

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

Mauro Costantini<sup>1</sup> and Robert M. Kunst<sup>2</sup>

Presented at the First Macroeconomic Forecasting Conference,  
Rome, March 27, 2009

---

<sup>1</sup>University of Vienna; [mauro.costantini@univie.ac.at](mailto:mauro.costantini@univie.ac.at)

<sup>2</sup>Institute for Advanced Studies, Vienna, and University of Vienna;  
[kunst@ihs.ac.at](mailto:kunst@ihs.ac.at)

- 1 Introduction
- 2 Multiple forecast encompassing
- 3 The data-based simulation design
- 4 Evaluating the prediction experiments
- 5 Summary and conclusion

# The research question for this paper

Does forecast averaging based on encompassing tests assist in improving predictions for real output growth—a main variable of interest in macroeconomic forecasting—in the framework of core models that additionally include CPI inflation and the unemployment rate?

# Main aspects

- Weights for forecast averaging are determined from an algorithm based on the forecast encompassing tests of HARVEY AND NEWBOLD (2000,2005).
- The potential benefits of the weighted average are evaluated by Monte Carlo experiments tuned to realistic designs that mimic the U.K. and French economies.

# Aspects of forecast averaging

- + An average of forecasts may pick up information from each forecast within a set that would be missed by a 'pure strategy'.
- + Averaging may mitigate the influence of forecast failure in individual models and tends to make the forecast more 'robust'.
- A single prediction based on a correctly specified model should be better than an average over several misspecified models.

For reviews on averaging, see BATES AND GRANGER (1969), CLEMENTS AND HENDRY (1998), TIMMERMAN (2006).

# Methods for determining averaging weights

Among other methods, one may consider:

- 1 uniform weights;
- 2 weights determined by information criteria;
- 3 weights determined by tests over training samples.

Here, we focus on the third option.

# The main idea of the algorithm

Suppose there are  $M$  (usually model-based) out-of-sample forecasts  $\hat{Y}_t^{(k)}$ ,  $k = 1, \dots, M$ , for a vector variable  $Y$ .

Compile all  $M$  forecasts over a training sample  $t = N - n + 1, \dots, N$ . These yield forecast errors  $e_t^{(k)} = Y_{jt} - \hat{Y}_{jt}^{(k)}$  for a component  $Y_j$  in focus.

Run encompassing regressions to check whether any of the  $M$  forecasts encompasses its rivals. Eliminate the encompassed models.

Construct a new forecast by uniform averaging over all models that remain.

# The encompassing regressions

Start with  $j = 1$ . Run the homogeneous regression

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

Consider the regression  $F$ -statistic. If the  $F$ -test accepts its null, model # 1 forecast-encompasses its rivals.

Then, run comparable regressions for  $e_t^{(j)}$ ,  $j = 2, \dots, M$ :

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

# The decision rule

- If all  $M$   $F$ -tests reject or all accept their null hypotheses, the forecast will be a uniformly weighted average of all models.
- If only some  $F$ -tests reject their null, only those models that encompass their rivals will be used in an otherwise uniform average.

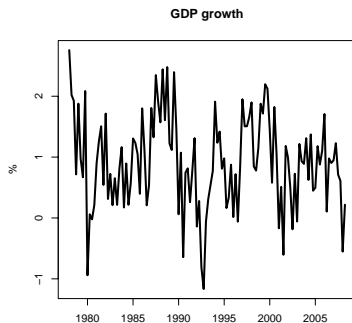
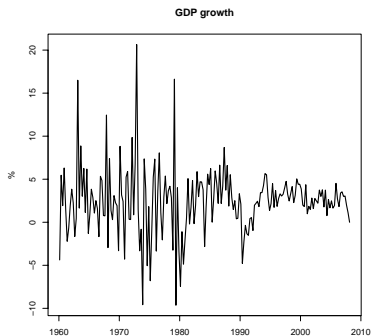
# The data-generation process

The DGP is a vector autoregression (VAR) fitted to quarterly data on three key macroeconomic variables—GDP growth, CPI inflation, unemployment—for the U.K. and for France: two different designs.

The VARs are fitted to the longest available time range: 1960–2008 for the U.K. and 1978–2008 for France.

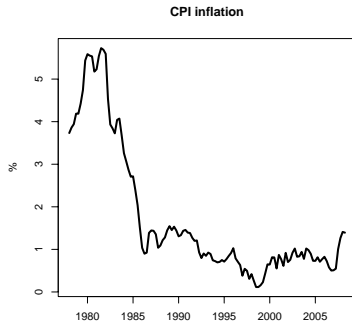
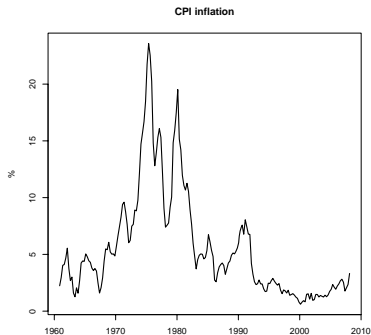
AIC yields a lag order  $p = 2$  for both cases: the DGP is a trivariate VAR with two lags and general errors covariance matrix  $\Sigma$ .

# The GDP growth rates



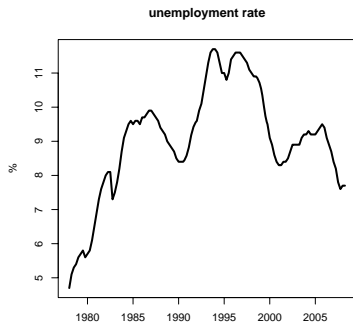
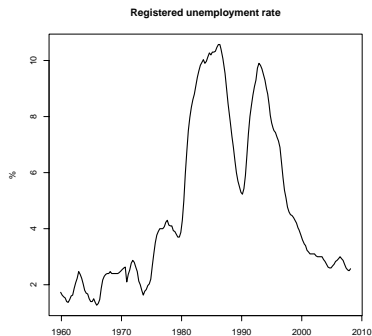
Left graph: U.K.; right graph: France

# The CPI inflation series



Left graph: U.K.; right graph: France

# The unemployment rates



Left graph: U.K.; right graph: France

# The U.K. data-based simulation design

$$Y_t = \mu + \sum_{j=1}^2 \Phi_j Y_{t-j} + \varepsilon_t, \varepsilon_t \sim N(0, \Sigma).$$

$$\mu = (1.414, 0.146, 0.047)',$$

$$\Phi_1 = \begin{pmatrix} -0.141 & -0.221 & -0.739 \\ 0.012 & 1.406 & -0.359 \\ -0.021 & 0.007 & 1.712 \end{pmatrix},$$

$$\Phi_2 = \begin{pmatrix} -0.025 & 0.062 & 0.785 \\ -0.020 & -0.428 & 0.342 \\ -0.020 & 0.005 & -0.718 \end{pmatrix}.$$

# The French data-based simulation design

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

# The errors covariance matrix

For the U.K. data:

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

For the French data:

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

## Details on the Monte Carlo design

- Sample size is varied from  $N = 40$  to  $N = 500$  (original U.K. data has  $N = 194$ , French data has  $N = 122$ ).
- Out-of-sample predictions for the last  $\frac{N}{4} - 1$  observations in each pseudo-sample. Expanding windows.
- Apply the encompassing algorithm to the training sample of  $\frac{N}{4} - 1$  observations and predict the last observation from a weighted model average.
- Number of replications is 10,000.

# A set that includes the generating model

Four model-based forecasts ( $M = 4$ ):

- 1 A trivariate VAR;
- 2 a bivariate VAR for GDP growth and inflation;
- 3 a bivariate VAR for GDP growth and unemployment;
- 4 a univariate AR model for GDP growth.

Within each model class, lag orders are determined via BIC. Note that the DGP is in the class # 1.

Table: 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
$\sigma^2$	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
$\sigma^2$	0.423			

# Main impression from the first experiment

- Univariate model dominates for very small  $N$ ;
- preferred bivariate model—with inflation for the U.K., with unemployment for France—takes over at  $N = 100$ ;
- 'true' trivariate model is in the lead for  $N = 500$ .

# The significance level for the encompassing tests

- Rigorous significance levels imply that each model appears to encompass all its rivals;
- small samples may suggest using a very loose level;
- consistency may demand the level to be taken to zero as  $N \rightarrow \infty$ .

Here, we set the level at 10% for all experiments.

Table: 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
$\sigma^2$	2.468		0.423	

# Main impression from model averaging in the first experiment

- The sophisticated encompassing-test procedure and uniform weighting yield almost identical results;
- there may be a slight relative advantage for the encompassing algorithm as  $N$  increases.

## A set that excludes the generating model

Four model-based forecasts ( $M = 4$ ):

- 1 A bivariate VAR for GDP growth and inflation that does not allow any dynamic feedback from GDP growth to inflation, with componentwise lag search;
- 2 a standard bivariate VAR for GDP growth and inflation;
- 3 a (standard) bivariate VAR for GDP growth and unemployment;
- 4 a univariate AR model for GDP growth.

Within each model class, lag orders are determined via BIC. Note that the generating class has now been excluded.

Table: 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
$\sigma^2$		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
$\sigma^2$		0.423		

## Main impression from the second experiment

- For the U.K. data, the sophisticated bivariate model with exogenous inflation forecasts best;
- For the French data, the bivariate model with unemployment forecasts best.

**Table:** MSE for weighted averages in the second experiment.

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
$\sigma^2$	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
$\sigma^2$	0.423					

# Main impression from averaging in the second experiment

- The univariate model tends to receive small weights as  $N$  increases for the U.K. design, it is 'encompassed';
- the bivariate model with exogenous inflation receives small weights for the French design, it is 'encompassed';
- encompassing tests and uniform weights yield very similar results, test-based weighting improves as  $N$  increases. Not much is gained by excluding the encompassed models.

# AIC instead of BIC?

Control simulations with AIC lag-order searches reveal that:

- Results are comparable;
- BIC is better, particularly for small samples.

# Multi-step prediction

We investigate two strategies for multi-step prediction:

- 1 Iterated prediction. Models are linear, so  $h$ -step conditional expectations  $E(Y_t | \mathcal{I}_{t-h})$  can be obtained by plugging in  $k$ -step forecasts for  $k < h$ .
- 2 Direct modeling. Search for the best model of the form

$$Y_t = \sum_{j=h}^p \Phi_j Y_{t-j} + \varepsilon_t.$$

Table:  $h$ -step MSE. U.K. design, iterated.

$N$	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	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

Table:  $h$ -step MSE. French design, iterated.

$N$	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	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

Table:  $h$ -step MSE. U.K. design, direct.

$N$	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	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

Table:  $h$ -step MSE. French design, direct.

$N$	$VAR_{2S} \pi$	$VAR_2 \pi$	$VAR_2 u$	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

# Main impression from multi-step prediction experiments

- In most cases, iterated prediction is better than direct modeling. In very small samples ( $N = 40$ ), direct modeling avoids instabilities of multivariate prediction models.
- The encompassing test yields better forecast combinations for the French-data design but not for the U.K. data.

# Summary results

- The test-based averaging algorithm fails to dominate simple uniform weighting convincingly.
- Pure strategies often outperform averages, even if none of the forecasts is based on a correctly specified model class.
- It pays to increase the model dimension with increasing  $N$ .

## Caveats and discussion

- **Caveat:** all models use similar information sets and often deliver similar forecasts—a realistic aspect of macroeconomic forecasting in practice. This feature affects the algorithm: there are many instances of equal weights.
- **Discussion:** In artificial designs, the algorithm may dominate convincingly. This investigation uses a ‘realistic’ design.

# Thank you for your attention