

A hierarchical procedure for the combination of forecasts

Mauro Costantini¹

Carmine Pappalardo²

¹University of Vienna, Dept. of Economics

²ISAE, Institute for Studies and Economic Analyses

Rome, Institute for Studies and Economic Analyses,

March 27, 2009

Outline of the talk

- ▶ Encompassing and combination
- ▶ ISAE models
- ▶ Algorithm procedure
- ▶ Empirical results
- ▶ Conclusions

Encompassing

In review method proposed for the evaluation, and in particular the concept of *forecast efficiency*, Granger and Newbold (1973) argue that the assessment in isolation of a set of forecast was of limited value. Rather, it is potentially more informative to compare one's forecast performance with that of a competitor. Ideally, not only should the competing forecast be outperformed, they should contain no useful information about the future not embodied in one's preferred forecasts.

Encompassing

One of the first test of forecast encompassing proposed is that of Chong and Hendry (1986, RES): "... Models which claim to congruently represent a data generation process must be able to account for the finding of rival models. A failure by one model M_1 to encompass some salient features of any rival model M_2 reveal the latter to incorporate information relevant to explaining the observed data, which information relevant is excluded from M_1 . This is so whether M_1 fails to encompass M_2 by (e.g) over or under predicting its goodness of fit since by knowing the data generating process, one could correctly deduce the fit of M_2"

(page. 676)

Forecast combination

Some reasons for forecast combination can be found in the forecasting literature (see Elliot, Granger and Timmermann, 2006)

1) A simple portfolio diversification argument motivates the idea of combining forecasts (Bates and Granger, 1969).

2) A second reason for using forecast combinations is that individual forecasts may be very differently affected by structural breaks caused.

3) A third and related reason for forecast combination is that individual forecasting models may be subject to misspecification bias of unknown form.

Encompassing and Combination: How are they related?

Let $(f_{1t}$ and $f_{2t})$ be two forecasts of the quantity y_t . Assume that these forecasts are unbiased in the sense that the errors:

$$e_{it} = y_t - f_{it}, \quad i = 1, 2 \quad (1)$$

then two errors have mean zero. Bates and Granger (1969) proposed to combine forecasts

$$f_{ct} = (1 - \lambda)f_{1t} + \lambda f_{2t} \quad (2)$$

which for $0 \leq \lambda \leq 1$ is a weighted average of the individual forecasts. The error of the combined forecast is then

$$\varepsilon_t = y_t - f_{ct} = (1 - \lambda)e_{1t} + \lambda e_{2t} \quad (3)$$

which has a zero mean.

Encompassing and Combination: How are they related?

Writing ε_t for the error of the combined forecast, it follows from (2) and (3) that

$$e_{1t} = \lambda(e_{1t} - e_{2t}) + \varepsilon_t \quad (4)$$

The methodology of combination forecast can be exploited to test the null hypothesis that one forecast encompass another, that is, given that the first is available, the second provides no further useful incremental information for prediction. If combination is to be based on a weighted average, a natural way to test the null hypothesis that forecast 1 encompassed forecast 2 is to test $\lambda = 0$ in the following regression (4)

Hierarchical procedure (1)

Step 1. Calculate the RMSFE of the out-of-sample forecast for each model using out-of-sample forecasts and realized values. Rank the models according to their past performance based on RMSFE;

Step 2. Select the best forecasting model (i.e. the model with the lowest RMSFE), and using the HLN statistics test sequentially whether the best forecasting model encompasses other models. If the best model encompasses the alternative model at some significance level α , delete the alternative model from the list;

Hierarchical procedure (2)

- Step 3.** Repeat step 2 with the second best model. The list of models includes the best model and those which are not encompassed by the best model;
- Step 4.** Continue with the third best model and so on, until no encompassed model remains in the list;
- Step 5.** Obtain the hierarchical forecast combination (HFC) with all the previously selected models using several forecast combining methods;

Hierarchical procedure (3)

Final step. Compare for each combining method the RMSFE for the hierarchical forecast combination ($\text{RMSFE}_{\text{HFC}}$) with the one obtained from the single best model (RMSFE_{BM}) and with that obtained from combining all models ($\text{RMSFE}_{\text{ALL}}$). When two relative RMSFE indices are computed ($\frac{\text{RMSFE}_{\text{HFC}}}{\text{RMSFE}_{\text{BM}}}$ and $\frac{\text{RMSFE}_{\text{HFC}}}{\text{RMSFE}_{\text{ALL}}}$), a ratio of less than one denotes that the hierarchical forecast outperforms the competing models.

Encompassing Tests used in the paper (1)

Harvey, Leybourne and Newbold (1998, JBES) proposed several encompassing tests. In the HLN test, if the forecasts from model 1 encompass the forecasts from model 2, then the covariance between e_{1t} and $e_{1t} - e_{2t}$ will be negative or zero, where e_{1t} and e_{2t} are the two sets of forecast errors obtained from using the two models. The alternative hypothesis is that the forecasts from model 1 do not encompass those from model 2, in which case the covariance between e_{1t} and $e_{1t} - e_{2t}$ will be positive.

Encompassing Tests used in the paper (2)

The following statistics is considered in our paper:

$$HLN = D \frac{\bar{d}}{\sqrt{n^{-1}2\pi\widehat{f_d(0)}}}, \quad (5)$$

where $D = n^{-1/2}[n + 1 - 2h + n^{-1}h(h - 1)]^{1/2}$,
 $\bar{d} = n^{-1} \sum_{t=T+1}^{T+n} d_t$, $d_t = e_{1t}(e_{1t} - e_{2t})$, $\widehat{f_d(0)}$ is a consistent estimate of the zero-frequency spectral density of d_t , n denotes the out-of-sample forecast observations and h is the number of steps ahead. In order to obtain a consistent estimate of $f_d(0)$, we follow the recommendations contained in Diebold and Mariano (1995) and Harvey *et al.* (1997).

Multiple encompassing tests used in the paper (1)

Harvey and Nebold (2000, JAE) generalizes the forecast encompassing approach to situations where comparisons of a forecast with more than one competitor are required. Let (f_{1t}, \dots, f_{Kt}) be K competing forecasts (taken to be unbiased or bias-corrected of the actual quantity Z_t) Consider testing the null hypothesis that one forecast, f_1 , encompasses its competitors. The joint testing procedure begins with:

$$f_{ct} = (1 - \lambda_1 - \lambda_2 - \dots - \lambda_{K-1})f_{1t} + \lambda_1 f_{2t} + \lambda_2 f_{3t} + \dots + \lambda_{K-1} f_{Kt} \quad (6)$$

Multiple encompassing tests used in the paper (2)

The equation (6) can alternatively written as

$$e_{1t} = \lambda_1(e_{1t} - e_{2t}) + \lambda_2(e_{1t} - e_{3t}) + \dots + \lambda_{K-1}(e_{1t} - e_{kt}) + \varepsilon_t, \quad (7)$$

where $0 \leq \lambda_i \leq 1$, $e_{it} = Z_t - f_{it}$ and ε_t is the error of the combined forecast. The null hypothesis that f_1 encompasses f_2, \dots, f_K is:

$$H_0 = \lambda_1 = \lambda_2 = \dots = \lambda_{K-1} = 0 \quad (8)$$

Harvey and Newbold refer to this concept as multiple forecast encompassing.

Method of combining forecast used in the paper (1)

The combining methods take the form of a linear combination of the individual forecast:

$$\hat{y}_{c,t+h|h} = w_{0,t} + \sum_{i=1}^n w_{i,t} \hat{y}_{i,t+h|t}, \quad (9)$$

where $\hat{y}_{c,t+h|t}^h$ is a given combination forecast whose weights, $\{w_{i,t}\}_{i=0}^n$ are computed using the individual out-of-sample forecast, y_{t+h}^h are the observations available from the start of the holdout out-of-sample period to time t and n is the number of the models.

Method of combing forecast used in the paper (2)

Three simple methods: the mean, the trimmed mean and the median. With regard to the mean, we set $w_{0,t} = 0$ and $w_{i,t} = \frac{1}{k}$ in equation (9); the trimmed mean uses $w_{0,t} = 0$ and $w_{i,t} = 0$ for the individual models that generate the smallest and largest forecasts at time t , while $w_{i,t} = \frac{1}{(k-2)}$ for the remaining individual models; with respect to the median (case not encompassed by equation (9)), the sample median of the forecasts set $\{\hat{y}_{i,t+h|t}^h\}_{i=1}^k$ is computed;

Method of combining forecast used in the paper (3)

The unrestricted OLS combining method (see Granger and Ramanathan, 1984). The combining weight are calculated using OLS regression; The WLS combining method proposed by Diebold and Pauly (1987). We applied the “t-lambda” method. It consist of a combining method with the combining weights calculated by WLS estimator. Diebold and Pauly (1987) suggested to use the weighting matrix $\Psi = \text{diag}[\Psi_{tt}] = [kt^\lambda]$, where $k, \lambda > 0, t = 1, \dots, T$ and T is the number of observations used in the WLS regression. In our empirical application, we use $\lambda = 1$ (weights that decrease at constant rate) and $\lambda = 3$ (weights that decrease at increasing rate).

Method of combing forecast used in the paper (4)

The DMSFE (Discount Mean Square Forecast Errors) combining methods. Following Stock and Watson (2004), the weights in equation (9) depends inversely on the historical forecasting performance of the individual models:

$$w_{i,t} = \frac{m_{it}^{-1}}{\sum_{j=1}^n m_{jt}^{-1}}, \quad (10)$$

where

$$m_{i,t} = \sum_{s=R}^{t-h} \delta^{t-h-s} (y_{s+h} - \hat{y}_{i,s+h|s}), \quad (11)$$

$w_{it} = 0$, and δ is a discount factor. When $\delta = 1$, there is no discounting; when $\delta < 1$, greater importance is attributed to the recent forecast performance of the individuals. We use $\delta = 0.9, 1.0$.

ISAE models

Seven time series models for forecasting the Italian industrial production (IPI) are used at ISAE: four single-equation models, a dynamic factor model, a VAR model, and an ARIMA model.

Empirical results: model estimation

All models presented in section 4 are estimated over a common sample 1997:7-2005:9. The forecasting exercise is carried out using both recursive and rolling schemes. The latter is generally used when there are concerns about turning points and biases from the use of older information. The rolling scheme is used for a sensitivity analysis with respect to the results of the combination obtained through the hierarchical procedure. For each model, a full set of 24 predictions at several steps ahead ($h=1,\dots,6$) is obtained.

Empirical results: Forecast error measures.

| Models (recursive) | RMSFE(1) | RMSFE(2) | RMSFE(3) | RMSFE(6) |
|--------------------|----------|----------|----------|----------|
| GW | 0.0192 | 0.0211 | 0.0388 | 0.0423 |
| Gas | 0.0205 | 0.0203 | 0.0214 | 0.0270 |
| GW _c | 0.0201 | 0.0168 | 0.0202 | 0.0222 |
| SE | 0.0180 | 0.0182 | 0.0399 | 0.0379 |
| VAR | 0.0236 | 0.0240 | 0.0243 | 0.0206 |
| ARIMA | 0.0297 | 0.0265 | 0.0391 | 0.0413 |
| Factor | 0.0279 | 0.0210 | 0.0372 | 0.0368 |

Notes: numbers in parentheses are the steps-ahead forecast.

Rank Classification. Recursive estimation

| | Recursive | | | |
|------|-----------------|-----------------|-----------------|-----------------|
| rank | h=1 | h=2 | h=3 | h=6 |
| 1 | SE | GW _c | GW _c | VAR |
| 2 | GW | SE | Gas | GW _c |
| 3 | GW _c | Gas | VAR | Gas |
| 4 | Gas | Factor | Factor | Factor |
| 5 | VAR | GW | GW | SE |
| 6 | Factor | VAR | SE | ARIMA |
| 7 | ARIMA | ARIMA | ARIMA | GW |

HLN Encompassing test results. Recursive estimation. (1)

| Recursive | h=1 | h=2 | h=3 | h=6 |
|------------------------|--------------------------------|------------------------------------|------------------------------------|---------------------------------|
| 1 ^o step | Best Model: SE | Best Model: GW _c | Best Model: GW _c | Best Model: VAR |
| p – values (Models) | 0.150 (SE/GW) | 0.004 (GW _c /GW) | 0.220 (GW _c /GW) | 0.278 (VAR/GW) |
| | 0.112 (SE/Gas) | 0.073 (GW _c /Gas) | 0.005 (GW _c /Gas) | 0.422 (VAR/Gas) |
| | 0.103 (SE/GW _c) | 0.061 (GW _c /SE) | 0.230 (GW _c /SE) | 0.286 (VAR/GW _c) |
| | 0.183 (SE/VAR) | 0.561 (GW _c /VAR) | 0.375 (GW _c /VAR) | 0.336 (VAR/SE) |
| | 0.426 (SE/ARIMA) | 0.794 (GW _c /ARIMA) | 0.009 (GW _c /ARIMA) | 0.381 (VAR/ARIMA) |
| | 0.052 (SE/Factor) | 0.880 (GW _c /Factor) | 0.465 (GW _c /Factor) | 0.261 (VAR/Factor) |

HLN Encompassing test results. Recursive estimation. (2)

| 2° step | Best Model: GW | Best Model: SE | Best Model: Gas | Best Model: - |
|------------------------|--------------------------------|-------------------|----------------------|---------------|
| p - values (Models) | | | | |
| $\alpha = 0.25$ | 0.167 (GW/Gas) | 0.508 (SE/GW) | 0.194 (Gas/GW) | |
| | 0.033 (GW/GW _c) | 0.409 (SE/Gas) | 0.190 (Gas/SE) | |
| | 0.031 (GW/VAR) | | 0.420 (Gas/ARIMA) | |
| | 0.075 (GW/Factor) | | | |
| $\alpha = 0.20$ | 0.167 (GW/Gas) | 0.508 (SE/GW) | 0.420 (Gas/ARIMA) | |
| | 0.033 (GW/GW _c) | 0.409 (SE/Gas) | | |
| | 0.031 (GW/VAR) | | | |
| | 0.075 (GW/Factor) | | | |
| $\alpha = 0.15$ | 0.167 (GW/Gas) | 0.508 (SE/GW) | 0.420 (Gas/ARIMA) | |
| | 0.033 (GW/GW _c) | 0.409 (SE/Gas) | | |
| | 0.075 (GW/Factor) | | | |
| $\alpha = 0.10$ | 0.075 (GW/Factor) | 0.508 (SE/GW) | 0.420 (Gas/ARIMA) | |
| $\alpha = 0.05, 0.01$ | - | 0.490 (SE/Gas) | 0.420 (Gas/ARIMA) | |
| | | 0.508 (SE/GW) | | |

HLN Encompassing test results. Recursive estimation. (2)

| | | | | |
|------------------------|----------------------------|---------------|-------------------|---------------|
| 3° step | Best Model: GW_c | Best Model: - | Best Model: VAR | Best Model: - |
| p - values (Models) | | | | |
| $\alpha = 0.25$ | 0.220 (GW_c /Gas) | | 0.373 (VAR/GW) | |
| | 0.322 (GW_c /VAR) | | 0.279 (VAR/SE) | |
| | 0.151 (GW_c /Factor) | | | |
| $\alpha = 0.20$ | 0.220 (GW_c /Gas) | | | |
| | 0.322 (GW_c /VAR) | | | |
| | 0.151 (GW_c /Factor) | | | |
| $\alpha = 0.15, 0.10$ | 0.151 (GW_c /Factor) | | | |
| 4° step | Best Model: Gas | Best Model: - | Best Model: - | Best Model: - |
| p - values (Models) | | | | |
| $\alpha = 0.25, 0.20$ | 0.273 (Gas/Factor) | | | |

Multiple encompassing results. Recursive estimation (1)

| Recursive | h=1 | h=2 | h=3 | h=6 |
|------------------------|-----------------|--------------------|--------------------|-----------------|
| 1 ^o step | Best Model: SE | Best Model: GW_c | Best Model: GW_c | Best Model: VAR |
| F – test (p-values) | 3.16 (0.069) | 3.96 (0.041) | 5.89 (0.018) | 0.94 (0.380) |
| 2 ^o step | Best Model: GW | Best Model: SE | Best Model: Gas | Best Model: - |
| F – test (p-values) | 4.12 (0.034) | 0.63 (0.543) | 2.65 (0.183) | |

Multiple encompassing results. Recursive estimation (2)

| Recursive | h=1 | h=2 | h=3 | h=6 |
|------------------------|-----------------------------|---------------|-----------------|---------------|
| 3 ^o step | Best Model: GW _c | Best Model: - | Best Model: VAR | Best Model: - |
| F - test (p-values) | 2.12 (0.182) | | 0.87 (0.452) | |
| 4 ^o step | Best Model: Gas | Best Model: - | Best Model: - | Best Model: - |
| F - test (p-values) | 1.45 (0.281) | | | |

Relative RMSFE results. Recursive estimation

| Combining | Recursive | | | | | |
|-----------------|-------------------------|---------|---------|--------------------------|---------|---------|
| | RMSFE _{HFC/BM} | | | RMSFE _{HFC/ALL} | | |
| | h=1 | h=2 | h=3 | h=1 | h=2 | h=3 |
| Mean | | | | | | |
| $\alpha = 0.25$ | 0.864** | 0.937** | 0.861** | 0.908** | 0.933** | 0.937** |
| $\alpha = 0.20$ | 0.864** | 0.937** | 0.861** | 0.908** | 0.933** | 0.937** |
| $\alpha = 0.15$ | 0.872** | 0.937** | 0.866** | 0.924** | 0.933** | 0.942** |
| $\alpha = 0.10$ | 0.872** | 0.937** | 0.866** | 0.924** | 0.933** | 0.942** |
| $\alpha = 0.05$ | - | 0.937** | 0.866** | - | 0.933** | 0.942** |
| $\alpha = 0.01$ | - | 0.937** | 0.866** | - | 0.933** | 0.942** |

What are the lessons to be learned from the paper? (1)

Can the superiority of the one forecasting model/method over the other be explained in one way or another?

The comparison of forecast accuracy is but one of many diagnostics that should be examined when comparing models. However, the superiority of a particular model in terms of forecast accuracy does not necessarily imply that forecasts from other models contain no additional information. One could opt for the best alternative, but this strategy discounts the possibility that the discarded forecast could embody useful information not contained in the preferred forecast. This strategy would be frequently sub-optimal (Newbold and Harvey, 2002).

What are the lessons to be learned from the paper? (2)

Which significance level of α ?

Given the number of steps ahead and the estimation scheme, the number of models selected for combination depends on the significance level. The lower the significance level α , the stronger the selection becomes between competing models. As α rises, a larger number of forecasts are selected for combination. The upper significance level is set at $\alpha = 0.25$, since no significant improvement in terms of forecast accuracy is found using the hierarchical combination for higher significance values. Similar results are also found in Kisinbay (2007).

What are the lessons to be learned from the paper? (3)

Can any methodological conclusions be drawn?

The hierarchical procedure described in the paper can be considered as an alternative to the optimal combination method and is capable of outperforming both the best single model and the combination of all models. Consistent with forecasting literature, the simpler averaging methods, as the mean, the trimmed mean and the median, perform better than other methods.

THE END

ISAE models: four single-equation models (2)

The general specification of single-equation multivariate models is:

$$\Delta_{12}y_t = \alpha + \gamma\Delta_{12}y_{t-h} + \sum_{j=h}^p \beta_j x_{t-j}^{m_h} + \delta d_t + \varepsilon_t^{m_h}, \quad (12)$$

where y_t is the log-transformed industrial production index, m denotes the models for each forecasting step ($h=1, \dots, 6$), $\Delta_{12} = (1 - L^{12})$, d_t denotes the deterministic components (month-on-month trading days variation up to 1 lag), x^{m_h} are not seasonally adjusted regressors and ε_t is the idiosyncratic error term. The regressors are log-transformed and seasonal differenced in order to obtain stationarity. All variables are considered at monthly frequency.

The four single-equation models (1)

The SE model includes the quantity of raw materials transported by rails (*TONN*) and the purchasing managers' index (*PMI*) as regressors. The *PMI* index is not differenced and rendered unbounded through the following transformation: $(PMI - 50)/100$. The GW model includes the following regressors: the supply of electric energy (*EE*), the lagged endogenous variable and the *PMI*.

The four single-equation models (2)

In the third model GW_c , the supply of electric energy (EE), PMI (both lagged by 1 period), the variable $\tilde{C}_{q,t}$ and a set of seasonal dummies (which take value equal to $\tilde{C}_{q,t}$ in the reference month and zero otherwise) are included. The $\tilde{C}_{q,t}$ variable is considered since some of the electricity components could be significantly affected by temperature patterns. The Gas model is based on the volume of natural gas required by the industrial sector ($Snam$) and PMI index.

The four single-equation models: reduced form (3)

For each single-equation model, a reduced form is obtained from a general unrestricted model (*GUM*) which is estimated over the period 1997:1-2005:9 using up to the 12th lag of the independent and dependent variables. The General-to-Specific approach is performed running Pc-Gets and the 'conservative' selection strategy is chosen. It delivers an overall significance level of approximately 1% (Hendry and Krolzig, 1999; Krolzig and Hendry, 2001). To get forecasts for more than one-step ahead without using any prediction of the selected indicators, each *GUM* is constructed discarding lower lagged regressors (Rünstler and Sèdillot, 2003).

The Factor model

Following Stock and Watson (1998, 2002), a dynamic factor model (Factor) is estimated:

$$\Delta_{12}y_t^{mh} = \beta_0 + \sum_{i=1}^4 B_i \hat{F}_{i,t-h} + \gamma \Delta_{12}y_{t-h} + \hat{\varepsilon}_t^{mh}, \quad (13)$$

where m denotes the models for each step ($h=1, \dots, 6$), and $i = 1, \dots, 4$ are the number of estimated factors (\hat{F}_{it}). Lagged values of the dependent variables also appear as predictors since the error term can be serially correlated. The factors are extracted from a large data-set of monthly ISAE business surveys regarding the manufacturing sector (current assessments on demand, production and inventories...), expressed in terms of *net balance*.

The end

The VAR model

$$\Delta\Delta_{12}y_t = \alpha\Delta_{12}y_{t-1} + \sum_{j=1}^{13} \beta_j\Delta\Delta_{12}y_{t-j} + \phi d_t + \varepsilon_t, \quad (14)$$

where $y_t = (IPI_t, TONN_t, PP_t)$, $\Delta = (1 - L)$, $\Delta_{12} = (1 - L^{12})$, PP denotes monthly ISAE production expectations which are rendered unbounded through the transformation $-\log(200/(PP + 100) - 1)$ and d_t represents the deterministic components (for the specification of the deterministic components see Bruno and Lupi, 2004).

The ARIMA model

The ARIMA model is included as a benchmark model which involves double differencing, both at regular and seasonal frequencies. According to the Schwarz information criterion for lag length selection, the final specification consists of an ARMA(2,3) polynomial for the regular part, $MA(1)_{12}$ for the seasonal frequencies