

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

Mauro Costantini* and Robert M. Kunst†

Abstract

We investigate whether and to what extent multiple encompassing tests may help to determine weights for forecast averaging in a standard vector autoregressive setting. An algorithm based on a multiple encompassing test assigns non-zero weights to candidate models that add information not covered by other models.

We explore the potential benefits of the algorithm in extensive Monte Carlo experiments. Simulations use realistic designs that are adapted to U.K. and to French macroeconomic data. The real economic growth rates of these two countries serve as the target series to be predicted. Generally, we find that the test-based averaging of forecasts yields a performance that is comparable to a simple uniform weighting of individual models. Furthermore, averaged forecasts seem to work only for small samples, where they successfully mitigate instabilities of estimated structures.

JEL Classification: C22, C53.

Keywords: Combining forecasts, encompassing tests, model selection, time series.

*Department of Economics, University of Vienna, BWZ, Vienna, Austria; E-mail address: mauro.costantini@univie.ac.at

†Corresponding author: Robert M. Kunst, Department of Economics and Finance, Institute for Advanced Studies, Vienna, Austria, and Department of Economics, University of Vienna, Vienna, Austria; E-mail address: kunst@ihs.ac.at, robert.kunst@univie.ac.at

1 Introduction

Forecasters are interested in getting the best forecast. The superiority of a particular model in terms of forecast accuracy does not necessarily imply that forecasts from other models do not contain additional information. Any strategy that picks a specific candidate model ignores the possibility that alternative forecasts may embody useful information not contained in the preferred forecast. Thus, as the literature on forecast combinations and encompassing outlines (see NEWBOLD AND HARVEY, 2002), pure strategies can be sub-optimal. In order to construct forecast combinations, encompassing tests can be used to check whether any extra relevant information is contained in forecasts from rival models. When several alternative forecasts of the same variable are available, then a combination of these forecasts may yield a more accurate performance than the individual one.

The idea of improving individual predictions by averaging has a long history, dating back at least to BATES AND GRANGER (1969) and summarized by CLEMENTS AND HENDRY (1998) and, more recently, by TIMMERMAN (2006). Several methods have been proposed to determine the individual weights for forecast combination, such as uniform weights, weights derived from information criteria, and weights based on regressions over training samples. The literature on model averaging proper is closely related (see HJORT AND CLAESKENS, 2003, or HANSEN, 2007), even though it does not necessarily target the prediction performance of the average.

While one could argue that using different models implies that most of them will be misspecified and that more emphasis should be put on searching the correct specification, this argument seems to miss a central issue of small-sample forecasting, as most empirical models will be misspecified in some sense. In addition, there is no guarantee that the correct model class yields the optimum forecasting workhorse following an estimation step over a comparatively short sample and it is easy to construct counter-examples for this feature (see also JUMAH AND KUNST, 2008).

In order to obtain combinations of model-based forecasts, we propose a new algorithm based on the multiple encompassing tests developed by HARVEY AND NEWBOLD (2000, 2005). The procedure discards those models that are encompassed by their competitors and then combines the retained models. By way of extensive Monte Carlo simulation, we investigate whether and the extent to which this procedure may help to determine the weights for forecasting averaging in a standard vector autoregressive setting. In this regard, a standard model is perceived to be a potential data-generating mechanism

for macroeconomic data rather than a simple but artificial design.

In detail, we use two simulation designs that are adapted to trivariate core systems for U.K. and French macroeconomic data. The real economic growth rates of these two countries serve as the target series to be predicted.

Although forecast combinations tend to considerably outperform pure forecasts in training samples, selected ‘optimal’ combinations rarely live up to their promises over the prediction interval. For this reason, in our experiments the training interval includes one quarter of the observations, and only one final data point is then used for the comparative evaluation. A high number of replications guarantees that the results can be viewed as significant.

In summary, the performance of the test-based algorithm does not seem to provide markedly better predictions than those obtained using simple uniformly weighted model averages. Only for some small and extremely large samples does the algorithm-based procedure gain advantages. These results appear to be typical for test-based algorithms that miss out on assigning proper dominance to the model that performs best with regard to forecasting performance. To some extent these differences in quality can be measured over training samples and they are informative, even if they fail to transgress boundaries of statistical significance.

The plan of the paper is as follows. Section 2 describes the proposed algorithm for combining forecasts. Section 3 outlines the simulation design and the backdrop data. Section 4 reports on the simulation results. Section 5 concludes.

2 Multiple forecast encompassing

Our algorithm for the construction of combinations of forecasts is based on the multiple encompassing tests developed by HARVEY AND NEWBOLD (2000,2005). In principle, it proceeds by keeping those models in the combination that are not encompassed by their competitors.

We consider M model-based forecasts that are based on specifying and estimating each of the M models. The aim is to forecast a specific component within the vector variable Y . The M candidate models yield series of out-of-sample forecasts $\hat{Y}_{jt}^{(k)}$ and of forecast errors $e_{jt}^{(k)} = Y_{jt} - \hat{Y}_{jt}^{(k)}$, $k = 1, \dots, M$ for any component j of the considered variables vector. In the following, we will only be interested in the prediction of a single specific variable in the vector Y that we without loss of generality can choose to be the first one

($j = 1$). This allows to restrict the evaluation of forecasts to the univariate mean-squared error criterion.

Suppressing the series index, denote the series of thus obtained forecast errors from model k for a given sample of length N as $e_t^{(k)}$ with $t = N - n + 1, \dots, N$, where n is the length of an evaluation sample such that $n \ll N$. For these series, we run M encompassing regressions. In the first regression, $e_t^{(1)}$ is the dependent variable:

$$e_t^{(1)} = \sum_{k=2}^M a_k (e_t^{(k)} - e_t^{(1)}) + u_t. \quad (1)$$

Then, $e_t^{(2)}$ is regressed on $M - 1$ differences:

$$e_t^{(2)} = \sum_{k=1, k \neq 2}^M a_k (e_t^{(k)} - e_t^{(2)}) + u_t,$$

and $M - 2$ analogous regressions follow with $e_t^{(j)}$, $j = 3, \dots, M$ on the left.

These homogeneous regressions yield regression F statistics that correspond to forecast-encompassing statistics according to HARVEY AND NEWBOLD (2005). A model j is said to forecast-encompass its rivals if the F statistic in the regression with dependent variable $e_t^{(j)}$ is insignificant at a specific level of significance.

Following the evidence of the forecast-encompassing tests, we form weighted average forecasts according to the following rule. If all M tests reject or all of them 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.

HARVEY AND NEWBOLD (2005), among others, demonstrate that the definition of encompassing is not trivial. Because of the limited power of the test procedures, a model may empirically encompass other models, although their prediction performance is perceptibly different. Such a perception may, for example, rely on the training sample that we use in our simulations.

3 The simulation experiment

3.1 The data

For the United Kingdom, we use three series from the OECD Main Economic Indicators database: a series on gross domestic product (GDP) in constant

prices that is actually a volume index and serves as the main variable to be predicted; the consumer price index (CPI); the registered unemployment rate. According to OECD, the first and third of these series have been seasonally adjusted. We prefer the registered unemployment rate to the conventional unemployment rate based on questionnaires, as it is the series with the longer time range. The data is quarterly and runs from 1960 to 2008:2.

For France, unfortunately a much shorter time range of comparable data is available. Whereas the CPI series would admit a longer range from 1960 to 2008:4, the other two variables impose a starting date of 1978:1. The definition of the GDP volume index is comparable to the U.K. equivalent, this data ends in 2008:3. A registered unemployment rate is not recorded in the OECD Main Economic Indicators database, so we use a harmonized unemployment rate instead, which also starts in 1978:1 and ends in 2008:4. Thus, the French data omits the OPEC–1 shock episode, in contrast to the U.K. data, which aspect may also be of interest with regard to enhancing the robustness of our results.

For the analysis, we transform the GDP series (X) and the CPI series (P) to growth rates and thus to variables that typically serve as indicators for business-cycle activity and are targeted by forecasting institutions. For GDP, we use first differences of the logarithms multiplied by four that correspond to quarterly indicators of annual real growth rates. Figure 1 shows that the resulting variable is quite volatile for the U.K. and much less so for France. However, due to its seasonally adjusted nature we do not think it is reasonable to consider $\log X_t - \log X_{t-4}$, which would implicitly de-seasonalize the series once more. By contrast, inflation is calculated as $\pi_t = \log P_t - \log P_{t-4}$ in the usual way. While inflation does not display major seasonal variation, this step also serves to eliminate potential seasonality. Finally, unemployment U_t is used without any further transformation. According to the source, unemployment rates have been seasonally adjusted.

In symbols, we use $Y_t = (4\Delta \log X_t, \pi_t, U_t)'$ or simply $Y_t = (Y_{1t}, Y_{2t}, Y_{3t})'$.

At least, 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. Usually, this property is

reproduced by models fitted to artificial data generated from the fitted model, except for some models fitted to very short samples.

We note that the U.K. and French data only serve as the basis for our simulation experiment. It will become clearer that 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.

3.2 The data-based simulation design

To the data, we fit trivariate vector autoregressive (VAR) models and we identify the lag order by the BIC criterion according to SCHWARZ. This results in a lag order of two for both economies, i.e. in a VAR(2) model of the form

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

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

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. 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 but some of its coefficient parameters may be statistically insignificant, such that simplification steps may be rewarding.

The corresponding estimates for the French data are:

$$\begin{aligned} \mu &= (-0.698, 0.617, 0.102)', \\ \Phi_1 &= \begin{pmatrix} 0.237 & 0.241 & 0.035 \\ -0.009 & 1.253 & 0.023 \\ -0.068 & 0.136 & 1.478 \end{pmatrix}, \end{aligned}$$

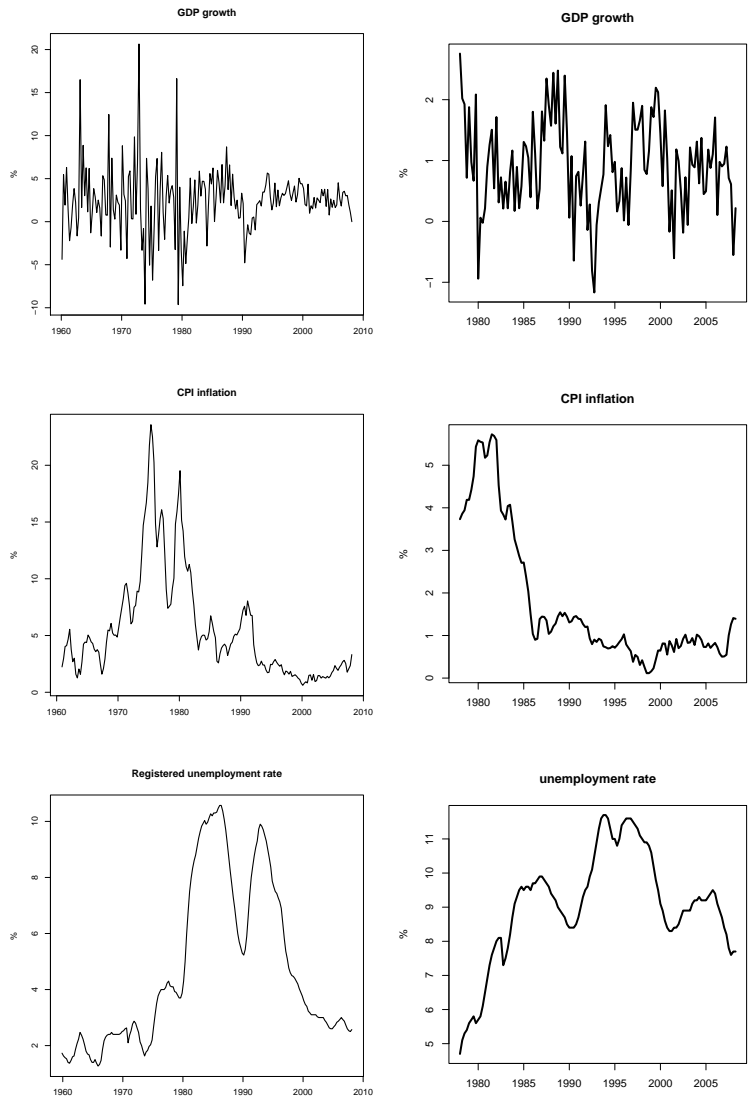


Figure 1: Growth rate of real GDP, CPI inflation, and unemployment rate, for the U.K. (left) and for France (right).

$$\Phi_2 = \begin{pmatrix} 0.292 & -0.191 & 0.077 \\ 0.026 & -0.318 & -0.082 \\ -0.037 & -0.113 & -0.482 \end{pmatrix}. \quad (3)$$

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). 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, that is in 2008, we now simulate artificial samples (‘pseudo-samples’) of given length by drawing errors from a normal distribution with variance matrix Σ , which, for the U.K. data, is given as

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

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. We note that the errors are indeed correlated across variables but that the correlation is comparatively weak. Note 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

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

which indicates that variation in the GDP growth rate is far lower in the,

concededly also shorter, French series that avoids the turbulence of the OPEC shocks in the 1970s.

The sample size is varied 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 $N = 500$ as some sort of evidence on what happens in large samples, where 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. 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 based on expanding windows of $t = 1, \dots, n$ 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 usage of a large number of replications, usually 10,000, mitigates such potentially disturbing effects. Note that the last observation is not contained in this stage of the prediction experiment. It is reserved for a later stage.

4 Evaluating prediction by sets of rival models and combinations

4.1 A set that includes the generating model

All our forecasts are model-based. They can be considered as versions of \hat{Y}_t defined by

$$\hat{Y}_t = \hat{\mu} + \sum_{j=1}^p \hat{A}_j Y_{t-j}, \quad (6)$$

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, $\hat{Y}_t = \hat{Y}_{t-1}(1)$.

In the first experiment, we use four model structures: the trivariate autoregression, bivariate autoregressions with the target GDP growth series and either of the other two supporting series included, and a univariate autoregression. These models can be expressed by respective restrictions on the matrices \hat{A}_j for all j : unrestricted matrices; elements at (1,3) equal 0; elements at (1,2) equal 0; elements at (1,2) and at (1,3) equal 0.

Note that the first model class contains the data-generating structure. 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.

Within each model class, lag orders are determined via BIC, with a maximum lag order of eight or 12 to represent the quarterly nature of the data, depending on N . These maxima are not often binding, as typically low lag orders are selected. The forecasts may be denoted by $\hat{Y}^{(k)}$ for $k = 1, \dots, 4$ to express dependence on the utilized prediction model.

The upper panel of Table 1 gives the prediction performance of these four models for the U.K. design. While large samples see the true structure clearly in the lead, the bivariate model that includes inflation comes in first for samples around $N = 100$, and very small samples even prefer the parsimonious univariate autoregression. Results for the French structure in the lower panel are comparable, with the preferred VAR_2u model substituting the $VAR_2\pi$ model.

In the notation of Section 2, $M = 4$. The four models yield series of forecast errors $e_{jt}^{(k)} = Y_{jt} - \hat{Y}_{jt}^{(k)}$ for $k = 1, \dots, 4$ and for a maximum of three of the considered variables (j). We are only interested in the prediction of the first variable that corresponds to GDP growth. As we have described in Section 2, we run $M = 4$ encompassing regressions for forecast errors for GDP growth $e_{1t} = e_t$. In the first regression, errors for the first model $e_t^{(1)}$ are the dependent variable, i.e. the regression

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

is run for $t = (3N/4) + 1, \dots, N - 1$. The second regression has the forecast errors for the second model $e_t^{(2)}$ as its dependent variable:

$$e_t^{(2)} = a_1(e_t^{(1)} - e_t^{(2)}) + a_2(e_t^{(3)} - e_t^{(2)}) + a_3(e_t^{(4)} - e_t^{(2)}) + u_t,$$

Table 1: Mean squared errors (MSE) for candidate models.

N	VAR_3	$VAR_2 \pi$	$VAR_2 u$	AR
U.K. design				
40	3.463	2.975	2.975	2.888*
100	2.759	2.663*	2.749	2.736
200	2.590	2.584*	2.642	2.699
500	2.516	2.540*	2.591	2.675
σ^2	2.468			
French design				
40	0.579	0.552	0.542	0.513*
100	0.478	0.479	0.457*	0.464
200	0.443	0.449	0.440*	0.445
500	0.431	0.439	0.432*	0.438
σ^2	0.423			

Note: N is the sample size. VAR_3 denotes the trivariate VAR; VAR_2 is the bivariate VAR, with its two versions, including GDP growth and π or u ; AR denotes the univariate autoregression. σ^2 is the theoretical error variance that serves as a lower bound. Asterisks mark the optimum among comparable predictions.

and this is followed by two more analogous regressions with $e_t^{(k)}$, $k = 3, 4$ on the left. A model is said to forecast-encompass its rivals if the corresponding regression F statistic is insignificant at a specific level of significance. For the standard approach, we fix this significance level at 10%. Because the candidate models are not too different from accurate specifications, the test tends to be conservative, so sharper significance levels are not too interesting.

Following the evidence of the forecast-encompassing tests, we form weighted average forecasts 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.

A comparison of the thus generated predictions for the last observations to a simple weighted average is presented in Table 2. In the U.K. design, the algorithm-based weighting beats the uniform weighting at $N = 500$ only. In the French design, results are identical for both procedures. Actually, average weights across the simulations remain close to $1/4$ for each of the models, which explains the extremely small differences between the two strategies. This is remarkable insofar as one might expect that the true generating model attains some lead in large samples and encompasses its rivals. This does not seem to be the case even at $N = 500$.

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. A potential source for this feature is the dependence within the replications.

Table 2: Mean squared errors (MSE) for weighted averages.

N	U.K. data		French data	
	10% rule	uniform	10% rule	uniform
40	2.906	2.904*	0.496	0.494*
100	2.617	2.614*	0.458*	0.458*
200	2.570	2.568*	0.438*	0.438*
500	2.621*	2.624	0.447*	0.447*
σ^2	2.468		0.423	

Note: Asterisks mark the optimum among comparable predictions.

4.2 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 that contains the target GDP growth rate and the rate of inflation. The difference to the basic VAR model $VAR_2\pi$ is, first, that lag orders are searched for ‘own’ lags and for ‘foreign’ lags independently. In restrictions on coefficient matrices, as introduced in the last subsection, this model corresponds to zero restrictions on the (1,3), (2,1), and (2,3) elements. This specification allows, for example, for a longer lag length in the diagonal of the VAR structure, as it was considered by SIMS (1972), among others in the literature. 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 structure

$$\begin{aligned} y_t &= \mu_1 + \sum_{j=1}^{p_1} a_j y_{t-j} + \sum_{j=1}^{p_2} b_j x_{t-j} + \varepsilon_{t,1}, \\ x_t &= \mu_2 + \sum_{j=1}^{p_3} c_j x_{t-j} + \varepsilon_{t,2}, \end{aligned} \tag{8}$$

with y denoting GDP growth and x denoting inflation. Lag orders p_j , $j = 1, \dots, 3$ are determined via BIC separately for the two equations.

The upper panel of Table 3 shows that, in the U.K. design, the univariate model still dominates at $N = 40$ but yields the top position to the bivariate model with the sophisticated lag search for $N = 100$ and larger samples. Table 4 shows that the weighted average based on encompassing tests is marginally worse than the uniformly weighted average that we use as a control. 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. Apparently, weights are approximately uniform for $N = 100$ and $N = 200$. At $N = 40$, the univariate model still has a markedly larger weight on average than its rivals. At $N = 500$, the univariate autoregression falls behind.

A technical problem in this simulation is that, in small and also in large samples, the selected lag orders often coincide for the sophisticated and the block search. This occurs in 19% of the cases for $N = 40$ but still in 4% for $N = 100$. For the French-data design, this feature re-increases for large N , and both searches lead to identical lag orders in almost all replications at

Table 3: Mean squared errors (MSE) for candidate models.

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR
U.K. design				
40	3.703	3.345	3.838	2.934*
100	2.638*	2.669	2.765	2.751
200	2.560*	2.584	2.642	2.699
500	2.530*	2.540	2.591	2.675
σ^2	2.468			
French design				
40	0.563	0.552	0.542	0.513*
100	0.487	0.479	0.457*	0.464
200	0.458	0.449	0.440*	0.445
500	0.440	0.439	0.432*	0.438
σ^2	0.423			

Note: $VAR_2\pi$ denotes the bivariate VAR model with GDP growth and inflation, VAR_{2S} is similar but uses exogenous inflation, VAR_2u is the bivariate VAR with GDP growth and the unemployment rate, and AR is the univariate AR model. σ^2 is the true errors variance that is given for comparison. Asterisks mark the optimum among comparable predictions.

$N = 500$. In those cases, we chose to exclude one of the two identical forecasts and to run the encompassing search over the remaining three models.

The lower panel of Table 3 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 appears less promising, and Table 3 confirms this apprehension. The univariate model is best for the small sample of $N = 40$ but the bivariate model with unemployment usurps the pole position for $N = 100$ and keeps it for larger N . In fact, the two bivariate variants with inflation yield identical forecasts with growing probability as N increases, and the two coincide in 97% of all replications for $N = 500$. It would be an obvious suggestion to perform the sophisticated lag search on the other bivariate combination, but we wanted to keep designs for the two countries comparable as much as possible.

Table 4: Mean squared errors (MSE) for weighted averages.

N	MSE weighted		weights			
	10% rule	uniform	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR
U.K. design						
40	2.857	2.847*	0.18	0.27	0.26	0.29
100	2.648	2.643*	0.23	0.26	0.26	0.26
200	2.567	2.566*	0.26	0.26	0.25	0.24
500	2.628*	2.629	0.28	0.28	0.25	0.19
σ^2	2.468					
French design						
40	0.495*	0.497	0.10	0.30	0.30	0.30
100	0.460*	0.462	0.16	0.28	0.29	0.28
200	0.441*	0.442	0.12	0.29	0.30	0.29
500	0.450*	0.451	0.01	0.33	0.33	0.33
σ^2	0.423					

Note: see Table 3.

Table 4 indicates that the prediction error re-increases as N increases from 200 to 500 (see also Table 2). This points to problems in the large-sample asymptotic behavior of the weighting search. Uniform weighting suffers from the fact that the comparatively poor univariate predictions are still weighted, while test-based weighting may demand for modifications in the significance level. For the U.K. design, the largest sample size is the only occasion where

the test-based weighting succeeds in beating the uniform control average. For the French data design, test-based weighting narrowly outperforms uniform weighting. The average weights reveal that the model VAR_{2S} 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, concededly arbitrarily, excluded from the race.

The MSE figures—averages over 10,000 replications—are not directly comparable to those of Table 3, which are averages over $10,000N/4$ squared errors, i.e. considerably more for large N . For example, for $N = 500$, the Table 3 values summarize predictions based on 375 up to 498 observations, while Table 4 uses independent samples of 499 observations. For this reason, the slightly larger numbers in Table 4 do not indicate that model averages are generally worse than pure models.

In all experiments, we ran control simulations, in which we substituted AIC lag-order searches for BIC lag-order searches. The counterpart to Table 3 is given as Table 5. Generally, AIC yields 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. We note that the critical issue is not the well known tendency to find too large lag orders in large samples, that is asymptotically. It is related more closely 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.

The outlined effect is extremely strong at $N = 40$, where the AIC-selected bivariate models attain mean squared errors that nearly twofold exceed the univariate model. Even at $N = 100$ and $N = 200$, all models predict worse than in the experiment reported in Tables 3 and 4 that is based on BIC selection. Only for $N = 500$ does the univariate model deliver better forecasts if its lag order is selected by AIC, giving a hint on asymptotic behavior. Results are comparable for both the U.K. and the French design. In the latter case, the preferred model is the bivariate model with unemployment, and this model needs larger samples to defeat the univariate benchmark than for the BIC-guided search.

Table 5: Mean squared errors (MSE) for candidate models selected by AIC.

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR
U.K. design				
40	5.987	6.639	7.235	3.133*
100	2.722*	2.742	2.825	2.797
200	2.620	2.610*	2.659	2.708
500	2.556	2.553*	2.597	2.662
σ^2	2.468			
French design				
40	0.597	0.590	0.579	0.516*
100	0.490	0.477	0.466	0.464*
200	0.456	0.452	0.443*	0.447
500	0.442	0.442	0.433*	0.441
σ^2	0.423			

Note: see Table 3.

4.3 Multi-step prediction

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 re-estimate the models by least squares, for example, thus essentially estimating models with the first few lags restricted to zero. This method is often called direct prediction (see, for example, MARCELLINO *et al.*, 2006).

We first focus on iterated prediction. Table 6 gives the results for horizons 2 to 4 for the U.K. design. For $N = 40$, all bivariate forecasts are useless, the enormous squared errors indicate massive forecast failure due to unstable estimated structures. The lesson to be learned appears to be to rely on univariate forecasts exclusively if degrees of freedom become a problem. For larger samples, forecast errors increase only moderately with the horizon, reflecting the strong autocorrelation in economic growth and the unconditional variance of 2.96. In other words, any prediction MSE beyond 2.96 is worse than a sample-mean benchmark, and this mark is attained only slowly as the horizon increases.

Table 7 gives an analogous evaluation for the French design. It differs insofar as $VAR_2 u$ is the preferred model and it continues to be that one in multi-step forecasting. Contrary to the British data, encompassing weighting

Table 6: Mean squared errors (MSE) for candidate models. U.K. design.

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR	enc	unif
horizon 2						
40	4.206	3.687	4.388	2.839*	2.789	2.784*
100	2.606*	2.636	2.735	2.715	2.645	2.644*
200	2.563*	2.581	2.643	2.684	2.582	2.580*
500	2.550*	2.556	2.608	2.672	2.656*	2.656*
horizon 3						
40	46.719	11.565	13.366	2.839*	2.789	2.777*
100	2.618*	2.629	2.728	2.719	2.654	2.652*
200	2.576*	2.581	2.646	2.688	2.594	2.592*
500	2.562*	2.564	2.617	2.676	2.659*	2.660
horizon 4						
40	4245	12104	59606	2.850*	2.832	2.820*
100	2.639*	2.650	2.751	2.721	2.666	2.665*
200	2.592*	2.596	2.662	2.689	2.604	2.603*
500	2.577*	2.578	2.631	2.677	2.672	2.671*

Note: see Table 3.

dominates uniform weights.

4.4 Direct modelling

As an alternative to the traditional plugging-in method of h -step forecasting, some authors consider ‘direct’ models of the form

$$Y_t = \mu + \sum_{j=h}^p \Phi_j Y_{t-j} + \varepsilon_t, \quad (9)$$

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

$$\hat{Y}_t(h) = \hat{\mu} + \sum_{j=h}^p \hat{\Phi}_j Y_{t+h-j-1} \quad (10)$$

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

Table 7: Mean squared errors (MSE) for candidate models. French design.

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR	enc	unif
horizon 2						
40	0.614	0.600	0.575	0.545*	0.524*	0.525
100	0.521	0.513	0.485*	0.493	0.493*	0.494
200	0.489	0.481	0.467*	0.476	0.472*	0.473
500	0.471	0.470	0.459*	0.469	0.480*	0.480*
horizon 3						
40	0.715	0.697	0.658	0.620*	0.587*	0.590
100	0.580	0.578	0.547*	0.552	0.550*	0.551
200	0.549	0.546	0.525*	0.537	0.530*	0.531
500	0.533	0.532	0.513*	0.530	0.526*	0.529
horizon 4						
40	0.759	0.737	0.686	0.645*	0.607*	0.613
100	0.595	0.595	0.560*	0.566	0.561*	0.563
200	0.563	0.561	0.536*	0.551	0.542*	0.544
500	0.546	0.546	0.524*	0.543	0.538*	0.543

Note: see Table 3.

Table 8 shows that, in the U.K. design, the direct modelling method is indeed better than iterated modelling for very short samples. In particular, it avoids the serious forecast failures at larger step sizes for $N = 40$. Conversely, predictions tend to deteriorate for larger samples as compared to iterated forecasting. The weights determined by the encompassing tests tend to be smaller for the most sophisticated model, the bivariate VAR with exogenous inflation, even though this model is in the lead for $N = 100$ but again converge to approximate uniformity as N increases.

Table 9 shows that the perspectives for direct modelling are not so good for the case of France. Just as for the iterative procedure, univariate AR is optimal for $N = 40$ but falls behind the $VARu$ specification as N increases. Encompassing dominates uniform weighting particularly for smaller samples. On the whole, direct modelling yields much weaker results than iterated modelling for the French design.

Tables 10 and 11 give the average weights that were determined by the model-encompassing step across the 10,000 replications for the same experiment as in Tables 8 and 9. The pattern is typical for most experiments that we run in this project. First consider the U.K. design. For $N = 40$,

Table 8: Mean squared errors (MSE) for candidate models by direct modelling. U.K. design

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR	enc	unif
horizon 2						
40	3.358	3.277	3.652	2.925*	2.868	2.855*
100	2.655*	2.691	2.781	2.732	2.679	2.673*
200	2.595*	2.617	2.674	2.686	2.604	2.601*
500	2.571*	2.578	2.630	2.667	2.662	2.661*
horizon 3						
40	3.349	3.366	3.573	2.984*	2.916*	2.919
100	2.706*	2.730	2.806	2.754	2.714	2.712*
200	2.637*	2.653	2.699	2.705	2.631	2.630*
500	2.611*	2.616	2.655	2.685	2.692	2.691*
horizon 4						
40	3.399	3.412	3.553	3.034*	2.959*	2.964
100	2.757*	2.776	2.835	2.766	2.755*	2.755*
200	2.677*	2.690	2.713	2.710	2.658	2.657*
500	2.648*	2.649	2.665	2.690	2.720	2.719*

Note: see Table 3.

Table 9: Mean squared errors (MSE) for candidate models by direct modelling. French design

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR	enc	unif
horizon 2						
40	0.652	0.648	0.608*	0.608*	0.574*	0.577
100	0.573	0.575	0.553*	0.564	0.560*	0.562
200	0.558	0.559	0.538*	0.555	0.549*	0.549*
500	0.552	0.551	0.531*	0.551	0.546*	0.548
horizon 3						
40	0.731	0.726	0.666	0.656*	0.615*	0.625
100	0.605	0.606	0.578*	0.584	0.574*	0.578
200	0.576	0.579	0.557*	0.570	0.559*	0.561
500	0.564	0.567	0.545*	0.563	0.559*	0.561
horizon 4						
40	0.750	0.750	0.679	0.672*	0.622*	0.637
100	0.622	0.623	0.594	0.593	0.589*	0.594
200	0.586	0.587	0.567*	0.575	0.567*	0.571
500	0.571	0.573	0.553*	0.568	0.570*	0.571

Note: see Table 3.

the bivariate model with exogenous inflation that is most vulnerable to instabilities due to its separate estimation of the marginal and the conditional parts obtains the smallest weight. Conversely, the univariate model has the largest weight. However, the preference for the univariate model does not achieve the obvious dominance of that model for this small sample size that is reported in the MSE tables. We ran the same experiment with looser significance levels, for example at 20%, with little change in the weighting. The dominance of the parsimonious univariate structure remains undetected in statistical comparisons. In larger samples, the weighting becomes more uniform, with a beginning downweighting of the univariate structure. Again, the statistical significance tests fail to exclude the univariate model, although it is not really competitive any more. Conversely, the MSE table shows a clear ranking of the rival models at the sample size of 500, with the univariate model trailing all other models.

Table 10: Test-based weights in the direct modelling experiment. U.K. design

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR
horizon 2				
40	0.133	0.286	0.281	0.299
100	0.183	0.274	0.271	0.273
200	0.245	0.255	0.252	0.249
500	0.264	0.261	0.248	0.227
horizon 3				
40	0.131	0.285	0.284	0.300
100	0.146	0.285	0.285	0.284
200	0.222	0.262	0.259	0.257
500	0.259	0.259	0.249	0.233
horizon 4				
40	0.125	0.287	0.288	0.301
100	0.121	0.293	0.292	0.294
200	0.193	0.270	0.268	0.268
500	0.253	0.254	0.252	0.240

Note: see Table 3.

Table 11: Test-based weights in the direct modelling experiment. French design

N	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	AR
horizon 2				
40	0.080	0.308	0.301	0.311
100	0.087	0.303	0.299	0.310
200	0.162	0.272	0.267	0.299
500	0.213	0.219	0.222	0.345
horizon 3				
40	0.076	0.312	0.298	0.314
100	0.071	0.311	0.300	0.318
200	0.126	0.286	0.273	0.315
500	0.199	0.224	0.205	0.373
horizon 4				
40	0.000	0.336	0.326	0.338
100	0.062	0.317	0.300	0.320
200	0.096	0.299	0.284	0.321
500	0.184	0.240	0.218	0.359

Note: see Table 3.

5 Summary and conclusion

There are several conclusions that we can draw from our extensive simulation experiments. The first and maybe most important is that the merits of the proposed forecast combination method are small in general. This conclusion is, of course, restricted to our simulation designs, which we however deem to be comparatively realistic. Sophisticated combination methods sure dominate if the rival forecasts are extremely different and are based on entirely different information sets. However, this situation is unlikely to be of major empirical relevance, as for example forecasts by competing forecasting institutions necessarily exploit similar information.

Forecast combination, whether based on simple uniform weights or on testing methods, shows its strength if the sample size is very small. Particularly with the aim of multi-step prediction, otherwise convenient time-series models can yield considerable incidences of forecast failure that are mitigated or even eliminated by forecast combinations. On average, sophisticated combinations based on multiple encompassing tests have more or less the same performance as uniform weights. It is to be expected that the advantage of the test-based procedure will show in very large samples, where it should asymptotically eliminate inferior models in the set. However, even for our largest samples with 500 observations, this advantage does not show in the simulation results.

In some experiments, we encountered a deterioration of forecasting performance of the test-based weighted average, as the sample size increases beyond $N = 200$. This counter-intuitive feature indicates that the significance level may deserve to be adapted to the sample size. Unfortunately, in some unreported control simulations with reduced significance level, we were unable to get more convincing results. The usual statistical recommendation to use sharper significance for larger samples, in order to reduce the type-I error risk, must be weighed against the inherent tendency of the test-based procedure to classify differences in forecasting performance as insignificant. A different conjecture is, for this reason, that some weighting of non-encompassed models may strengthen the performance of the procedure in larger samples. This direction will be an issue for further research.

Furthermore, some more conclusions can be drawn, those however with an even more pronounced caveat with regard to sample-specific effects. First, BIC tends to dominate AIC-based selection, particularly in small samples but also for our longer series. While 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, this just means that the asymptotic effect kicks in if the sample sizes exceed the usual macroeconomic sets by far and that the known distortions of AIC dominate. This is another instance for the general recommendation to stick to parsimony in prediction.

In the last subsection, we study another aspect, the competition of direct modelling versus traditional iterated forecasting. It appears that direct modelling deserves attention if the sample size is small, as it avoids the instabilities due to iterating explosive estimates. In larger samples, direct modelling ceases to be interesting and it generally falls behind traditional iterations.

Another issue for further research is to check the robustness of our results against the backdrop of other macroeconomic data sets. We motivated in the beginning that shaping designs on actual data is potentially more relevant to forecasting practitioners than relying on artificial designs.

References

- [1] BATES, J.M., AND GRANGER, C.W.J. (1969) ‘The combination of forecasts,’ *Operations Research Quarterly* **20**, 451–468.
- [2] CHATFIELD, C. (2001) *Time-series forecasting*. Chapman& Hall.
- [3] CLEMENTS, M., AND HENDRY, D.F. (1998) *Forecasting economic time series*. Cambridge University Press.
- [4] HANSEN, B. (2007). Least squares model averaging. *Econometrica*, 75, 1175–1189.
- [5] HARVEY, D.I., AND NEWBOLD, P. (2000) ‘Tests for Multiple Forecast Encompassing,’ *Journal of Applied Econometrics* **15**, 471–482.
- [6] HARVEY, D.I., AND NEWBOLD, P. (2005) ‘Forecast Encompassing and Parameter Estimation,’ *Oxford Bulletin of Economics and Statistics* **67**, 815–835.
- [7] HJORT, N.L., AND CLAESKENS, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98, 879–899.
- [8] JUMAH, A., AND KUNST, R.M. (2008). Seasonal Prediction of European Cereal Prices: Good Forecasts Using Bad Models? *Journal of Forecasting*, 27, 391–406.

- [9] JUSELIUS, K. (2006) *The cointegrated VAR model*. Oxford University Press.
- [10] MARCELLINO, M., STOCK, J.H., AND WATSON, M.W. (2006) ‘A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series,’ *Journal of Econometrics* **135**, 499–526.
- [11] NEWBOLD, P. AND HARVEY, I. H. (2002) ‘Forecast Combination and Encompassing’. In: M.P. Clements & D.F. Hendry *A Companion to Economic Forecasting*. Blackwell Publishers, pp. 268-283.
- [12] MCQUARRIE, A.D.R., AND TSAI, C.-L. (1998) *Regression & time series model selection*, World Scientific.
- [13] SCHORFHEIDE, F. (2004) ‘VAR forecasting under misspecification,’ *Journal of Econometrics* **128**, 99–136.
- [14] SIMS, C.A. (1972) ‘Money, Income, and Causality,’ *American Economic Review* **62**, 540–552.
- [15] TIMMERMANN, A. (2006). Forecast combinations. In: ELLIOTT, G., GRANGER, C.W.J., AND TIMMERMANN, A. *Handbook of Economic Forecasting*, Elsevier.