Combining forecasts based on multiple encompassing tests in a macroeconomic core system

MAURO COSTANTINI[1] and ROBERT KUNST[2*]

[1]Department of Economics, University of Vienna, Vienna
[2]Department of Economics and Finance, Institute for Advanced Studies, and Department of Economics, University of Vienna

Abstract

This paper investigates whether and to what extent multiple encompassing tests may help determine weights for forecast averaging in a standard vector autoregressive setting. To this end we consider a new test-based procedure, which assigns non-zero weights to candidate models that add information not covered by other models. The potential benefits of this procedure are explored in extensive Monte Carlo simulations using realistic designs that are adapted to U.K. and to French macroeconomic data, to which trivariate vector autoregressions (VAR) are fitted. Thus simulations rely on potential data-generating mechanism for macroeconomic data rather than on simple but artificial designs. We run two types of forecast 'competitions'. In the first one, one of the model classes is the trivariate VAR, such that it contains the generating mechanism. In the second specification, none of the competing models contains the true structure. The simulation results show that the performance of test-based averaging is comparable to uniform weighting of individual models. In one of our role-model economies, test-based averaging achieves advantages in small samples. In larger samples, pure prediction models outperform forecast averages.

KEY WORDS Combining forecasts, encompassing tests, model selection, time series.

*Correspondence to: Department of Economics and Finance, Institute for Advanced Studies, Stumpergasse 56, 1060 Vienna, Austria, E-mail: kunst@ihs.ac.at

INTRODUCTION

It is now well established that forecast combinations often improve forecast accuracy. That is, a linear combination of two or more predictions may yield more accurate forecasts than any of the individual candidates if it successfully extracts useful and independent information from the component forecasts. The remark by Flores and White (1989) that 'there is still concern as to the preferred method of combination', however, continues to be valid today.

Starting from the seminal contribution by Bates and Granger (1969), researchers have considered various methods to determine the individual weights in forecast averages, such as uniform weights, weights derived from information criteria, or weights based on regression over training samples. Several studies found a remarkably strong performance for 'naive' uniform weighting, while de Menezes and Bunn (1993) show that the relative performance of more sophisticated forecast combinations may be sensitive to the stability of the correlation structure. For recent surveys on the voluminous literature, see Clements and Hendry (1998) and Timmermann (2006), whereas Clemen (1989) provides a good survey on the early contributions. Generally, as Kisinbay (2007) pointed out, two methodological strands can be distinguished: combinations that retain all candidates, and combinations that use an elimination step to discard inferior candidates. While the majority of contributions adhere to the former option, we consider a strategy of the latter kind that assigns weights for forecast combinations on the basis of forecast encompassing tests.

Methods that incorporate a selection step before combining the forecasts were considered by Hallman and Kamstra (1989), Chen and Anandalingam (1990), Chandrasekharan *et al.* (1994), Swanson and Zeng (2001), and recently Kisinbay (2007), among others. This option is often supported by arguments of the cost of keeping additional candidates and of increased predictive accuracy due to faster elimination of inferior rivals, as the sample size increases. To our knowledge, Hallman and Kamstra (1989) were the first to use encompassing tests in combining forecasts. Swanson and Zeng (2001) build on the regression technique of Granger and Ramanathan (1984) and use model selection based on information criteria to optimize the forecast combination. The approach most closely related to ours is that of Kisinbay (2007) who discards encompassed rival forecasts on the basis of the forecast-encompassing test of Harvey *et al.* (1998) applied in one direction and then averages the remaining set. We use a comparable procedure that, however, attains complete symmetry by running a multiple encompassing test of Harvey and Newbold (2000) in all directions and accounts explicitly for cases where all or none of the rivals are encompassed.[1]

As Granger (1989) pointed out, encompassing tests are a more adequate tool than simple tests on improved accuracy in combining forecasts, as the dominance of one forecast by another one is neither sufficient nor necessary for settling the question of whether or not to combine. In short, a forecast is said to encompass a rival forecast if the rival forecast does not contribute to a reduction in forecast loss and hence the encompassed forecast receives

[1]Costantini and Pappalardo (forthcoming) provide a formal proof on the consistency of the use of the encompassing test in one direction proposed by Kisinbay (2007).

a zero weight in a forecast combination. In our procedure models that are encompassed by their competitors are first discarded and then a new combined forecast as the arithmetic mean of the predicted values from the retained models is formed.

In order to investigate whether and to what extent this procedure may help determine the weights for forecast averaging in a standard vector autoregressive setting we use extensive Monte Carlo simulations. In addition to the forecast combination procedure, our simulation design is another contribution of this paper. Most forecast comparisons are based either on Monte Carlo with simple and artificial designs or on limited data. We combine the two approaches by simulating time-series structures that closely resemble actual data. Specifically, we adapt our simulation designs to trivariate core systems for U.K. and French macroeconomic data, to which vector autoregressions (VAR) are fitted. We study all procedures in two types of simulation experiments. In the first experiment, one of the model classes is the trivariate VAR that contains the generating mechanism, albeit the parameters are treated as unknown. Harvey and Newbold (2005) have demonstrated that the data-generating process does not necessarily forecast-encompass its misspecified rivals in small samples. In the second experiment, none of the competing models contains the true structure. We view this specification as the more realistic one, as in typical empirical applications the true data-generating process will be more complex than any of the utilized prediction models.

Generally, we find that the performance of the test-based averaging of forecasts is comparable to a simple uniform weighting of individual models. The experiment based on the French role-model economy reveals some advantages for test-based averages in small samples. Benefits of averaging are strongest for the smallest investigated samples of $N = 40$. In large samples, pure prediction models considerably outperform forecast averages, as the encompassing tests do not eliminate poor prediction models fast enough for increasing sample size.

The plan of this paper is as follows. The next section describes the new forecast averaging procedure for combining forecasts. The third section outlines the simulation design and the backdrop data. The fourth section reports on the simulation results. The fifth section concludes.

THE ENCOMPASSING TEST PROCEDURE

This section presents the encompassing test procedure used to determine the weights of combination forecast. The procedure is based on the multiple forecast encompassing F–test developed by Harvey and Newbold (2000).

Consider M model-based forecasts formed from estimated structures within M model classes. The aim is to forecast a specific component within a given vector variable Y. The M candidate models yield series of out-of-sample forecasts $\hat{Y}_{jt}^{(k)}$ and of forecast errors $e_{jt}^{(k)} = Y_{jt} - \hat{Y}_{jt}^{(k)}$, $k = 1, \ldots, M$, for any component j of the considered vector variable.

The simulation experiment studies the prediction of a single specific variable in the vector Y that without loss of generality can be chosen as the first, Y_{1t}. This allows restricting the evaluation of forecasts to the univariate mean-squared error criterion. Suppressing the series index, denote the forecast errors series from model k for a given sample of length N as $e_t^{(k)}$ with $t = N - n + 1, \ldots, N$, where n is the length of an evaluation sample such that $n << N$.

The encompassing test procedure is based on M encompassing regressions:

$$\begin{aligned}
e_t^{(1)} &= a_1(e_t^{(1)} - e_t^{(2)}) + a_2(e_t^{(1)} - e_t^{(3)}) + \ldots + a_{M-1}(e_t^{(1)} - e_t^{(M)}) + u_t^{(1)}, \\
e_t^{(2)} &= a_1(e_t^{(2)} - e_t^{(1)}) + a_2(e_t^{(2)} - e_t^{(3)}) + \ldots + a_{M-1}(e_t^{(2)} - e_t^{(M)}) + u_t^{(2)}, \\
&\ldots \\
e_t^{(M)} &= a_1(e_t^{(M)} - e_{(t)}^{(1)}) + a_2(e_t^{(M)} - e_t^{(2)}) + \ldots + a_{M-1}(e_t^{(M)} - e_t^{(M-1)}) + u_t^{(M)}.
\end{aligned} \tag{1}$$

These homogeneous regressions yield M regression F statistics. A model k is said to forecast-encompass its rivals if the F statistic in the regression with dependent variable $e_t^{(k)}$ is insignificant at a specific level of significance. Following the evidence of the forecast-encompassing tests, weighted average forecasts are obtained according to the following rule. If F–tests reject or accept the null hypotheses in all M regressions, a new forecast will be formed as a uniformly weighted average of all model-based predictions $M^{-1} \sum_{k=1}^{M} \hat{Y}_t^{(k)}$. If some F–tests reject their null, only those models that encompass their rivals are combined in a uniform average forecast.

THE SIMULATION EXPERIMENT

The data

As pointed out in the introduction, we consider two simulation designs. These designs are adapted to trivariate systems for U.K. and French macroeconomic data. All data used in the experiment is taken from the OECD Main Economic Indicators.

We selected the U.K. and France, as these are—together with Germany, which fails to offer long series due to the unification episode—the two largest and most important European economies. The systems include gross domestic product (GDP), the consumer price index (CPI), and the unemployment rate. Forecasting institutions customarily use such low-dimensional purely data-based core systems to obtain extrapolation benchmarks for their econometric or judgmental official forecasts. The choice of the variables is guided by the fact that real economic growth, CPI inflation, and the unemployment rate are the three variables that are most often reported in the mass media and are also maybe the only economic quantities that are known to a general audience. Our prediction experiments focus on real economic growth—i.e. the growth rate of real GDP—which is often regarded as the most important target variable of economic policy.

With regard to the U.K., we use GDP at constant price (volume level), the CPI, and the registered unemployment rate. All series are available at a quarterly frequency and cover the period 1960:1 to 2008:2. According to the source, GDP and the registered unemployment rate have been seasonally adjusted. We prefer the registered unemployment rate to the conventional unemployment rate based on questionnaires, as it covers a much longer time period.

With respect to French data, availability of comparable data restricts the analysis to a much shorter time range, 1978:1-2008:3. The definition of the GDP volume index is comparable to its U.K. equivalent. Because the OECD Main Economic Indicators database does not provide a registered unemployment rate, we use the harmonized unemployment rate. According to the source, GDP and the unemployment rate have been seasonally adjusted. It should be noted that the French data does not include the first price oil shock episode, which may be of interest with regard to enhancing the robustness of our results.

The empirical analysis uses the growth rates of GDP (X) and of CPI (P). The GDP data are transformed using the first differences of logarithms multiplied by four—a quarterly indicator of annual real growth rates. Figure 1 shows that this variable is quite volatile for the U.K. and much less for France. Due to its seasonally adjusted nature, this transformation is preferable to the difference $\log X_t - \log X_{t-4}$, which would imply a repeated de-seasonalization of the series. By contrast, inflation is calculated as $\pi_t = \log P_t - \log P_{t-4}$ in order to eliminate potential seasonality. Finally, unemployment U_t is used without any further transformation. In symbols, we use $Y_t = (4\Delta \log X_t, \pi_t, U_t)'$ or simply $Y_t = (Y_{1t}, Y_{2t}, Y_{3t})'$.

Inflation and the unemployment rate are often subjected to statistical unit-root tests that fail to reject their null, such that both variables are often considered I(1). They are admittedly borderline cases, and for short-term forecasting not too much is lost by viewing these variables as stationary, as long as the implied multivariate time-series models are stable. Generally, we found that structures fitted to the data, such as our backdrop trivariate second-order vector autoregression, are indeed stable in the sense that all their roots are outside the unit circle.

Figure 1 about here

It should be noted that the U.K. and French data only serve as the basis for our simulation experiment. We do not assume that we identify the true data-generating process for these series nor do we intend to really forecast the British or French economies.

The data-based simulation design

For the design of the simulation experiments, trivariate vector autoregressive (VAR) models are fitted to the data. To identify the lag order of the VAR models, we apply the BIC criterion according to Schwarz (1978). This results in a VAR(2) model of the form:

$$Y_t = \mu + \sum_{j=1}^{2} \Phi_j Y_{t-j} + \varepsilon_t,$$

for $t = 3, \ldots, N$.

Parameter estimates for the U.K. data are:

$$\begin{aligned}
\mu &= (1.414, 0.146, 0.047)', \\
\Phi_1 &= \begin{pmatrix} -0.141 & -0.221 & -0.739 \\ 0.012 & 1.406 & -0.359 \\ -0.021 & 0.007 & 1.712 \end{pmatrix}, \\
\Phi_2 &= \begin{pmatrix} -0.025 & 0.062 & 0.785 \\ -0.020 & -0.428 & 0.342 \\ -0.020 & 0.005 & -0.718 \end{pmatrix}.
\end{aligned} \tag{2}$$

In (2), all numbers have been rounded to three decimal digits, while the actual simulation design uses estimates at the machine precision. For these numbers at highest precision, the estimated VAR model has six polynomial roots, two real roots at 0.43 and at 0.95, a complex root pair with a small imaginary part at 0.90, and a mainly imaginary root pair with low modulus of 0.22. In summary, the estimated VAR structure is stable. Some of its coefficient parameters may be statistically insignificant, such that simplification steps may be rewarding.

The corresponding estimates for the French data are:

$$\begin{aligned}
\mu &= (-0.698, 0.617, 0.102)', \\
\Phi_1 &= \begin{pmatrix} 0.237 & 0.241 & 0.035 \\ -0.009 & 1.253 & 0.023 \\ -0.068 & 0.136 & 1.478 \end{pmatrix}, \\
\Phi_2 &= \begin{pmatrix} 0.292 & -0.191 & 0.077 \\ 0.026 & -0.318 & -0.082 \\ -0.037 & -0.113 & -0.482 \end{pmatrix}.
\end{aligned} \tag{3}$$

This model has four real roots at the locations $-0.426, 0.253, 0.543, 0.835$, and an almost real complex root pair at $0.882 \pm 0.019i$. There are several noteworthy differences to the U.K. model. First, evidence on cycles is much weaker, excepting the semi-annual cycle imposed by the negative root. Second, dynamic dependence between GDP growth and inflation is less pronounced, while the connection of GDP growth and unemployment is stronger than in the British case. These subtle aspects are not so easy to recognize from the coefficient structure but they will become obvious in the prediction experiments.

Note that a lag order of two is common or even 'recommended' for role-model macroeconomic systems (see, for example, Juselius, 2006; Lütkepohl, 2005). Alternatively, the popular AIC would yield a much higher lag order, which may indicate that linear VAR models do not capture the dynamics of the observed data too well. The visual correspondence of simulated trajectories with the actual data is satisfactory. From starting values at the end of the actual data, 2008, we now simulate artificial samples ('pseudo-samples') of given length by drawing errors from a normal distribution with variance-covariance matrix Σ for both countries' data. With regard to the U.K. data, the variance-covariance matrix takes the following form:

$$\Sigma = \begin{pmatrix} 2.468 & -0.137 & -0.076 \\ -0.137 & 0.169 & 0.007 \\ -0.076 & 0.007 & 0.015 \end{pmatrix}, \tag{4}$$

which corresponds to the maximum-likelihood estimate from the VAR residuals. At the same time, the diagonal entries of Σ serve as lower boundaries for mean square forecast errors. It should be noted that restricting all simulations to Gaussian random variables ensures that the robustness issues studied by Harvey and Newbold (2000) do not arise.

The analogous matrix for the French data is

$$\Sigma = \begin{pmatrix} 0.423 & 0.015 & -0.015 \\ 0.015 & 0.035 & -0.004 \\ -0.015 & -0.004 & 0.021 \end{pmatrix}, \tag{5}$$

which indicates that variation in the GDP growth rate is far lower in the French series that avoids the turbulence of the OPEC shocks in the 1970s.

The sample size varies from $N = 40$ to $N = 500$, such that it covers the typical sample sizes of economic interest. We note that the sample of the original U.K. data has $N = 194$ and that of the French data has $N = 122$. This may already be at the upper bound of usual macroeconomic analysis, as many empirical researchers tend to consider the possibility of structural breaks and institutional change and focus on shorter samples. We wish to keep the long samples of $N = 500$ to obtain some evidence on large-sample performance, i.e. when estimates get close to their true values or at least asymptotic limits.

From each pseudo-sample, we keep the last $N/4$ observations for evaluating predictions. All predictions are based on time-series models with estimated coefficients. For 40 observations, the lower bound of 30 observations appears to be a binding constraint for useful estimation. The last $N/4 - 1$ observations are then predicted from expanding windows of $t = 1, \ldots, n$ with n varying from $3N/4$ to $N-2$. Thus, the last forecast is based on a more precisely estimated structure than the first, and performance within one pseudo-sample may be dependent. The comparatively large number of 10,000 replications mitigates such potentially disturbing effects. It should be noted that the last observation is not contained in this stage of the prediction experiment. It is reserved for the second stage.

Each experiment considers four rival prediction models, $M = 4$ in the notation of the second section. The four models yield series of forecast errors $e_{jt}^{(k)} = Y_{jt} - \hat{Y}_{jt}^{(k)}$ for $k = 1, \ldots, 4$ and $j = 1, 2, 3$. In our analysis, we are interested only in the prediction of the first variable ($j = 1$), GDP growth. In order to determine the weights of the combination forecast, the encompassing test procedure described in the second section is applied. We run $M = 4$ encompassing regressions for the GDP growth forecast errors, $e_{1t} = e_t$. In the first encompassing regression, $e_t^{(1)}$ is the dependent variable:

$$e_t^{(1)} = a_1(e_t^{(2)} - e_t^{(1)}) + a_2(e_t^{(3)} - e_t^{(1)}) + a_3(e_t^{(4)} - e_t^{(1)}) + u_t. \tag{6}$$

The dependent variable of the second regression is the forecast error of the second model $e_t^{(2)}$:

$$e_t^{(2)} = a_1(e_t^{(1)} - e_t^{(2)}) + a_2(e_t^{(3)} - e_t^{(2)}) + a_3(e_t^{(4)} - e_t^{(2)}) + u_t. \qquad (7)$$

In all these regressions, t runs from $1+3N/4$ to $N-1$. These two encompassing regressions are followed by two more analogous regressions with $e_t^{(k)}$, $k = 3, 4$ on the left side. When the corresponding regression F statistic in (6) is insignificant at a specific level of significance, the first model is said to forecast-encompass its rivals. In our analysis we evaluate the procedure at the customary significance levels of 1%, 5%, and 10%.

Following the evidence of the forecast-encompassing tests procedure, weighted average forecasts are then formed according to the following rule: if all four tests reject or all accept their null hypotheses, the forecast will be a uniformly weighted average of all models; if some F–tests reject their null, only those models that encompass their rivals will be used in an otherwise uniform average. The encompassing tests are applied to the $N/4-1$ predictions that were generated in the first stage. They determine a weighted prediction average for the observation at position N. To assess the performance of the encompassing test procedure, we compare the mean square errors derived from the forecast combination based on the simple uniform weights and those based on the weights selected by the encompassing rule.

EVALUATING PREDICTION BY SETS OF RIVAL MODELS AND COMBINATIONS

This section reports two simulation experiments that investigate the performance of the encompassing test procedure in a realistic environment. In the first experiment, one model class contains the data-generating structure (see Harvey and Newbold, 2005, for a simulation experiment in the case of two competing models). In the second one, all models are 'misspecified'.

A set that includes the generating model

All our forecasts are model-based. They are versions of $\hat{Y}_t$ defined by

$$\hat{Y}_t = \hat{\mu} + \sum_{j=1}^{p} \hat{A}_j Y_{t-j}, \qquad (8)$$

where $\hat{A}$ denotes an estimate of a coefficient matrix. In the following, we use two forms of notation to denote predictions. If no confusion about the prediction horizon can arise, $\hat{Y}_t$ denotes a forecast for Y_t using data up to $t-1$. Alternatively, $\hat{Y}_{t-h}(h)$ is an h–step prediction using information until and including time point $t-h$ for the time point t. This latter notation corresponds to the one used by Chatfield (2001). Note that, for one-step forecasts, $\hat{Y}_t = \hat{Y}_{t-1}(1)$.

Our first experiment uses four model structures: the trivariate autoregression; two bivariate autoregressions, one for the target GDP growth series and inflation ($VAR_{2\pi}$), and one for GDP growth and unemployment (VAR_{2u}); and a univariate autoregression. These models can be expressed by respective restrictions on the matrices $\hat{A}_j$ for all j as follows: unrestricted matrices; elements at (1,3) equal 0; elements at (1,2) equal 0; elements at (1,2) and at (1,3) equal 0.

The lag structures are empirically determined by minimizing BIC, where a maximum lag order $p_{\max}$ is set depending on N. In detail, $p_{\max} = 4$ for $N = 40$, $p_{\max} = 8$ for $N = 100, 200$, and $p_{\max} = 12$ for $N = 500$. These maxima are not often binding, as the BIC search typically finds low lag orders.

In all experiments, we ran unreported control simulations, in which AIC replaced BIC in the lag-order searches. The AIC criterion yields generally worse results. In small samples, AIC tends to identify too large lag orders, and this tendency is even more pronounced in multivariate rather than univariate models. For this reason, the univariate AR model dominates all its rival models convincingly. The critical issue may be related to the approximation in small samples that has given rise to 'corrected' versions, such as AIC_u and AIC_c (see McQuarrie and Tsai, 1998). While such modifications mitigate the underlying problem somewhat, we feel that the stronger penalty of BIC is the better choice in our modelling environment. This is seemingly in contradiction to the traded wisdom that AIC is to optimize asymptotic forecast performance at the cost of over-estimating lag orders.

The first model class contains the DGP. All other models are in the strict sense 'misspecified', as the univariate or bivariate marginal models of a trivariate VAR are ARMA rather than autoregressive and would typically impose an infinite lag order for autoregressive approximations. Clearly, in small samples such approximations can be helpful for prediction, and this presumption will generally be corroborated in the experiments.

Due to the BIC search for lag orders, the four models are non-nested. Thus the anomalies described, for example, by Clark and McCracken (2001) should not arise. In nested models, forecasts based on different models coincide in large samples, which invalidates the standard distributions of encompassing statistics. In our model set with different dimensions, however, the higher-dimensional prediction models have lower lag orders than the lower-dimensional models.

The upper panel of Table 1 reports the mean squared errors (MSE) for evaluating the prediction performance of these four models for the U.K. design. While in the large samples ($N = 500$) the data-generating model class outperforms all its rivals, the bivariate model that includes inflation shows a better performance for moderate samples ($N = 100$), and the parsimonious univariate autoregression is preferred for very small samples. The lower panel gives the results for the France design. These are comparable, with the preferred VAR_{2u} model substituting the $VAR_{2\pi}$ model.

Table 2 reports the performance of the forecast combinations based on the simple uniform weights and of those based on the weights selected by the encompassing rule. In both designs, differences in terms of forecast accuracy are very small. The encompassing test-based weighting outperforms the uniform weighting at $N = 500$ only.

While accuracy smoothly improves as T rises from 40 to 200, there is a drop in accuracy for the largest sample of $T = 500$. Note that it has no parallel in the performance of the pure models reported in Table 1. We also note that for $N = 100$ and $N = 200$ weighted model averages are slightly better than pure models—notwithstanding the mentioned limited comparability between pure and weighted predictions—while this order is reversed for $N = 500$. Potential sources for this feature are the dependence within the replications and also the fact that even the test-based weighting eliminates poor forecasting models rather slowly as N increases (see Table 3).

Typically, the weights are almost uniform for the significance level of 1% and become more specific, as the significance becomes looser. In the case of 10% level, i.e. for the specification with the strongest deviation from uniformity, Table 3 gives the average weights. For small samples, even these weights are close to the uniform distribution with 0.25 allotted to each model. Even at the largest sample size $N = 500$, the 'true' model class achieves less than 40% but starts dominating its rivals.

In the U.K. design, the univariate model is often encompassed and its average weight drops below 10% for $N = 500$. The unsatisfactory performance of the implied average (see Table 2) shows that it is still too often in the set of non-encompassed models and thus deteriorates the prediction MSE relative to the pure trivariate model. In the French design, a similar remark applies to the bivariate model with inflation, whose performance as a pure model tends to be considerably poorer than that of the rival models. Nonetheless, it still receives a weight of almost 20% .

A set that excludes the generating model

In our second experiment, we omit the generating trivariate model from the forecasting structures. We replace it with a bivariate model $VAR_{2\pi,S}$ that contains the target GDP growth rate and the rate of inflation. There are two differences with respect to the basic VAR model $VAR_{2\pi}$. First, lag orders are searched for 'own' lags and for 'foreign' lags independently. In terms of the restrictions on coefficient matrices, this model corresponds to zero restrictions on the (1,3), (2,1), and (2,3) elements of $\hat{A}$. This specification allows for a longer lag length in the diagonal of the VAR structure (see Sims, 1972). Second, the inflation rate is modelled as a fully 'exogenous' variable in the sense that it is modelled univariately and the potential dynamic feedback from output to inflation is ignored. This implies the following structure

$$\begin{aligned} Y_{1,t} &= \mu_1 + \sum_{j=1}^{p_1} a_j Y_{1,t-j} + \sum_{j=1}^{p_2} b_j Y_{2,t-j} + \varepsilon_{t,1}, \\ Y_{2,t} &= \mu_2 + \sum_{j=1}^{p_3} c_j Y_{1,t-j} + \varepsilon_{t,2}, \end{aligned} \tag{9}$$

where Y_1 denotes GDP growth and Y_2 indicates inflation. Lag orders p_j, $j = 1, \ldots, 3$, are separately determined by BIC minima for the two equations.

Table 4 reports results on the forecast accuracy for the four basic models. In the U.K. design, the univariate model dominates for very small samples ($N = 40$), but the bivariate model with the sophisticated lag search outperforms all other models for $N = 100$ and larger samples. The lower panel of Table 4 gives parallel results for the French data design. We already noted that the link between inflation and GDP growth is weaker than in the British case, and that the link between growth and unemployment is much stronger. Thus, the sophisticated model $VAR_{2\pi,S}$ appears less promising, and this conjecture is confirmed by Table 4. As in the U.K. design, the univariate model outperforms its rivals for the small sample of $N = 40$, while the bivariate model with unemployment achieves the best accuracy for $N = 100$ and larger N. It would be an obvious suggestion to perform the sophisticated lag search on the other bivariate combination VAR_{2u}, but we wanted to keep designs for the two countries comparable as much as possible.

The upper panel of Table 5 shows that the weighted average based on encompassing tests is marginally worse than the uniformly weighted average for the U.K. design. For $N = 40$, both types of model averages beat even the best individual model, which corroborates the idea of model averaging, in the sense that each model picks up some dynamics that others miss, such that each of them contributes to improving the prediction. The lower panel of Table 5 gives comparable results for the French design. Test-based averages are better than simple uniform averages in this case, and the local optimum appears to be at the 1% significance level, excepting the largest sample size. Only for $N = 40$, do the averaged forecasts outperform the pure strategies of Table 4.

Table 6 shows the corresponding average weights for all models at the significance level of 10%. Apparently, weights are approximately uniform for $N = 100$ and $N = 200$. At $N = 40$, the univariate model still has a larger weight on average than its rivals. At $N = 500$, the univariate autoregression falls behind for the U.K. design, while it still receives a sizeable weight for France. In both designs, the relative weight on the 'preferred' model increases monotonically, as N rises. It is only the preferred model that differs: for the U.K. $VAR_{2\pi,S}$, for France $VAR_{2u,S}$. For reasons of space, we do not report the weights for the other significance levels (results are available upon request from the authors). By construction, these tend to be more uniform than the 10% weights.

In this simulation experiment, a technical problem arises in small and in large samples: the selected lag orders often coincide for the model $VAR_{2\pi}$ and $VAR_{2\pi,S}$ with respective sophisticated and block search. This occurs in 19% of the U.K. design cases for $N = 40$ and still in 4% for $N = 100$. For the French-data design, where the link between output growth and inflation is weaker, this feature re-increases for large N, and both searches lead to identical lag orders in 97% of all replications at $N = 500$. In these cases, we chose to exclude one of the two identical forecasts, say $VAR_{2\pi,S}$, and to run the encompassing search over the remaining three models.

Table 5 indicates that the prediction error re-increases as N increases from 200 to 500, in analogy to our first experiment reported in Table 2. These considerations hold for both designs and points to problems in the large-sample asymptotic behavior of the weighting search. Uniform weighting suffers from the large weight given to the comparatively poor univariate predictions, and also test-based weighting may gain from modifications in the significance level. Contrary to typical statistical recommendations, increasing the significance level beyond 10% in larger samples may help to drop inferior prediction models from the weighted average. In further unreported experiments, we found that the performance of test-based weighted forecasts improves considerably at extremely loose significance levels for $N = 500$.

At conventional significance levels, test-based weighting does not outperform the uniform control average for the U.K. design. With respect to the French data design, however, test-based weighting outperforms uniform weighting. The average weights (see Table 6) reveal that the model $VAR_{2\pi,S}$ is selected slightly less often in small samples than the other candidates. In large samples, it gives identical forecasts to $VAR_{2\pi}$ and is excluded from the experiment.

It should be noted that the comparability of the MSE reported in Table 4 and 5 is limited, as the former values are averages over $10,000(N/4-1)$ squared errors, while the latter values just average 10,000 replications. This discrepancy is strongest for large N. When $N = 500$, the values in Table 4 summarize predictions based on 375 up to 498 observations, while in Table 5 independent samples of 499 observations are used. For this reason, the slightly larger numbers in Table 5 do not provide evidence that model averages are generally worse than pure models.

Iterated multi-step prediction

This subsection extends the previous analysis for the one-step horizon to multiple-step ahead forecasts. The focus is now exclusively on the simulation design that excludes the generating model class, as we feel it is the more realistic one and therefore of more practical relevance. In most empirical applications, it is plausible to assume that the data-generating model is far more complex than any of the utilized prediction models.

Traditionally, there are two ways to tackle the problem of multi-step prediction using linear time-series models. The first one is to plug in the predictions at smaller step sizes for the unknown data. This method is often called iterative prediction. The second one is to gauge model selection specifically to the task of multi-step prediction by restricting the first few lags to zero. This method is often called direct prediction (see for example Marcellino *et al.*, 2006), and we will report on it in the next subsection.

This subsection focuses on iterated prediction. Table 7 gives the results for horizons 2 to 4 for both designs. As N increases, the emphasis shifts from the univariate AR model to the preferred structures for both countries, i.e. to the $VAR_{2\pi,S}$ model for the U.K. and the VAR_{2u} for France. Forecast errors increase only moderately with the horizon, reflecting the strong autocorrelation in economic growth.

Table 8 summarizes the corresponding statistics for combined forecasts. Generally, the multi-step evidence conforms qualitatively to the single-step results reported above. With respect to the U.K. design, uniform weighting dominates test-based weighting at the investigated significance levels at the smaller sample sizes. At $N = 500$, test-based weighting is on a par with uniform weights, whereas its performance again deteriorates relative to smaller samples and pure models. For the French design, test-based weights tend to outperform the uniform benchmark but performance is flat across significance levels, giving no recommendation on behalf of risk levels.

Multi-step prediction by direct modelling

As an alternative to the traditional plug-in method of h–step forecasting, some authors consider 'direct' models of the form

$$Y_t = \mu + \sum_{j=h}^{p} \Phi_j Y_{t-j} + \varepsilon_t, \tag{10}$$

which are subset models of the ordinary VAR(p) with the restriction $\Phi_j = 0$ for $j < h$. Among these models, an optimum lag order p can again be determined by information criteria, and the value $\hat{Y}_t(h)$ calculated as

$$\hat{Y}_t(h) = \hat{\mu} + \sum_{j=h}^{p} \hat{\Phi}_j Y_{t+h-j} \tag{11}$$

serves as an h–step predictor of Y_{t+h}. The evidence on the relative advantages of this method is fragile, and many studies appear to give some preference to the plug-in method (see Marcellino *et al.*, 2006; Schorfheide, 2005).

Tables 9 and 10 show that the direct modelling method is less efficient than iterated forecasting at all horizons for both the U.K. and French design. The differences between the two approaches, however, are not homogeneous across sample sizes, and direct modelling shows its relatively best performance at $N = 40$. Again, MSE values for the averaged predictions provide uncertain recommendations with regard to the optimum significance levels.

CONCLUSION

This paper considers a method of forecast averaging that determines the weights for forecast combinations in accordance with the rejection/acceptance decision of a multiple encompassing test developed by Harvey and Newbold (2000). Using simulation designs that are adapted to trivariate systems for U.K. and French data, we investigate the implications of this method on the accuracy of forecasts in a vector autoregressive framework. While one simulation design considers a model that contains the data-generating mechanism, in a second design all models are 'misspecified'.

In the design that includes the generating model class, univariate models dominate at the smallest sample size, while only at the largest sample, $N = 500$, does the trivariate structure outperform its rival models. This result seems to be relevant, as the three variables in our core models are known to have relatively strong dynamic interdependence. Regarding the forecast combination, model averaging shows its strength when the sample size is small, while in larger samples model averages become less attractive. By construction, naive uniform averaging assigns considerable weights to inferior rivals, and even the test-based weighting procedure discards the inferior model quite slowly.

When the simulation design excludes the generating model, averaging again gains the best performance in small samples, while in larger samples averaging becomes unattractive and even leads to a deterioration in performance as the sample size grows. In the experiment based on French data, the test-based weighting scheme outperforms naive averages at all sample sizes.

All simulation experiments consider three customary significance levels for the encompassing test in the averaging procedure. Unfortunately, our results do not provide any clear recommendation regarding the optimum significance level. The performance of our procedure in different data-based designs and in even larger samples than $N = 500$ may be of interest in this regard. We leave such experiments for our future research.

REFERENCES

Bates, JM, Granger, CWJ. The combination of forecasts. *Operations Research Quarterly* 1969; **20**: 451–468.

Chandrasekharan, R, Moriarty, MM, and Wright, GP. Testing for unreliable estimators and insignificant forecasts in combined forecasts. *Journal of Forecasting* 1994; **13**: 611-624.

Chatfield, C. *Time-series Forecasting*. Chapman & Hall, 2001.

Chen, L, Anandalingam, G. Optimal selection of forecasts. *Journal of Forecasting* 1990; **9**: 283-297.

Clark, TE, McCracken, MW. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 2001; **105**: 85–110.

Clemen, RT. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* 1989; **5**: 559-583.

Clements, M, Hendry, DF. *Forecasting Economic Time Series*. Cambridge University Press, 1998.

Costantini, M, Pappalardo, C. A hierarchical procedure for the combination of forecasts. *International Journal of Forecasting*, forthcoming.

de Menezes, L, Bunn, DW. Diagnostic tracking and model specification in combined forecasts of U.K. inflation. *Journal of Forecasting* 1993; **12**: 559-572.

Flores, BE, White, EM. Subjective versus objective combining of forecasts: an experiment. *Journal of Forecasting* 1989; **8**: 331-341.

Granger, CWJ. Combining forecasts—twenty years later. *Journal of Forecasting* 1989; **8**: 167–73.

Granger, CWJ, Ramanathan, R. Improved methods of combining forecasts. *Journal of Forecasting* 1984; **3**: 197-204.

Hallman, J, Kamastra, M. Combining algorithms based on robust estimation techniques and co-integrating restrictions. *Journal of Forecasting* 1989; **8**: 189-198.

Harvey, DI, Leybourne, S, Newbold, P. Tests for forecast encompassing. *Journal of Business and Economic Statistics* 1998; **16**: 254–259.

Harvey, DI, Newbold, P. Tests for multiple forecast encompassing. *Journal of Applied Econometrics* 2000; **15**: 471–482.

Harvey, DI, Newbold, P. Forecast encompassing and parameter estimation. *Oxford Bulletin of Economics and Statistics* 2005; **67**: 815–835.
Juselius, K. *The Cointegrated VAR Model.* Oxford University Press, 2006.
Kisinbay, T. 2007. The use of encompassing tests for forecast combinations. Working paper, International Monetary Fund, 2007; **264**: 1-21.
Lütkepohl, H. *New Introduction to Multiple Time Series.* Springer, 2005.
Marcellino, M, Stock, JH, Watson, MW. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* 2006; **135**: 499–526.
McQuarrie, ADR, Tsai, CL. *Regression & Time Series Model Selection.* World Scientific, 1998.
Schorfheide, F. VAR forecasting under misspecification. *Journal of Econometrics* 2005; **128**: 99–136.
Swanson, NR, Zeng, T. Choosing among competing econometric forecasts: regression-based forecast combination using model selection. *Journal of Forecasting* 2001; **20**: 425–440.
Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**: 461–464.
Sims, CA. Money, income, and causality. *American Economic Review* 1972; **62**: 540–552.
Timmermann, A. Forecast combinations, in Elliott, G, Granger, CWJ, and Timmermann, A. (ed.), *Handbook of Economic Forecasting.* Elsevier, 2006.

Table 1: *MSE for candidate models.*

N	VAR_3	$VAR_{2\pi}$	VAR_{2u}	AR
U.K. design				
40	3.5014	3.0063	3.3283	2.9148*
100	2.7667	2.6689*	2.7648	2.7508
200	2.5905	2.5842*	2.6416	2.6988
500	2.5155*	2.5397	2.5908	2.6746
σ^2		2.468		
France design				
40	0.5794	0.5523	0.5421	0.5125*
100	0.4777	0.4786	0.4573*	0.4639
200	0.4434	0.4493	0.4400*	0.4450
500	0.4310*	0.4394	0.4320	0.4379
σ^2		0.423		

Notes: N is the sample size; VAR_3 denotes the trivariate VAR; $VAR_{2\pi}$ is the bivariate VAR with GDP growth and π; VAR_{2u} is the bivariate VAR with GDP growth u; AR denotes the univariate autoregression; σ^2 is the true error variance that serves as a lower bound. Asterisks indicate the optimum among comparable predictions.

Table 2: *MSE for weighted averages.*

N	uniform	1%	5%	10%
U.K. design				
40	2.8955*	2.8959	2.8985	2.9094
100	2.6587*	2.6650	2.6714	2.6748
200	2.5681*	2.5727	2.5788	2.5802
500	2.6242	2.6231*	2.6267	2.6293
σ^2	2.468			
France design				
40	0.4943*	0.4944	0.4958	0.4977
100	0.4577*	0.4579	0.4587	0.4603
200	0.4378*	0.4380	0.4387	0.4391
500	0.4474	0.4473	0.4473*	0.4473
σ^2	0.423			

Notes: N is the sample size; σ^2 is the true error variance; asterisks denote the optimum among comparable predictions.

Table 3: *Test-based procedure weights for models at 10% significance level.*

N	VAR_3	$VAR_{2\pi}$	VAR_{2u}	AR
U.K. design				
40	0.229	0.261	0.239	0.271
100	0.238	0.274	0.238	0.250
200	0.276	0.290	0.235	0.199
500	0.393	0.320	0.195	0.092
France design				
40	0.237	0.246	0.251	0.265
100	0.232	0.229	0.276	0.263
200	0.256	0.221	0.272	0.250
500	0.324	0.182	0.290	0.204

Notes: see Table 1.

Table 4: *MSE for candidate models.*

N	$VAR_{2\pi,S}$	$VAR_{2\pi}$	VAR_{2u}	AR
U.K. design				
40	2.9723	3.0063	3.3238	2.9148*
100	2.6209*	2.6689	2.7648	2.7508
200	2.5545*	2.5842	2.6416	2.6988
500	2.5281*	2.5397	2.5908	2.6748
σ^2		2.468		
France design				
40	0.5627	0.5523	0.5421	0.5125*
100	0.4867	0.4786	0.4573*	0.4639
200	0.4578	0.4493	0.4400*	0.4450
500	0.4396	0.4394	0.4320*	0.4379
σ^2		0.423		

Notes: $VAR_{2\pi}$ denotes the bivariate VAR model with GDP growth and inflation; $VAR_{2\pi,S}$ is similar to the $VAR_{2\pi}$, but uses exogenous inflation; VAR_{2u} is the bivariate VAR with GDP growth and the unemployment rate; AR is the univariate AR model. σ^2 is the true errors variance. Asterisks denote the optimum among comparable predictions.

Table 5: *MSE for weighted averages.*

N	uniform	1%	5%	10%
U.K. design				
40	2.8177*	2.8236	2.8300	2.8360
100	2.6403*	2.6456	2.6532	2.6562
200	2.5652*	2.5673	2.5790	2.5802
500	2.6276*	2.6319	2.6356	2.6375
σ^2	2.468			
France design				
40	0.4970	0.4939*	0.4953	0.4967
100	0.4618	0.4603*	0.4605	0.4608
200	0.4419	0.4413*	0.4414	0.4415
500	0.4511	0.4501	0.4499	0.4497*
σ^2	0.423			

Notes: see Table 2.

Table 6: *Average weights for rival prediction models at 10% level.*

N	$VAR_{2\pi,S}$	$VAR_{2\pi}$	VAR_{2u}	AR
U.K. design				
40	0.139	0.291	0.273	0.297
100	0.238	0.269	0.248	0.246
200	0.297	0.271	0.242	0.190
500	0.355	0.307	0.234	0.103
France design				
40	0.096	0.292	0.303	0.308
100	0.132	0.261	0.314	0.292
200	0.081	0.274	0.347	0.298
500	0.005	0.279	0.418	0.299

Notes: see Table 4.

Table 7: *MSE for candidate models.*

N	$VAR_{2S\pi}$	$VAR_{2\pi}$	VAR_{2u}	AR	$VAR_{2\pi,S}$	$VAR_{2\pi}$	VAR_{2u}	AR
		U.K. design				France design		
horizon 2								
40	2.8588	2.8724	3.1710	2.8195*	0.6140	0.6003	0.5755	0.5454*
100	2.6061*	2.6356	2.7349	2.7148	0.5208	0.5129	0.4850*	0.4932
200	2.5633	2.5808*	2.6434	2.6846	0.4889	0.4809	0.4669*	0.4755
500	2.5495*	2.5564	2.6082	2.6722	0.4707	0.4705	0.4586*	0.4688
horizon 3								
40	2.8799	2.8752	3.1982	2.8208*	0.7151	0.6965	0.6583	0.6201*
100	2.6185*	2.6287	2.7284	2.7190	0.5798	0.5775	0.5474*	0.5520
200	2.5762*	2.5808	2.6462	2.6878	0.5488	0.5459	0.5246*	0.5368
500	2.5621*	2.5645	2.6169	2.6759	0.5325	0.5323	0.5133*	0.5298
horizon 4								
40	2.9220	2.9180	3.3131	2.8285*	0.7588	0.7371	0.6864	0.6448*
100	2.6392*	2.6496	2.7514	2.7210	0.5949	0.5947	0.5601*	0.5659
200	2.5921*	2.5956	2.6621	2.6888	0.5625	0.5610	0.5361*	0.5505
500	2.5769*	2.5781	2.6309	2.6773	0.5465	0.5463	0.5242*	0.5434

Notes: see Table 4.

Table 8: *MSE for averaged prediction.*

N	uniform	1%	5%	10%	uniform	1%	5%	10%
		U.K. design				France design		
horizon 2								
40	2.7750*	2.7761	2.7795	2.7817	0.5252	0.5251	0.5244*	0.5255
100	2.6441*	2.6454	2.6493	2.6539	0.4938	0.4932*	0.4933	0.4941
200	2.5800*	2.5828	2.5862	2.5885	0.4721	0.4721	0.4712*	0.4716
500	2.6559*	2.6559*	2.6633	2.6667	0.4804	0.4789	0.4786	0.4783*
horizon 3								
40	2.7648*	2.7667	2.7684	2.7745	0.5900	0.5885	0.5877	0.5866*
100	2.6524*	2.6544	2.6598	2.6574	0.5509	0.5505	0.5498	0.5494*
200	2.5917*	2.5945	2.5994	2.6032	0.5313	0.5298*	0.5299	0.5304
500	2.6603	2.6601*	2.6643	2.6655	0.5293	0.5255*	0.5264	0.5268
horizon 4								
40	2.7944*	2.7945	2.7978	2.8054	0.6134	0.6094	0.6076	0.6053*
100	2.6648*	2.6665	2.6700	2.6742	0.5635	0.5612	0.5606*	0.5617
200	2.6029*	2.6059	2.6093	2.6100	0.5438	0.5422	0.5413	0.5412*
500	2.6714*	2.6732	2.6770	2.6778	0.5429	0.5388*	0.5405	0.5404

Notes: see Table 4.

Table 9: *MSE for candidate models by direct modelling.*

N	$VAR_{2\pi,S}$	$VAR_{2\pi}$	VAR_{2u}	AR	$VAR_{2\pi,S}$	$VAR_{2\pi}$	VAR_{2u}	AR
		U.K. design				France design		
horizon 2								
40	3.358	3.277	3.652	2.925*	0.652	0.648	0.608*	0.608
100	2.655*	2.691	2.781	2.732	0.573	0.575	0.553*	0.564
200	2.595*	2.617	2.674	2.686	0.558	0.559	0.538*	0.555
500	2.571*	2.578	2.630	2.667	0.552	0.551	0.531*	0.551
horizon 3								
40	3.349	3.366	3.573	2.984*	0.731	0.726	0.666	0.656*
100	2.706*	2.730	2.806	2.754	0.605	0.606	0.578*	0.584
200	2.637*	2.653	2.699	2.705	0.576	0.579	0.557*	0.570
500	2.611*	2.616	2.655	2.685	0.564	0.567	0.545*	0.563
horizon 4								
40	3.399	3.412	3.553	3.034*	0.750	0.750	0.679	0.672*
100	2.757*	2.776	2.835	2.766	0.622	0.623	0.594	0.593*
200	2.677*	2.690	2.713	2.710	0.586	0.587	0.567*	0.575
500	2.648*	2.649	2.665	2.690	0.571	0.573	0.553*	0.568

Notes: see Table 4.

Table 10: *MSE for averaged prediction by direct modelling.*

N	uniform	1%	5%	10%	uniform	1%	5%	10%
		U.K. design				France design		
horizon 2								
40	2.847*	2.856	2.860	2.860	0.577	0.573*	0.574	0.575
100	2.673*	2.678	2.680	2.684	0.562	0.560	0.561	0.560*
200	2.601*	2.603	2.609	2.610	0.549	0.548	0.547*	0.547
500	2.661*	2.663	2.667	2.669	0.548	0.545	0.544	0.544*
horizon 3								
40	2.897	2.894	2.892*	2.899	0.625	0.616	0.613	0.612*
100	2.712*	2.714	2.716	2.721	0.579	0.575*	0.575	0.575
200	2.630*	2.632	2.638	2.639	0.561	0.558*	0.559	0.559
500	2.691*	2.694	2.693	2.699	0.561	0.559	0.559*	0.559
horizon 4								
40	2.899	2.888*	2.891	2.888	0.637	0.622	0.619	0.617*
100	2.755*	2.755	2.756	2.762	0.592	0.583*	0.584	0.585
200	2.657*	2.659	2.662	2.665	0.571	0.567*	0.569	0.569
500	2.719*	2.721	2.729	2.733	0.571	0.569*	0.569	0.571

Notes: see Table 4.