Pooling of Forecasts

DAVID F. HENDRY[†], MICHAEL P. CLEMENTS[††]
[†] DEPARTMENT OF ECONOMICS, UNIVERSITY OF OXFORD, OX1 1NF.
[††] DEPARTMENT OF ECONOMICS, UNIVERSITY OF WARWICK,
COVENTRY, CV4 7AL.

E-mail: `david.hendry@economics.ox.ac.uk`
E-mail: `m.p.clements@warwick.ac.uk`

Summary   We consider forecasting using a combination, when no model coincides with a non-constant data generation process (DGP). Practical experience suggests that combining forecasts adds value, and can even dominate the best individual device. We show why this can occur when forecasting models are differentially mis-specified, and is likely to occur when the DGP is subject to deterministic shifts. Moreover, averaging may then dominate over estimated weights in the combination. Finally, it cannot be proved that only non-encompassed devices should be retained in the combination. Empirical and Monte Carlo illustrations confirm the analysis.

1. INTRODUCTION

In the third of a century since Bates and Granger (1969), the combination of individual forecasts of the same event has frequently been found to outperform the individual forecasts, in the sense that the combined forecast delivers a smaller mean-squared forecast error (MSFE) – see *inter alia* Diebold and Lopez (1996) and Newbold and Harvey (2001) for recent surveys, and Clemen (1989) for an annotated bibliography. Studies such as Newbold and Granger (1974) provided early evidence consistent with that claim. Moreover, simple rules for combining forecasts, such as averages (i.e., equal weights), often work as well as more elaborate rules based on the relative past performance of the forecasts to be combined: see Stock and Watson (1999) and Fildes and Ord (2001). Nevertheless, despite some potential explanations (such as Granger (1989)), precisely why forecast combinations should work well does not appear to be fully understood. This paper addresses that issue.

There are a number of potential explanations. First, if two models provide partial, but incompletely overlapping, explanations, then some combination of the two might do better than either alone. In particular, if two forecasts are differentially biased (one upwards, one downwards), it is easy to see why combining could be an improvement over either. However, it is unclear why investigators would construct systematically biased models; and there are other solutions to

forecast biases than pooling. Moreover, it is less easy to see why a combination need improve over the best of a group, particularly if there are some decidedly poor forecasts in that group.

Secondly, in non-stationary time series, most forecasts will fail in the same direction when forecasting over a period within which a break unexpectedly occurs. Combination is unlikely to provide a substantial improvement over the best individual forecasts in such a setting. However, what will occur when forecasting after a deterministic shift depends on the extent of model mis-specification, data correlations, the size of breaks and so on, so combination may help. Since a theory of forecasting allowing for model mis-specification interacting with intermittent deterministic shifts has explained many other features of the empirical forecasting literature (see Clements and Hendry (1999)), we explore the possibility that it can also account for the benefits from pooling.

Thirdly, averaging reduces variance to the extent that separate sources of information are used. Since we allow all models to be differentially mis-specified, such variance reduction remains possible. Nevertheless, we will ignore sample estimation uncertainty below to focus on specification issues, so any gains from averaging reducing that source of variance will be additional to those we delineate.

Next, an alternative interpretation of combination is that, relative to a 'baseline' forecast, additional forecasts act like intercept corrections (ICs). It is well known that appropriate ICs can improve forecasting performance not only if there are structural breaks, but also if there are deterministic mis-specifications. Indeed, Clements and Hendry (1999) present eight distinct interpretations of the role that ICs can play in forecasting, and (e.g.) interpret the cross-country pooling in Hoogstrate, Palm, and Pfann (1996) as a specific form of IC.

Finally, pooling can also be viewed as an application of Stein–James 'shrinkage' estimation (see e.g., Judge and Bock (1978)). If the unknown future value is viewed as a 'meta-parameter' of which all the individual forecasts are estimates, then averaging may provide a 'better' estimate thereof.

Thus, we evaluate the possible benefits of combining forecasts in light of the nature of the economic system and typical macroeconomic models thereof, to discern the properties of the system and models – and the relationships between the two – that result in forecast combination reducing MSFEs. In particular, given that a general theory of economic forecasting which allows for structural breaks and mis-specified models has radically different implications from one that assumes stationarity and well-specified models (see Clements and Hendry (1999) and Hendry and Clements (2001a)), we explore the role of forecast combinations in the former.

Section 2 confirms that combinations of forecasts are ineffective when forecasting using the correct conditional expectation in a weakly-stationary process. Thus, departures from 'optimality', due to mis-specification, mis-estimation, or non-stationarities are necessary to explain gains from combination. Section 3 considers whether combination could deliver gains in a weakly-stationary process when forecasting models are differentially mis-specified by using only subsets of the relevant information. We show there is a range of values of the parameters of the data generation process (DGP) where this can occur, but gains are not guaranteed. Nevertheless, the logic of why gains ensue in such a setting points to why combination might work in general, partly by providing 'insurance' against obtaining the worst forecasts. Section 4 notes alternative ways of implementing forecast combinations, then 5 considers the role of encompassing—which is violated by the need to pool—and discusses whether only non-encompassed models are worth pooling. If the weights used in any combination are estimated, then they directly reflect a lack of encompassing; however, if pre-fixed weights, such as the average, are used, encompassed models may lower rather than raise the efficiency of the combined forecast. Section 6 extends the anal-

ysis to processes subject to deterministic shifts, where the combination can dominate in MSFE. Moreover, previously encompassed models may later become dominant, so averaging across all contenders cannot be excluded as a sensible strategy. Section 7 provides an empirical illustration based on the data set originally used by Bates and Granger (1969), and by demonstrating the efficacy of ICs, suggests that combination works there because of deterministic shifts of the form underlying our theoretical approach. The Monte Carlo study of the behaviour in finite samples of our theoretical approximations in section 8 supports their applicability in practice. Section 9 concludes.

2. FORECASTING BY THE CONDITIONAL EXPECTATION

Consider a weakly-stationary n-dimensional stochastic process $\{\mathbf{x}_t\}$ with density $\mathsf{D}_\mathsf{x}(\mathbf{x}_t|\mathbf{X}_{t-1}, \boldsymbol{\theta})$, which is a function of past information $\mathbf{X}_{t-1} = (\ldots \mathbf{x}_1 \ldots \mathbf{x}_{t-1})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq R^k$. Forecasts of $\mathbf{x}_{T+h}$ based on the conditional expectation given information up to period T:

$$\hat{\mathbf{x}}_{T+h|T} = \mathsf{E}[\mathbf{x}_{T+h} \mid \mathbf{X}_T], \tag{1}$$

are conditionally unbiased:

$$\mathsf{E}\left[\mathbf{x}_{T+h} - \hat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T\right] = \mathsf{E}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T\right] - \mathsf{E}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T\right] = \mathbf{0}, \tag{2}$$

and no other predictor conditional on only $\mathbf{X}_T$ has a smaller MSFE matrix:

$$\mathsf{M}\left[\hat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T\right] = \mathsf{E}\left[\left(\mathbf{x}_{T+h} - \hat{\mathbf{x}}_{T+h|T}\right)\left(\mathbf{x}_{T+h} - \hat{\mathbf{x}}_{T+h|T}\right)' \mid \mathbf{X}_T\right]. \tag{3}$$

Moreover, both (2) and (3) hold for all h. Consequently, on a MSFE basis for forecasting $\mathbf{x}_{T+h}$, the conditional expectation cannot be beaten, as is well known. However, the empirical evidence that combination is useful clearly indicates that the above framework is inappropriate as an analytic basis.

There are several possible explanations for the empirical outcome. First, forecasts $\widetilde{\mathbf{x}}_{T+h|T}$ are used that are based on only subsets of the available information $\mathbf{X}_T$. Secondly, the functions of past data used to form those forecasts do not coincide with the conditional expectation. Thirdly, parameter estimation uncertainty is sufficiently large that averaging is advantageous. Finally, the underlying data density $\mathsf{D}_\mathsf{x}(\mathbf{x}_t|\mathbf{X}_{t-1}, \boldsymbol{\theta})$ is not constant, in which case, the first two mistakes are almost bound to occur as well, particularly if deterministic shifts are the source of the non-constancy.[1] The proliferation of competing forecasting methods and models is also evidence for the first two potential explanations. Here, we first explore the implications of combining the forecasts from mis-specified models when $\mathsf{D}_\mathsf{x}(\cdot)$ is constant, then consider what happens when the DGP is subject to intermittent breaks.

3. FORECASTS FROM MIS-SPECIFIED CONSTANT MODELS

To articulate our approach, we approximate the DGP $\mathsf{D}_\mathsf{x}(\mathbf{x}_t|\mathbf{X}_{t-1}, \boldsymbol{\theta})$ by the constant-parameter first-order vector autoregression (VAR):

$$\mathbf{x}_t = \boldsymbol{\gamma} + \boldsymbol{\Gamma}\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{4}$$

---

[1] We do not consider combination to offset measurement errors in preliminary data: see Gallo and Mariano (1994).

where $\epsilon_t \sim \mathsf{IN}_n\left[\mathbf{0}, \mathbf{\Omega}_\epsilon\right]$. Section 6 considers the impacts of breaks due to deterministic shifts. We focus on 1-step ahead forecasts for $T+1$ from time T purely to simplify the algebra; no issue of principle seems involved in generalizing to multi-step forecasts. Also, we restrict attention to forecasting the scalar y_t, which is one element of $\mathbf{x}_t$, and in this section, assume that, in the absence of structural breaks, $\mathbf{x}_t$ in (4) has been reduced to weak stationarity by appropriate transformations. Thus, partitioning $\mathbf{x}_t' = (\mathbf{x}_{1,t}' : \mathbf{x}_{2,t}')$, the model determining y_t is given by:

$$y_t = \boldsymbol{\beta}_1' \mathbf{x}_{1,t-1} + \boldsymbol{\beta}_2' \mathbf{x}_{2,t-1} + e_t, \tag{5}$$

where $e_t \sim \mathsf{IN}\left[0, \sigma_e^2\right]$, independently of $\mathbf{x}_{t-1}$. Since the processes are all weakly stationary, intercepts are set to zero.

Two investigators unaware of the nature of the process in (5), fit separate models of the form:

$$y_t = \mathbf{a}' \mathbf{x}_{1,t-1} + u_t = \mathbf{a}' \mathbf{w}_t + u_t, \tag{6}$$

and:

$$y_t = \mathbf{b}' \mathbf{x}_{2,t-1} + v_t = \mathbf{b}' \mathbf{z}_t + v_t. \tag{7}$$

Each model is mis-specified by omitting the components which the other includes – the absence of overlapping variables seems an inessential simplification (the switch to $\mathbf{w}_t$ and $\mathbf{z}_t$ is to ease notation below, but note that $\mathbf{w}_{T+1}$ and $\mathbf{z}_{T+1}$ are known at the forecast origin). Moreover, as we believe the explanation for any benefits from combination derive from specification—rather than estimation—issues, we further simplify by neglecting sampling variability in the coefficients $\mathbf{a}$ and $\mathbf{b}$. The assumption that the partial models span the information set is to simplify the algebra, and does not seem consequential: section 8 provides a Monte Carlo illustration.

It must be stressed that in such a constant-parameter framework, pooling the information will produce the optimal forecast, as the resulting model coincides with the DGP, whereas pooling the forecasts will not in general (but see Granger (1989) for an example). However, that implication need not generalize to non-constant DGPs.

Let:

$$\left(\begin{array}{c} \mathbf{w}_t \\ \mathbf{z}_t \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\phi}_{w,t} \\ \boldsymbol{\phi}_{z,t} \end{array} \right) + \left(\begin{array}{c} \boldsymbol{\xi}_{w,t} \\ \boldsymbol{\xi}_{z,t} \end{array} \right), \tag{8}$$

where $\boldsymbol{\phi}_{w,t}$ and $\boldsymbol{\phi}_{z,t}$ are fixed functions of past variables, and:

$$\left(\begin{array}{c} \boldsymbol{\xi}_{w,t} \\ \boldsymbol{\xi}_{z,t} \end{array} \right) \sim \mathsf{IN}_n\left[\left(\begin{array}{c} \mathbf{0} \\ \mathbf{0} \end{array} \right), \left(\begin{array}{cc} \mathbf{\Omega}_{ww} & \mathbf{\Omega}_{wz} \\ \mathbf{\Omega}_{wz} & \mathbf{\Omega}_{zz} \end{array} \right) \right]. \tag{9}$$

Our interest is in comparing the accuracy of the forecasts from the models in (6) and (7) against that of a pooled forecast, based on MSFEs (as that is the criterion most frequently applied in practice: but see Clements and Hendry (1993)). We set $\boldsymbol{\phi}_{w,t} = \boldsymbol{\phi}_{z,t} = \mathbf{0}$, so both dynamics and deterministic factors are ignored, and this is known to the investigators, so intercepts and further lags are omitted: section 8 investigates dynamics via Monte Carlo simulations.

The 1-step ahead forecast from (6) is denoted $\widehat{y}_{T+1} = \widehat{\mathbf{a}} \mathbf{w}_{T+1}$, so the forecast error is:

$$\widehat{u}_{T+1} = y_{T+1} - \widehat{y}_{T+1} = \left(\boldsymbol{\beta}_1 - \widehat{\mathbf{a}}\right)' \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \mathbf{z}_{T+1} + e_{T+1}.$$

The corresponding forecast from (7) uses $\widetilde{y}_{T+1} = \widetilde{\mathbf{b}}' \mathbf{z}_{T+1}$ with:

$$\widetilde{v}_{T+1} = y_{T+1} - \widetilde{y}_{T+1} = \boldsymbol{\beta}_1' \mathbf{w}_{T+1} + \left(\boldsymbol{\beta}_2 - \widetilde{\mathbf{b}}\right)' \mathbf{z}_{T+1} + e_{T+1}.$$

Neither forecast should encompass the other. Section 5 considers testing for non-encompassing before forecast combining.

Next, we derive the conditional biases and variances of the forecast errors. First:

$$\mathsf{E}\left[\widehat{u}_{T+1} \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right] = \left(\boldsymbol{\beta}_1 - \mathsf{E}\left[\widehat{\mathbf{a}}\right]\right)' \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \mathbf{z}_{T+1},$$

and similarly for $\mathsf{E}\left[\widetilde{v}_{T+1} \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right]$. Let $\widehat{\mathbf{a}} = \mathsf{E}\left[\widehat{\mathbf{a}}\right] + \boldsymbol{\delta}_{\widehat{\mathbf{a}}}$, where

$$
\begin{aligned}
\mathsf{E}\left[\widehat{\mathbf{a}}\right] &= \mathsf{E}\left[\left(\sum \mathbf{w}_t \mathbf{w}_t'\right)^{-1} \sum \mathbf{w}_t y_t\right] \\
&= \boldsymbol{\beta}_1 + \mathsf{E}\left[\left(\sum \mathbf{w}_t \mathbf{w}_t'\right)^{-1} \sum \mathbf{w}_t \mathbf{z}_t'\right] \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \boldsymbol{\Pi}_{zw}' \boldsymbol{\beta}_2,
\end{aligned}
\tag{10}
$$

where:

$$\boldsymbol{\Pi}_{zw}' = \mathsf{E}\left[\left(\sum \mathbf{w}_t \mathbf{w}_t'\right)^{-1} \sum \mathbf{w}_t \mathbf{z}_t'\right] = \boldsymbol{\Omega}_{ww}^{-1} \boldsymbol{\Omega}_{wz},$$

using:

$$\mathbf{z}_t = \boldsymbol{\Pi}_{zw} \mathbf{w}_t + \boldsymbol{\eta}_{zw,t} \quad \text{where} \quad \mathsf{E}\left[\mathbf{w}_t \boldsymbol{\eta}_{zw,t}'\right] = \mathbf{0}. \tag{11}$$

Notice that:

$$\mathsf{V}\left[\mathbf{z}_t\right] = \boldsymbol{\Pi}_{zw} \mathsf{V}\left[\mathbf{w}_t\right] \boldsymbol{\Pi}_{zw}' + \mathsf{V}\left[\boldsymbol{\eta}_{zw,t}\right],$$

where $\mathsf{V}\left[\cdot\right]$ denotes a variance, so:

$$\mathsf{V}\left[\boldsymbol{\eta}_{zw,t}\right] = \boldsymbol{\Omega}_{\eta_{zw}} = \boldsymbol{\Omega}_{zz} - \boldsymbol{\Omega}_{zw} \boldsymbol{\Omega}_{ww}^{-1} \boldsymbol{\Omega}_{wz},$$

and:

$$\mathsf{V}\left[\widehat{\mathbf{a}}\right] = T^{-1} \sigma_e^2 \boldsymbol{\Omega}_{ww}^{-1}.$$

Similarly:

$$\mathsf{E}\left[\widetilde{\mathbf{b}}\right] = \boldsymbol{\Pi}_{wz}' \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2,$$

where:

$$\boldsymbol{\Pi}_{wz}' = \mathsf{E}\left[\left(\sum \mathbf{z}_t \mathbf{z}_t'\right)^{-1} \sum \mathbf{z}_t \mathbf{w}_t'\right] = \boldsymbol{\Omega}_{zz}^{-1} \boldsymbol{\Omega}_{zw}.$$

Thus:

$$
\begin{aligned}
\widehat{u}_{T+1} &= \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1 - \boldsymbol{\Pi}_{zw}' \boldsymbol{\beta}_2 - \boldsymbol{\delta}_{\widehat{\mathbf{a}}}\right)' \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \boldsymbol{\Pi}_{zw} \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \boldsymbol{\eta}_{zw,T+1} + e_{T+1} \\
&= -\boldsymbol{\delta}_{\widehat{\mathbf{a}}}' \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \boldsymbol{\eta}_{zw,T+1} + e_{T+1},
\end{aligned}
$$

with:

$$\mathsf{E}\left[\widehat{u}_{T+1} \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right] = \boldsymbol{\beta}_2 \boldsymbol{\eta}_{zw,T+1}.$$

Letting $\mathsf{M}\left[\cdot\right]$ denote MSFE:

$$
\begin{aligned}
\mathsf{E}\left[\widehat{u}_{T+1}^2 \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right] &= \mathsf{M}\left[\widehat{u}_{T+1} \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right] \\
&= \mathbf{w}_{T+1}' \mathsf{V}\left[\widehat{\mathbf{a}}\right] \mathbf{w}_{T+1} + \sigma_e^2 + \boldsymbol{\beta}_2' \boldsymbol{\Omega}_{\eta_{zw}} \boldsymbol{\beta}_2 \\
&\simeq \sigma_e^2 + \boldsymbol{\beta}_2' \boldsymbol{\Omega}_{\eta_{zw}} \boldsymbol{\beta}_2
\end{aligned}
$$

where the final expression ignores terms of $\mathsf{O}_p\left(T^{-1}\right)$. Similarly:

$$\widetilde{v}_{T+1} = -\boldsymbol{\delta}'_{\widehat{\mathbf{b}}}\mathbf{z}_{T+1} + \boldsymbol{\beta}'_1\boldsymbol{\eta}_{xz,T+1} + e_{T+1},$$

with:

$$\mathsf{E}\left[\widetilde{v}^2_{T+1} \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right] = \mathbf{z}'_{T+1}\mathsf{V}\left[\widehat{\mathbf{b}}\right]\mathbf{z}_{T+1} + \sigma^2_e + \boldsymbol{\beta}'_1\boldsymbol{\Omega}_{\eta_{wz}}\boldsymbol{\beta}_1$$
$$\simeq \sigma^2_e + \boldsymbol{\beta}'_1\boldsymbol{\Omega}_{\eta_{wz}}\boldsymbol{\beta}_1.$$

To order the outcome accuracy, we assume $\mathsf{E}\left[\widehat{u}^2_{T+1}|w_{T+1}, z_{T+1}\right] < \mathsf{E}\left[\widehat{v}^2_{T+1}|w_{T+1}, z_{T+1}\right]$, so $\boldsymbol{\beta}'_2\boldsymbol{\Omega}_{\eta_{zw}}\boldsymbol{\beta}_2 < \boldsymbol{\beta}'_1\boldsymbol{\Omega}_{\eta_{wz}}\boldsymbol{\beta}_1$. Consequently, $\widehat{y}_{T+1}$ would transpire on average to be the more accurate forecast here: equivalent results hold for the opposite ranking.

A combined forecast is:

$$\widehat{\widetilde{y}}_{T+1} = (1 - \lambda)\,\widehat{y}_{T+1} + \lambda\widetilde{y}_{T+1} = \widehat{y}_{T+1} + \lambda\left(\widetilde{y}_{T+1} - \widehat{y}_{T+1}\right),$$

where the last expression relates pooling to intercept correction, with error:

$$\widehat{\widetilde{e}}_{T+1} = (y_{T+1} - \widehat{y}_{T+1}) + \lambda\left(\widehat{y}_{T+1} - \widetilde{y}_{T+1}\right) = \widehat{u}_{T+1} + \lambda\left(\widetilde{v}_{T+1} - \widehat{u}_{T+1}\right)$$
$$= -\boldsymbol{\delta}'_{\widehat{\mathbf{a}}}\mathbf{w}_{T+1} + \boldsymbol{\beta}'_2\boldsymbol{\eta}_{zw,T+1} + e_{T+1}$$
$$+ \lambda\left(\boldsymbol{\delta}'_{\widehat{\mathbf{a}}}\mathbf{w}_{T+1} + \boldsymbol{\beta}'_1\boldsymbol{\eta}_{wz,T+1} - \boldsymbol{\beta}'_2\boldsymbol{\eta}_{zwx,T+1} - \boldsymbol{\delta}'_{\widehat{\mathbf{b}}}\mathbf{z}_{T+1}\right),$$

so:

$$\mathsf{E}\left[\widehat{\widetilde{e}}_{T+1} \mid w_{T+1}, z_{T+1}\right] = \lambda\boldsymbol{\beta}'_1\boldsymbol{\eta}_{wz,T+1} + (1 - \lambda)\,\boldsymbol{\beta}'_2\boldsymbol{\eta}_{zw,T+1}.$$

Also (ignoring terms of $\mathsf{O}_p(T^{-1})$):

$$\mathsf{E}\left[\widehat{\widetilde{e}}^2_{T+1} \mid \mathbf{w}_{T+1}, \mathbf{z}_{T+1}\right]$$
$$\simeq \sigma^2_e + \lambda^2\boldsymbol{\beta}'_1\boldsymbol{\Omega}_{\eta_{wz}}\boldsymbol{\beta}_1 + (1 - \lambda)^2\boldsymbol{\beta}'_2\boldsymbol{\Omega}_{\eta_{zw}}\boldsymbol{\beta}_2 + 2\lambda\left(1 - \lambda\right)\boldsymbol{\beta}'_1\mathsf{E}\left[\boldsymbol{\eta}_{wz,T+1}\boldsymbol{\eta}'_{zw,T+1}\right]\boldsymbol{\beta}_2$$

where:

$$\mathsf{E}\left[\boldsymbol{\eta}_{wz,T+1}\boldsymbol{\eta}'_{zw,T+1}\right] = \mathsf{E}\left[(\mathbf{z}_t - \boldsymbol{\Pi}_{zw}\mathbf{w}_t)\,(\mathbf{w}_t - \boldsymbol{\Pi}_{wz}\mathbf{z}_t)'\right]$$
$$= -\boldsymbol{\Omega}_{zz}\boldsymbol{\Omega}^{-1}_{zz}\boldsymbol{\Omega}_{zw} + \boldsymbol{\Omega}_{zw}\boldsymbol{\Omega}^{-1}_{ww}\boldsymbol{\Omega}_{wz}\boldsymbol{\Omega}^{-1}_{zz}\boldsymbol{\Omega}_{zw}$$
$$= -\boldsymbol{\Omega}_{zw}\left(\mathbf{I}_{n_1} - \boldsymbol{\Pi}'_{zw}\boldsymbol{\Pi}'_{wz}\right).$$

The last line is the matrix analogue of $(1 - \mathsf{R}^2_{wz})$, and has a negative sign: intuitively, if the regression of $\mathbf{z}_t$ on $\mathbf{w}_t$ over- (under-) estimates, the reverse regression will do the opposite.

Stock and Watson (1999) find that a combination obtained by pooling forecasts across many methods does well, using either the mean or median forecast, so we focus on the case where $\lambda = 0.5$. Then:

$$\mathsf{M}\left[\widehat{\widetilde{e}}_{T+1} \mid w_{T+1}, z_{T+1}\right]$$
$$\simeq \sigma^2_e + 0.25\left[\boldsymbol{\beta}'_1\boldsymbol{\Omega}_{\eta_{wz}}\boldsymbol{\beta}_1 + \boldsymbol{\beta}'_2\boldsymbol{\Omega}_{\eta_{zw}}\boldsymbol{\beta}_2 - 2\boldsymbol{\beta}'_1\boldsymbol{\Omega}_{zw}\left(\mathbf{I}_{n_1} - \boldsymbol{\Pi}'_{zw}\boldsymbol{\Pi}'_{wz}\right)\boldsymbol{\beta}_2\right], \qquad (12)$$

as against the smaller of the two individual forecast errors:

$$\mathsf{M}\left[\widehat{u}_{T+1} \mid w_{T+1}, z_{T+1}\right] \simeq \sigma^2_e + \boldsymbol{\beta}'_2\boldsymbol{\Omega}_{\eta_{zw}}\boldsymbol{\beta}_2.$$

So:

$$\mathsf{M}\left[\widehat{\widehat{e}}_{T+1} \mid w_{T+1}, z_{T+1}\right] < \mathsf{M}\left[\widehat{u}_{T+1} \mid w_{T+1}, z_{T+1}\right],$$

if and only if:

$$\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{zw} \left(\mathbf{I}_{n_1} - \boldsymbol{\Pi}_{zw}' \boldsymbol{\Pi}_{wz}'\right) \boldsymbol{\beta}_2 < 3\boldsymbol{\beta}_2' \boldsymbol{\Omega}_{\eta_{zw}} \boldsymbol{\beta}_2.$$

Let $\boldsymbol{\beta}_2' \boldsymbol{\Omega}_{\eta_{zw}} \boldsymbol{\beta}_2 = k\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1$ where $k < 1$, then combination dominance requires:

$$\left(1 - 3k\right) \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{zw} \left(\mathbf{I}_{n_1} - \boldsymbol{\Pi}_{zw}' \boldsymbol{\Pi}_{wz}'\right) \boldsymbol{\beta}_2 < 0.$$

This is more likely to hold if the marginal effects of $\mathbf{w}$ and $\mathbf{z}$ on y in the DGP are of the same sign and 'match' the sign of $\boldsymbol{\Omega}_{zw}$.

In the special case that $\boldsymbol{\Omega}_{zw} = \mathbf{0}$, combination dominance requires:

$$1 < 3k,$$

so an improvement over the better individual forecast by averaging is possible within that range (and similarly for the alternative ranking). However, the larger forecast error was:

$$\sigma_e^2 + \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1,$$

as against (12), so when $\boldsymbol{\Omega}_{zw} = \mathbf{0}$, dominance requires:

$$k < 3,$$

which is bound to hold. Thus, averaging guarantees 'insurance', and may provide dominance when the models are differentially mis-specified for a constant DGP.

3.1. Scalar case

In the scalar case when $n_1 = n_2 = 1$, somewhat more transparent results can be obtained. Denote the correlation between w and z by r_{wz} and their variances by σ_w^2 and σ_z^2, then domination by the average over the best requires:

$$\left(1 - 3k\right) \beta_1^2 \rho - 2\beta_1 \beta_2 \mathsf{r}_{wz} < 0,$$

for $\rho = \sigma_w / \sigma_z > 0$ with $\beta_2^2 \sigma_z^2 = k\beta_1^2 \sigma_w^2$. Normalizing such that $\beta_1 = \beta_2 = 1$, then $k = 1/\rho^2$ so $\rho > 1$ and dominance requires:

$$\rho^2 - 2\mathsf{r}_{wz}\rho - 3 < 0 \ \text{ subject to } \ \rho > 1.$$

This is bound to hold when ρ is close to unity, and also for $\rho < 3$ when r_{wz} is close to $+1$.

Also, against the larger forecast error (again using the normalized parameter values):

$$3\rho^2 + 2\rho\mathsf{r}_{wz} > 1,$$

which must always hold even when $\mathsf{r}_{wz} < 0$. Thus, combination—even by averaging—seems likely to be advantageous here.

4. IMPLEMENTING FORECAST COMBINATIONS

Forecast combination can be implemented in many different ways: see Granger and Ramanathan (1984), Diebold (1988), Wall and Correia (1989) and Coulson and Robins (1993) among others. Potential approaches range from simple averaging to more complex schemes designed to give optimal combination weights. In this last case, the weights are often estimated to optimize some criterion (e.g., minimizing the MSFE of the combined forecast) on a post-model-estimation 'training sample' for which the realizations are available, prior to undertaking genuine out-of-sample forecasting. Sometimes the individual models' explanatory variables will be assumed known, and the true values can be conditioned on, at either training or forecasting stages, or alternatively these may themselves be forecast.

   Forecasting is seldom a 'one-off' venture, and typically forecasts will be made at a number of successive forecast origins. The individual models may be re-specified and/or re-estimated at each origin, as may the combination weights – one can imagine the training window moving through the sample as the forecast origin progresses. The estimation windows may be of fixed length so that early observations are dropped, or may expand indefinitely. The success (or otherwise) of forecast combining is likely to depend in part on how it is implemented, so that explanations of its efficacy will be multi-faceted. Nevertheless, given a careful articulation of the context in which forecasting is undertaken, it should be possible to determine which factors are likely to play a key role.[2]

4.1. Forecast combination as a bias correction

Suppose $\{\widehat{y}_{T+i}, \widetilde{y}_{T+i}\}$ denotes a set of forecasts over a training period $i = 1, \ldots, R$, where $\widehat{y}_{T+i}$ is the 1-step ahead forecast of y at $T + i$ based on $T + i - 1$, etc., and the parameter estimates are based on a sample over $1, \ldots, T$. We allow the forecasts to be biased, possibly because they are generated from assumed constant-parameter models in the presence of structural breaks: Granger (1989) recommends 'unbias(ing) the component forecasts' prior to combination. Thus, $\mathsf{E}\left[y_{T+i} - \widehat{y}_{T+i}\right] \neq 0$ and $\mathsf{E}\left[y_{T+i} - \widetilde{y}_{T+i}\right] \neq 0$ for $i = 1, \ldots, R$, and this is reflected in non-zero values of the corresponding sample moments. Suppose the weights are calculated to minimize the MSFE of the combined forecast, imposing the restriction that the weights sum to unity, and allowing for bias by including an intercept. Letting $\mathbf{y}$, $\widehat{\mathbf{y}}$ and $\widetilde{\mathbf{y}}$ denote the vectors of observations over $T + 1$ to $T + R$, the weight α is estimated from:

$$\mathbf{y} = \delta\mathbf{i} + \alpha\widehat{\mathbf{y}} + (1 - \alpha)\widetilde{\mathbf{y}} + \varepsilon \tag{13}$$

where $\mathbf{i}$ is an R-dimensional vector of 1s, or:

$$\mathbf{y} - \widetilde{\mathbf{y}} = \delta\mathbf{i} + \alpha\left[(\mathbf{y} - \widetilde{\mathbf{y}}) - (\mathbf{y} - \widehat{\mathbf{y}})\right] + \varepsilon.$$

By the Frisch–Waugh theorem (see Frisch and Waugh (1933)), one can equivalently run the regression of $\mathbf{M_i}(\mathbf{y} - \widetilde{\mathbf{y}})$ on $\mathbf{M_i}(\widetilde{\mathbf{y}} - \widehat{\mathbf{y}})$ where $\mathbf{M_i} = \mathbf{I}_R - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'$:

$$\mathbf{M_i}(\mathbf{y} - \widetilde{\mathbf{y}}) = \widehat{\alpha}\left[\mathbf{M_i}(\mathbf{y} - \widetilde{\mathbf{y}}) - \mathbf{M_i}(\mathbf{y} - \widehat{\mathbf{y}})\right] + \widehat{\varepsilon}.$$

---

[2]As an example of a possible factor, consider the early successes based on the combination of time-series models and (largely) static economic models. The failure to model the dynamics in the latter and the absence of causal factors in the former constituted important sources of model mis-specification.

Using:

$$\mathbf{M_i}\left(\mathbf{y}-\widetilde{\mathbf{y}}\right) = \mathbf{y} - \left[\widetilde{\mathbf{y}} + \mathbf{i}R^{-1}\mathbf{i}'\left(\mathbf{y}-\widetilde{\mathbf{y}}\right)\right] = \mathbf{y} - \left[\widetilde{\mathbf{y}} + \widetilde{\theta}\mathbf{i}\right],$$

where $\widetilde{\theta}$ is the sample estimate of the bias in $\widetilde{\mathbf{y}}$, and $\widetilde{\mathbf{y}}_{bc} = \widetilde{\mathbf{y}} + \widetilde{\theta}\mathbf{i}$ is the bias-corrected forecast. Similarly, $\mathbf{M_i}\left(\mathbf{y}-\widehat{\mathbf{y}}\right) = \mathbf{y} - \widehat{\mathbf{y}}_{bc}$, where

$$\widehat{\mathbf{y}}_{bc} = \widehat{\mathbf{y}} + \mathbf{i}R^{-1}\mathbf{i}'\left(\mathbf{y}-\widehat{\mathbf{y}}\right) = \widehat{\mathbf{y}} + \widehat{\theta}\mathbf{i},$$

so:

$$\mathbf{y} = \widehat{\alpha}\widehat{\mathbf{y}}_{bc} + \left(1-\widehat{\alpha}\right)\widetilde{\mathbf{y}}_{bc} + \widehat{\varepsilon}.$$

The combination forecast is

$$\widehat{\widetilde{y}}_{bc,T+R+i} = \widehat{\alpha}\widehat{y}_{bc,T+R+i} + \left(1-\widehat{\alpha}\right)\widetilde{y}_{bc,T+R+i}, \quad i \geq 1,$$

that is, a combination of the bias-corrected forecasts. Bias correction should account for a reduction in the MSFE, so that the appropriate benchmarks for the combined forecast should be $\widetilde{y}_{bc}$ and $\widehat{y}_{bc}$ rather than $\widetilde{y}$ and $\widehat{y}$. In practice, the combined forecast is usually only compared to the uncorrected individual forecasts.

An alternative interpretation of the role of δ in (13) is as an 'intercept correction' for the forecast given by $\widehat{\widetilde{y}}_{bc,T+R+i}$. This interpretation is clearer if we assume there is just a single forecast $\widetilde{\mathbf{y}}$, so that the problem is simply to calculate δ in:

$$\mathbf{y} = \delta\mathbf{i} + \widetilde{\mathbf{y}} + \varepsilon \tag{14}$$

or:

$$\mathbf{y} - \widetilde{\mathbf{y}} = \delta\mathbf{i} + \varepsilon$$

so that $\widetilde{\delta} = \widetilde{\theta}$, namely the sample estimate of the bias. If $\widetilde{\delta} > 0$ because of a tendency to under-predict, the intercept-corrected forecasts $\widetilde{y}_{bc,T+R+i} = \widetilde{y}_{T+R+i} + \widetilde{\delta}$ are revised up by that amount.

5. THE ROLE OF ENCOMPASSING

When fixed weights are used (as in an average), it is easy to illustrate a case where only non-encompassed models are worth pooling. In particular, when (5) is one of the forecasting equations, averaging with any subset model or models will produce systematically poorer forecasts. This should hold more generally for weakly-stationary processes—since all other forecasts are then inferentially redundant—and suggests testing for forecast encompassing prior to averaging: see Harvey, Leybourne, and Newbold (1998) and Diebold (1989), who relate encompassing to forecast combinations. Ericsson and Marquez (1993) and Andrews, Minford, and Peel (1995) provide empirical examples of forecast-encompassing tests. However, section 6.4 provides a counter example in processes subject to deterministic shifts where an encompassed model may later dominate: since breaks seem pandemic in macroeconomics, no general result can be established.

5.1. Estimated weights

Two forces operate here. First, under weak stationarity, there is the detrimental effect of the uncertainty added by estimation of the weights. Secondly, there is an offset from the benefit of choosing the best weights. Overall, we suspect estimation probably does not explain much of the success of pooling: whether or not the weights are estimated, combining must be better than the worst of the individual forecasts, and could beat the best. Section 8 shows that this occurs in the Monte Carlo.

When the weights are estimated by regression, then any forecast which contributes to a combination is not encompassed by the others (see Chong and Hendry (1986)). Thus, estimated weights assign little role to encompassed forecasts, as their weights will be insignificant. While the need to pool violates encompassing (see Lu and Mizon (1991), and Ericsson (1992)), and so reveals non-congruence, congruence *per se* cannot be established as a necessary feature for good forecasting: see Hendry and Clements (2001a). Indeed, the next section suggests that averaging might be preferable when unanticipated breaks can occur. Section 8 confirms that estimated weights need not dominate over fixed.

6. COMBINING UNDER EXTRANEOUS STRUCTURAL BREAKS

Hendry and Doornik (1997) and Hendry (2000) establish that deterministic shifts are the problematic class of structural breaks in a forecasting context, so we focus on those. We consider a DGP where the regressor processes $\mathbf{x}_{1,t-1}$ and $\mathbf{x}_{2,t-1}$ in (5) experience breaks at different times, but the forecasting model remains unchanged. Thus, $\phi_{w,t}$ and $\phi_{z,t}$ in (8) are non-constant, beyond being functions of past variables. The DGP for the y process in terms of $\mathbf{w}_t$ and $\mathbf{z}_t$ remains:

$$y_t = \boldsymbol{\beta}_1'\mathbf{w}_t + \boldsymbol{\beta}_2'\mathbf{z}_t + e_t, \tag{15}$$

where $e_t \sim \mathsf{IN}\left[0, \sigma_e^2\right]$. As before, dynamics and intercepts are assumed absent merely to simplify the algebra, so prior to forecasting, $\phi_{z,t} = \phi_{w,t} = \mathbf{0}$, whereas in-sample:

$$\left(\begin{array}{c} \mathbf{w}_t \\ \mathbf{z}_t \end{array} \right) \sim \mathsf{IN}_n \left[\left(\begin{array}{c} \mathbf{0} \\ \mathbf{0} \end{array} \right), \left(\begin{array}{cc} \boldsymbol{\Omega}_{ww} & \boldsymbol{\Omega}_{wz} \\ \boldsymbol{\Omega}_{wz} & \boldsymbol{\Omega}_{zz} \end{array} \right) \right]. \tag{16}$$

Again, the investigators fit separate models of the form:

$$y_t = a_0 + \mathbf{a}_1'\mathbf{w}_t + u_t, \tag{17}$$

$$y_t = b_0 + \mathbf{b}_1'\mathbf{z}_t + v_t. \tag{18}$$

Now intercepts are included, to offset any mean values induced by deterministic shifts. We first allow only the $\mathbf{z}$ process to shift by $\phi_{z,T+1} = \boldsymbol{\tau}_z$ (redefined to simplify notation) which is in fact a change at the end of the estimation sample, influencing the forecast-period behaviour of y. Since the shifts occur in the processes determining the regressors, we refer to these as extraneous breaks.

The 1-step ahead forecast from (17) is:

$$\widehat{y}_{T+1} = \widehat{a}_0 + \widehat{\mathbf{a}}_1'\mathbf{w}_{T+1},$$

so the forecast error $\widehat{u}_{T+1} = y_{T+1} - \widehat{y}_{T+1}$ is:

$$\begin{aligned}
\widehat{u}_{T+1} &= (\boldsymbol{\beta}_1 - \widehat{\mathbf{a}}_1)' \, \mathbf{w}_{T+1} - \widehat{a}_0 + \boldsymbol{\beta}_2' \mathbf{z}_{T+1} + e_{T+1} \\
&= (\boldsymbol{\beta}_2' \boldsymbol{\tau}_z - \widehat{a}_0) + (\boldsymbol{\beta}_1 - \widehat{\mathbf{a}}_1)' \, \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \boldsymbol{\xi}_{z,T+1} + e_{T+1},
\end{aligned} \tag{19}$$

using (8), where we have placed the changed term first. The corresponding forecast from (18) uses $\widetilde{y}_{T+1} = \widetilde{b}_0 + \widetilde{\mathbf{b}}_1' \mathbf{z}_{T+1}$ with $\widetilde{v}_{T+1} = y_{T+1} - \widetilde{y}_{T+1}$:

$$\begin{aligned}
\widetilde{v}_{T+1} &= \boldsymbol{\beta}_1' \mathbf{w}_{T+1} + \left(\boldsymbol{\beta}_2 - \widetilde{\mathbf{b}}_1 \right)' z_{T+1} - \widetilde{b}_0 + e_{T+1} \\
&= -\widetilde{b}_0 + \left(\boldsymbol{\beta}_2 - \widetilde{\mathbf{b}}_1 \right)' \mathbf{z}_{T+1} + \boldsymbol{\beta}_1' \boldsymbol{\xi}_{w,T+1} + e_{T+1}.
\end{aligned}$$

Next, we derive the conditional biases and variances of the forecast errors. This requires the relationship equations between the regressors, of which the first is given by:

$$\mathbf{z}_{T+1} = \boldsymbol{\psi} + \boldsymbol{\Pi}_{zw} \mathbf{w}_{T+1} + \boldsymbol{\eta}_{zw,T+1} \quad \text{where} \quad \mathsf{E}\left[\boldsymbol{\eta}_{zw,T+1}\right] = \mathbf{0} \quad \mathsf{E}\left[\mathbf{w}_{T+1}\boldsymbol{\eta}_{zw,T+1}'\right] = \mathbf{0} \quad (20)$$

so:

$$\begin{pmatrix} \boldsymbol{\psi}' \\ \boldsymbol{\Pi}_{zw}' \end{pmatrix} \simeq \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{ww} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\tau}_z' \\ \boldsymbol{\Omega}_{wz} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\tau}_z' \\ \boldsymbol{\Omega}_{ww}^{-1}\boldsymbol{\Omega}_{wz} \end{pmatrix}. \tag{21}$$

Thus, from the estimation sample, prior to any shifts, and assuming least-squares estimates of in-sample parameters, from (10):

$$\mathsf{E}\left[\widehat{a}_0\right] = \mathbf{0} \quad \text{and} \quad \mathsf{E}\left[\widehat{\mathbf{a}}_1\right] = \boldsymbol{\beta}_1 + \boldsymbol{\Pi}_{zw}' \boldsymbol{\beta}_2,$$

so:

$$\begin{aligned}
\mathsf{E}\left[\widehat{u}_{T+1} \mid w_{T+1}, z_{T+1}\right] &= -\mathsf{E}\left[\widehat{a}_0\right] + (\boldsymbol{\beta}_1 - \mathsf{E}\left[\widehat{\mathbf{a}}_1\right])' \, \mathbf{w}_{T+1} + \boldsymbol{\beta}_2' \mathbf{z}_{T+1} \\
&= \boldsymbol{\beta}_2' \left(\mathbf{z}_{T+1} - \boldsymbol{\Pi}_{zw} \mathbf{w}_{T+1} \right) \\
&= \boldsymbol{\beta}_2' \boldsymbol{\tau}_z + \boldsymbol{\beta}_2' \boldsymbol{\eta}_{zw,T+1},
\end{aligned}$$

using (21). Again we ignore $\mathsf{O}_p\left(T^{-1}\right)$ terms arising from estimation, so:

$$\widehat{u}_{T+1} \simeq \boldsymbol{\beta}_2' \boldsymbol{\tau}_z + \boldsymbol{\beta}_2' \boldsymbol{\eta}_{zx,T+1} + e_{T+1},$$

with:

$$\mathsf{E}\left[\widehat{u}_{T+1}^2 \mid w_{T+1}, z_{T+1}\right] \simeq \sigma_e^2 + \boldsymbol{\beta}_2' \left[\boldsymbol{\Omega}_{\eta_{zw}} + \boldsymbol{\tau}_z \boldsymbol{\tau}_z'\right] \boldsymbol{\beta}_2.$$

However, a break may also be induced in the other forecasting model when $\mathbf{z}_{T+1}$ shifts because:

$$\mathbf{w}_{T+1} = \boldsymbol{\kappa} + \boldsymbol{\Pi}_{wz} \mathbf{z}_{T+1} + \boldsymbol{\eta}_{wz,T+1} \quad \text{where} \quad \mathsf{E}\left[\boldsymbol{\eta}_{wz,T+1}\right] = \mathbf{0} \quad \text{and} \quad \mathsf{E}\left[\mathbf{z}_{T+1}\boldsymbol{\eta}_{wz,T+1}'\right] = \mathbf{0},$$

so $\boldsymbol{\kappa} = -\boldsymbol{\Pi}_{wz}\boldsymbol{\tau}_z$, whereas $\boldsymbol{\Pi}_{wz}' = \boldsymbol{\Omega}_{zz}^{-1}\boldsymbol{\Omega}_{zw}$, leading to a forecast error of:

$$\widetilde{v}_{T+1} \simeq \boldsymbol{\beta}_1' \boldsymbol{\eta}_{wz,T+1} - \boldsymbol{\beta}_1' \boldsymbol{\Pi}_{wz}\boldsymbol{\tau}_z + e_{T+1}.$$

Then the squared error is:

$$\mathsf{E}\left[\widetilde{v}_{T+1}^2 \mid w_{T+1}, z_{T+1}\right] \simeq \sigma_e^2 + \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_1' \boldsymbol{\Pi}_{wz}\boldsymbol{\tau}_z \boldsymbol{\tau}_z' \boldsymbol{\Pi}_{wz}' \boldsymbol{\beta}_1.$$

We continue to assume that, in the absence of the break, the model including $\mathbf{w}$ is the more accurate, that is, $\boldsymbol{\beta}_2' \boldsymbol{\Omega}_{\eta_{zw}} \boldsymbol{\beta}_2 = k\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1$ for $k < 1$. Then, to the approximations involved:

$$\mathsf{E}\left[\widehat{u}_{T+1}^2 \mid w_{T+1}, z_{T+1}\right] \simeq \sigma_e^2 + k\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 + \left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2$$

$$\mathsf{E}\left[\widetilde{v}_{T+1}^2 \mid w_{T+1}, z_{T+1}\right] \simeq \sigma_e^2 + \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 + \left(\boldsymbol{\beta}_1' \boldsymbol{\Pi}_{wz} \boldsymbol{\tau}_z\right)^2 .$$

Consequently, $\widetilde{y}_{T+1}$ could be the more accurate forecast here, despite being less accurate prior to the break. This is more likely the larger $\boldsymbol{\tau}_z$ and the less correlated are $\mathbf{z}$ and $\mathbf{w}$ – in the limit, when $\boldsymbol{\Pi}_{wz} = \mathbf{0}$, $\widehat{\mathbf{b}}_1$ is a consistent estimator of $\boldsymbol{\beta}_2$, and the term involving $\boldsymbol{\tau}_z$ drops out of the MSFE for $\widetilde{y}_{T+1}$.

The average forecast is:

$$\widehat{\widetilde{y}}_{T+1} = \frac{1}{2}\left(\widehat{y}_{T+1} + \widetilde{y}_{T+1}\right),$$

with error:

$$\widehat{\widetilde{e}}_{T+1} = (y_{T+1} - \widehat{y}_{T+1}) + \frac{1}{2}\left(\widehat{y}_{T+1} - \widetilde{y}_{T+1}\right) = \widehat{u}_{T+1} + \frac{1}{2}\left(\widetilde{v}_{T+1} - \widehat{u}_{T+1}\right)$$

$$= \frac{1}{2}\left(\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1' \boldsymbol{\Pi}_{wz}\right)\boldsymbol{\tau}_z + \frac{1}{2}\left(\boldsymbol{\beta}_1' \boldsymbol{\eta}_{wz,T+1} + \boldsymbol{\beta}_2' \boldsymbol{\eta}_{zw,T+1}\right) + e_{T+1},$$

so:

$$\mathsf{E}\left[\widehat{\widetilde{e}}_{T+1} \mid w_{T+1}, z_{T+1}\right] = \frac{1}{2}\left(\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1' \boldsymbol{\Pi}_{wz}\right)\boldsymbol{\tau}_z + \frac{1}{2}\left(\boldsymbol{\beta}_1' \boldsymbol{\eta}_{wz,T+1} + \boldsymbol{\beta}_2' \boldsymbol{\eta}_{zw,T+1}\right).$$

Again ignoring terms of $\mathsf{O}_p(T^{-1})$:

$$\mathsf{E}\left[\widehat{\widetilde{e}}_{T+1}^2 \mid w_{T+1}, z_{T+1}\right]$$
$$\simeq \sigma_e^2 + 0.25\left[(1+k)\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{zw}\left(\mathbf{I}_{n_1} - \boldsymbol{\Pi}_{zw}' \boldsymbol{\Pi}_{wz}'\right)\boldsymbol{\beta}_2 + \left[\left(\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1' \boldsymbol{\Pi}_{wz}\right)\boldsymbol{\tau}_z\right]^2\right].$$

Thus, the combined forecast could beat both individual forecasts depending on the relative sizes of the unmodelled shift in the $\mathbf{z}$ process to the error variances.

To illustrate this, we consider two simplifications: first $\boldsymbol{\Omega}_{wz} = \mathbf{0}$, then a scalar case in section 6.1. Against $\widehat{y}_{T+1}$ (the more accurate forecast in the absence of breaks) in the first simplification, the average forecast dominates when:

$$(1 - 3k)\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 < 3\left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2 ,$$

which is bound to hold for $k > 1/3$ and could hold even for small k. Against the second forecast:

$$(k - 3)\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 + \left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2 < \mathbf{0}.$$

If we approximate by $k = 1$, then both hold when:

$$\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 > \frac{1}{2}\left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2 > -\frac{3}{2}\left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2$$

where the last inequality must be true. If instead, k is small, then:

$$\frac{1}{3}\left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2 < \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\eta_{wz}} \boldsymbol{\beta}_1 < 3\left(\boldsymbol{\beta}_2' \boldsymbol{\tau}_z\right)^2 .$$

Thus, irrespective of whether k is large or small, the average can 'win' against both mis-specified forecasting devices when the DGP experiences deterministic shifts.

6.1. Scalar illustration

In the scalar case when $n_1 = n_2 = 1$, using the approach in section 3.1:

$$\mathsf{E}\left[\widehat{\widetilde{e}}^2_{T+1} \mid w_{T+1}, z_{T+1}\right]$$
$$\simeq \sigma_e^2 + 0.25 \left\{ \left(1 - \mathsf{r}^2_{wz}\right) \left[(1 + k)\,\beta_1^2 \sigma_w^2 - 2\beta_1 \beta_2 \sigma_{wz}\right] + (\beta_2 - \beta_1 \pi_{wz})^2\, \tau_z^2 \right\},$$

with:

$$\mathsf{E}\left[\widehat{u}^2_{T+1} \mid w_{T+1}, z_{T+1}\right] \simeq \sigma_e^2 + \beta_2^2 \sigma_z^2 \left(1 - \mathsf{r}^2_{wz}\right) + \beta_2^2 \tau_z^2$$
$$\mathsf{E}\left[\widetilde{v}^2_{T+1} \mid w_{T+1}, z_{T+1}\right] \simeq \sigma_e^2 + \beta_1^2 \sigma_w^2 \left(1 - \mathsf{r}^2_{wz}\right) + \beta_1^2 \tau_z^2 \pi_{wz}^2.$$

Against $\widehat{y}_{T+1}$, the average outperforms in the normalized case if (as $\mathsf{r}_{wz} = \rho \pi_{wz}$ and $k\rho^2 = 1$):

$$- \left(1 - \mathsf{r}^2_{wz}\right)\left(3\rho + 2\mathsf{r}_{wz} - \rho^3\right) < \frac{\tau_z^2}{\sigma_z^2}\left(3\rho + 2\mathsf{r}_{wz} - k\rho\mathsf{r}^2_{wz}\right).$$

When ρ is close to unity and r_{wz} is large, this reduces to:

$$- \left(1 - \mathsf{r}^2_{wz}\right) < \frac{\tau_z^2}{\sigma_z^2} \tag{22}$$

which must hold. Alternatively, if $\mathsf{r}_{wz} = 0$, then:

$$\rho^2 < 3\left(1 + \frac{\tau_z^2}{\sigma_z^2}\right),$$

which will hold when the relative break is sufficiently large. Against $\widetilde{y}_{T+1}$, the average dominates if:

$$- \left(1 - \mathsf{r}^2_{wz}\right)\left(3\rho^4 + 2\rho\mathsf{r}_{wz} - \rho^2\right) < \frac{\tau_z^2}{\sigma_z^2}\left(3\mathsf{r}^2_{wz} + 2\mathsf{r}_{wz} - \rho^2\right).$$

As before, when ρ is close to unity and r_{wz} is large, we replicate (22). And if $\mathsf{r}_{wz} = 0$, dominance requires

$$3\rho^2 > 1 + \frac{\tau_z^2}{\sigma_z^2}.$$

Thus, dominance over both individual models simultaneously requires:

$$\frac{1}{3}\left(1 + \frac{\tau_z^2}{\sigma_z^2}\right) < \rho^2 < 3\left(1 + \frac{\tau_z^2}{\sigma_z^2}\right).$$

We conclude that there is a wide range over which averaging will dominate.

6.2. Later breaks

If, in a later forecast period, there is a break in the other process, then a similar analysis applies with the rankings of the individual models reversed. The algebra naturally becomes tedious, but

the outcome must depend on both the absolute and relative sizes of the breaks, whether earlier breaks were modelled or not, the robustness of devices to breaks, and the sizes of the signal-noise ratios. There must exist combinations in which the average dominates over individual forecasting devices, on average over repeated forecasting episodes, because other devices swing from good to bad performance. Such later breaks may also vitiate estimation of weights: when a method is doing well because it had not previously suffered forecast failure, estimation will attribute an above-average weight to it. Any later shift in that 'current best' device would induce poorer performance than just the average.

6.3. Breaks in falsely-included variables

If some of the variables that are included with non-zero cofficients in forecasting models are in fact irrelevant, then an analogous derivation is feasible to show that the effects of breaks favour combination. When such variables experience a deterministic shift, the forecasts from that model will be poor, since the dependent variable will not have been affected. Any average will attribute a smaller weight than unity to such a set of forecasts, and so outperform it. Later breaks in other variables in rival models will similarly worsen their performance, leaving the average as the 'winner'.

6.4. Within-equation breaks

Finally, a break in the y process introduces further complications, depending on the class of models under analysis. When a break occurs after forecasts are announced, all devices will fail, usually in the same direction, so averaging will neither resolve nor exacerbate that problem. However, some methods will continue to fail for many later periods – especially equilibrium-correction models (EqCMs) – again usually in the same direction (see e.g., Clements and Hendry (1999)). If the EqCMs were previously the dominant approach, then we have the analogue of the conditions in section 6, namely a switch in ranking between methods pre and post break, precisely the situation when averaging can dominate on average. Now, however, in the sub-periods, the average may or may not dominate. Moreover, estimated weights would emphasize the near encompassing of an EqCM over (say) a first-differenced autoregression, so could do less well than the average. Indeed, when simple – but robust – forecasting devices are encompassed by the EqCM, and so excluded from the pooling, we have a counter example to any claim that only non-encompassed models should be included in the average.

6.5. Pooling information

In the present context, pooling of information should prove more successful than pooling forecasts for all extraneous breaks, but not for breaks in the equation of interest. Since there are usually many variables involved, the former type of break should be far more frequent than the latter, supporting pooling. In Hendry and Clements (2001b), we explore this idea to explain the success of 'factor forecasts', or diffusion indices, as in Stock and Watson (1999) and Forni, Hallin, Lippi, and Reichlin (2000). Moreover, extraneous breaks become endogenous in a system, so our approach also suggests an explanation for why multi-step (or dynamic) estimation

may be advantageous: see Chevillon (2000). However, when different transformations (e.g., log and linear) of the same variable are involved, pooling information is less likely to dominate.

7. EMPIRICAL ILLUSTRATION

Bates and Granger (1969) provide an example of the usefulness of combining forecasts from linear and exponential trend models of output. Table 1 records an output index for the UK gas, electricity and water sectors for the years 1948 to 1965, along with forecast errors from linear and exponential trend models of output $\{y_t\}$, given by $y_t = \alpha + \beta t + error_t$ and $\ln(y_t) = a + bt + error_t$, where t is a linear time trend. The forecast error in period t ($t = 1950, \dots, 1965$) is calculated from a forecast based on estimating the model on data up to $t - 1$. The results in the table show that although the exponential model forecasts have a much smaller sum of squared errors (SSE) than the linear model, nevertheless, a combination which attaches a small weight to the linear forecasts has a smaller SSE. For example, for a fixed weight of 0.16 on the linear forecasts, the combined forecast SSE is 78.8.[3] This clearly supports combination, but it is of interest to interpret how the gain comes about given our analysis.

The forecast errors from the linear model become large and positive from around 1961 onwards, indicating that the constant absolute increase model is inappropriate. On average, the exponential model over-predicts (negative errors), albeit to a lesser extent. Combination is seen to work because the two sets of forecasts are biased in different directions. This view is supported by the SSEs of the bias-corrected forecast errors (see the last two columns of the table), and the results of combining the bias-corrected forecasts. The bias-corrected forecast of period t is calculated by adding the sample mean of the forecast errors up to period $t - 1$ to the forecast of period t. Because the bias term is calculated from past forecast errors up to that point, it adapts only slowly to the run of positive errors in the linear forecasts of the 1960s.[4] The SSE of the bias-corrected exponential forecasts is 77, less than the combined forecast SSE of 78.8 (with a weight of 0.16), but more pertinently, we find that any fixed-weight combination of the bias-corrected forecasts, with weights in the interval $(0, 1)$, has a larger SSE than that of the exponential model forecasts.[5] Of course, the fixed-weight combination forecasts discussed are not feasible, in the sense that they are based on knowledge of the full set of forecast errors, and they can also be improved upon by varying-weight schemes, as shown by Bates and Granger (1969). This example shows that gains from combination may disappear if individual forecasts are first corrected, consistent with the derivation when there are no breaks that combination exploits offsetting biases.

A final implication, given the autocorrelated forecast errors, is that intercept correction or differencing should improve the forecasts. For the latter, the SSEs become 73.9 and 59.0 for the linear and exponential models respectively, providing a dramatic improvement for the former,

[3]The figures we report are based on our own calculations. We reproduce the forecasts, and forecast errors etc., based only on the actual series. Some small differences were observed relative to Bates and Granger's figures, presumably because of improved precision.

[4]If we were to estimate combination weights for the original forecasts based on the whole sample, and include an intercept in the combination, a much smaller SSE of 60.1 results, partly because the bias-corrections are now calculated based on the full-sample, and the sample biases of the individual forecasts will be zero. However, the optimal combination weights that sum to unity are now -0.23 and 1.23, and difficult to interpret.

[5]The optimal combination for the bias-corrected forecasts, imposing the constraint that they sum to unity (and with a zero intercept in the combination) was -0.22 on the linear forecasts, delivering an SSE of 72.61.

Table 1. Forecasts of output indices, 1950–65

| | Actual | 1-step forecast errors | | | | |
| | | Linear | Exponential | Combination | Linear | Exponential |
					Bias-corrected	Bias-corrected
1948	58.0					
1949	62.0					
1950	67.0	1.0	0.7	0.77	1.0	0.7
1951	72.0	0.7	0.1	0.21	-0.3	-0.6
1952	74.0	-2.5	-3.4	-3.24	-3.3	-3.8
1953	77.0	-2.2	-3.3	-3.11	-1.9	-2.4
1954	84.0	2.1	0.8	0.99	2.8	2.2
1955	88.0	1.0	-0.6	-0.37	1.2	0.4
1956	92.0	0.4	-1.7	-1.33	0.4	-0.7
1957	96.0	0.0	-2.5	-2.08	-0.0	-1.4
1958	100.0	-0.2	-3.2	-2.71	-0.3	-2.0
1959	103.0	-1.3	-4.8	-4.28	-1.4	-3.4
1960	110.0	1.9	-2.1	-1.47	2.0	-0.3
1961	116.0	3.2	-1.4	-0.71	3.1	0.4
1962	125.0	7.0	1.8	2.60	6.7	3.5
1963	133.0	8.8	2.8	3.74	8.0	4.3
1964	137.0	6.1	-0.9	0.26	4.7	0.3
1965	145.0	8.0	-0.0	1.26	6.3	1.1
Sample bias		2.1	-1.1	-0.6	1.8	-0.1
Sum of squared errors		263.3	84.4	78.8	211.9	77.0

The output series is the output index for the gas, electricity and water sector, given in (Bates and Granger 1969, Table A1, p. 462). The combination forecast has fixed weights of 0.16 and 0.84 on the (uncorrected) linear and exponential forecasts

and a smaller – but worthwhile – gain for the latter, which now does better than any combination. Clements and Hendry (1999) treat inappropriate specification or estimation of deterministic terms as near equivalents of shifts in those terms, so that interpretation is also consistent with the present gains from combination and differencing.

8. A MONTE CARLO STUDY

We consider a range of settings. The first set include extraneous shifts in white noise processes to match the theory derivations and check their applicability in finite samples (section 8.1). We then allow for dynamic models (section 8.2), breaks in the DGP equation itself (section 8.3), and situations where some of the explanatory variables are absent from all of the models (section 8.4).

8.1. Shifts in an extraneous variable

Forecast period shift. Table 2 reports a selection of results from a Monte Carlo study of the usefulness of combination in small samples for constant DGPs and when there is a shift in the mean of an extraneous variable. The DGP is as given in section 6 with w_t and z_t scalar variables. The models are estimated without intercepts on the sample up to T, and used to forecast $T + 1$. The means $\phi_z = \phi_w = 0$ in-sample, but we allow $\phi_{z,T+1} = \tau_z$ to be non-zero in some experiments. Results are given for three combination schemes: simple averaging, the 'optimal' combination, and the use of relative MSFE weights.[6] We set $\beta_1 = \beta_2 = 1$, with a DGP disturbance variance of 0.16. The table records the Monte Carlo estimates of the biases and MSFEs over $50,000$ replications, for a number of sample sizes T, and different values of σ_w, σ_z and ρ. We also record the Monte Carlo estimates of $\widehat{b}_1$ and $\widehat{a}_1$. The columns headed M_z, M_w and M_c relate to forecasts from the models including z, w, and the simple average of the two, whereas the columns headed M_λ and M_{λ_M} show the optimal and MSFE weight combinations respectively.

For the first three rows of the table, $\tau_z = 0$, so show the effects of combination when there are no structural shifts. The model including w (M_w) is the more accurate of the individual models, because with $\beta_1 = \beta_2 = 1$, the higher variability of w ($\sigma_w^2 > \sigma_z^2$) means that it explains more of the variation in the dependent variable. Nevertheless, the simple average of the two forecasts yields a smaller MSFE. The optimal combination assigns a weight of just over 0.6 to M_w (a little higher when relative MSFE weights are used), and the combined forecast is then a little smaller than in the case of averaging. Monte Carlo estimates of these weights are shown in the table under the columns headed $\widehat{\lambda}$ and $\widehat{\lambda}_M$. Notice that the individual forecasts (and therefore the combinations) have a zero bias (to two decimal places) in the absence of deterministic non-stationarities. The high value of ρ entails that the effects of z and w in the individual models (estimated as $\widehat{b}_1$ and $\widehat{a}_1$) are quite different from their effects in the DGP. Our analytic derivations

---

[6] The optimal combination, as derived by Bates and Granger (1969), chooses the weights to minimise the MSFE of the combined forecast (subject to the weights on the individual forecasts summing to unity). This involves covariance terms between the models' forecast errrors. When these are ignored, the optimal weight is given by the relative MSFEs alone. For simplicity, we substitute the in-sample estimated residuals for the 1-step in-sample $(1, \ldots, T)$ forecast errors in calculating the weights, so that the period t 'forecast error' is based on parameter estimates obtained on data up to T, rather than $t - 1$.

	T	$\phi_{z,T+1}$	σ_z	$\phi_{w,T}$	σ_w	ρ	Forecast bias					MSFE								
							M_z	M_u	M_C	M_λ	$M_{\lambda M}$	M_z	M_u	M_C	M_λ	$M_{\lambda M}$	λ	λ_M	b_1	a_1
							No shift													
1	100	0.0	1.0	0.0	1.5	0.75	0.00	-0.00	0.00	0.00	0.00	1.14	0.60	0.27	0.24	0.25	0.61	0.66	2.13	1.50
2	20	0.0	1.0	0.0	1.5	0.75	0.01	0.00	0.00	0.01	0.01	1.20	0.63	0.29	0.27	0.28	0.61	0.65	2.13	1.50
3	10	0.0	1.0	0.0	1.5	0.75	0.00	-0.00	-0.00	-0.00	-0.00	1.28	0.67	0.33	0.32	0.34	0.61	0.64	2.12	1.50
							$\phi_{z,T+1} = \sigma_z$													
4	20	1.0	1.0	0.0	1.5	0.75	-1.12	1.00	-0.06	0.19	0.27	2.52	1.64	0.31	0.34	0.40	0.65	0.65	2.13	1.50
5	20	1.0	1.0	0.0	1.0	0.00	0.00	1.01	0.51	0.50	0.50	1.28	2.24	0.95	0.99	0.99	0.50	0.50	1.00	1.00
6	20	1.5	1.0	0.0	1.5	0.75	-0.75	1.50	0.38	0.13	0.04	1.21	3.47	0.44	0.32	0.33	0.39	0.35	1.50	2.12
7	20	1.0	1.0	0.0	1.5	-0.75	1.13	2.01	1.07	0.94	1.04	2.54	1.64	1.94	1.49	1.79	0.65	0.65	-0.12	0.50
8	20	1.5	1.5	0.0	1.0	-0.75	0.75	1.51	1.13	0.67	1.01	1.23	3.48	2.07	1.11	1.74	-0.11	0.35	0.50	-0.13
							$\phi_{z,T+1} = 2 \times \sigma_z$													
9	20	2.0	1.0	0.0	1.5	0.75	-2.25	2.00	-0.12	0.37	0.53	6.50	4.64	0.37	0.52	0.73	0.61	0.65	2.13	1.50
10	20	2.0	1.0	0.0	1.0	0.00	0.00	2.01	1.00	0.99	1.00	1.47	5.25	1.76	1.83	1.82	0.50	0.50	1.00	1.00
11	20	3.0	1.0	0.0	1.5	0.75	-1.50	3.00	0.75	0.25	0.08	3.00	10.23	0.89	0.46	0.48	0.39	0.35	1.50	2.12
12	20	2.0	1.0	0.0	1.5	-0.75	2.25	2.01	2.13	1.87	2.07	6.52	4.65	5.38	4.13	5.02	0.65	0.65	-0.12	0.50
13	20	3.0	1.5	0.0	1.0	-0.75	1.50	3.01	2.26	1.32	2.01	3.01	10.26	5.90	2.62	4.83	-0.11	0.35	0.50	-0.13
							$\phi_{z,T+1} = \sigma_z,\ \phi_{w,T:T+1} = 2 \times \sigma_w$													
14	20	1.0	1.0	3.0	1.5	0.75	1.73	-0.18	0.78	0.52	0.43	4.39	0.84	1.03	0.68	0.63	0.64	0.69	2.13	1.42
15	20	1.0	1.0	2.0	1.0	0.00	1.91	1.01	1.46	1.40	1.41	5.01	2.49	2.97	2.84	2.87	0.55	0.54	1.00	1.00
16	20	1.5	1.0	2.0	1.5	0.75	1.16	-0.27	0.44	0.59	0.64	2.05	1.63	0.65	0.76	0.84	0.41	0.38	1.50	1.93
17	20	1.0	1.0	3.0	1.5	-0.75	3.98	2.19	3.09	2.00	2.71	17.25	5.60	10.37	4.88	8.16	0.69	0.69	-0.12	0.58
18	20	1.5	1.5	2.0	1.0	-0.75	2.66	3.29	2.97	2.58	2.86	7.77	12.34	9.69	7.43	8.95	0.07	0.37	0.50	0.06
							$\phi_{z,T+1} = 2 \times \sigma_z,\ \phi_{w,T:T+1} = -2 \times \sigma_w$													
19	20	2.0	1.0	-3.0	1.5	0.75	-5.09	3.19	-0.95	0.23	0.63	27.61	10.95	1.37	0.68	1.22	0.64	0.69	2.13	1.42
20	20	2.0	1.0	-2.0	1.0	0.00	-1.89	2.01	0.06	0.22	0.20	5.19	5.53	0.76	1.12	1.05	0.55	0.54	1.00	1.00
21	20	3.0	1.0	-2.0	1.5	0.75	-3.39	4.78	0.69	-0.05	-0.32	12.38	24.38	0.93	0.65	0.99	0.41	0.38	1.50	1.94
22	20	2.0	1.0	-3.0	1.5	-0.75	-0.59	0.82	0.11	0.82	0.38	2.01	1.47	0.92	1.50	1.00	0.69	0.69	-0.12	0.58
23	20	3.0	1.5	-2.0	1.0	-0.75	-0.39	1.23	0.42	0.11	-0.20	1.01	3.05	1.07	1.07	0.97	0.07	0.37	0.50	0.06

Table 2. Simulation results: extraneous shifts

M_z and M_u refer to the forecasts from the models including z and u, respectively. M_C assigns equal weights to each forecast, M_λ assigns 'optimal' weights based on the formula in Bates and Granger, and $M_{\lambda M}$ assigns optimal weights but omitting the covariance terms between the models' forecast errors.

	T	$\phi_{z,T+1}$	σ_z	σ_w	ρ	Forecast bias					MSFE					$\hat\lambda$	$\hat\lambda_M$	$\hat b_1$	$\hat a_1$
						M_z	M_w	M_C	M_Λ	M_{Λ_M}	M_z	M_w	M_C	M_Λ	M_{Λ_M}				
								No shift											
1	100.00	0.00	1.00	1.50	0.75	0.01	-0.00	0.00	0.00	0.00	5.36	2.47	0.83	0.66	0.74	0.61	0.68	2.13	1.50
2	20.00	0.00	1.00	1.50	0.75	-0.01	0.01	0.00	0.00	0.00	5.15	2.35	1.06	0.82	0.92	0.61	0.66	2.13	1.50
3	10.00	0.00	1.00	1.50	0.75	0.00	-0.00	-0.00	-0.00	-0.00	4.79	2.22	1.17	0.85	0.98	0.61	0.65	2.12	1.50
								$\phi_{z,T+1} = \sigma_z$											
4	20.00	1.00	1.00	1.50	0.75	-1.13	1.01	-0.06	0.21	0.33	6.85	3.37	1.17	0.92	1.10	0.61	0.66	2.13	1.50
5	20.00	1.00	1.00	1.00	0.00	-0.01	1.00	0.49	0.47	0.50	5.65	6.11	3.00	2.50	2.72	0.50	0.50	1.00	1.00
6	20.00	1.50	1.50	1.00	0.75	-0.75	1.51	0.38	0.15	0.06	3.14	7.36	1.24	0.89	1.02	0.39	0.34	1.50	2.12
7	20.00	1.00	1.00	1.50	-0.75	1.12	1.00	1.06	0.75	0.99	6.82	3.34	4.07	2.45	3.35	1.12	0.66	-0.12	0.50
8	20.00	1.50	1.50	1.00	-0.75	0.75	1.49	1.12	0.64	0.96	3.13	7.30	4.13	2.44	3.39	-0.12	0.34	0.50	-0.12
								$\phi_{z,T+1} = 2 \times \sigma_z$											
9	20.00	2.00	1.00	1.50	0.75	-2.26	2.01	-0.13	0.42	0.66	11.90	6.38	1.49	1.20	1.61	0.61	0.66	2.13	1.50
10	20.00	2.00	1.00	1.00	0.00	-0.01	2.00	0.99	0.93	1.00	6.90	9.10	4.05	3.42	3.76	0.50	0.50	1.00	1.00
11	20.00	3.00	1.50	1.00	0.75	-1.51	3.01	0.75	0.31	0.11	5.39	14.15	1.81	1.10	1.29	0.39	0.34	1.50	2.12
12	20.00	2.00	1.00	1.50	-0.75	2.24	2.00	2.12	1.50	1.97	11.81	6.34	7.74	4.31	6.39	1.12	0.66	-0.12	0.50
13	20.00	3.00	1.50	1.00	-0.75	1.49	2.99	2.24	1.28	1.93	5.35	14.04	8.05	4.24	6.52	-0.12	0.34	0.50	-0.12

Table 3. Simulation results: extraneous shift when z and w are $AR(1)$ processes

Legend as for Table 2

| | T | τ | δ_0 | δ_1 | ρ | Forecast bias | | | | | MSFE | | | | | $\hat{\lambda}$ | $\hat{\lambda}_M$ | $\hat{b}_1$ | $\hat{a}_1$ |
						M_z	M_w	M_C	M_λ	M_{λ_M}	M_z	M_w	M_C	M_λ	M_{λ_M}				
1	20.00	15	.8	0	0.75	0.61	0.60	0.61	0.61	0.61	1.64	1.04	0.68	0.67	0.68	0.61	0.63	2.13	1.50
2	20.00	18	.8	0	0.75	0.73	0.72	0.73	0.73	0.73	1.80	1.19	0.84	0.82	0.83	0.61	0.64	2.13	1.50
3	20.00	15	0	1	0.75	0.01	0.01	0.01	0.01	0.01	1.85	2.36	1.31	1.38	1.41	0.55	0.56	2.38	1.62
4	20.00	18	0	1	0.75	0.01	0.01	0.01	0.01	0.01	2.08	2.47	1.56	1.65	1.68	0.59	0.61	2.23	1.55
5	20.00	15	.8	0	0.00	0.61	0.61	0.61	0.61	0.61	3.04	1.67	1.44	1.33	1.35	0.69	0.66	1.00	1.00
6	20.00	18	.8	0	0.00	0.73	0.73	0.73	0.73	0.73	3.20	1.82	1.60	1.49	1.50	0.69	0.66	1.00	1.00
7	20.00	15	0	1	0.00	0.02	0.01	0.01	0.02	0.02	3.24	4.38	2.74	2.90	2.89	0.59	0.58	1.25	1.00
8	20.00	18	0	1	0.00	0.02	0.01	0.01	0.01	0.01	3.47	4.33	2.93	3.14	3.13	0.65	0.63	1.10	1.00
9	20.00	15	.8	0	-0.75	0.61	0.61	0.61	0.61	0.61	1.64	1.04	1.20	1.01	1.13	1.10	0.63	-0.12	0.50
10	20.00	18	.8	0	-0.75	0.73	0.73	0.73	0.73	0.73	1.80	1.19	1.36	1.17	1.28	1.11	0.64	-0.12	0.50
11	20.00	15	0	1	-0.75	0.01	0.01	0.01	0.01	0.01	1.85	2.34	1.95	2.11	1.98	0.83	0.57	0.13	0.37
12	20.00	18	0	1	-0.75	0.01	0.01	0.01	0.01	0.01	2.08	2.45	2.12	2.34	2.17	1.01	0.61	-0.02	0.45

Table 4. Simulation results: shifts in y-equation

Legend as for Table 2

	T	$\phi_{z,T+1}$	σ_z	σ_w	ρ	Forecast bias					MSFE					$\hat{\lambda}$	$\hat{\lambda}_M$	$\hat{b}_1$	$\hat{a}_1$
						M_z	M_w	M_C	M_λ	$M_{\lambda M}$	M_z	M_w	M_C	M_λ	$M_{\lambda M}$				
1	100.00	0.00	1.00	1.50	0.75	1.00	0.99	1.00	1.00	1.00	3.31	2.61	2.05	2.02	2.02	0.59	0.59	2.62	1.83
2	20.00	0.00	1.00	1.50	0.75	0.95	0.94	0.94	0.94	0.94	3.40	2.65	2.03	2.06	2.04	0.59	0.58	2.63	1.83
3	10.00	0.00	1.00	1.50	0.75	0.90	0.90	0.90	0.90	0.90	3.64	2.81	2.11	2.24	2.18	0.59	0.57	2.63	1.83
4	20.00	1.00	1.00	1.50	0.75	-0.68	1.94	0.63	0.86	0.84	3.11	5.53	1.57	2.06	1.97	0.59	0.58	2.63	1.83
5	20.00	1.00	1.00	1.00	0.00	0.45	1.95	1.20	1.18	1.19	3.59	7.04	3.43	3.53	3.50	0.50	0.50	1.50	1.50
6	20.00	1.50	1.50	1.00	0.75	-0.31	2.44	1.07	0.81	0.84	1.95	8.46	2.30	1.99	1.99	0.41	0.42	1.83	2.63
7	100.00	0.00	1.00	1.50	0.75	1.99	1.98	1.99	1.99	1.99	6.28	5.56	5.00	4.98	4.98	0.59	0.58	2.62	1.83
8	20.00	0.00	1.00	1.50	0.75	1.90	1.89	1.89	1.89	1.89	6.11	5.35	4.73	4.77	4.74	0.59	0.57	2.63	1.83
9	10.00	0.00	1.00	1.50	0.75	1.80	1.80	1.80	1.80	1.80	6.11	5.27	4.57	4.74	4.63	0.59	0.56	2.63	1.83
10	20.00	1.00	1.00	1.50	0.75	0.27	2.89	1.58	1.81	1.77	2.73	10.13	3.68	4.63	4.42	0.59	0.57	2.63	1.83
11	20.00	1.00	1.00	1.00	0.00	1.40	2.90	2.15	2.13	2.14	5.37	11.65	6.62	6.69	6.66	0.50	0.50	1.50	1.50
12	20.00	1.50	1.50	1.00	0.75	0.64	3.39	2.02	1.76	1.81	2.29	14.01	5.24	4.47	4.56	0.41	0.42	1.83	2.63

Table 5. Simulation results: shift in variable excluded from both models

Legend as for Table 2

ignore terms of $O\left(T^{-1}\right)$: the Monte Carlo suggests that the qualitative results are the same for $T = 100$ and $T = 10$ (compare the first and third results), suggesting that these terms are indeed unimportant.

The next set of rows report results for a shift ϕ_z equal to one standard deviation of the z-equation disturbance term, namely $\tau_z = \sigma_z$ (equalling one standard deviation of z in the absence of explanatory variables in the z-equation). Consider row 4. This suggests that the relative percentage reductions in MSFE can be much larger when there are structural shifts. The bias in the forecasts from M_w is approximately the value of the shift. By including z, M_z picks up the value of the shift, but because the coefficient on z is approximately double that in the DGP, this model over-predicts by approximately the amount of the shift. Now the combination based on optimal weights (M_λ) no longer delivers the smallest MSFE: just as the best model in-sample may not yield the most accurate forecasts when there are structural changes, so the optimal combination in-sample may no longer be optimal for out-of-sample forecasting. When $\rho = 0$ (row 5), $\widehat{b}_1$ is an unbiased estimator of β_1, so that M_z is unbiased. Nevertheless, combination is still better (averaging is optimal): it pays to combine with the biased predictor. Rows 7 and 8 illustrate the results of combination when $\rho < 0$, so that both individual models are biased in the same direction, and averaging leads to a worse outcome than the best, but still outperforms the worst individual forecast. The optimal combination remains dominant, but the weights are outside $(0, 1)$, and relative MSFE weights give similar results to averaging. The third set of rows are for $\phi_{z,T+1} = 2\sigma_z$. Row 9 illustrates a greater proportionate reduction in MSFE from combination. Row 10 ($\rho = 0$) indicates that for shifts of this size the bias induced in M_w is large enough to counteract the benefits to combination, and M_z has the smallest MSFE.

Forecast and estimation period shifts. The fourth and fifth panels of table 2 replicate the second and third, but with a shift in the intercept of the w process of two and minus two times the standard error of its disturbance, respectively, taking effect in periods T and $T + 1$ ($\phi_{w,T} = \phi_{w,T+1} = \pm 2\sigma_z$). We allow intercepts in the M_z and M_w models, but otherwise proceed as above. Note that the impact of the single observation T on the estimation of the models' parameters is relatively minor, so that the results for consecutive shifts in the forecast period would be qualitatively similar. From the bottom panel, it is apparent that combination can yield large percentage reductions in MSFE when the explanatory variables undergo shifts in different directions, and the variables are positively correlated ($\rho > 0$, rows 19 and 21). Then, the upward bias in the coefficient estimates exacerbates the forecast biases of M_z and M_w. When $\rho < 0$, the models' slope parameter estimates are biased towards zero, and the forecasts from both individual models are closer to the actual value of y_{T+1} that results from the largely offsetting shifts in the two explanatory variables. When the shifts are in the same direction (rows 14 – 18) the deterioration in the individual models' performances is less pronounced, but depending on the relative sizes of the shifts and the importance of the individual explanatory variables, combination can either beat the best individual forecast or guard against inadvertently choosing the worst. For example, in row 14 the size of the shift in w relative to that in z is such that M_w is better than averaging, but a scheme that assigns a higher weight (λ) to M_w delivers a smaller MSFE (the table presents results for average values of 0.64 and 0.69). When the size of the shifts is more comparable (e.g., row 16), averaging is again beneficial.

8.2. Autocorrelated explanatory variables

Table 3 reports results for a subset of these experiments, except that z and w now follow AR(1) processes, with an autoregressive coefficient of 0.9. Keeping the same values of the disturbance variances as before, the variances of z and w increase by a factor of approximately five, so that the costs to omitting either in terms of MSFE is now larger: see M_z and M_w in the first three rows of the table, for example. The proportionate gains to combination are correspondingly greater. Now, combination pays even when $\rho = 0$ and $\tau_z = 2$ (row 10), but note that the size of the shift relative to the standard error of z has fallen.

8.3. Shifts in the y equation

Table 4 reports the results for the three combination schemes and the individual model forecasts when there are shifts in the y equation. We chose the parameter values corresponding to row 2 of table 2, so $\tau_z = 0$, $\sigma_z = 1$ and $\sigma_w = 1.5$, and $\rho = 0.75$ in the first panel. The shifts in the y equations are defined by: τ, the time of the shift, whereby the new values take effect from $\tau + 1$ onwards, and $\tau = 15$ or 18 for $T = 20$; δ_0, the shift in the (hitherto zero-valued) intercept of 0.8 (twice the standard deviation of the y-equation disturbance term); and δ_1, the shift in the coefficient on z, where $\delta_1 = 1$ so the coefficient doubles in size. The second and third sets of four rows repeat the first, but with $\rho = 0$, and $\rho = -0.75$. For both the M_z and M_w models, intercepts are estimated to accommodate the shifts in the y equation.

The results suggest the following. The individual forecasts (and therefore combinations) remain unbiased when δ_1 is not equal to zero, because z is a mean-zero variable. Nevertheless, for both types of shift, combination proves to be efficacious for $\rho = 0.75$ and $\rho = 0$, but, as in the absence of such shifts, is generally less so when ρ is negative.

8.4. Completely omitted variables

Our analytic derivations assume that the variables in the models span the explanatory variables in the DGP, so each model only excludes variables which the other contains. The condition that all the variables in the DGP are included in at least one of the models would appear to be unimportant to our explanations of why pooling works, but we checked that aspect in a further Monte Carlo study reported in table 5. There we report experiments based on rows 1 to 6 of table 2, but allowing an additional variable $\{q_t\}$ to enter the DGP with a unit coefficient. This variable is mean-zero white noise, with a variance of unity, but with a shift to a mean of unity (rows 1 to 6) or 2 (rows 7 to 12) for periods T and $T + 1$, i.e., $\phi_q = 0$ but $\phi_{q,T:T+1} = 1$ or 2. If q were uncorrelated with the explanatory variables and $\phi_{q,T:T+1} = 0$, our analytical calculations would be unaffected, since q could be subsumed into the disturbance term so only affect the equation error variance. Maintaining the interrelatedness assumption, a shift in ϕ_q is equivalent to a shift in the intercept of the y equation. The interesting cases are when q is correlated with one or both of z and w. In our experiments, both correlations are one half. We also estimate intercepts in both the M_w and M_z models.

The first three rows of table 5 (and rows 7 to 9) show that the individual model forecasts (and therefore the combined forecast) are approximately biased by an amount equal to the size of the shift in q. Nevertheless, combination reduces the MSFEs. When z also shifts (rows 4 to 6), then

because q and z shift in the same direction, and because the 'omitted variable bias' in M_z causes the coefficient on z to be upward biased, the forecast biases of M_z are smaller than either M_w or the combination. When, in addition, $\sigma_z > \sigma_w$ so that z is the more important determinant of y (row 6), combination is worse than M_z (but only marginally so). For larger shifts in ϕ_q, M_z is relatively better than the combined forecast.

8.5. Summary

These simulations confirm the analytical results, and explore a number of extensions. The qualitative nature of the conclusions based on the analytical work hold up, so that model mis-specification and parameter non-constancy are seen to explain why combination, and especially averaging, often works in practice. When a DGP variable which is not included in either of the individual models undergoes a shift in mean, at the same time that other variables shift, a range of outcomes is possible depending on the exact design, that is, the relative sizes of the shifts, their relative contributions to the total variation in the dependent variable, and the signs of the cross-correlations, etc. In general, allowing for variables that do not enter any of the models could strengthen or weaken the case for combination when there are shifts.

9. CONCLUSION

Practical experience shows that combining forecasts has value added and can dominate even the best individual device. Thus, we considered selecting forecasting methods by pooling several individual devices when no model coincides with a non-constant data generation process (DGP).

　　We first show that averaging guarantees 'insurance', and may provide dominance, when the models are differentially mis-specified even for a constant DGP. While such a result can occur in weakly-stationary processes, we suspect that empirical findings are better explained by the intermittent occurrence of deterministic shifts in unmodelled explanatory variables. Consequently, we demonstrate that when forecasting time series that are subject to deterministic shifts, the average of a group of forecasts from differentially mis-specified models can outperform them all on average over repeated forecasting episodes. Moreover, averaging may well then dominate over estimated weights in the combination. Finally, it cannot be proved that only non-encompassed devices should be retained in the combination.

　　In practice, trimmed means, or perhaps medians, might be needed to exclude 'outlying' forecasts, since otherwise, one really poor forecast would worsen the combination needlessly.

　　Both the empirical and Monte Carlo simulation illustrations confirmed the theoretical analysis. The average of the levels forecasts outperformed the best individual forecast in both settings, sometimes spectacularly. However, in the empirical example, bias correcting the forecasts removed much of the benefit of averaging, and other devices for robustifying forecasts to breaks did even better. Thus, although we have established that combination can be beneficial in our theoretical framework, comparisons with other approaches are merited.

　　Hendry and Clements (2001a) present ten cases where well-known empirical phenomena in economic forecasting can be explained by the use of mis-specified models of processes that experience intermittent deterministic shifts. The present paper extends that list to eleven. We believe that the related results on forecasting using 'factor models' can be accounted for by the same general theory, and are also investigating multi-step estimation within that framework.

ACKNOWLEDGEMENTS

REFERENCES

Andrews, M. J., P. Minford, and D. Peel (1995). Forecast encompassing tests of the Liverpool model. *Oxford Bulletin of Economics and Statistics 61*, 221–256.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operations Research Quarterly 20*, 451–468.

Chevillon, G. (2000). Multi-step estimation for forecasting non-stationary processes. MPhil Thesis, Economics Department, University of Oxford.

Chong, Y. Y. and D. F. Hendry (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies 53*, 671–690.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting 5*, 559–583.

Clements, M. P. and D. F. Hendry (1993). On the limitations of comparing mean squared forecast errors. *Journal of Forecasting 12*, 617–637.

Clements, M. P. and D. F. Hendry (1999). *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.

Clements, M. P. and D. F. Hendry (Eds.) (2001). *A Companion to Economic Forecasting*. Oxford: Blackwells.

Coulson, N. F. and R. P. Robins (1993). Forecast combination in a dynamic setting. *Journal of Forecasting 12*, 63–68.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts. *Journal of Business and Economic Statistics 6*, 105–111.

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting 5*, 589–592.

Diebold, F. X. and J. A. Lopez (1996). Forecast evaluation and combination. In G. S. Maddala and C. R. Rao (Eds.), *Handbook of Statistics*, Volume 14, pp. 241–268. Amsterdam: North–Holland.

Ericsson, N. R. (1992). Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling 14*, 465–495.

Ericsson, N. R. and J. Marquez (1993). Encompassing the forecasts of U.S. trade balance models. *Review of Economics and Statistics 75*, 19–31.

Fildes, R. and K. Ord (2001). Forecasting competitions – their role in improving forecasting practice and research. See Clements and Hendry (2001), pp. 322–253.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized factor model: Identification and estimation. *Review of Economics and Statistics 82*, 540–554.

Frisch, R. and F. V. Waugh (1933). Partial time regression as compared with individual trends. *Econometrica 1*, 221–223.

Gallo, G. M. and R. S. Mariano (1994). Combining provisional data and forecasts in nonlinear models. Working papers n.47, Dipartimento Statistico, Universita' Degli Studi Di Firenze.

Granger, C. W. J. (1989). Combining forecasts - Twenty years later. *Journal of Forecasting 8*, 167–173.

Granger, C. W. J. and R. Ramanathan (1984). Improved methods of combining forecasts. *Journal of Forecasting 3*, 197–204.

Harvey, D., S. Leybourne, and P. Newbold (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics 16*, 254–259.

Hendry, D. F. (2000). On detectable and non-detectable structural change. *Structural Change and Economic Dynamics 11*, 45–65.

Hendry, D. F. and M. P. Clements (2001a). Economic forecasting: Some lessons from recent research. Mimeo, Economics Department, University of Oxford.

Hendry, D. F. and M. P. Clements (2001b). Forecasting using factor models. Mimeo, Economics Department, University of Oxford.

Hendry, D. F. and J. A. Doornik (1997). The implications for econometric modelling of forecast failure. *Scottish Journal of Political Economy 44*, 437–461.

Hoogstrate, A. J., F. C. Palm, and G. A. Pfann (1996). To pool or not to pool: Forecasting international output growth rates. Research Memorandum 96/025, Meteor, University of Limburg, Maastricht.

Judge, G. G. and M. E. Bock (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.

Lu, M. and G. E. Mizon (1991). Forecast encompassing and model evaluation. In P. Hackl and A. H. Westlund (Eds.), *Economic Structural Change, Analysis and Forecasting*, pp. 123–138. Berlin: Springer-Verlag.

Newbold, P. and C. W. J. Granger (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society A 137*, 131–146.

Newbold, P. and D. I. Harvey (2001). Forecasting combination and encompassing. See Clements and Hendry (2001), pp. 268–283.

Stock, J. H. and M. W. Watson (1999). A comparison of linear and nonlinear models for forecasting macroeconomic time series. In R. F. Engle and H. White (Eds.), *Cointegration, Causality and Forecasting*, pp. 1–44. Oxford: Oxford University Press.

Wall, K. D. and C. Correia (1989). A preference-based method for forecast combination. *Journal of Forecasting 8*, 269–292.