Combining Point Forecasts: The Simple Average Rules, OK?

Jeremy Smith and Kenneth F. Wallis

Department of Economics
University of Warwick
Coventry CV4 7AL, UK
[K.F.Wallis@warwick.ac.uk]

Revised March 2005

Abstract This paper explores a possible explanation of the forecast combination puzzle, that simple combinations of point forecasts are repeatedly found to outperform sophisticated weighted combinations in empirical applications. The explanation lies in the effect of finite-sample error in estimating the combining weights. A small Monte Carlo study and a reappraisal of an empirical study by Stock and Watson (2003a) support this explanation. The Monte Carlo evidence, together with a large-sample approximation to the variance of the combining weight, also supports the popular recommendation to ignore forecast error covariances in estimating the weight.

Key Words Forecast pooling; combination forecast comparisons; choice of weights

Acknowledgments The helpful comments of Mark Watson and Kenneth West, and access to the database of James Stock and Mark Watson, are gratefully acknowledged.

1. INTRODUCTION

The idea that combining different forecasts of the same event might be worthwhile has gained wide acceptance since the seminal article of Bates and Granger (1969). Twenty years later, Clemen (1989) provided a review and annotated bibliography containing over 200 items, which he described as "an explosion in the number of articles on the combination of forecasts". These mostly concerned point forecasts of the future realisation of a random variable, which has been a continuing focus of attention of the forecast combination literature; a companion paper by Wallis (2004) considers extensions to the combination of interval and density forecasts. Despite the explosion of activity, Clemen found a variety of issues that remained to be addressed, the first of which was "What is the explanation for the robustness of the simple average of forecasts?" (1989, p.566). That is, why is it that, in comparisons of combinations of point forecasts, a simple average, with equal weights, often outperforms more complicated weighting schemes. This empirical finding continually reappears, for example in several recent papers by Stock and Watson (1999, 2003a, 2004), and remains a puzzle. A possible explanation is explored in this paper.

The explanation rests on the observation that, in comparing weighted and unweighted forecast combinations in real or pseudo out-of-sample testing, the underlying statistical problem shares two key features with a model comparison problem recently addressed by Clark and West (2004), and is amenable to similar analysis. The first common feature is that the models being compared are nested, and in this situation many forecast evaluation procedures available in the literature, which apply to non-nested comparisons, are inadequate. The second feature is that the optimal combining weights are not known a priori, but must be estimated, and in finite samples the resulting estimation error may cause a reversal of the underlying theoretical optimality result.

The paper proceeds as follows. Section 2 presents the general framework for analysis, beginning with the case of two competing forecasts and then considering generalisations to combinations of many forecasts. Section 3 contains two empirical applications, the first a small Monte Carlo study of combinations of two forecasts with equal error variances, the second a reappraisal of a study of combination forecasts of US output growth by Stock and Watson (2003a). Section 4 concludes.

2. WEIGHTED AND UNWEIGHTED COMBINATIONS OF FORECASTS

2.1 Optimal weights

We first follow Bates and Granger (1969) and consider the case of two competing point forecasts, f_{1t} and f_{2t}, made h periods earlier, of the quantity y_t. The forecast errors are

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

It is usually assumed that the forecasts are unconditionally unbiased, or "unbiased on average" in Granger and Newbold's (1986, p.144) term, so that

$$E\left(e_{it}\right) = 0, \quad i = 1, 2 .$$

We denote the forecast error variances as σ_i^2, $i = 1, 2$, and their covariance as σ_{12}. The combined forecast is the weighted average

$$f_{Ct} = k f_{1t} + (1-k) f_{2t} , \tag{1}$$

which is also unbiased in the same sense. Its error variance is minimised by setting the weight k equal to

$$k_{opt} = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} , \tag{2}$$

noting a sign error in Bates and Granger's equation (1). This expression can also be recognised as the coefficient in a regression of e_{2t} on $(e_{2t} - e_{1t})$, which suggests a way of estimating k_{opt} from data on forecasts and outcomes. A further interpretation is that this is equivalent to the extended realisation-forecast regression

$$y_t = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t} + u_t \tag{3}$$

subject to the restrictions $\alpha = 0, \ \beta_1 + \beta_2 = 1$. Although weights outside the (0,1) interval might be thought to be hard to justify, all these interpretations admit this possibility. An estimate based on (2) is negative whenever sample moments satisfy $s_{12} > s_2^2$, and exceeds one if $s_{12} > s_1^2$.

The minimised error variance of the combined forecast is no greater than the smaller of the two individual forecast error variances, hence in general there is a gain from combining using the optimal weight. Equality occurs if the smaller variance is that of a forecast which is already the minimum mean square error (mmse) forecast; there is then no gain in combining it with an inferior forecast. If the mmse forecast is f_{2t}, say, with error variance σ_2^2, then it also

holds that $\sigma_{12} = \sigma_2^2$ for any other forecast f_{1t} based on the same information set, whereupon $k_{opt} = 0$. In this case, and if h=1, the error term in (3) is non-autocorrelated. In all other cases the error term is expected to exhibit autocorrelation, hence $\hat{k}_{opt}$ or its regression equivalent is not in general fully efficient.

The simple average, with equal weights, is the case $k = \frac{1}{2}$. This is optimal if $\sigma_1^2 = \sigma_2^2$, that is, the two competing forecasts are equally good (or bad), irrespective of any covariance between their errors. A further possibility suggested by Bates and Granger is to neglect any covariance term, or assume it to be zero, and use the expression

$$k' = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \ .$$

This also gives equal weights if the error variances are equal, and restricts the weights to the (0,1) interval. However, if f_{2t} is the mmse forecast and f_{1t} is any other forecast this does not deliver weights of 0 and 1 as in the previous paragraph. That k' is the weight attached to the first forecast, f_{1t}, can be emphasised by expressing it in the alternative form

$$k' = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \ . \qquad (4)$$

This makes it clear that the weights are inversely proportional to the corresponding forecast error variances, and gives an expression which is more amenable to generalisation below.

2.2 Pseudo out-of-sample comparisons of combined forecasts

The general approach to forecast evaluation that is followed in the literature is called pseudo out-of-sample forecasting by Stock and Watson in their textbook (2003b, §12.7) and empirical studies cited above, because it mimics real-time out-of-sample forecasting yet the "future" outcomes are known, and so forecast performance can be assessed. To this end, a sample of available data, real or artificial, is divided into two subsamples: the first is used for estimating the forecasting relationships; the second for evaluating their forecast performance. Forecasting with a constant lead time, say one step ahead, implies that the information set on which the forecast is based is updated as the forecast moves through the evaluation subsample, and it is an open question whether and, if so, how the estimated relationships should also be updated. The three possibilities that figure prominently in the literature are

referred to as fixed, recursive and rolling schemes (see McCracken and West, 2002, for a recent review).

In the fixed scheme, the forecast relationships are estimated only once, using the original estimation subsample, and are not re-estimated as forecasting moves forward in time. In the recursive scheme re-estimation takes place as the forecast origin advances and more "past" data become available, with the estimation dataset gradually increasing in size. The rolling scheme also updates the estimates but keeps the size of the estimation dataset constant, by removing an observation from the beginning as each new observation is added to the end of the dataset. In the present context of comparisons of different combinations of forecasts the estimation subsample is sometimes divided further, into a first portion used for the estimation of the individual forecasting relationships and a second portion used for a preliminary evaluation of their performance from which combination weights can be estimated. The performance of the combined forecasts is then assessed in the evaluation subsample as above. Which of the three estimation schemes is used has a bearing on the asymptotics of the various available tests (Clark and McCracken, 2001). However many studies of combined forecasts are based on informal comparisons of mean squared forecast errors (MSFEs) over the evaluation subsample rather than formal inference procedures, whereupon the choice of updating scheme is immaterial. The comparisons developed below remain in this informal mode.

We adopt the Clark-McCracken-West notational conventions and denote, in one-step-ahead forecasting, the size of the available sample as T+1. This is divided into the initial estimation ("regression") subsample of R observations and the second evaluation ("prediction") subsample of P observations, with $R+P=T+1$. The first forecast is made at time R of observation y_{R+1}, and the last (the P^{th}) is made at time T of observation y_{T+1}. The sample MSFE of the combined forecast (1) is then

$$\hat{\sigma}_C^2 = \frac{1}{P}\sum_{t=R+1}^{T+1}\left(y_t - f_{Ct}\right)^2 .$$

We consider two such statistics: one, denoted $\hat{\sigma}_s^2$, is the MSFE of the simple average of forecasts, f_{st} say, with $k = \frac{1}{2}$ in (1); the other, denoted $\hat{\sigma}_w^2$, is the MSFE of a weighted average f_{wt} using some estimate $\hat{k}$ in (1).

In comparing the simple average with the weighted average of forecasts we note that the models being compared are nested. The null model has $k = \frac{1}{2}$; under the alternative, $k \neq \frac{1}{2}$. We follow Clark and West (2004) and consider the difference in MSFEs, $\hat{\sigma}_s^2 - \hat{\sigma}_w^2$, under the null. Expanding the t^{th} term in this difference gives

$$\begin{aligned}(y_t - f_{st})^2 - (y_t - f_{wt})^2 &= (y_t - f_{st})^2 - (y_t - f_{st} - (f_{wt} - f_{st}))^2 \\ &= 2(y_t - f_{st})(f_{wt} - f_{st}) - (f_{wt} - f_{st})^2 \\ &= 2(y_t - f_{st})\left(\left(\hat{k} - \tfrac{1}{2}\right) f_{1t} + \left(1 - \hat{k} - \tfrac{1}{2}\right) f_{2t}\right) - (f_{wt} - f_{st})^2 \ . \qquad (5)\end{aligned}$$

Assuming that an unbiased estimator of *k* is used, and neglecting any correlation between its sampling error, which is a function of the estimation subsample, and the evaluation subsample, the first term has expected value zero. However the second term is strictly positive so in sum, under the null, we expect to find

$$\hat{\sigma}_s^2 - \hat{\sigma}_w^2 \approx -\frac{1}{P}\sum_{t=R+1}^{T+1} (f_{wt} - f_{st})^2 < 0 \ .$$

Thus the simple average is expected to outperform the weighted average systematically, in a situation in which they are theoretically equivalent. Estimating *k* gives the weighted average a better fit in the estimation period, but this amounts to overfitting in the present circumstances, and does not carry through to the forecast period.

We note that the approximate discrepancy in the above inequality can be calculated directly, and consider Clark and West's (2004) suggestion that comparisons be based on an adjusted MSFE

$$\hat{\sigma}_w^2 - \text{adj} \equiv \hat{\sigma}_w^2 - \frac{1}{P}\sum_{t=R+1}^{T+1} (f_{wt} - f_{st})^2 \ . \qquad (6)$$

Before discussing estimation of *k*, we consider generalisations to more than two competing forecasts, since the number of forecasts being combined may be relevant to the choice of estimator.

2.3 Combining many forecasts

The general framework presented above readily extends to more than two competing forecasts, although some practical issues arise. With *n* competing point forecasts $f_{it},\ i = 1,...,n,$ the combined forecast is

$$f_{Ct} = \sum_{i=1}^{n} k_i f_{it} \ ,$$

with $\Sigma k_i = 1$ if the individual forecasts are unbiased and this is also desired for the combined forecast. Granger and Ramanathan (1984) consider estimation of the corresponding generalisation of regression equation (3), namely

$$y_t = \alpha + \beta_1 f_{1t} + ... + \beta_n f_{nt} + u_t \ , \qquad (7)$$

and the question of whether or not the coefficient restrictions $\alpha = 0$ and/or $\Sigma\beta_i = 1$ should be imposed. The unconstrained regression clearly achieves the smallest error variance *ex post*, and gives an unbiased combined forecast even if individual forecasts are biased. If the practical objective is to improve *ex ante* forecast performance, however, then the imposition of the restrictions improves forecast efficiency, as shown by Clemen (1986), for example.

Estimation of the regression equation (7) runs into difficulty if the number of individual forecasts being combined, *n*, is close to the number of observations in the regression subsample, *R*. This is a feature of the applications by Stock and Watson in the three articles referred to in the Introduction. The first article (Stock and Watson, 1999) analyses the performance of 49 linear and nonlinear univariate forecasting methods, in combinations with weights estimated over 60 or 120 months; this is done for 215 different series. The second article (2003a) considers combinations of up to 37 forecasts of U.S. output growth based on individual leading indicators, with weights estimated recursively, with an initial sample of 68 quarterly observations. The third article (2004) extends the growth forecasts to the G-7 countries, and the number of individual leading indicators considered for each country ranges between 56 and 75. (A further article (Stock and Watson, 2003c) similarly studies the role of asset prices as indicators of future inflation and output growth in the G-7 countries, but is not relevant for our present purposes, because no weighted combinations of forecasts are considered.)

In these circumstances Stock and Watson abandon estimation of the optimal combining weights by regression or, as they put it, abandon estimation of the large number of covariances among the different forecast errors. They follow the suggestion of Bates and Granger (1969) noted in the final paragraph of Section 2.1 and base estimated weights on the generalisation of expression (4), thus

$$\hat{k}_i' = \frac{\frac{1}{s_i^2}}{\sum_{j=1}^{n} \frac{1}{s_j^2}}, \; i = 1, ..., n, \tag{8}$$

where $s_i^2, \; i = 1, ..., n,$ is the MSFE of f_{it} over an estimation dataset. Earlier empirical studies summarised by Clemen (1989, p.562) also support the suggestion "to ignore the effects of correlations in calculating combining weights." Stock and Watson use several variants of this estimator, of which two are of particular interest. The first is to raise each MSFE term in the above expression to the power ω. With $0 < \omega < 1$ this shrinks the weights towards equality, the case $\omega = 0$ corresponding to the simple average with $k_i = 1/n$. Or with $\omega > 1$ more weight is placed on the better performing forecasts than is indicated by the inverse MSFE weights. The second variant is to calculate MSFEs as discounted sums of past squared forecast errors, so that forecasts that have been performing best most recently receive the greatest weight.

The common finding of the three cited studies in respect of comparisons of different combined forecasts is described as the "forecast combination puzzle – the repeated finding that simple combination forecasts outperform sophisticated adaptive combination methods in empirical applications" (Stock and Watson, 2004, p.428). The differences are not necessarily large, for example in the second article the MSFE improvement of the simple average does not exceed four percent (2003a, Table 4), but there are no reversals.

The explanation advanced in Section 2.2 carries over to the case of $n > 2$ competing forecasts. Under the null the simple average is expected to outperform the weighted average in terms of their MSFEs, in the absence of the adjustment defined in equation (6). Given $n \times 1$ vectors of forecasts and estimated weights, the weighted combination forecast is $f_t' \hat{k}$. Taking expectations in the *R*-sample, and conditioning on the *P*-sample, the expected difference in MSFE is equal to $trace\left(\Sigma_{ff} V_{\hat{k}}\right)$, where Σ_{ff} is the *P*-sample moment matrix of the forecasts and $V_{\hat{k}}$ is the mean square error matrix of $\hat{k}$ around the true value of equal weights. Thus the expected discrepancy is smaller the more accurate, in this sense, are the estimates of the weights. This effect diminishes as *R* increases, whereas the discrepancy remains of the same order of magnitude as *P* increases.

3. EMPIRICAL APPLICATIONS

3.1 A Monte Carlo study

Our first experiment describes the behaviour of the MSFE discrepancy $\hat{\sigma}_s^2 - \hat{\sigma}_w^2$ analysed in Section 2.2 in the same circumstances, namely under the null that $k = \frac{1}{2}$. Thus the two competing forecasts have equal error variances, and their simple average is compared to a weighted average with an estimated weight. The data generating process is the Gaussian AR(2) process

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right),$$

subject to the stationarity conditions $\phi_2 < 1 + \phi_1$, $\phi_2 < 1 - \phi_1$, $-1 < \phi_2 < 1$. The first two autocorrelation coefficients are then

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}, \quad \rho_2 = \phi_1 \rho_1 + \phi_2 \ .$$

We set up two cases of competing one-step-ahead forecasts with equal error variances.

Case 1. In the first case both forecasts are based only on the most recent observation. The first forecast is the naïve "no-change" forecast, and the second forecast is a first-order autoregression with the same forecast error variance. Thus

$$f_{1t} = y_{t-1}, \quad f_{2t} = \left(2\rho_1 - 1\right) y_{t-1}, \quad \sigma_i^2 = 2\left(1 - \rho_1\right)\sigma_y^2, \quad i = 1, 2 .$$

The contemporaneous correlation between the two forecast errors is equal to ρ_1. We consider values of ϕ_1 of 0.4 and 0.8, with ϕ_2 taking values in the range $-1 < \phi_2 < 1 - \phi_1$, hence the forecast errors are positively correlated, with ρ_1 lying in the range $0.2 < \rho_1 < 1$ or $0.4 < \rho_1 < 1$ respectively. In our experiments the ϕ_2-values are varied by steps of 0.1, except that the non-stationary boundaries are avoided by taking a minimum value of –0.98 and a maximum value of 0.58 or 0.18 respectively.

Case 2. In the second case the two forecasts are again based on only a single observation, but now with either a one-period or a two-period lag; each forecast is unbiased, conditional on its limited information set. Thus

$$f_{it} = \rho_i y_{t-i}, \quad \sigma_i^2 = \left(1 - \rho_i^2\right)\sigma_y^2, \quad i = 1, 2 .$$

To equate the error variances we choose parameter values such that $\rho_1^2 = \rho_2^2$, specifically $\phi_1 = \phi_2$ to deliver $\rho_1 = \rho_2$, or $\phi_1 = -\phi_2$ to deliver $\rho_1 = -\rho_2$. With these restrictions stationarity requires that $-1 < \phi_2 < 0.5$, with $\phi_1 = \pm\phi_2$ as appropriate. If $\phi_1 = \phi_2 = 0$ there is an obvious singularity: the series is white noise, the forecasts are equal to the mean of zero and have the same error, and k_{opt} is indeterminate.

In each case 1000 artificial time series samples are generated for each parameter combination. After discarding a "start-up" portion, each sample is divided into an estimation subsample of *R* observations and an evaluation subsample of *P* observations. The forecast parameter values are assumed known, and the estimation subsample is used simply to estimate the combining weight, via the above expressions for k_{opt} and k'. Estimates based on equation (2) need not satisfy $0 \le \hat{k}_{opt} \le 1$, as noted in Section 2.1, especially if the correlation between the competing forecast errors is large and positive, and we consider two possibilities. One is to use the actual point estimate, whatever value materialises; the second, following widespread practice endorsed by Granger and Newbold (1986, §9.2), is to replace an estimate outside this range by the nearest boundary value, 0 or 1 as appropriate. There are then four combined forecasts whose MSFEs are calculated over the last *P* observations: three weighted averages, in turn using $\hat{k}'$ and the initial and truncated $\hat{k}_{opt}$, and the simple average using $k = \frac{1}{2}$. The estimation cost of each weighted average is expressed as the percentage increase in MSFE above that of the simple average, namely $100\left(\hat{\sigma}_w^2 - \hat{\sigma}_s^2\right)\Big/\hat{\sigma}_s^2$.

Results for Case 1. Figure 1 shows the mean (over 1000 replications) percentage increase in MSFE over the simple average for the three combined forecasts based on estimated weights, with subsample sizes $R = 30$ and $P = 6$. The cost of estimating *k* is in general positive, as anticipated. However the different estimates give substantially different performance, for both values of ϕ_1 used in these examples. First, estimating the optimal weight, including the covariance term, increases the MSFE of the weighted average by a rather greater amount than using the estimate that neglects the forecast error correlation. Second, restricting the point estimate of the optimal weight to the (0,1) interval makes little difference when the correlation between the forecast errors is low, but improves the performance of the combined forecast as this correlation increases. The forecast error correlation increases with ϕ_2, and is equal to 0.4

at the point at which the two upper plots begin to diverge in panel (a) of Figure 1; its value is 0.57 at the equivalent point in panel (b). In each panel the last points plotted refer to an almost degenerate case in which the first-order autocorrelation coefficient of the data is close to 1 and the two competing forecasts, and hence their errors, are close to equality.

The key to the differences shown in Figure 1 is the sampling distribution of the different estimates of the weight. These are shown in Figure 2 for a parameter combination at which the two upper plots in panel (b) of Figure 1 are clearly separated, but not extremely so: the values are $\phi_1 = 0.8,\ \phi_2 = -0.1$, at which the forecast error correlation, neglected in the k' formula, is equal to 0.73. The distribution of the initial estimate of the optimal weight is shown in panel (a) of Figure 2. In our sample of 1000 there are 44 negative estimates, and these are set equal to zero in panel (b), which results in an improvement in relative MSFE of some 0.5%, as shown in Figure 1(b). Note the changes in scale between the panels of Figure 2: in particular, panels (b) and (c) have the same horizontal scale, to emphasise the much smaller dispersion of the estimate $\hat{k}'$, which in turn calls for a change in vertical scale. The better performance of $\hat{k}'$ in this experiment is striking, and supports the recommendation to ignore the error covariances noted above, based on practical applications.

Further support for a preference for $\hat{k}'$ over $\hat{k}_{opt}$ is provided by the asymptotic approximations to the variances of these estimators calculated in the Appendix. These obey the relation

$$\text{asy var}\left(\hat{k}'\right) = \left(1-\rho\right)^2 \text{asy var}\left(\hat{k}_{opt}\right)$$

where ρ is the forecast error correlation coefficient. This correlation is positive in our experiments, and can be expected to be positive more generally, since the innovation ε_t is common to the competing forecast errors. The formulae developed in the Appendix are seen to provide a good approximation to the simulation variance of the estimates obtained from samples of *R*=30 observations. More generally these results offer an explanation of the relatively poor performance of combinations based on the optimal weight.

Results for Case 2. The results shown in Figure 3 are qualitatively similar to those reported for Case 1. The cost of estimating *k* is in general positive. Again the different estimates of the weights yield substantially different performance, the ranking of the different estimates

remaining as seen in Case 1, with the performance of $\hat{k}'$ again markedly superior to that of $\hat{k}_{opt}$, thanks to the much smaller dispersion of its sampling distribution. Comparing the performance of the two estimates of k_{opt}, Figure 3(a) shows that at $\phi_1 = \phi_2 = -0.5$, at which the forecast error correlation, neglected in the k' expression, is 0.833, truncating the original estimate gives an improvement in relative MSFE of 0.4%: this is the result of setting 39 negative estimates equal to zero and 34 estimates equal to one, in our sample of 1000. For $\phi_1 = -\phi_2 = 0.5$ (see Figure 3(b)) there is an improvement in relative MSFE of 0.35%; here the error correlation is slightly smaller, at 0.805, and slightly fewer $\hat{k}_{opt}$ values are set to the boundary values, 35 and 23 respectively. An example of the sampling distributions of the weights presented in Figure 4 shows the truncation effects diagrammatically, also that $\hat{k}'$ again has much smaller dispersion. In both panels of Figure 3 the truncation effect increases as ϕ_1 approaches zero from above or below, when the correlation between the forecast errors increases towards 1 and we approach the singularity at $\phi_1 = \phi_2 = 0$ noted above.

The behaviour of the estimates of the optimal weight differs between the examples of Case 1 and Case 2 discussed above. In the first case truncation of the initial estimate is necessary on only one side, thus in Figure 2(b) there is a pile-up at zero but not at one. In the second case the (0,1) interval is breached on both sides, as shown in Figure 4. Recalling that $\hat{k}_{opt} < 0$ if $s_2^2 < s_{12}$ and $\hat{k}_{opt} > 1$ if $s_1^2 < s_{12}$, and that the experiment is designed with $\sigma_1^2 = \sigma_2^2$, the explanation of the asymmetry lies in the sampling distribution of the variance estimates. In the example of Case 1 shown in Figure 2 the sample variance of s_2^2 is some 35% greater than that of s_1^2, resulting in a tail of the distribution such that $s_2^2 < s_{12}$ on 44 out of 1000 occasions, whereas the equivalent condition for $\hat{k}_{opt} > 1$ never occurs in this sample. In neither Case 1 nor Case 2 does estimating an intercept term in the extended realisation-forecast regression equation (3) improve the MSFE of the associated combined forecast. In those experiments in which estimating α has a noticeable effect the result is an increase in the combined forecast MSFE, associated with greater imprecision in the estimate of k_{opt}.

The effects under discussion are finite-sample estimation effects, as noted initially in general terms and more precisely in the closing paragraph of Section 2. These effects relate to

the size of the "regression" sample, *R*, not the "prediction" sample, *P*. Increasing *P* in our experiments has little effect on the mean of the MSFE costs, such as those plotted in Figures 1 and 3, although their sampling variance falls, as expected. Increasing *R*, however, reduces the MSFE cost of the weighted average, due to increased accuracy of the estimated weight, and at *R*=1200 there is essentially no gain in using the simple average, and hence no puzzle.

Departures from equal weights. These examples show that if the optimal combining weights are equal, then the simple average beats estimated combinations. A practical next question might be, but what if the optimal weights are not equal? How different must the optimal weights be for the bias effect from assuming them equal to dominate the estimation variance effect, so that the combination with estimated weights beats the simple average? A little evidence on this question is contained in Figures 5 and 6. For particular parameter combinations of Case 1 and Case 2 respectively, we change the coefficient on the second forecast in each case, in order to change its forecast error variance and hence the required combining weight. This is indicated by the associated value of k', using the expression without the covariance term that the preceding experiments clearly indicate is to be preferred. Otherwise the experimental design remains unaltered, 1000 replications being undertaken to estimate the percentage MSFE cost at each of a range of values of k'. When this is equal to 0.5 the results correspond to those given above; departures from 0.5 are indicated by referring to the experiments as Case 1* and Case 2* respectively.

The results show that, in these examples, large departures from equal weights are not required in order to turn the comparison around. The MSFE cost of the weighted estimate using $\hat{k}_{opt}$ is greater than that using $\hat{k}'$, hence positive costs persist for greater departures from equality. But in all the cases presented, the cost has turned negative, that is, the simple average has lost its advantage, before the weights are as different as (0.4, 0.6). This combination is appropriate if one forecast has MSFE 50% greater than its competitor, and how relevant this is in practice is for the user to judge.

3.2 Forecasts of US output growth during the 2001 recession

For a practical application we revisit the study of Stock and Watson (2003a), which evaluates the performance of leading indicator forecasts during the 2001 recession in the United States. This recession differed in many ways from its predecessors, and Stock and Watson find that

individual leading indicators also performed differently before and during this recession. Some previously reliable leading indicators provided little or no indication of the slowdown.

Of particular interest for our present purpose is their Table 4, which reports relative MSFEs of various combination forecasts of annual growth rates of real GDP and the Index of Industrial Production over the period 1999Q1-2002Q3, at lead times of h=2 and h=4 quarters. As noted in Section 2.3, the simple average of individual leading indicator-based forecasts dominates a weighted average using inverse MSFE weights as in equation (8). The weights are based on MSFEs calculated as discounted sums of past squared forecast errors, with a quarterly discount factor of 0.95. From their programs and database we recreate the combined forecast MSFEs on which the relative MSFEs reported in their table are based. (Throughout their article Stock and Watson report the MSFEs of individual and combined forecasts relative to the MSFE of a benchmark autoregressive forecast, whereas we need the numerators of these ratios.) We then undertake further calculations following the algebra developed above, with results as shown in Table 1. These examples use n=35 component forecasts, while P=13 when h=2 and P=11 when h=4.

Table 1. MSFEs of combined quarterly forecasts of output growth (annual percentage rates)

	RGDP		IP	
	h=2	h=4	h=2	h=4
$\hat{\sigma}_s^2$	1.0078	3.8730	4.4663	23.0034
$\hat{\sigma}_w^2$	1.0319	4.0194	4.4926	23.2114
$\hat{\sigma}_w^2 - \text{adj}$	1.0314	4.0165	4.4894	23.2032
$\hat{\sigma}_s^2 - \hat{\sigma}_w^2$	–0.0241	–0.1464	–0.0263	–0.2080
adj	0.0005	0.0029	0.0032	0.0082
remainder	–0.0236	–0.1435	–0.0231	–0.1998

The first two rows of the table give the MSFEs of the simple and weighted average forecasts and show that the simple average does better in all four cases. In the third row we make the adjustment defined in equation (6) and given in the fifth row, but in no case does this change the ranking. Decomposing the MSFE difference, given in the fourth row, into the

two components developed in equation (5), shows that the (sample mean of) the first component, which has expected value zero under the equal-weight null, heavily dominates the second, "adjustment", component. This second component is the average squared difference between the two combination forecasts, or equivalently, between their forecast errors, and its small size relative to the separate MSFEs indicates that these forecasts are very close to one another. From equation (5), second line, the remainder term is equal to $2\left(\hat{\sigma}_s^2 - \hat{\sigma}_{sw}\right)$, and if this is negative in a situation in which the two forecast MSFEs are close to one another then the forecast error correlation must be close to 1. Calculation of the forecast error correlation coefficient from the data in the first column of Table 1 gives 0.99981.

We plot the forecast errors for all four cases under consideration in Figure 7, which confirms these interpretations of the data in Table 1. The two combination forecast errors are virtually indistinguishable, and the difference between the two combination forecasts is unlikely to be of any importance in the context of a practical decision problem. The dates shown correspond to the date of the forecast, the final outcome available being that for 2002Q3. Stock and Watson note that at the time of writing the NBER had not yet dated the trough: on 17 July 2003 this was announced as November 2001 which, with the previous peak at March 2001, gave a contraction duration of 8 months. Figure 7 shows that the combination forecasts substantially overestimated growth throughout this period, starting from the quarter before the peak and, in the year-ahead forecasts, extending well into the recovery phase. We beg to differ from Stock and Watson's conclusion that the combination forecast performance is "encouraging".

In this example the forecast combination puzzle is of no importance from a practical point of view. From a statistical point of view it is an example of the gain in efficiency that can be obtained by imposing, rather than estimating, a restriction that is approximately true. The distribution of estimated weights at the start of the prediction period for the example in column 1 of Table 1 is presented in Figure 8, and this shows rather little variation around the value of $1/n$=0.029 used by the simple average. The performance of the individual indicators varies over time, hence so do their optimal combining weights, but when the relative weights are small this variation is also likely to have little practical significance.

4. CONCLUSIONS

Three main conclusions emerge from the foregoing analysis.

(1) If the optimal combining weights are equal or close to equality, a simple average of competing forecasts is expected to be more accurate, in terms of MSFE, than a combination based on estimated weights.

(2) However if estimated weights are to be used, then it is better to neglect any covariances between forecast errors and base the estimates on inverse MSFEs alone, than to use the optimal formula originally given by Bates and Granger for two forecasts, or its regression generalisation for many forecasts.

(3) When the number of competing forecasts is large, so that under equal weighting each has a very small weight, the simple average can gain in efficiency by trading off a small bias against a larger estimation variance. Nevertheless, in an example from Stock and Watson (2003a), we find that the "forecast combination puzzle" rests on a gain in MSFE that has no practical significance.

APPENDIX: THE VARIANCES OF THE COMBINING WEIGHTS

We calculate large-sample approximations to the variances of the two estimators of the combining weight, in the case of two forecasts with equal error variances $\sigma_1^2 = \sigma_2^2 = \sigma_e^2$, say. We first consider the "inverse MSFE" coefficient, which neglects the covariance between the forecast errors, namely

$$\hat{k}' = \frac{\frac{1}{R}\Sigma e_{2t}^2}{\frac{1}{R}\Sigma e_{1t}^2 + \frac{1}{R}\Sigma e_{2t}^2} = \frac{s_2^2}{s_1^2 + s_2^2} = \frac{1}{1 + s_1^2 / s_2^2},$$

retaining *R* to denote the estimation sample size. Standard results on the variance of functions of random variables (Stuart and Ord, 1994, §10.6) give, for the nonlinear transformation,

$$\text{var}\left(1+x\right)^{-1} \approx \left(\frac{\partial \left(1+x\right)^{-1}}{\partial x} \right)^2 \text{var}(x),$$

and evaluating the derivative at the mean of 1 gives

$$\text{var}\left(\hat{k}'\right) \approx \frac{1}{16}\text{var}\left(\frac{s_1^2}{s_2^2}\right).$$

Also using the expression for the variance of a ratio of positive random variables, we have

$$\text{var}\left(\hat{k}'\right) \approx \frac{1}{16}\left(\frac{E\left(s_1^2\right)}{E\left(s_2^2\right)}\right)^2 \left(\frac{\text{var}\left(s_1^2\right)}{E^2\left(s_1^2\right)} + \frac{\text{var}\left(s_2^2\right)}{E^2\left(s_2^2\right)} - \frac{2\text{cov}\left(s_1^2, s_2^2\right)}{E\left(s_1^2\right).E\left(s_2^2\right)} \right)$$

$$= \frac{1}{16\sigma_e^4}\left[\text{var}\left(s_1^2\right) + \text{var}\left(s_2^2\right) - 2\,\text{cov}\left(s_1^2, s_2^2\right) \right]. \qquad \text{(A.1)}$$

Turning to the optimal weight given in equation (2), the estimate is

$$\hat{k}_{opt} = \frac{\frac{1}{R}\Sigma e_{2t}^2 - \frac{1}{R}\Sigma e_{1t}e_{2t}}{\frac{1}{R}\Sigma e_{1t}^2 + \frac{1}{R}\Sigma e_{2t}^2 - \frac{2}{R}\Sigma e_{1t}e_{2t}} = \frac{s_2^2 - s_{12}}{s_1^2 + s_2^2 - 2s_{12}} = \frac{1}{1 + \frac{s_1^2 - s_{12}}{s_2^2 - s_{12}}}.$$

This last expression is of the same form as $\hat{k}'$, with $s_i^2 - s_{12}$ replacing s_i^2, $i = 1, 2$. To follow the same development as above we first note that

$$E\left(s_i^2 - s_{12}\right) = \left(1-\rho\right)\sigma_e^2,$$

where ρ is the forecast error correlation coefficient. On expanding expressions for the variances and covariance of $s_i^2 - s_{12}$, $i = 1, 2,$ and collecting terms, we then obtain

$$\text{var}\left(\hat{k}_{opt}\right) \approx \frac{1}{16(1-\rho)^2 \sigma_e^4}\left[\text{var}\left(s_1^2\right) + \text{var}\left(s_2^2\right) - 2\,\text{cov}\left(s_1^2, s_2^2\right)\right]$$

$$= \frac{1}{(1-\rho)^2}\text{var}\left(\hat{k}'\right). \qquad \text{(A.2)}$$

Or, more directly, we observe that the expression in square brackets in (A.1) is the variance of $s_1^2 - s_2^2$, and that this is equal to the variance of $\left(s_1^2 - s_{12}\right) - \left(s_2^2 - s_{12}\right)$. Since the errors of competing forecasts are in general positively correlated, the estimation variance of the optimal weight can be expected to exceed that of the inverse MSFE weight.

To implement expressions (A.1) and (A.2) further development of the terms in square brackets is needed. The asymptotic variance of the sample variance of a normally distributed autocorrelated series with autocorrelation coefficients ρ_j is

$$\text{var}\left(s^2\right) \approx \frac{2\sigma^4}{R}\sum_{-\infty}^{\infty}\rho_j^2 .$$

This result is given in many time-series texts, whose authors usually cite Bartlett (1946). In this article Bartlett stated a more general result, for the covariance of two sample autocovariances, and he subsequently gave an outline derivation in his book (1955, §9.1). Following the same approach gives the covariance term we require as

$$\text{cov}\left(s_1^2, s_2^2\right) \approx \frac{2}{R}\sum_{-\infty}^{\infty}\gamma_{12}^2(j) ,$$

where $\gamma_{12}(j)$ is the cross-lagged covariance function of e_1 and e_2. So altogether we have

$$\text{var}\left(\hat{k}'\right) \approx \frac{1}{8R}\left(ACS_1 + ACS_2 - 2CCS\right)$$

where ACS_i is the (doubly infinite) sum of squared autocorrelation coefficients of series e_{it}, i=1,2, and CCS is the corresponding sum of their squared cross-lagged correlation coefficients. In the set-up of our Monte Carlo experiments these coefficients are obtained from the autocovariance generating function of the AR(2) process for y_t and the related generating functions for the filtered series $e_{it} = h_i(L)y_t$. The infinite sums are truncated once the individual terms are sufficiently small, and possible sensitivity of the final result to the truncation point is checked.

The resulting "theoretical" standard deviations of the distributions of the two estimators are shown in Table A.1 for a selection of the parameter values used in our experiments, alongside their "empirical" counterparts calculated from the simulation sample distributions. With 1000 independent replications at each parameter combination, the standard error of the estimated standard deviation (sd) is equal to $\text{sd}/\sqrt{2000} = 0.022\text{sd}$. With this in mind, the large-sample approximation is seen to provide reliable guidance to the simulation results for sample sizes as small as 30, for most of the parameter combinations considered. The approximation becomes less good as the autocorrelation of the forecast error series increases, and in these circumstances somewhat larger sample sizes are required before the equivalence is restored. Nevertheless the theoretical analysis of this Appendix provides more general support for a preference for $\hat{k}'$ over $\hat{k}_{opt}$: neglecting the covariances of the competing forecast errors can be expected to lead to improved performance of combined forecasts based on estimated weights.

Table A.1. Empirical and theoretical standard deviations of $\hat{k}'$ and $\hat{k}_{opt}$

		$\hat{k}'$		$\hat{k}_{opt}$	
ϕ_2	ρ	Empirical	Theoretical	Empirical	Theoretical
		Case 1: $\phi_1 = 0.4$			
0.4	0.67	0.086	0.104	0.323	0.243
0.2	0.50	0.087	0.097	0.193	0.174
0.0	0.40	0.079	0.084	0.139	0.131
-0.2	0.33	0.069	0.070	0.107	0.102
-0.4	0.29	0.059	0.057	0.083	0.078
-0.6	0.25	0.048	0.044	0.065	0.058
-0.8	0.22	0.036	0.030	0.047	0.038
		Case 1: $\phi_1 = 0.8$			
0.1	0.89	0.040	0.046	0.549	0.416
-0.1	0.73	0.053	0.057	0.228	0.208
-0.3	0.62	0.051	0.053	0.146	0.137
-0.5	0.53	0.046	0.045	0.103	0.096
-0.7	0.47	0.037	0.034	0.072	0.064
-0.9	0.42	0.027	0.019	0.046	0.032
		Case 2: $\phi_1 = \phi_2$			
-0.9	0.57	0.043	0.037	0.106	0.086
-0.7	0.71	0.051	0.049	0.184	0.172
-0.5	0.83	0.046	0.046	0.285	0.274
-0.3	0.93	0.033	0.032	0.480	0.468
-0.1	0.99	0.012	0.012	1.374	1.347
0.1	0.99	0.014	0.014	1.242	1.219
0.3	0.87	0.044	0.044	0.351	0.343
		Case 2: $\phi_1 = -\phi_2$			
0.3	0.87	0.042	0.044	0.347	0.343
0.1	0.99	0.013	0.014	1.205	1.219
-0.1	0.99	0.011	0.012	1.315	1.347
-0.3	0.93	0.030	0.032	0.456	0.468
-0.5	0.83	0.043	0.046	0.270	0.274
-0.7	0.71	0.048	0.049	0.175	0.172
-0.9	0.57	0.041	0.037	0.103	0.086

REFERENCES

Bartlett, M.S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society Supplement*, 8, 27-41.

Bartlett, M.S. (1955). *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*. Cambridge: Cambridge University Press.

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

Clark, T.E. and McCracken, M.W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105, 85-110.

Clark, T.E. and West, K.D. (2004). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Research Working Paper 04-03, Federal Reserve Bank of Kansas City.

Clemen, R.T. (1986). Linear constraints and the efficiency of combined forecasts. *Journal of Forecasting*, 5, 31-38.

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

Granger, C.W.J. and Newbold, P. (1986). *Forecasting Economic Time Series*, 2nd ed. London: Academic Press.

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

McCracken, M.W. and West, K.D. (2002). Inference about predictive ability. In *A Companion to Economic Forecasting* (M.P. Clements and D.F. Hendry, eds), pp.299-321. Oxford: Blackwell.

Stock, J.H. and Watson, M.W. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger* (R.F. Engle and H.White, eds), pp.1-44. Oxford: Oxford University Press.

Stock, J.H. and Watson, M.W. (2003a). How did leading indicator forecasts perform during the 2001 recession? *Federal Reserve Bank of Richmond Economic Quarterly*, 89/3, 71-90.

Stock, J.H. and Watson, M.W. (2003b). *Introduction to Econometrics*. Boston: Addison Wesley.

Stock, J.H. and Watson, M.W. (2003c). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, 41, 788-829.

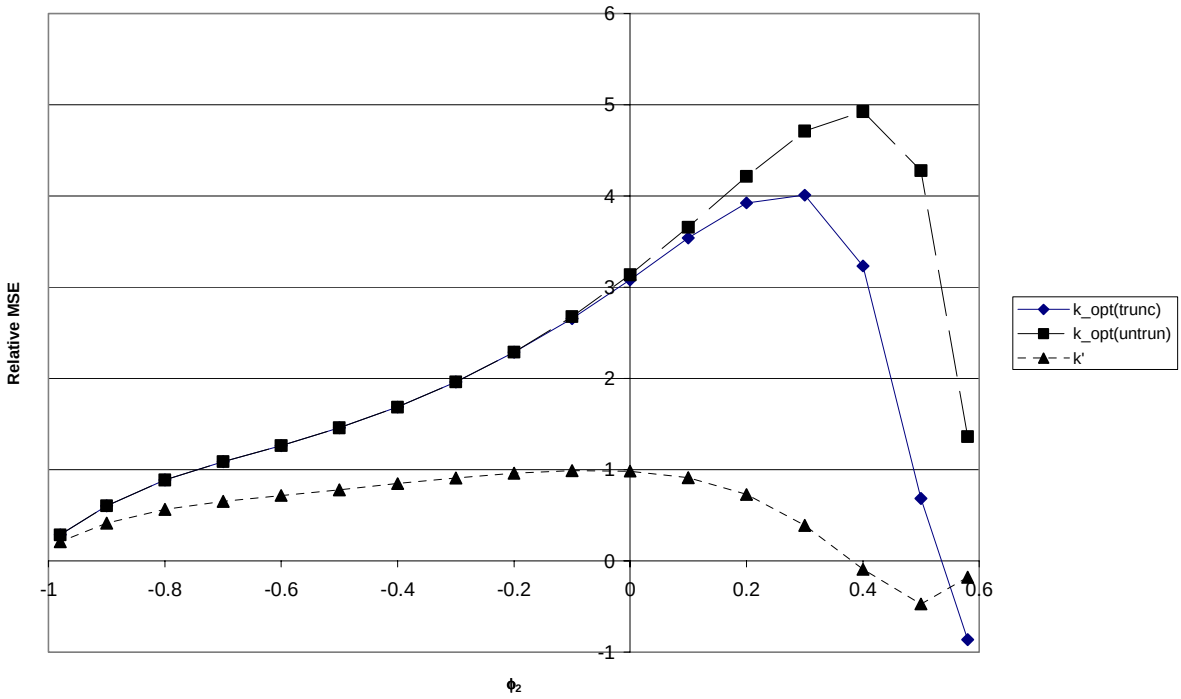
Stock, J.H. and Watson, M.W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23, 405-430.

Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, 6th ed., vol. 1. London: Edward Arnold.

Wallis, K.F. (2004). Combining density and interval forecasts: a modest proposal. Unpublished paper, University of Warwick.

Figure 1. Percentage MSFE cost of weighted combination forecasts; Case 1.

(a) ϕ_1 =0.4, R=30, P=6



(b) ϕ_1 =0.8, R=30, P=6

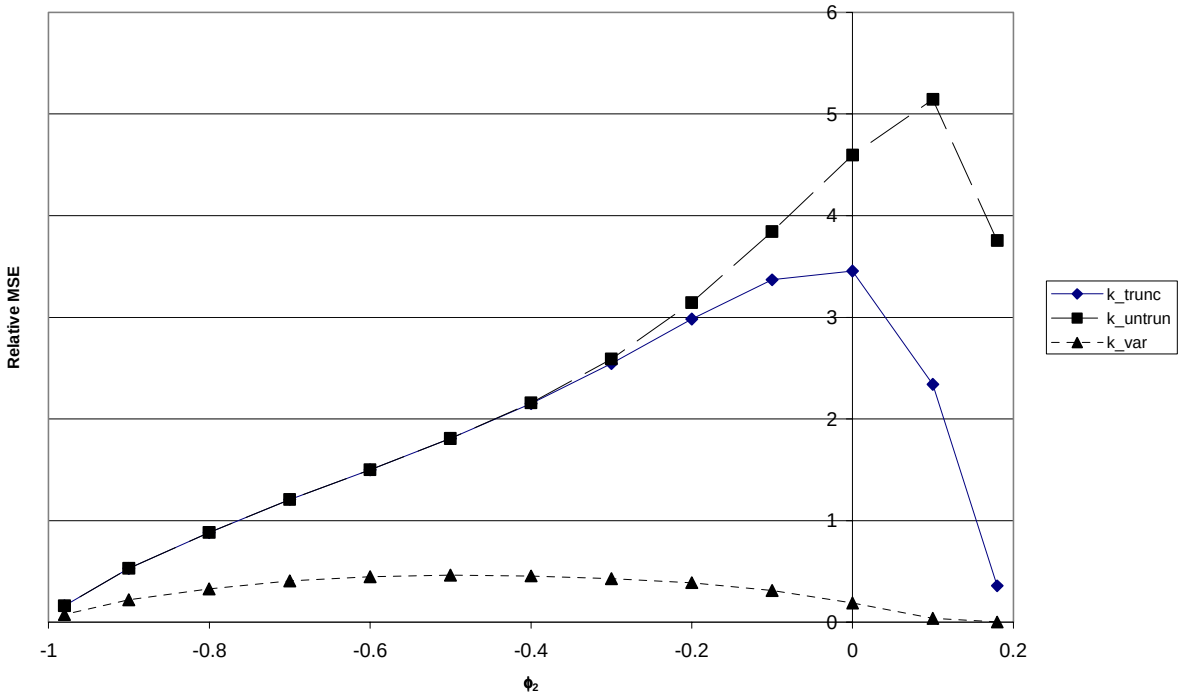
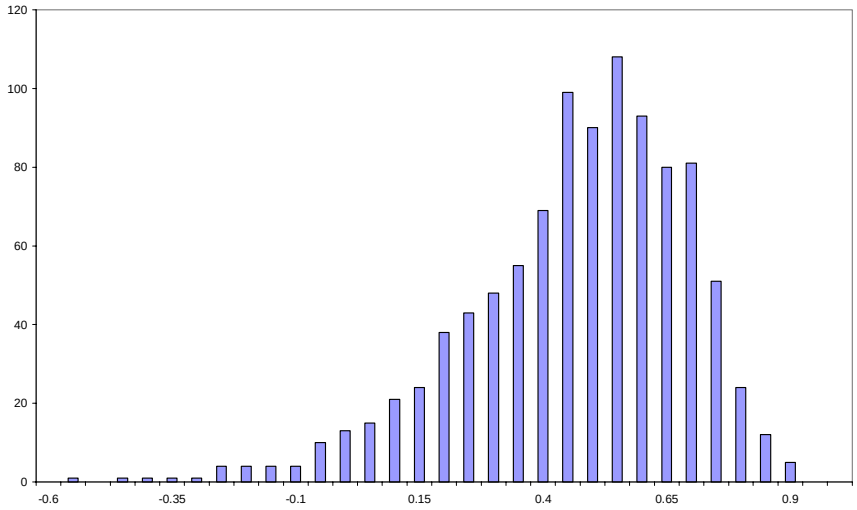
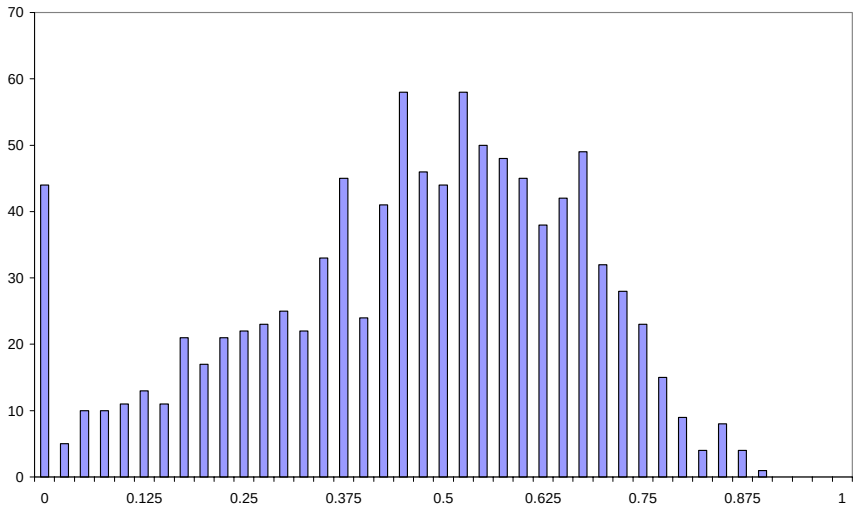


Figure 2. Histograms of estimated weights; Case 1, $\phi_1 = 0.8,\ \phi_2 = -0.1$, R=30.

(a) $\hat{k}_{opt}$, untruncated



(b) $\hat{k}_{opt}$, truncated



(c) $\hat{k}'$

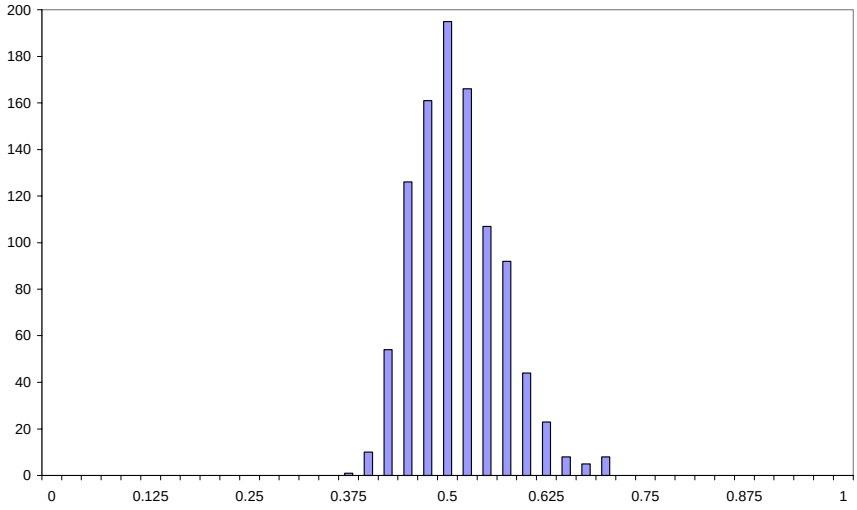
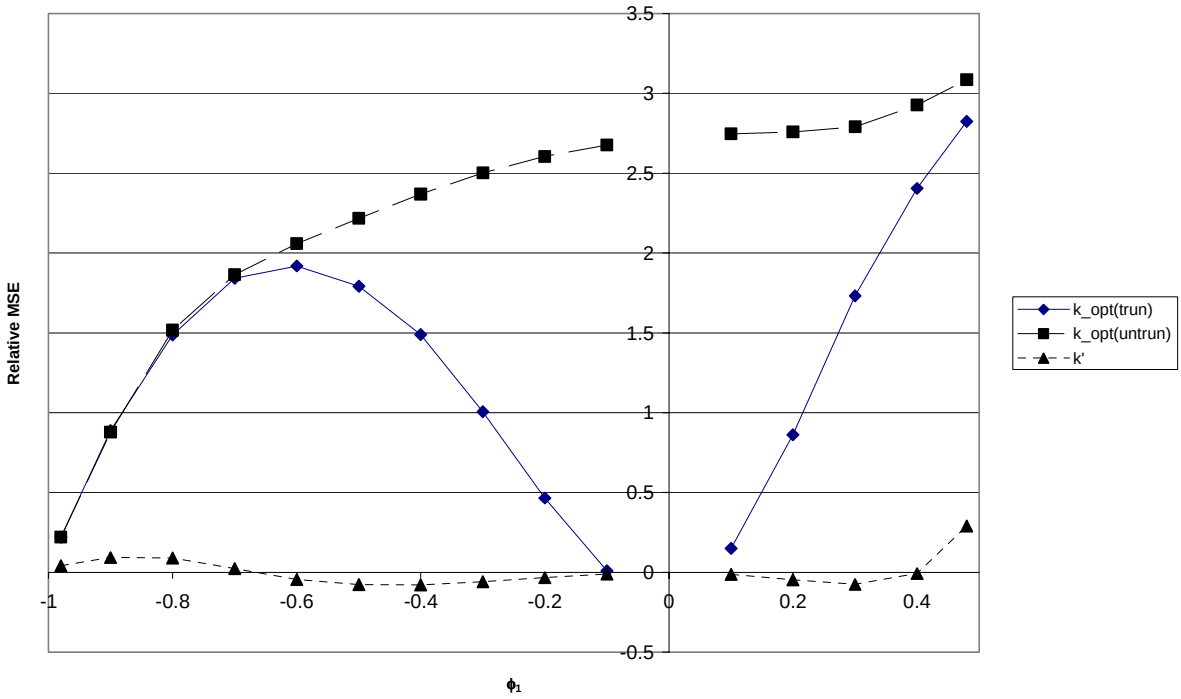


Figure 3. Percentage MSFE cost of weighted combination forecasts; Case 2.

(a) $\phi_1 = \phi_2$, R=30, P=6



(b) $\phi_1 = -\phi_2$, R=30, P=6

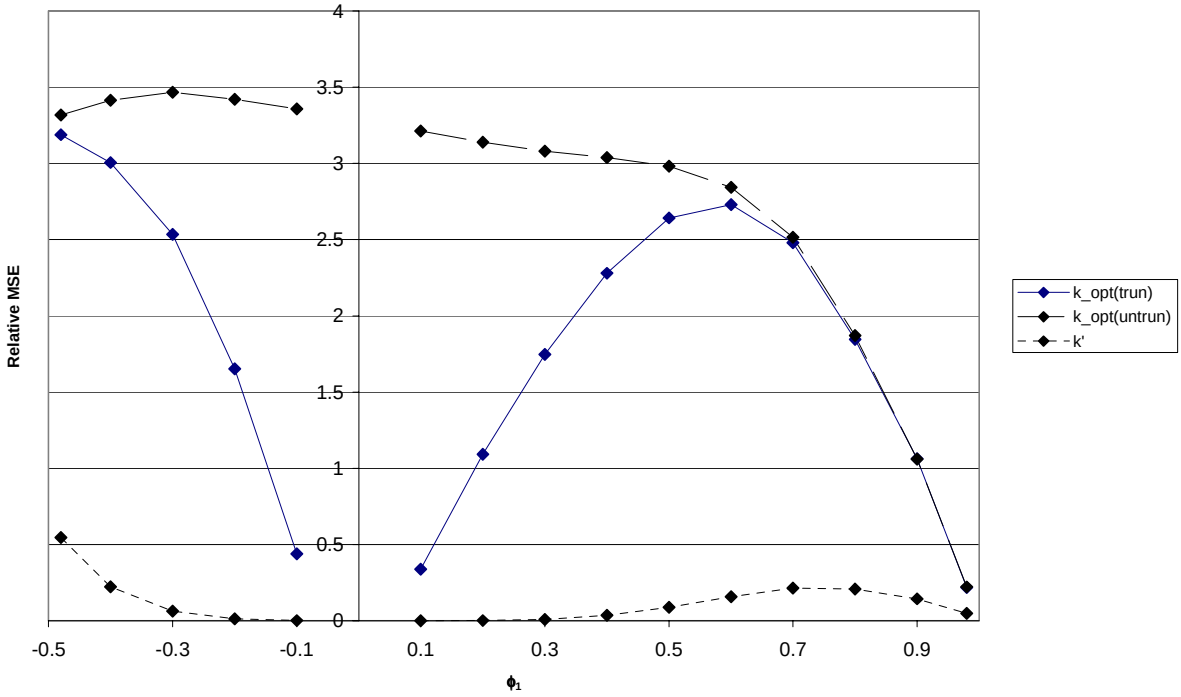
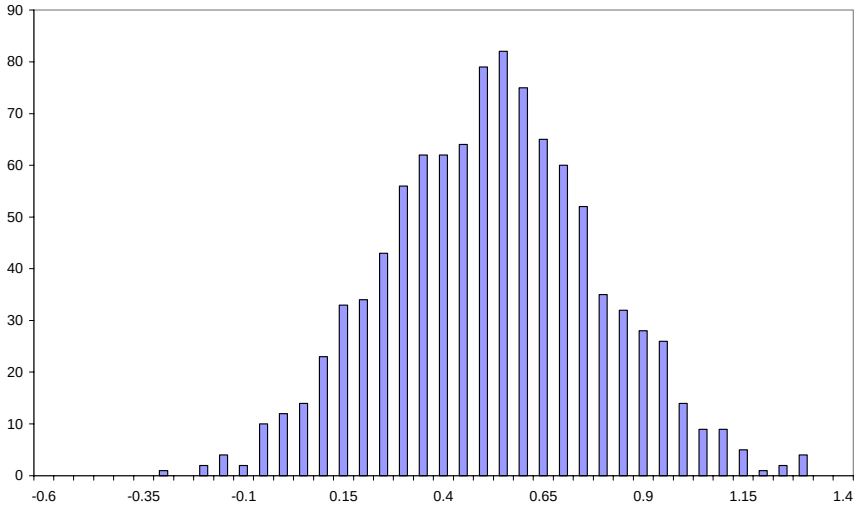
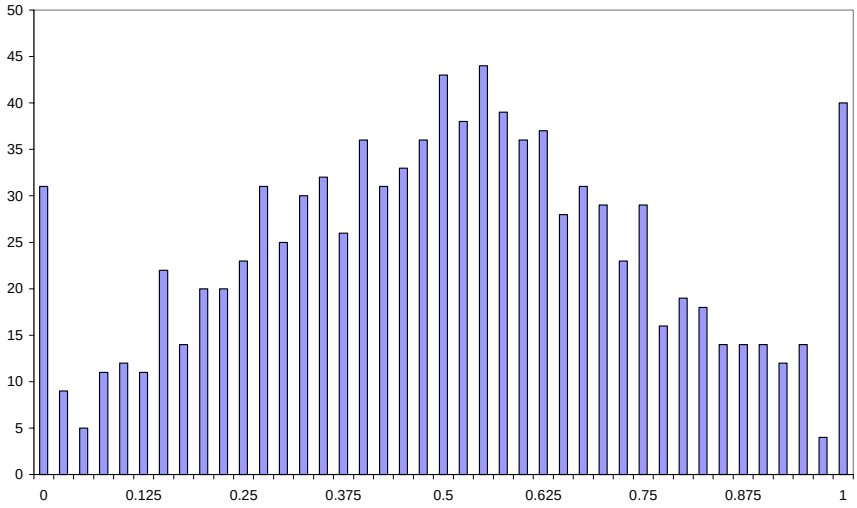


Figure 4. Histograms of estimated weights; Case 2, $\phi_1 = -\phi_2 = 0.5$, R=30.

(a) $\hat{k}_{opt}$, untruncated



(b) $\hat{k}_{opt}$, truncated



(c) $\hat{k}'$

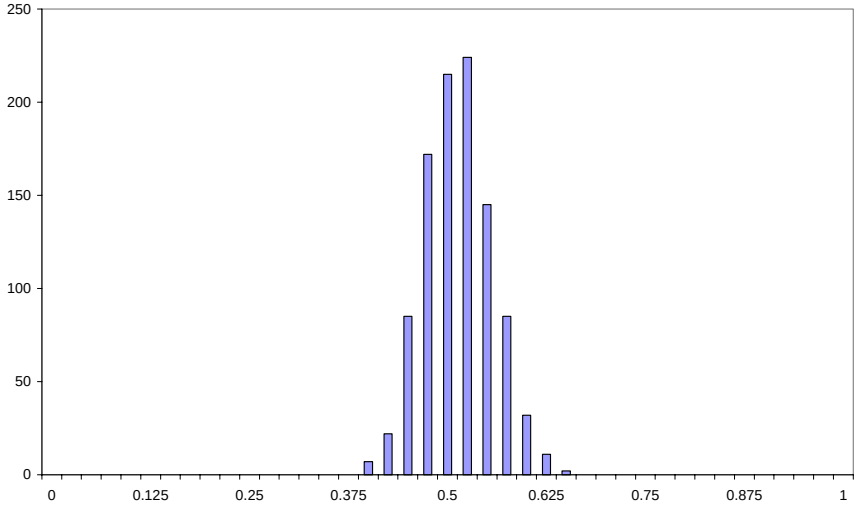
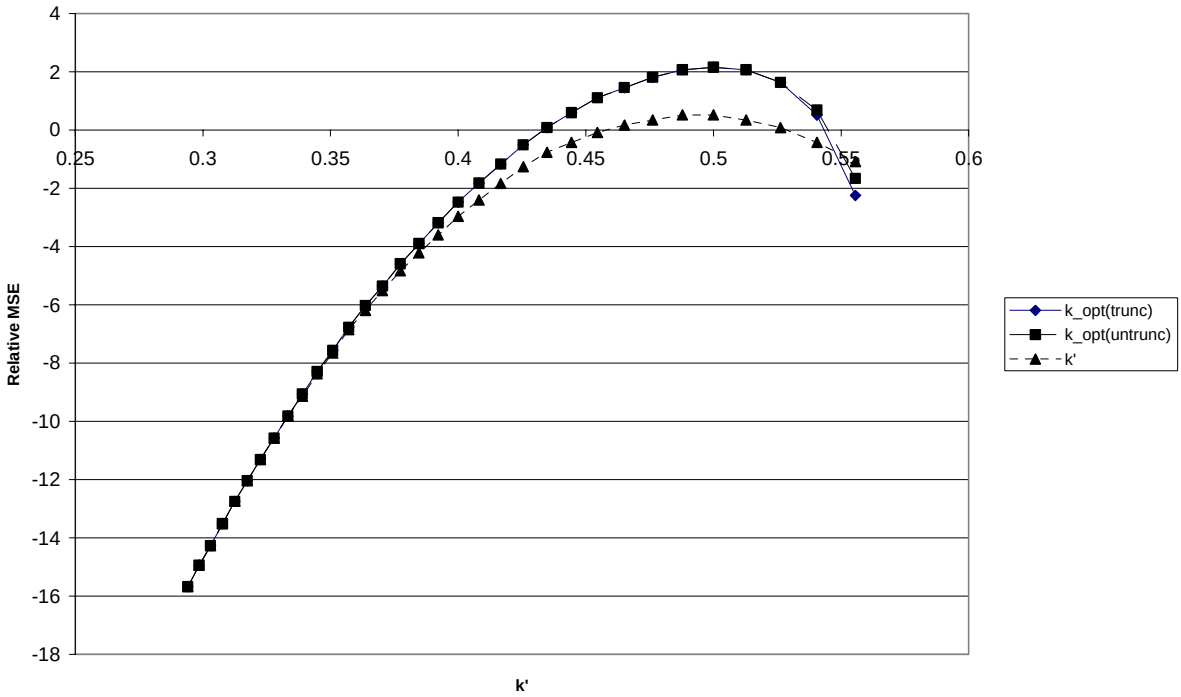


Figure 5. Percentage MSFE cost of weighted combination forecasts; Case 1*.

(a) $\phi_1 = 0.8, \phi_2 = -0.4$, R=30, P=6



(b) $\phi_1 = 0.4, \phi_2 = -0.2$, R=30, P=6

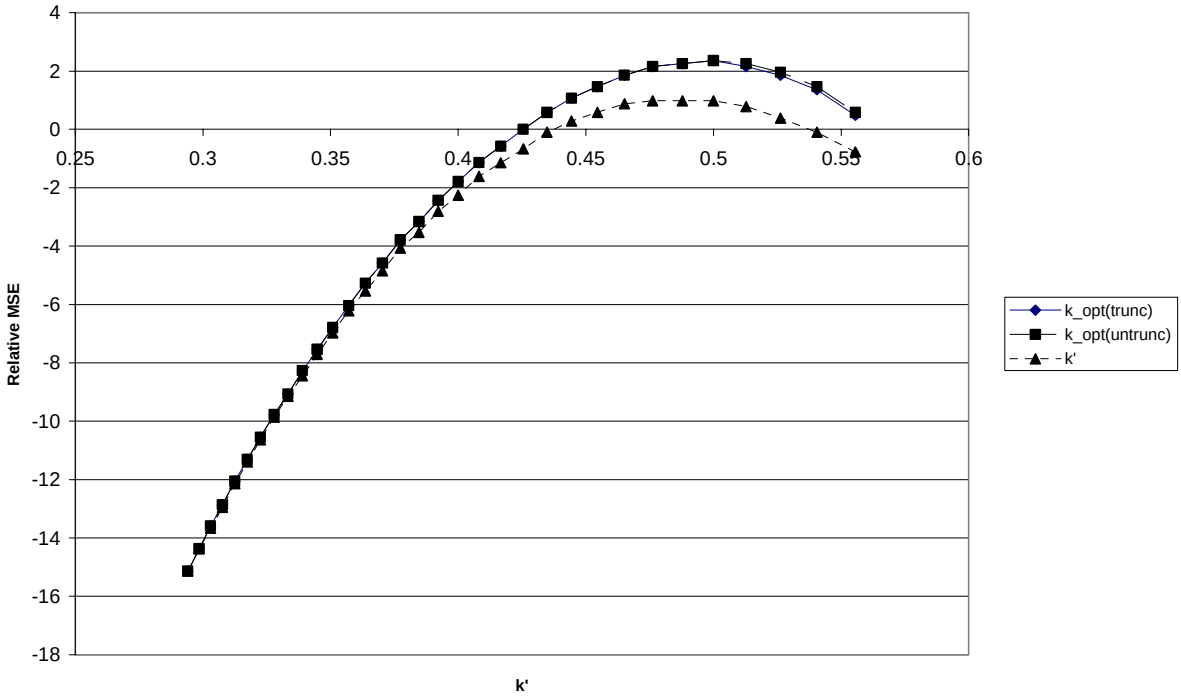
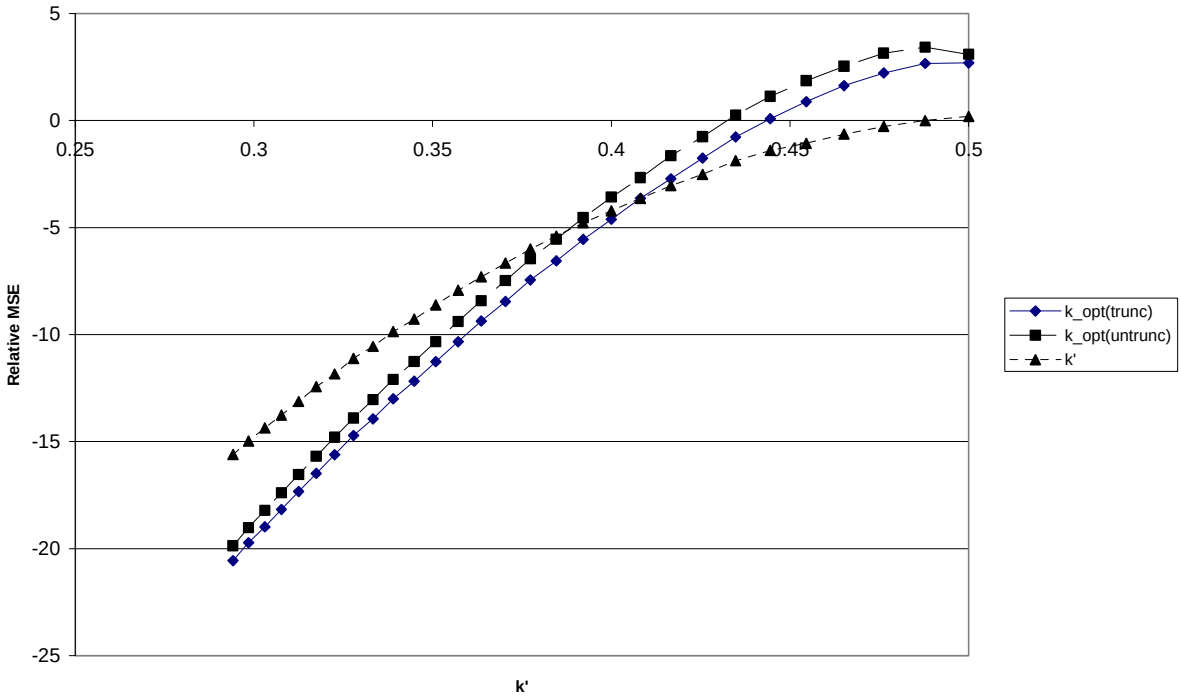


Figure 6. Percentage MSFE cost of weighted combination forecasts; Case 2*.

(a) $\phi_1 = \phi_2 = 0.45$, R=30, P=6



(b) $\phi_1 = -\phi_2 = 0.7$, R=30, P=6

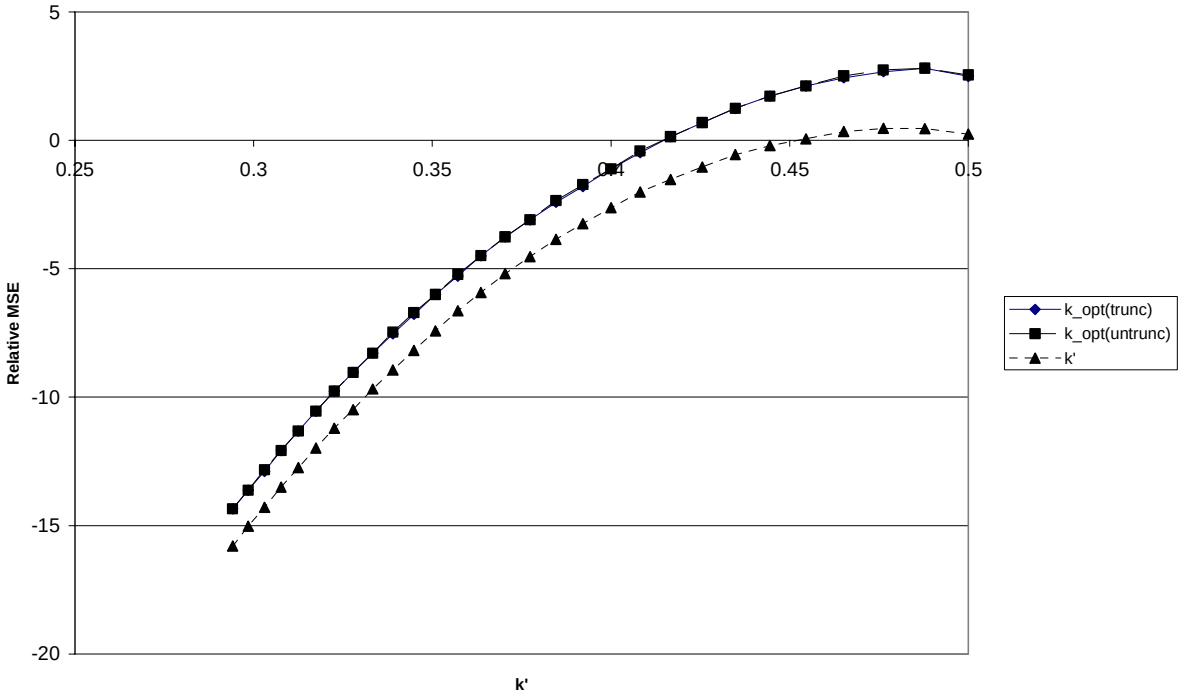
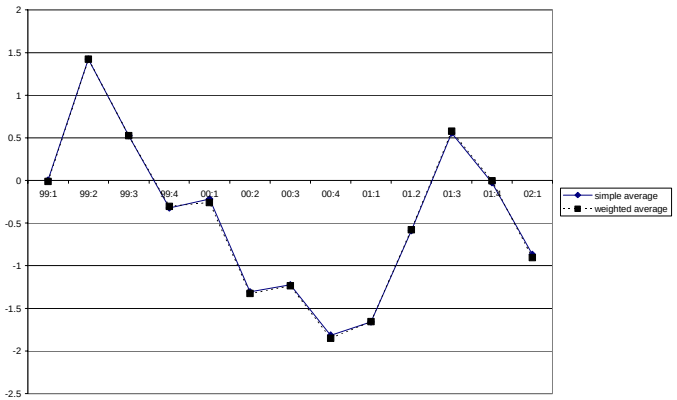
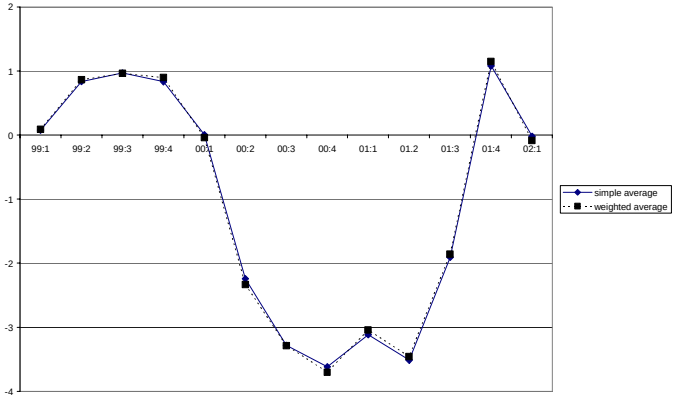


Figure 7. Errors of combined quarterly forecasts of output growth (annual percentage rates)

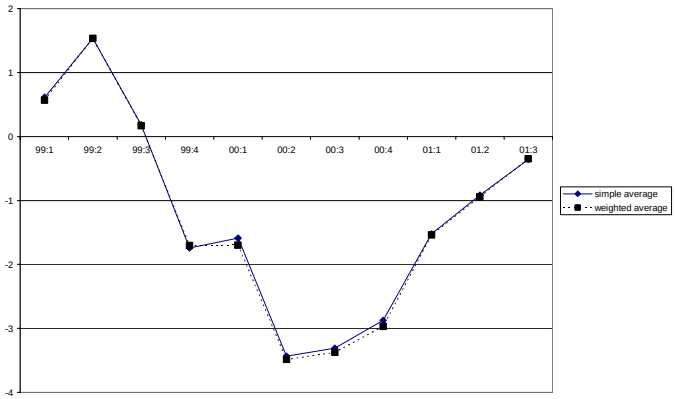
RGDP, *h*=2



IP, *h*=2



RGDP, *h*=4



IP, *h*=4

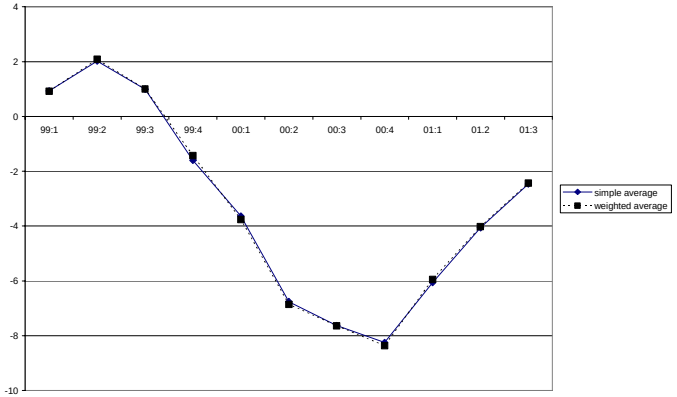


Figure 8. Distribution of weights in weighted average RGDP forecast (*n*=35, *h*=2, first period)