \eta_t. \tag{6}$$

To test the null hypothesis that neither model encompasses the other we perform two regressions. Regress the out-of-sample forecast error from Model j on the out-of-sample forecast from Model k, as in (6), regression "jk." Call the resulting estimate of the θ_1 coefficient $\hat{\theta}_1(jk)$. Call $\hat{\theta}_1(kj)$ the θ_1 estimate that results from the analogous regression kj. If $\hat{\theta}_1(jk)$ is *not* significant, but $\hat{\theta}_1(kj)$ *is* significant, then we reject the null hypothesis that neither model encompasses the other in favor of the alternative hypothesis that Model j encompasses Model k. We say that Model k encompasses Model j if, conversely, $\hat{\theta}_1(kj)$ is *not* significant, but $\hat{\theta}_1(jk)$ *is* significant. We fail to reject the null hypothesis that neither model encompasses the other in forecast if both $\hat{\theta}_1(kj)$ and $\hat{\theta}_1(jk)$ are significant, or if both $\hat{\theta}_1(kj)$ and $\hat{\theta}_1(jk)$ are not significant. Nonoverlapping information sets may lead to both estimated coefficients being significant, and multicollinearity may lead to both estimated coefficients being insignificant.

Columns 2–10 of Table III (columns 1–9 of the data entries) contain p-values associated with the heteroskedasticity robust t-statistics on θ_1 for all possible $j-k$ comparisons.[16] P-values less than 0.01 reveal that the out-of-sample forecast from the model listed along the top of the table explains, at the 1% significance level, the out-of-sample forecast error from the model listed down the left side of the table and thus that the model listed down the side cannot encompass the model listed along the top, at the 1% level of significance.

These results provide strong evidence of the superiority of the SEP-NN method over the competing linear methods. The SEP-NN out-of-sample forecasts encompass all the competing linear methods as well as the Kernel method at the 5% level of significance or better. The SEP-NN θ_1 coefficient [refer to (6)] is significant at the 0.1% level in the MAV out-of-sample forecast error regression, at the 0.1% level in the average out-of-sample forecast error regression, at the 1.3% level in the GARCH out-of-sample forecast error regression, at the 1.7% level in the Kernel out-of-sample forecast error regression, at the 4.9% level in OLS out-of-sample forecast error regression, while never having the SEP-NN out-of-sample forecast error explained at better than the 9.4% level. With respect to the other ANN combination methods, SEP-NN encompasses EP-NN(B) and EP-NN(W) at the 5% level as well, but not the EP-NN(M), suggesting that EP-NN(M) and SEP-NN are possibly picking up slightly different effects. No other method does nearly this well. The EP-NN(M) and OLS encompass

[16] These statistics make use of a modification of the heteroskedasticity-robust covariance matrix estimator [51] termed "HC3" in [35]. We expect to have heteroskedastic errors in this regression, so it is important to account for heteroskedasticity. A simple demonstration of this is as follows. The residual ϵ_t can be rewritten as $\sigma_t \eta_t$ where η_t is i.i.d., with mean zero and variance one. The out-of-sample forecast error in predicting ϵ_t^2 (an object we do not actually observe, but rather estimate) making use of σ_t^2, is $\epsilon_t^2 - \sigma_t^2 = \sigma_t^2(\eta_t^2 - 1)$ so that the forecast is unbiased, $E[\epsilon_t^2 - \sigma_t^2] = 0$, but the forecast error, of even the optimal forecast, is heteroskedastic.

only five of the eight alternative models at the 5% level, and OLS does not encompass these as strongly as does SEP-NN. EP-NN(W) encompasses four, the Kernel only one, and GARCH encompasses none. The MAV out-of-sample forecast is routinely encompassed at the 0.1% level or better, the average at the 0.5% level or better, neither encompassing a single other model.

Also of interest is that all of the EP models we looked at performed well, suggesting that the pseudorandom numbers used to evolve the networks were not critical in finding a model which would perform well on out-of-sample criteria. The EP-NN model which had the best in-sample MSE, EP-NN(B), did perform somewhat worse than the other ANN models on the out-of-sample encompassing tests. This suggests that there may be some problems with overfitting even if the neural network architecture includes only three nodes as all our EP models do.

Of further and related interest is the fact that we also looked at performance statistics for several SEP-NN models, those of the best evolved network, the median, and worst, though we report only that of the best. Interestingly, the SEP procedure showed no symptoms of overfitting in such comparisons, with each network performing similarly well. This indicates that our choice of architecture and training algorithm successfully compromised between functional and parametric flexibility and the tendency to overfit.[17]

VIII. Discussion

As the forecasts from all the models are a function of only two variables—the MAV and GARCH inputs—we may plot the three-dimensional surface that maps these inputs into the combined forecast. Fig. 3(a)–(d) plots the SEP-NN and OLS combining surfaces estimated in-sample together with the data points used in this estimation. Fig. 3(a) shows only the SEP forecasting surface, (b) shows only the OLS forecasting surface, (c) shows only the data points, and (d) shows all the surfaces and data in a single plot.

The forecasts are normalized so that the average value is equal to one, as in Figs. 1 and 2. The curved surface of diamonds plots the functional relationship—estimated on the in-sample data April 1969 to December 1979—between the MAV and GARCH individual forecasts and the SEP-NN combined forecast. The nonlinearity of the SEP-NN produces the curved relationship shown in Fig. 3(a) and (d). Similarly, in Fig. 3(b) and (d), the flat surface of circles shows the OLS combined forecast. The pillars rising up from the floor of Fig. 3(c) and (d) indicate actual in-sample data points which were used in estimating the OLS and EP-NN and SEP-NN models (the height of the pillars is slightly raised so that they are easily visible through the surfaces). The flagged pillars (roughly half of all the data points) are data points for which the SEP-NN combination produces a smaller error than does the OLS combination. The unflagged pillars are data points for which the OLS combination produces a smaller error than does the SEP-NN. The surfaces in Fig. 3(a) and (b) indicate the difference in response we can expect from the SEP-NN and OLS models for various pairs of GARCH-MAV forecast inputs, while the pillars in Fig. 3(c) and (d) provide some indication as to whether or not the differences in response are relevant for the in-sample period. We can produce analogous surfaces for the out-of-sample period but these are less interesting. This is because the models are re-estimated as we "move" through the data with the rolling window updating procedure outlined in Section VI, and hence the surface changes over time in the out-of-sample period.

As displayed in Fig. 3(d), most of the data points occur in the area where the SEP-NN and OLS forecasting functions intersect, thus the similarity of the SEP-NN and OLS forecast summary statistics of Table I(a). We also see, however, that numerous data points occur away from the curve intersections, in particular when one or both of the MAV and GARCH forecasts are large in magnitude. In these areas we get some intuition for what the SEP-NN combination is doing that is different than the OLS combination. For instance, the SEP-NN combination makes very little adjustment in its forecast as the GARCH falls below the value of the MAV forecast when they are both large (greater than four) and in fact increases its forecast when the MAV forecast is close to four and the GARCH forecast is greater than two but falling. In this area we are riding up the curve in the SEP-NN surface. We see such a coincidence of forecasts in Fig. 1 in June of 1970 and in October 1974 displayed in Fig. 2. Very short periods of quiet in the midst of high volatility will produce such patterns with sharply falling GARCH forecasts but little change in the MAV forecast. The OLS forecast is constrained, because of its linearity, to treat such periods the same way it treats all periods, moving up and down primarily with changes in the GARCH forecast (the slope in the MAV axis is quite small). It is in these periods, where the SEP-NN forecast reacts little to changes in the GARCH forecast, that the SEP-NN forecasts also dominate the OLS forecasts. This is displayed by the predominance of flagged data points in the back middle of Fig. 3(c) and (d) where both GARCH and MAV forecasts are large.

The SEP-NN forecast also reacts quite strongly to large values of the MAV forecast, in particular if they are associated with moderate values (3-5) of the GARCH forecast, as seen in the steeply rising section of the SEP-NN surface in Fig. 3(a) and (d). In this steeply rising section of the SEP-NN surface, the back left of Fig. 3(d), the OLS forecasts were as often associated with smaller errors as were the SEP-NN forecasts, perhaps indicating that the architecture of the SEP-NN model, with three nodes, was insufficient to allow it to moderate reaction to falling GARCH forecasts and stable MAV forecasts when the MAV forecast was greater than four and the GARCH forecast was between three and five. An example of this is seen in Fig. 2 in November of 1974.

IX. Conclusions

We have described an experiment which evolves single hidden-layer perceptrons to combine forecasts of stock price volatility, demonstrating the utility of evolutionary programming in a concrete application. Such forecasting is important in its own right, in particular in the context of financial

[17] Full details of evolved parameters are available from the authors.

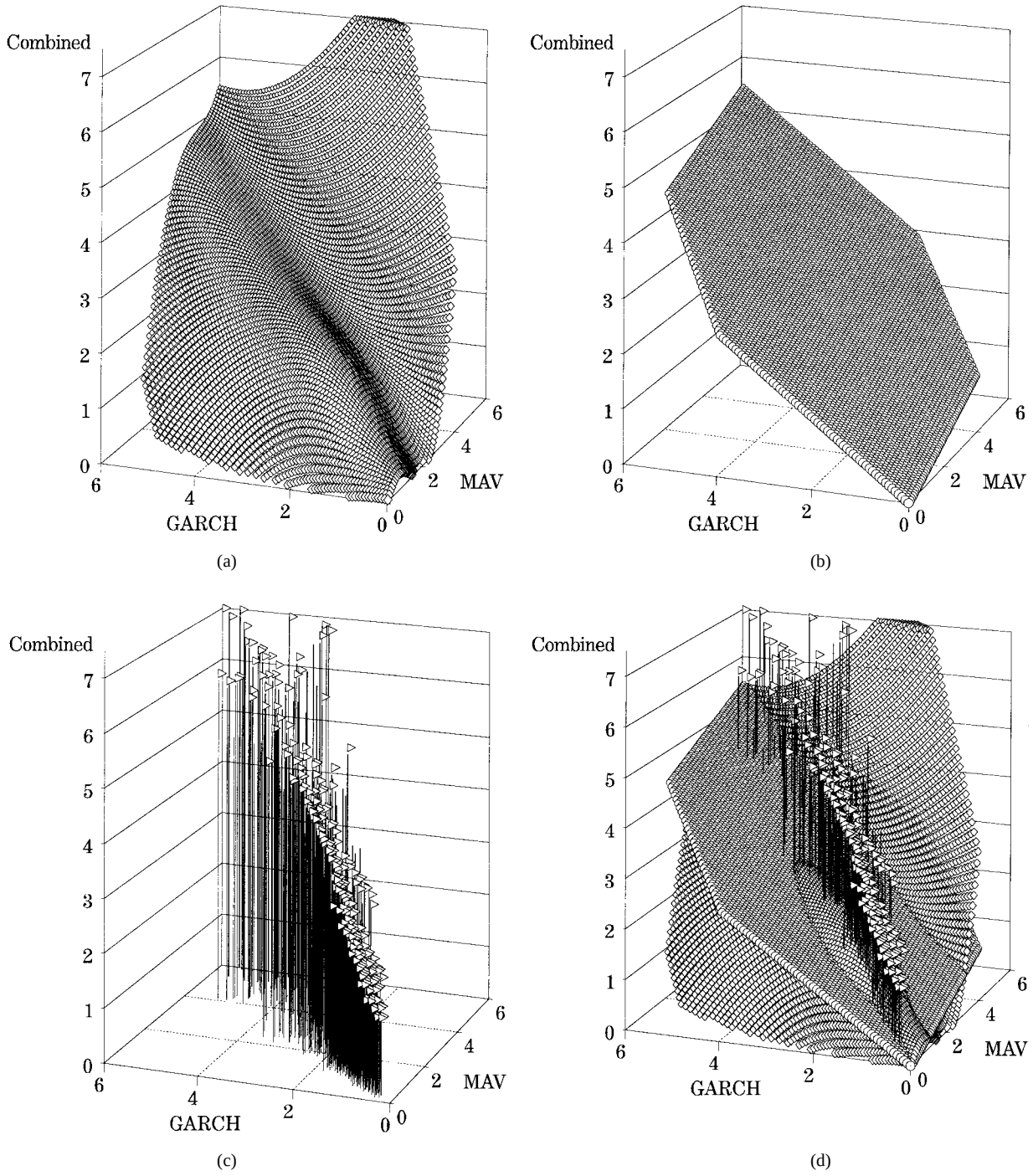

Fig. 3. Plots the SEP-NN and OLS combining surfaces with data points from 1969–1979. The forecasts are normalized so that the average value is equal to one. The curved surface of diamonds plots the functional relationship between the MAV and GARCH individual forecasts and the SEP-NN forecast. The flat surface of circles shows the OLS combined forecast. (a) shows only the SEP forecasting surface (diamonds), (b) shows only the OLS forecasting surface (balloons), (c) shows only the data points (pillars), and (d) shows all the surfaces and data in a single plot with SEP forecasts superior to OLS flagged.

risk management. Evolved ANN models were shown to have a strong advantage over simple linear models and a nonparametric Kernel method. A self-adaptive scheme was also shown to yield some benefits over more simple evolutionary programming schemes.

Careful examination based on statistical tests revealed that the evolved networks were significantly superior to linear combination methods described in the forecast combination literature. It should be emphasized, however, that the results presented cover only one example, and it remains to be seen if they are generally more applicable. It is also hoped that further research might discover methods of dynamic parameter adjustment, so that the evolutionary programs are more sensitive in monitoring their own performance throughout a trial. Evolving the nonlinear ANN weights as new information flows in, rather than just the linear weights as done here, is not computationally feasible with our large data set, but is a natural next step to take. This application also does not exhaust

the pool of applications for which these combining techniques may be used, in finance and beyond.

ACKNOWLEDGMENT

The authors thank D. Fogel, R. G. Donaldson, P. Gomme, L. Kramer, and in particular the anonymous referees for their help.

REFERENCES

[1] J. M. Bates and C. W. J. Granger, "The combination of forecasts," *Oper. Res. Quart.*, vol. 20, pp. 1–68, 1969.
[2] H. J. Bierens, "Comment on artificial neural networks: An econometric perspective," *Econometric Rev.*, vol. 13, pp. 93–97, 1994.
[3] T. Bollerslev, R. T. Chou, and K. F. Kroner, "ARCH modeling in finance: a review of the theory and empirical results," *J. Econometrics*, vol. 54, pp. 1–30, 1991.
[4] W. S. Chan and H. Tong, "On tests for nonlinearity in time series analysis," *J. Forecasting*, vol. 5, pp. 217–228, 1986.
[5] Y. Y Chong and D. F. Hendry, "Econometric evaluation of linear macroeconomic models," *Rev. Economic Stud.*, vol. 53, pp. 671–690, 1986.
[6] R. T. Clemen, "Combining forecasts: a review and annotated bibliography," *Int. J. Forecasting*, vol. 5, pp. 559–583, 1989.
[7] R. G. Donaldson and M. Kamstra, "Using artificial neural networks to combine financial forecasts," *J. Forecasting*, vol. 15, pp. 49–61, 1996.
[8] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, pp. 987–1007, 1986.
[9] R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss, "Semiparametric estimates of the relationship between weather and electricity sales," *J. Amer. Statist. Assoc.*, vol. 81, pp. 310–320, 1986.
[10] L. Fisher, "Some new stock market indices," *J. Business*, vol. 29, pp. 191–225, 1966.
[11] M. B. Fishman, D. S. Barr, and W. L. Loick, "Using neural nets in market analysis," *Tech. Analysis of Stocks and Commodities*, Apr., pp. 18–22, 1991.
[12] D. B. Fogel, "Applying evolutionary programming to selected traveling salesmen problems," *Cybern. Syst.*, vol. 24, pp. 27–36, 1993.
[13] ——, "Applying evolutionary programming to selected control programs," *Comp. Math. Appl.*, vol. 27, pp. 89–104, 1994.
[14] ——, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. Piscataway, NJ: IEEE Press, 1995.
[15] D. B. Fogel, L. J. Fogel, and W. Atmar, "Meta-evolutionary programming," in *Proc. 25th Asilomar Conf. on Signals, Systems and Computers*. Pacific Grove, CA: Maple, 1991, pp. 540–545.
[16] D. B. Fogel, L. J. Fogel, and V. W. Porto, "Evolving neural networks," *Biological Cybern.*, vol. 63, pp. 487–493, 1990.
[17] L. J. Fogel, A. J. Owens, and M. J. Walsh, *Artificial Intelligence through Simulated Evolution*. New York: Wiley, 1966.
[18] D. K. Gehlhaar and D. B. Fogel, "Tuning evolutionary programming for conformationally flexible modular docking," in *Evolutionary Programming V*. Cambridge, MA: MIT Press, 1997, pp. 419–432.
[19] C. W. J. Granger and R. Ramanathan, "Improved methods of combining forecasts," *J. Forecasting*, vol. 3, pp. 197–204, 1984.
[20] C. W. J. Granger, "Invited review: Combining forecasts—Twenty years later," *J. Forecasting*, vol. 8, pp. 167–173, 1989.
[21] J. Hallman and M. Kamstra, "Combining algorithms based on robust estimation techniques and cointegrating restrictions," *J. Forecasting*, vol. 8, pp. 189–198, 1989.
[22] W. Härdle, *Applied Nonparametric Regression*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[23] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computing*. Redwood City, CA: Addison-Wesley, 1991.
[24] U. Hjorth and L. Holmqvist, "On model selection based on validation with applications to pressure and temperature prognosis," *Appl. Statist.*, vol. 30, pp. 264–274, 1981.
[25] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251–257, 1991.
[26] K. Hornik, M. Stinchcombe, and H. White, "Multi-layer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.
[27] ——, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, pp. 551–560, 1990.
[28] M. Jarque and A. Bera, "Efficient tests for normality, homoskedasticity and serial independence of regression residuals," *Econ. Lett.*, vol. 6, pp. 255–259, 1980.
[29] L. Kavalieris, "The estimation of the order of an autoregression using recursive residuals and cross-validation," *J. Times Series Anal.*, vol. 10, pp. 271–28, 1989.
[30] D. M. Keenan, "A Tukey nonadditive type test for time series nonlinearity," *Biometrica*, vol. 72, pp. 39–44, 1985.
[31] H. Kitano, "Designing neural networks using genetic algorithms with graph generation system," *Complex Syst.*, vol. 4, 1990.
[32] C. Kuan and H. White, "Artificial neural networks: an econometric perspective," *Econometric Rev.*, vol. 13, pp. 1–91, 1994.
[33] T. Lee, C. W. J. Granger, and H. White, "Testing for neglected nonlinearity in time series models: a comparison of neural network methods and alternative tests," *J. Econometrics*, vol. 56, pp. 269–290, 1993.
[34] J. R. McDonnell and D. Waagen, "Neural network structure design by evolutionary programming," in *Proc. Second Annu. Conf. Evolutionary Programming*, La Jolla, CA: Evolutionary Programming Society, 1993, pp. 79–89.
[35] J. G. MacKinnon and H. White, "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *J. Econometrics*, vol. 48, pp. 817–838, 1985.
[36] G. F. Miller, M. Todd, and S. U. Hegde, "Designing neural networks using genetic algorithms," in *Proc. Third Int. Conf. Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann, 1989, pp. 379–384.
[37] C. Min and A. Zellner, "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates," *J. Econometrics*, vol. 56, pp. 89–118, 1993.
[38] D. Montana and L. Davis, "Training feedforward neural networks using genetic algorithms," in *Proc. 11th Int. Joint Conf. Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann, 1989, pp. 762–767.
[39] A. Pagan and W. Schwert, "Alternative models for conditional stock volatility," *J. Econometrics*, vol. 45, pp. 267–290, 1990.
[40] P. M. Robinson, "Automatic frequency domain inference on semiparametric and nonparametric models," *Econometrica*, vol. 59, pp. 1329–1364, 1991.
[41] N. Saravanan, "Evolving neural networks: application to a prediction problem," in *Proc. 2nd Annu. Conf. Evolutionary Programming*, La Jolla, CA: Evolutionary Programming Society, 1993, pp. 72–78.
[42] N. Saravanan and D. B. Fogel, "Learning strategy parameters in evolutionary programming: An empirical study," in *Proc. Third Annu. Conf. Evolutionary Programming*. River Edge, NJ: World Scientific, 1994, pp. 269–280.
[43] N. Saravanan, D. B. Fogel, and K. M. Nelson, "A comparison of methods of self-adaptation in evolutionary algorithms," *BioSystems*, vol. 36, pp. 157–166, 1995.
[44] M. Scholes and J. Williams, "Estimating betas from nonsynchronous data," *J. Finan. Econ.*, vol. 5, pp. 309–327, 1977.
[45] H-P. Schwefel, *Numerical Optimization of Computer Models*. Chichester, U.K.: Wiley, 1981.
[46] M. Stinchcombe and H. White, "Using feedforward networks to distinguish multivariate populations," in *Proc. Int. Joint Conf. Neural Networks*, ICANN'94, Sorrento, Italy, May 26-29, 1994, pp. 7–16.
[47] J. H. Stock, "Nonparametric policy analysis: An application to estimating hazardous waste cleanup benefits," in *Nonparametric and Semiparametric Methods in Economics and Statistics*, W. Barnett, J. Powell, and G. Tauchen, Eds. Cambridge: Cambridge Univ. Press, 1991.
[48] P. Stoica, P. Eykhoff, P. Janssen, and T. Soderstrom, "Model selection by cross-validation," *Int. J. Contr.*, vol. 43, pp. 1841–1878, 1986.
[49] T. Subba Rao and M. Gabr, "A test for linearity of stationary time series," *J. Time Series Anal.*, vol. 1, pp. 145–158, 1980.
[50] R. S. Tsay, "Non-linearity tests for time series," *Biometrica*, vol. 73, pp. 461–466, 1986.
[51] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, pp. 817–838, 1980.
[52] ——, "Neural network learning and statistics," *AI Expert*, pp. 48–50, 1989.
[53] ——, "Some asymptotic results for learning in single hidden layer feedforward network models," *J. Amer. Statist. Assoc.*, vol. 84, pp. 1003–1013, 1989.
[54] ——, "Connectionist nonparametric regression multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, vol. 3, pp. 535–549, 1990.
[55] D. Wolpert, "Combining generalizers using partitions of the learning set," Santa Fe Institute, Santa Fe, NM, Work Paper Series 93-02-009, 1993.

Paul G. Harrald was educated in the United Kingdom and Canada.

He is currently a Lecturer in business economics at the Manchester School of Management, Manchester, U.K. His recent interests are in numerical optimization, and in particular in evolutionary techniques. His other research interests include artificial life models, evolutionary games, and computable economic models.

Mark Kamstra received the Ph.D. degree in economics from the University of California at San Diego.

He is currently an Assistant Professor of Economics at Simon Fraser University, Burnaby, B.C., Canada. Recent research interests include empirical asset pricing models, conditional variance forecasting, and the combination of forecasts, both quantitative and qualitative. A common thread through much of his work is the application of artificial neural networks.