KU LEUVEN
FACULTY OF ECONOMICS
AND BUSINESS

The forecast combination puzzle: A simple theoretical explanation

Claeskens G, Magnus J, Vasnev A, Wang W.

KBI_1604

The forecast combination puzzle: A simple theoretical explanation

Gerda Claeskens
KU Leuven, Belgium

Jan R. Magnus
Vrije Universiteit Amsterdam and Tinbergen Institute, The Nertherlands

Andrey L. Vasnev[1]
University of Sydney, New South Wales, Australia

Wendun Wang
Econometric Institute, Erasmus University Rotterdam
and Tinbergen Institute, The Netherlands

Abstract:
This paper offers a theoretical explanation for the stylized fact that forecast combinations with estimated optimal weights often perform poorly in applications. The properties of the forecast combination are typically derived under the assumption that the weights are fixed, while in practice they need to be estimated. If the fact that the weights are random rather than fixed is taken into account during the optimality derivation, then the forecast combination will be biased (even when the original forecasts are unbiased) and its variance is larger than in the fixed-weights case. In particular, there is no guarantee that the 'optimal' forecast combination will be better than the equal-weights case or even improve on the original forecasts. We provide the underlying theory, some special cases, and a numerical illustration.

Key words: forecast combination, optimal weights

JEL Codes: C53, C52

[1] Corresponding author

1 Introduction

When several forecasts of the same event are available, it is natural to try and find a (linear) combination of these forecasts which is 'best' in some sense. If we define 'best' in terms of the mean squared error and if the variances of the forecasts and their covariances are known, then optimal weights can be derived. In practice, these (co)variances are not known and need to be estimated. This leads to estimated optimal weights and an estimated optimal forecast combination. Empirical evidence and extensive simulations show that the estimated optimal forecast combination typically does not perform well, and that the arithmetic mean often performs better. This empirical fact has become known as the 'forecast combination puzzle'.

The history of the puzzle is elegantly summarized in Graefe et al. (2014, Section 4), and a rigorous attempt to explain it, using simulations and an empirical example, was undertaken by Smith and Wallis (2009) who show that the effect of the error in estimating the weights can be large, thus providing an empirical explanation of the forecast puzzle. Smith and Wallis use the words 'finite-sample' error, which suggests that this error may vanish asymptotically. But it is not so easy to find an asymptotic justification for ignoring the noise generated by estimating the weights. To begin with it is not clear what 'asymptotic' means here. What goes to infinity? The number of forecasts? If so, then the number of weights also goes to infinity. The number of observations underlying the total (but finite) set of forecasts? That would make more sense, but it would be difficult to analyze.

In this paper we provide a theoretical explanation for the empirical and simulation results of Smith and Wallis (2009) and others. The key ingredient in our approach is to acknowledge explicitly that the optimal weights should be derived by explicitly taking the estimation step into account. In other words, the derivation and estimation of optimal weights are viewed as a joint effort, not as two separate efforts. This approach differs from (almost) all previous research, not only Bates and Granger (1969), but also later contributions, important and insightful as they may be, such as Hansen (2008), Elliott (2011), Liang et al. (2011), and Hsiao and Wan (2014). The separation of mathematical derivation and statistical estimation can be quite dangerous. Such separations are still quite common in econometrics, although their disadvantages have been highlighted, specifically in the model-averaging literature which explicitly attempts to combine model selection and estimation, so that uncertainty in the model selection procedure is not ignored when reporting properties of the estimates; see for example Magnus and De Luca (2014).

In order to highlight our main findings we first provide graphical illustrations for the case of two forecasts, as analyzed in Bates and Granger (1969). We thus linearly combine two forecasts of an event μ:

$$y_c = wy_1 + (1-w)y_2. \tag{1}$$

If the weight w is considered to be fixed, then the forecast combination is unbiased ($\mathrm{E}\, y_c = \mu$) if the original forecasts are unbiased, and the variance of the combination will be

$$\operatorname{var}(y_c) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\rho\sigma_1\sigma_2, \tag{2}$$

where σ_1^2 and σ_2^2 are the variances of y_1 and y_2 respectively and $\rho = \operatorname{corr}(y_1, y_2)$ denotes the correlation.

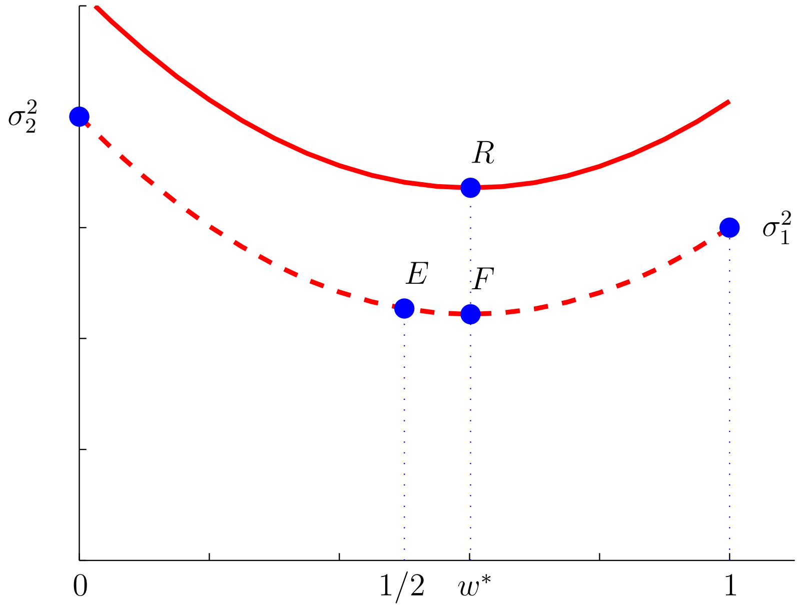


Figure 1: Variance of forecast combination, two dimensions: fixed weights (dashed) and random weights under normality (solid)

The variance is a quadratic function of w, as plotted in Figure 1 (dashed line). At $w = 0$ we obtain σ_2^2; at $w = 1$ we obtain σ_1^2; and at $w = 1/2$ we obtain point E. The optimum F is reached at $w = w^*$, the optimal weight giving the smallest variance of the forecast combination.

Now suppose that the weights are estimated, so that they are random rather than fixed. In the special case where (y_1, y_2, w) follows a trivariate normal distribution, the combination is biased (even when the original forecasts are unbiased), since

$$\mathrm{E}\, y_c = \mu + \operatorname{cov}(w, y_1 - y_2), \tag{3}$$

and the variance is given by

$$\begin{aligned}\operatorname{var}(y_c) &= (\operatorname{E} w)^2\sigma_1^2 + (1-\operatorname{E} w)^2\sigma_2^2 + 2(\operatorname{E} w)(1-\operatorname{E} w)\rho\sigma_1\sigma_2 \\ &\quad + \operatorname{var}(w)\operatorname{var}(y_1-y_2) + (\operatorname{cov}(w, y_1-y_2))^2. \qquad (4)\end{aligned}$$

In another special case where w is independent of (y_1, y_2), the combination is unbiased and

$$\begin{aligned}\operatorname{var}(y_c) &= (\operatorname{E} w)^2\sigma_1^2 + (1-\operatorname{E} w)^2\sigma_2^2 + 2(\operatorname{E} w)(1-\operatorname{E} w)\rho\sigma_1\sigma_2 \\ &\quad + \operatorname{var}(w)\operatorname{var}(y_1-y_2). \qquad (5)\end{aligned}$$

In either case the variance is shifted upwards, as shown in Figure 1 (solid line). The solid line gives the variance as a function of $\operatorname{E} w$ and the optimum is reached at the same point w^* as before, but leading to a higher variance of the forecast combination. We see that the equal-weights point at $w = 1/2$ (point E), though not optimal with fixed weights, has a variance which is smaller than the optimum with estimated weights (point R). This figure provides the essence of our answer to the forecast combination puzzle.

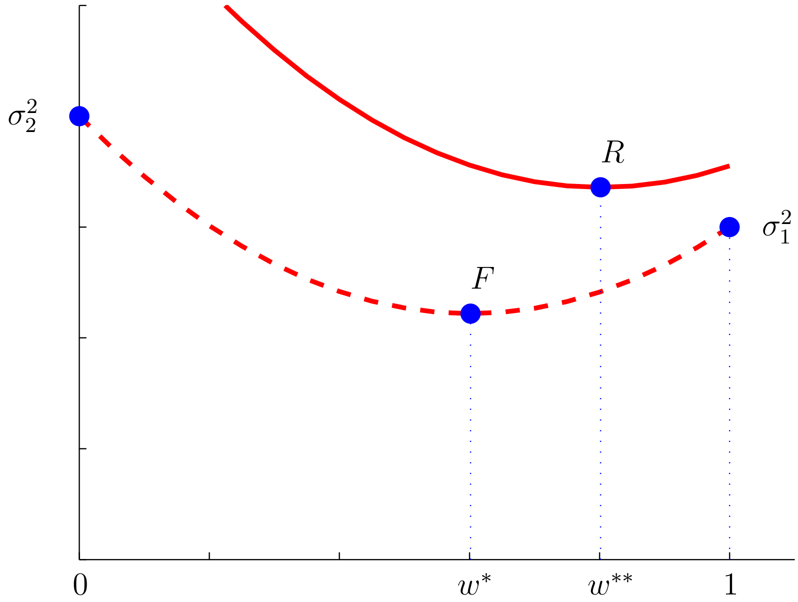


Figure 2: Variance of forecast combination, two dimensions: random weights, general case

The expressions in (4) and (5) concern special cases (normality and independence, respectively). In general, when the weights are estimated, the combined forecast will be biased, as given in (3), with its variance given by

$$\begin{aligned}\operatorname{var}(y_c) = (\mathrm{E}\,w)^2\sigma_1^2 + (1-\mathrm{E}\,w)^2\sigma_2^2 + 2(\mathrm{E}\,w)(1-\mathrm{E}\,w)\rho\sigma_1\sigma_2 \\ + \mathrm{E}\,[(w-\mathrm{E}\,w)(y_1-y_2)\,((\mathrm{E}\,w)y_1 + (1-\mathrm{E}\,w)y_2 - \mu)] \\ + \mathrm{E}[(w-\mathrm{E}\,w)^2(y_1-y_2)^2] - (\operatorname{cov}(w, y_1-y_2))^2. \qquad (6)\end{aligned}$$

Compared to (4) and (5) there are now additional terms that shift and distort the fixed-weights curve of Figure 1, and this is illustrated in Figure 2. The optimal weight is now given by w^{**} rather than by w^*. Note that if we would plot the mean squared error rather than the variance, the conclusions would not be affected. The three curves in Figures 1 and 2 provide the essence of this paper. The underlying formulae will be derived in m rather than in two dimensions, but the story remains the same.

The simplified setup presented above assumes that the event μ is nonrandom, and it also does not include a constant term w_0 in the combined forecast. Both assumptions can be criticized, so we briefly address them here. If μ is random, we define the forecast errors $e_1 = y_1 - \mu$ and $e_2 = y_2 - \mu$. Including a constant term in the combined forecast gives

$$y_c = w_0 + wy_1 + (1-w)y_2.$$

The forecast error of y_c is then

$$e_c = y_c - \mu = w_0 + we_1 + (1-w)e_2.$$

Assume that the forecasts are unbiased, so that $\mathrm{E}\,e_1 = \mathrm{E}\,e_2 = 0$. Then we find, in the case of fixed weights,

$$\mathrm{E}\,e_c = w_0, \qquad \operatorname{var}(e_c) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\rho\sigma_1\sigma_2,$$

as in (2), except that σ_1^2 and σ_2^2 now denote the variances of e_1 and e_2 and $\rho = \operatorname{corr}(e_1, e_2)$. The mean squared error of e_c is minimized for $w_0 = 0$ and $w = w^*$, so nothing changes. Now consider the case of random weights. Then,

$$\mathrm{E}\,e_c = \mathrm{E}\,w_0 + \operatorname{cov}(w, y_1 - y_2),$$

which vanishes for $\mathrm{E}\,w_0 = -\operatorname{cov}(w, y_1 - y_2)$. Including an intercept thus absorbs the bias. Regarding $\operatorname{var}(e_c)$, this will be an expression like (6), but more complicated because the variance and covariances involving w_0 need to be included. Since the essence of the story is not affected, we shall continue to assume that the event μ is nonrandom and that the combined forecast does not include a constant term. Only in our small Monte Carlo experiment in Section 6 we assume that μ is random. We shall however work in m rather than in two dimensions.

The remainder of this paper is organized as follows. In Section 2 we reiterate the classical forecast combination problem in a multivariate setting assuming that the weights are fixed. In Section 3 we analyze the properties of the forecast combination when the weights are random and the estimation is explicitly taken into account. Some special cases are considered in Section 4. Our explanation of the puzzle is summarized in Section 5. Section 6 provides a numerical illustration, and some concluding remarks are offered in Section 7.

2 Moments of the forecast combination: fixed weights

Thus motivated, let $y = (y_1, \dots, y_m)'$ be a vector of unbiased forecasts so that $\mathrm{E}\, y_j = \mu$ for all j, and let $w = (w_1, \dots, w_m)'$ be a vector of fixed (nonrandom) weights constrained by $\sum_j w_j = 1$. Assuming that y has a finite variance Σ_{yy}, we obtain the mean and variance of the forecast combination $y_c = w'y$ as

$$\mathrm{E}\, y_c = \mu, \qquad \operatorname{var}(y_c) = w'\Sigma_{yy}w. \tag{7}$$

It is easy to show that the variance is minimized (as a function of w, under the constraint $\sum_j w_j = 1$) when $w = w^*$, where

$$w^* = \frac{\Sigma_{yy}^{-1}\imath}{\imath'\Sigma_{yy}^{-1}\imath} \tag{8}$$

and $\imath$ denotes the vector of m ones. The optimal forecast is then $y_c^* = w^{*\prime}y$ and its variance is

$$\operatorname{var}(y_c^*) = \frac{1}{\imath'\Sigma_{yy}^{-1}\imath}. \tag{9}$$

These are well-established results; see Bates and Granger (1969) for the bivariate case and Elliott (2011) for its multivariate extension.

Denote the diagonal elements of Σ_{yy} by $\sigma_1^2, \dots, \sigma_m^2$. Then, for each j,

$$\operatorname{var}(y_c^*) \leq \sigma_j^2. \tag{10}$$

This follows by considering the vectors $a_j = \Sigma_{yy}^{1/2}e_j$ and $b = \Sigma_{yy}^{-1/2}\imath$, where e_j denotes the m-dimensional vector with one in its j-th position and zeros elsewhere. Then, by Cauchy-Schwarz,

$$1 = (e_j'\imath)^2 = (a_j'b)^2 \leq (a_j'a_j)(b'b) = (e_j'\Sigma_{yy}e_j)(\imath'\Sigma_{yy}^{-1}\imath) = \sigma_j^2 / \operatorname{var}(y_c^*).$$

Hence the optimally combined forecast has smaller variance than each of the individual forecasts. Equality can occur for at most one of the individual forecasts, because Σ_{yy} is assumed to remain positive definite. Equality for the j-th forecast occurs if and only if a_j and b are linearly dependent, that is, if and only if $\operatorname{cov}(y_i, y_j) = \operatorname{var}(y_j)$ for $i = 1, \dots, m$.

We note that we imposed the restriction that the weights add up to one, but not that each weight lies between zero and one. If all covariances are zero so that Σ_{yy} is diagonal, then the optimal weights are given by $(1/\sigma_j^2)/\sum_i(1/\sigma_i^2)$ $(j = 1, \dots, m)$, and these clearly lie between zero and one. But this holds only if Σ_{yy} is a diagonal matrix. Even in the case where only one covariance is not zero, say $\operatorname{cov}(y_i, y_j) = \operatorname{cov}(y_j, y_i) \neq 0$ for some i and j, the optimal weights w_i^* and w_j^* do not necessarily lie between zero and one; they do if and only if

$$\operatorname{corr}(y_i, y_j) < \frac{\min(\sigma_i, \sigma_j)}{\max(\sigma_i, \sigma_j)}.$$

Apparently, the combination of a high positive correlation with a high variation in reliability forces the optimal weights outside the $(0, 1)$ interval. Of course, it is possible to choose a positive definite matrix, say V, such that the components of $V^{-1}\imath$ are all positive, for example the diagonal matrix $V = \operatorname{diag}(\sigma_1^2, \dots, \sigma_m^2)$. An alternative set of weights can then be defined as

$$w^\dagger = \frac{V^{-1}\imath}{\imath' V^{-1}\imath}, \tag{11}$$

and these weights lie between zero and one, but they are — in general — not optimal. The forecast combination $y_c^\dagger = w^{\dagger\prime} y$ is still unbiased, but its variance is now

$$\operatorname{var}(y_c^\dagger) = \frac{\imath' V^{-1}\Sigma_{yy}V^{-1}\imath}{(\imath' V^{-1}\imath)^2}. \tag{12}$$

Letting $x = V^{-1/2}\imath$ and $P = V^{-1/2}\Sigma_{yy}V^{-1/2}$, we obtain

$$\frac{\operatorname{var}(y_c^\dagger)}{\operatorname{var}(y_c^*)} = \frac{x'Px}{x'x} \cdot \frac{x'P^{-1}x}{x'x}$$

and hence, by Kantorovich's inequality (Abadir and Magnus, 2005, Exercise 12.17),

$$1 \leq \frac{\operatorname{var}(y_c^\dagger)}{\operatorname{var}(y_c^*)} \leq \frac{(\lambda_1 + \lambda_m)^2}{4\lambda_1\lambda_m}, \tag{13}$$

where λ_1 and λ_m denote the largest and smallest eigenvalue of P, respectively. This provides an estimate of the possible loss of precision caused by choosing $w^\dagger$ instead of w^*. In the most common case where we choose $V = \operatorname{diag}(\sigma_1^2, \dots, \sigma_m^2)$, we note that P is the correlation matrix associated with Σ_{yy}. Although important, the issue of optimal weights outside the $(0,1)$ interval is not considered further in the current paper.

When weights are fixed, the optimal forecast combination y_c^* is an improvement over individual forecasts, because it remains unbiased and has smaller variance. In applications, however, the weights will typically be random and we now turn to this more realistic case.

3 Moments of the forecast combination: random weights

As in the previous section, let $y = (y_1, \dots, y_m)'$ be a vector of unbiased forecasts with $\mathrm{E}\, y_j = \mu$, and let $w = (w_1, \dots, w_m)'$ be a vector of weights constrained by $\sum_j w_j = 1$, but now random rather than fixed. Let $\Delta y_j = y_j - \mathrm{E}\, y_j$ and $\Delta y = (\Delta y_1, \dots, \Delta y_m)'$. Assuming that y and w are jointly distributed with finite fourth-order moments, and writing

$$\operatorname{var}\begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yw} \\ \Sigma_{wy} & \Sigma_{ww} \end{pmatrix},$$

we have

$$y_c = w'y = \mu + w'\Delta y,$$

and hence

$$\mathrm{E}\, y_c = \mu + \mathrm{E}(w'\Delta y) = \mu + \operatorname{tr} \Sigma_{wy},$$

so that y_c is in general a biased forecast. Also,

$$\operatorname{var}(y_c) = \operatorname{var}(w'\Delta y), \qquad \operatorname{MSE}(y_c) = \operatorname{var}(w'\Delta y) + (\operatorname{tr} \Sigma_{wy})^2.$$

This is not yet very informative. To gain more insight we let $\Delta w_j = w_j - \mathrm{E}\, w_j$ and $\Delta w = (\Delta w_1, \dots, \Delta w_m)'$. Then, $w = \mathrm{E}\, w + \Delta w$ and hence

$$w'\Delta y = (\mathrm{E}\, w)'(\Delta y) + (\Delta w)'(\Delta y),$$

so that

$$\operatorname{var}(w'\Delta y) = (\mathrm{E}\, w)'\Sigma_{yy}(\mathrm{E}\, w) + 2(\mathrm{E}\, w)'\, \mathrm{E}[(\Delta y)(\Delta y)'(\Delta w)] + \operatorname{var}[(\Delta w)'(\Delta y)].$$

This leads to the following proposition.

Proposition 3.1. *The mean, variance, and mean squared error of the forecast combination $y_c = w'y$ are given by*

$$\mathrm{E}\, y_c = \mu + \operatorname{tr} \Sigma_{wy},$$

$$\operatorname{var}(y_c) = (\mathrm{E}\, w)'\Sigma_{yy}(\mathrm{E}\, w) + 2(\mathrm{E}\, w)'d + \delta - (\operatorname{tr}\Sigma_{wy})^2,$$

and

$$\operatorname{MSE}(y_c) = (\mathrm{E}\, w)'\Sigma_{yy}(\mathrm{E}\, w) + 2(\mathrm{E}\, w)'d + \delta,$$

where the vector d and the scalar δ denote third- and fourth-order moments respectively, and are defined as

$$d = \mathrm{E}\left[(\Delta y)(\Delta y)'(\Delta w)\right], \qquad \delta = \mathrm{E}\left[(\Delta w)'(\Delta y)\right]^2.$$

We note the generality of this proposition. The only two things assumed (apart from the existence of moments) are that each individual forecast is unbiased and that the weights add up to one, and it is precisely the combination of these two assumptions that leads to the simplicity of the formulas. It is *not* assumed that the weights lie between zero and one. There is no problem in deriving the counterpart of Proposition 3.1 for biased forecasts, but the formulae become cumbersome and they are not needed for the story we wish to tell.

The distribution of the weights w is given by their location ($\mathrm{E}\, w$) and by their shape (moments of Δw). We can choose the location optimally by minimizing $\operatorname{MSE}(y_c)$ with respect to $\mathrm{E}\, w$ under the restriction that the weights add up to one, and this leads to $\mathrm{E}\, w = w^{**}$, where

$$w^{**} = \left(\frac{1 + \imath'\Sigma_{yy}^{-1}d}{\imath'\Sigma_{yy}^{-1}\imath}\right)\Sigma_{yy}^{-1}\imath - \Sigma_{yy}^{-1}d.$$

It is important to note that the 'optimal' weights w^* given in Equation (8) are no longer optimal in the random-weights case, unless $d = 0$ which occurs for example when $\Sigma_{ww} = 0$ (so that $\Delta w = 0$, the fixed-weights case) or if the joint distribution is not skewed (for example symmetric) so that third-order moments vanish. With $\mathrm{E}\, w$ chosen optimally as w^{**}, the variance of y_c is given by

$$\operatorname{var}(y_c) = \frac{1 + 2\imath'\Sigma_{yy}^{-1}d - [(\imath'\Sigma_{yy}^{-1}\imath)(d'\Sigma_{yy}^{-1}d) - (\imath'\Sigma_{yy}^{-1}d)^2]}{\imath'\Sigma_{yy}^{-1}\imath} + \delta - (\operatorname{tr}\Sigma_{wy})^2.$$

When weights are random rather than fixed the analysis and the conclusions are less straightforward. First, the forecast combination y_c will generally have a larger variance when weights are random, because of the additional randomness in the weights, but this is not always so. Second, it is no longer the case that the variance of y_c is necessarily smaller than the variance of each individual forecast, even when we choose the weights 'optimally', say $\mathrm{E}\, w = w^*$ or $\mathrm{E}\, w = w^{**}$. Some special cases will be instructive and highlight these differences.

4 Special cases

We consider three special cases.

No skewness. If the joint distribution of (y, w) is not skewed, then the mean and variance of the forecast combination $y_c = w'y$ are given by

$$\operatorname{E} y_c = \mu + \operatorname{tr} \Sigma_{wy}$$

and

$$\operatorname{var}(y_c) = (\operatorname{E} w)'\Sigma_{yy}(\operatorname{E} w) + \delta - (\operatorname{tr} \Sigma_{wy})^2.$$

No skewness occurs, for example, when the joint distribution is symmetric, whatever definition of multivariate symmetry one employs. If the joint distribution is not skewed then the third-order moments $d = \operatorname{E}[(\Delta y)(\Delta y)'(\Delta w)]$ all vanish, so that $w^* = w^{**}$ and hence

$$\operatorname{MSE}(y_c) = (\operatorname{E} w)'\Sigma_{yy}(\operatorname{E} w) + \delta$$

contains only two terms. In this case, the combined forecast does not necessarily have smaller variance than each individual forecast. The first term is smaller than the individual variance σ_j^2, see Equation (10), but $\delta = \operatorname{E}[(\Delta w)'(\Delta y)]^2$ is positive and, if it is large enough, then $\operatorname{MSE}(y_c) > \sigma_j^2$.

Normality. The variance of the weights Σ_{ww} plays a key role in the variance of the combination. This is why it may be good to select an estimator with small variation in weights even when this is not the optimal estimator. For example, the estimator based on $w^\dagger$ may be 'better' than the estimator based on w^*.

The effect of Σ_{ww} is well brought out in the case of joint normality. The mean and variance of the forecast combination $y_c = w'y$ are then given by

$$\operatorname{E} y_c = \mu + \operatorname{tr} \Sigma_{wy}$$

and

$$\operatorname{var}(y_c) = (\operatorname{E} w)'\Sigma_{yy}(\operatorname{E} w) + \operatorname{tr}(\Sigma_{ww}\Sigma_{yy}) + \operatorname{tr}(\Sigma_{wy}\Sigma_{yw}).$$

This follows from the fact that multivariate normality implies no skewness, so that $d = 0$, and also, using Anderson (1958, p. 39),

$$\begin{aligned}\delta_{ij} \equiv \operatorname{E}[(\Delta w_i)(\Delta y_i)(\Delta w_j)(\Delta y_j)] &= \operatorname{cov}(w_i, y_i)\operatorname{cov}(w_j, y_j) \\ &\quad + \operatorname{cov}(w_i, w_j)\operatorname{cov}(y_i, y_j) + \operatorname{cov}(w_i, y_j)\operatorname{cov}(y_i, w_j),\end{aligned}$$

so that

$$\delta = \sum_{ij} \delta_{ij} = (\operatorname{tr} \Sigma_{wy})^2 + \operatorname{tr}(\Sigma_{ww}\Sigma_{yy}) + \operatorname{tr}(\Sigma_{wy}\Sigma_{yw}).$$

The result then follows from Proposition 3.1.

Independence. One naturally expects the estimated weights w and the forecasts y to be correlated, because they are typically estimated from the same data set. In some cases, however, it may be possible to estimate the weights independently from the forecasts. When this happens, that is, when y and w are independent with finite second-order moments, then the forecast combination $y_c = w'y$ is unbiased,

$$\mathrm{E}\, y_c = \mu,$$

and its variance and mean squared error are given by

$$\mathrm{var}(y_c) = \mathrm{MSE}(y_c) = (\mathrm{E}\, w)'\Sigma_{yy}(\mathrm{E}\, w) + \mathrm{tr}(\Sigma_{ww}\Sigma_{yy}).$$

5 Discussion

In their 'simple explanation of the forecast puzzle' Smith and Wallis (2009) offer three main conclusions in terms of the mean squared error of the forecast (MSFE). We now analyze these conclusions in the context of the theory developed in Section 3. Their first conclusion is that

> '[...] a simple average of competing forecasts is expected to be more accurate, in terms of MSFE, than a combination based on estimated weights.'

This is the situation illustrated for two dimensions in Figures 1 and 2. The combination with equal weights is unbiased and its variance has only one component: $\imath'\Sigma_{yy}\imath/m^2$. In many situations this leads to a smaller mean squared error than a biased combination with additional components d and δ, as given in Proposition 3.1 for the case when the weights are estimated.

The second conclusion is that

> '[...] 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 generalization for many forecasts.'

Apart from the fact that including covariances may lead to negative weights, we have seen that estimating the covariances increases the variance of the weights, as also illustrated by Figures 2 and 4 in Smith and Wallis (2009). For fixed weights the relationship between the two variances (with and without covariances) is given by (13), but the additional terms from Proposition 3.1 are likely to be larger for the optimal weights based on estimated covariances. The special cases in Section 4 emphasize this point by showing explicitly how the variance of the weights, Σ_{ww}, appears in the formulae.

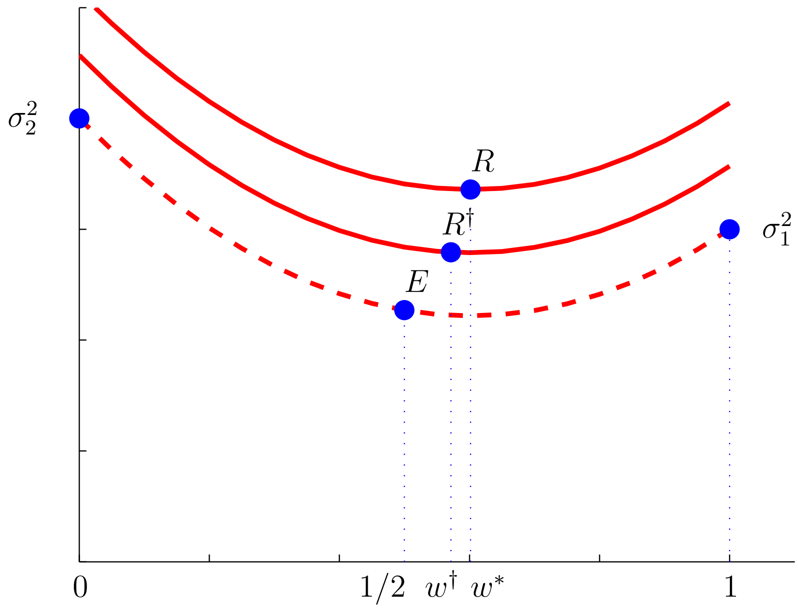


Figure 3: Variance of forecast combination, two dimensions: random weights under normality with and without covariances

Figure 3 provides a stylized illustration in two dimensions. The figure is identical to Figure 1, except that the middle curve has been added and the minimum point F on the lowest curve has been removed. It gives the variance of the forecast combination as a function of $\mathrm{E}\, w$. The bottom curve plots the variance when the weights are nonrandom; the point E on the curve (not the minimum) gives the variance when $w = 1/2$: equal weights. The top curve plots the variance according to Proposition 3.1 and the minimum of the curve is in R, representing the point where the optimal choice for $\mathrm{E}\, w$ is estimated. The middle curve represents the restricted case without covariances, where $\mathrm{E}\, w$ is an estimate of $\sigma_2^2/(\sigma_1^2+\sigma_2^2)$, as in (11). The minimum on the middle curve does not occur at $R^\dagger$, but because the three variance curves move parallel to each other and fewer parameters are required to estimate the variance in the middle curve than in the top curve, $R^\dagger$ is typically smaller than R.

The third conclusion of Smith and Wallis (2009) is:

> ‘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 (2003), [. . .] the forecast combination puzzle rests on a gain in MSFE that has no practical significance.'

This statement is based on simulations and empirical findings, but now it can be assessed in any situation by comparing the variance of the combination with equal weights, $\imath'\Sigma_{yy}\imath/m^2$, with the variance of the combination with estimated weight $w^\dagger$, given by the general formula in Proposition 3.1.

6 Numerical illustration

Our Figures 1–3 are stylized in order to provide a simple explanation of the puzzle. In actual applications the shift and distortion of the dashed curve in Figure 1 will vary in accordance with our theoretical results in (6) (for two dimensions) and Proposition 3.1 (for m dimensions). To support our theoretical results and better understand these shifts and distortions we now present a simple simulation study.

We closely follow the experimental design of Smith and Wallis (2009, Section 3.1), in particular their case 2. We draw a sequence of $T + 1$ observations from a strictly stationary AR(2) process

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \epsilon_t \qquad (t = 1, \dots, T + 1),$$

where the $\{\epsilon_t\}$ are independent and identically distributed standard-normal variates, and ϕ_1 and ϕ_2 are given parameters subject to the stationarity conditions $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$, and $|\phi_2| < 1$. The variance of the process is given by

$$\sigma_z^2 = \operatorname{var}(z_t) = \frac{1 - \phi_2}{(1 + \phi_2)[(1 - \phi_2)^2 - \phi_1^2]}$$

and the first two autocorrelation coefficients are

$$\rho_1 = \operatorname{corr}(z_t, z_{t-1}) = \frac{\phi_1}{1 - \phi_2}, \qquad \rho_2 = \operatorname{corr}(z_t, z_{t-2}) = \phi_1\rho_1 + \phi_2.$$

Our aim is to forecast the final observation z_{T+1}. Two forecasts are available,

$$y_1 = \rho_1 z_T, \qquad y_2 = \rho_2 z_{T-1},$$

and we are interested in the properties of various forecast combinations $y_c = wy_1 + (1 - w)y_2$ for different values of ϕ_1 and ϕ_2. We let $T = 30$ and use the thirty observations $(z_1, \dots, z_T)$ to estimate the weight w.

Since the forecast z_{T+1} is random rather than fixed, we define $e_{1t} = z_t - \rho_1 z_{t-1}$ and $e_{2t} = z_t - \rho_1 z_{t-2}$, and consider the forecast errors

$$e_1 = e_{1,T+1} = z_{T+1} - y_1, \qquad e_2 = e_{2,T+1} = z_{T+1} - y_2.$$

Their variances are

$$\sigma_1^2 = \mathrm{var}(e_1) = \sigma_z^2(1-\rho_1^2), \qquad \sigma_2^2 = \mathrm{var}(e_1) = \sigma_z^2(1-\rho_2^2),$$

and their correlation is given by $\rho = \mathrm{cov}(e_1, e_1)/(\sigma_1\sigma_2)$, where

$$\mathrm{cov}(e_1, e_2) = \sigma_z^2(1-\rho_2)(1-\rho_1^2+\rho_2).$$

Letting $\bar{e}_1 = (1/(T-2))\sum_{t=2}^{T-1} e_{1,t+1}$ and $\bar{e}_2 = (1/(T-2))\sum_{t=2}^{T-1} e_{2,t+1}$ we obtain unbiased estimates of the second-order moments as

$$\begin{pmatrix} \hat{\sigma}_1^2 & \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 \\ \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 \end{pmatrix} = \frac{1}{T-3}\sum_{t=2}^{T-1}\begin{pmatrix} (e_{1,t+1}-\bar{e}_1)^2 & (e_{1,t+1}-\bar{e}_1)(e_{2,t+1}-\bar{e}_2) \\ (e_{1,t+1}-\bar{e}_1)(e_{2,t+1}-\bar{e}_2) & (e_{2,t+1}-\bar{e}_2)^2 \end{pmatrix}.$$

Three weights are considered: the arithmetic mean $w = 1/2$, the estimated optimal weight

$$w^* = \frac{\hat{\sigma}_2^2 - \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2}$$

according to (8), and the estimated simplified weight (with $\rho = 0$)

$$w^\dagger = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

according to (11). When the weight is fixed at $w = 1/2$, we do not estimate it and simply compute its exact variance as

$$\mathrm{var}[(e_1+e_2)/2] = \frac{\sigma_z^2}{4}\left(4 - 3\rho_1^2 - 3\rho_2^2 + 2\rho_1^2\rho_2\right).$$

But when the weights are either w^* or $w^\dagger$ we estimate them, because our purpose is to better understand the uncertainty caused by the estimation of weights. For the same reason, we do *not* estimate the parameters ϕ_1 and ϕ_2; these are set to their true values. Any uncertainty shown in the simulations is therefore caused by weight estimation.

This experiment is repeated 1,000,000 times, which suffices to control the simulation error. For given ϕ_1 and ϕ_2, each run produces a value of w and of the two forecast errors $e_1 = z_{T+1} - y_1$ and $e_2 = z_{T+1} - y_2$. Since the forecast z_{T+1} is random rather than fixed, Equation (6) needs to be written in terms of e_1 and e_2

as discussed at the end of Section 1. The variance of the error $e_c = we_1 + (1-w)e_2$ of the combined forecast is

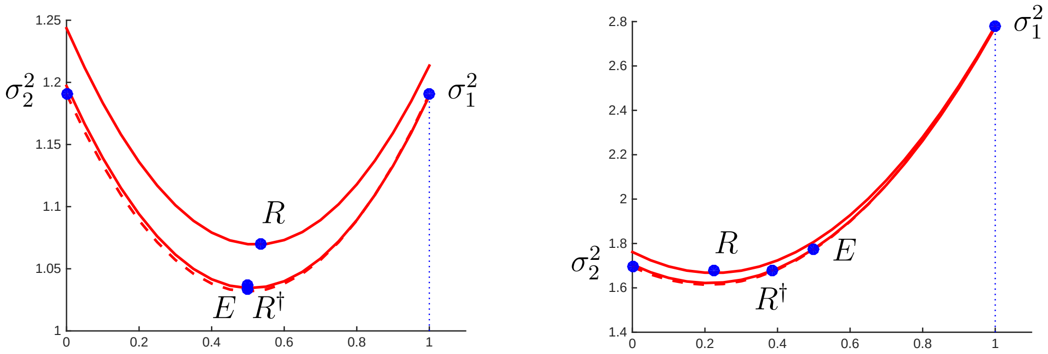
$$\begin{aligned}\operatorname{var}(e_c) = {} & \underbrace{(\mathrm{E}\, w)^2\sigma_1^2}_{\text{term 1}} + \underbrace{(1-\mathrm{E}\, w)^2\sigma_2^2}_{\text{term 2}} + \underbrace{2(\mathrm{E}\, w)(1-\mathrm{E}\, w)\rho\sigma_1\sigma_2}_{\text{term 3}} \\ & + \underbrace{\mathrm{E}\left[(w-\mathrm{E}\, w)(e_1-e_2)\left((\mathrm{E}\, w)e_1 + (1-\mathrm{E}\, w)e_2\right)\right]}_{\text{term 4}} \\ & + \underbrace{\mathrm{E}[(w-\mathrm{E}\, w)^2(e_1-e_2)^2]}_{\text{term 5}} - \underbrace{(\operatorname{cov}(w, e_1-e_2))^2}_{\text{term 6}}, \end{aligned} \tag{14}$$

which has six terms each of which can be calculated from the simulations. For both $w = w^*$ and $w = w^\dagger$ we compute $\operatorname{var}(e_c)$ and its six components for various values of ϕ_1 and ϕ_2.

The results are presented in Tables 1 and 2. In Table 1 we let $\phi_1 = \phi_2$ for values ranging between -0.9 and 0.4. In Table 2 we fix $\phi_1 = 0.5$ and let ϕ_2 range between -0.9 to 0.4. The first three terms of (14) are present whether or not randomness of w is taken into account. Terms 4 and 5 account for randomness of w caused by the estimation. Term 6 represents the squared bias which is negligible in all cases, so we omit this term in the tables.

Our results are in general agreement with those of Smith and Wallis (2009). In Table 1, where $\phi_1 = \phi_2$, the variance of e_c is larger for the estimated weight $w^\dagger$ than for the fixed weight $w = 1/2$, but not much. However, the variance of e_c is much larger (3–4%) for the estimated weight w^* than for the fixed weight. In Table 2, where $\phi_1 \neq \phi_2$, the variance of e_c is generally smaller, sometimes substantially smaller (up to about 15%), for the estimated weights $w^\dagger$ and w^* than for the fixed weight. This is because when the optimal weight deviates much from one-half, the gain from estimating the optimal weight is larger than the loss caused by estimation error; see Elliott (2011) for a detailed discussion of this issue. When $w = w^\dagger$ terms 4 and 5 are close to zero, but when $w = w^*$ term 5 (the fourth-order moments) can be substantial.

Figures are particularly informative as they show the relative position of the forecasts and the corresponding curves. We consider two special cases, one representing each table. Figure 4a shows the case where $\phi_1 = \phi_2 = 0.4$, while Figure 4b shows the case where $\phi_1 = 0.5$ and $\phi_2 = -0.8$. In Figure 4a we have $\phi_1 = \phi_2$ and hence $\rho_1 = \rho_2$ and $\operatorname{var}(y_1) = \operatorname{var}(y_2)$. This implies that σ_1^2 and σ_2^2 are close, so the expected values of the estimated $w^\dagger$ and w^* are both close to 1/2. Since the estimation of $w^\dagger$ hardly affects the properties of the forecast combination, the points E and $R^\dagger$ (and the corresponding curves) are almost identical. The estimation of w^*, on the other hand, increases the variance of the combination, so point R is much higher and its corresponding curve is shifted up by terms 4 and 5 reported in Table 1.



(a) $\phi_1 = \phi_2 = 0.4$

(b) $\phi_1 = 0.5$, $\phi_2 = -0.8$

Figure 4: Relative position of the forecasts and the corresponding curves. Point E represents the combination with equal weights, point $R^\dagger$ the combination with estimated $w^\dagger$, and point R the combination with estimated w^*. The original forecasts are labeled by their variances σ_1^2 and σ_2^2.

In Figure 4b the original forecasts y_1 and y_2 have different variances and hence σ_1^2 and σ_2^2 are not close. Again, the estimation of $w^\dagger$ hardly affects the corresponding curve, but point $R^\dagger$ slides along the curve yielding a smaller variance than the equal-weight combination E. The estimation of w^* distorts the corresponding curve, but the minimum point R offsets this distortion and produces a variance similar to point $R^\dagger$, as reported in Table 2.

7 Concluding remarks

In analyzing the properties of a combined forecast we have followed an integrated approach where the estimation of the weight is explicitly accounted for from the start. Weight estimation always affects the variance of the combination. In some situations the effect may be small, but in the case where the optimal weight is estimated the influence is substantial. This is our explanation of the forecast combination puzzle.

In this paper we have concentrated on the bias, variance, and mean squared error of the combined forecast. These are the moments that scientists are typically interested in. Other (functions of) moments can be similarly analyzed, for example the absolute percentage error, mean absolute deviation, or directional accuracy.

Acknowledgements

The authors are grateful the editor and two reviewers for their constructive comments, and to Rob J. Hyndman for stimulating conversations. Earlier versions of this paper were presented at the Econometric Theory Research Seminar and the SERG meeting in Sydney (November 2013); and at Curtin University in Perth, the University of Tasmania, AMES2014 in Taipei, the Chinese Academy of Sciences in Beijing, ISF2014 in Rotterdam, and ESEM2014 in Toulouse — all in 2014. We thank the participants for their positive comments. Claeskens acknowledges the support of the Fund for Scientific Research Flanders, KU Leuven Grant GOA/12/14, and of the IAP Research Network P7/06 of the Belgian Science Policy.

Table 1: Detailed numerical analysis of Equation (14) for fixed weight $w = 1/2$ and for estimated $w^\dagger$ and w^* when $\phi_1 = \phi_2 \in [-0.9, 0.4]$. Simulation error is the difference between $\mathrm{var}(e_c)$ and the sum of terms 1 to 5.

	$w = 1/2$	w is estimated by $w^\dagger$								
$\phi_1 = \phi_2$	$\mathrm{var}(e_c)$	$\mathrm{var}(e_c)$	$\mathrm{E}(w)$	$\mathrm{E}(e_c)$	Term 1	Term 2	Term 3	Term 4	Term 5	Simul Err
−0.9	4.1413	4.1914	0.5069	0.00009	1.3525	1.2796	1.5094	0.0199	0.0065	0.0236
−0.8	2.2840	2.3012	0.5051	0.00006	0.7088	0.6802	0.8950	0.0070	0.0042	0.0060
−0.7	1.6782	1.6856	0.5034	−0.00003	0.4968	0.4836	0.6978	0.0031	0.0027	0.0016
−0.6	1.3867	1.3905	0.5019	−0.00003	0.3936	0.3876	0.6055	0.0014	0.0016	0.0007
−0.5	1.2222	1.2235	0.5011	−0.00001	0.3348	0.3319	0.5556	0.0006	0.0009	−0.0002
−0.4	1.1224	1.1222	0.5005	0.00001	0.2982	0.2970	0.5272	0.0002	0.0004	−0.0008
−0.3	1.0609	1.0594	0.5002	0.00000	0.2749	0.2745	0.5114	0.0001	0.0002	−0.0017
−0.2	1.0243	1.0278	0.5001	−0.00001	0.2605	0.2604	0.5035	0.0000	0.0000	0.0034
−0.1	1.0055	1.0058	0.5001	0.00000	0.2526	0.2525	0.5005	0.0000	0.0000	0.0003
0.1	1.0045	1.0043	0.4999	0.00000	0.2524	0.2526	0.4994	0.0000	0.0000	−0.0002
0.2	1.0156	1.0179	0.4997	−0.00001	0.2601	0.2607	0.4948	0.0000	0.0001	0.0022
0.3	1.0283	1.0278	0.4997	−0.00005	0.2744	0.2751	0.4788	0.0000	0.0006	−0.0012
0.4	1.0317	1.0335	0.4996	0.00007	0.2971	0.2981	0.4365	0.0000	0.0028	−0.0010

	$w = 1/2$	w is estimated by w^*								
$\phi_1 = \phi_2$	$\mathrm{var}(e_c)$	$\mathrm{var}(e_c)$	$\mathrm{E}(w)$	$\mathrm{E}(e_c)$	Term 1	Term 2	Term 3	Term 4	Term 5	Simul Err
−0.9	4.1413	4.2911	0.5176	0.00017	1.4100	1.2248	1.5078	0.0502	0.0444	0.0540
−0.8	2.2840	2.3681	0.5152	0.00015	0.7374	0.6527	0.8942	0.0217	0.0414	0.0207
−0.7	1.6782	1.7405	0.5106	−0.00010	0.5111	0.4697	0.6975	0.0120	0.0397	0.0103
−0.6	1.3867	1.4387	0.5040	−0.00009	0.3969	0.3844	0.6054	0.0069	0.0389	0.0061
−0.5	1.2222	1.2676	0.4963	−0.00007	0.3285	0.3382	0.5555	0.0040	0.0382	0.0031
−0.4	1.1224	1.1632	0.4853	0.00002	0.2804	0.3153	0.5268	0.0019	0.0380	0.0008
−0.3	1.0609	1.0993	0.4688	0.00002	0.2415	0.3101	0.5094	0.0011	0.0378	−0.0006
−0.2	1.0243	1.0651	0.4384	−0.00023	0.2002	0.3285	0.4958	−0.0003	0.0378	0.0029
−0.1	1.0055	1.0431	0.3571	−0.00006	0.1288	0.4175	0.4596	−0.0002	0.0374	0.0000
0.1	1.0045	1.0415	0.6676	0.00010	0.4503	0.1116	0.4433	−0.0006	0.0378	−0.0007
0.2	1.0156	1.0549	0.5845	−0.00016	0.3559	0.1798	0.4807	−0.0008	0.0380	0.0011
0.3	1.0283	1.0649	0.5546	−0.00039	0.3381	0.2180	0.4731	−0.0007	0.0384	−0.0023
0.4	1.0317	1.0674	0.5358	0.00026	0.3418	0.2565	0.4343	−0.0011	0.0381	−0.0019

Table 2: Detailed numerical analysis of Equation (14) for fixed weight $w = 1/2$ and for estimated $w^\dagger$ and w^* when $\phi_1 = 0.5$ and $\phi_2 \in [-0.9, 0.4]$. Simulation error is the difference between $\mathrm{var}(e_c)$ and the sum of terms 1 to 5.

	$w = 1/2$	w is estimated by $w^\dagger$								
ϕ_2	$\mathrm{var}(e_c)$	$\mathrm{var}(e_c)$	$\mathrm{E}(w)$	$\mathrm{E}(e_c)$	Term 1	Term 2	Term 3	Term 4	Term 5	Simul Err
-0.9	2.7064	2.3201	0.3235	-0.00017	0.5508	1.0599	0.7105	-0.0099	0.0169	-0.0082
-0.8	1.7724	1.6749	0.3857	0.00000	0.4132	0.6395	0.6201	-0.0026	0.0074	-0.0027
-0.7	1.4637	1.4408	0.4309	0.00005	0.3640	0.4827	0.5894	-0.0003	0.0035	0.0015
-0.6	1.3115	1.3095	0.4656	-0.00001	0.3387	0.3972	0.5705	0.0004	0.0017	0.0011
-0.5	1.2222	1.2243	0.4917	0.00001	0.3224	0.3444	0.5554	0.0005	0.0009	0.0007
-0.4	1.1645	1.1650	0.5117	-0.00002	0.3117	0.3094	0.5422	0.0004	0.0005	0.0007
-0.3	1.1251	1.1199	0.5264	0.00002	0.3045	0.2860	0.5302	0.0004	0.0003	-0.0014
-0.2	1.0972	1.0903	0.5366	0.00001	0.2999	0.2707	0.5189	0.0003	0.0002	0.0003
-0.1	1.0771	1.0676	0.5428	0.00000	0.2976	0.2618	0.5077	0.0003	0.0003	-0.0001
0.1	1.0516	1.0407	0.5444	0.00003	0.2994	0.2599	0.4821	0.0004	0.0006	-0.0018
0.2	1.0430	1.0378	0.5398	-0.00002	0.3035	0.2670	0.4645	0.0005	0.0011	0.0011
0.3	1.0345	1.0294	0.5311	0.00000	0.3099	0.2803	0.4394	0.0004	0.0022	-0.0030
0.4	1.0228	1.0276	0.5177	-0.00011	0.3190	0.3019	0.4003	0.0004	0.0045	0.0014

	$w = 1/2$	w is estimated by w^*								
ϕ_2	$\mathrm{var}(e_c)$	$\mathrm{var}(e_c)$	$\mathrm{E}(w)$	$\mathrm{E}(e_c)$	Term 1	Term 2	Term 3	Term 4	Term 5	Simul Err
-0.9	2.7064	2.2844	0.1868	-0.00025	0.1837	1.5314	0.4932	0.0110	0.0535	0.0114
-0.8	1.7724	1.6724	0.2252	-0.00007	0.1408	1.0173	0.4566	0.0055	0.0467	0.0055
-0.7	1.4637	1.4636	0.2733	0.00008	0.1465	0.7869	0.4774	0.0038	0.0434	0.0057
-0.6	1.3115	1.3494	0.3421	-0.00004	0.1829	0.6019	0.5161	0.0032	0.0415	0.0039
-0.5	1.2222	1.2692	0.4380	0.00002	0.2558	0.4211	0.5470	0.0033	0.0386	0.0035
-0.4	1.1645	1.2016	0.5706	-0.00016	0.3877	0.2392	0.5317	0.0039	0.0350	0.0040
-0.3	1.1251	1.1373	0.7314	0.00016	0.5879	0.0919	0.4178	0.0052	0.0311	0.0036
-0.2	1.0972	1.0841	0.8794	0.00014	0.8055	0.0183	0.2213	0.0047	0.0298	0.0046
-0.1	1.0771	1.0488	0.9520	0.00002	0.9154	0.0029	0.0935	0.0024	0.0321	0.0024
0.1	1.0516	1.0405	0.8437	0.00028	0.7191	0.0306	0.2563	-0.0010	0.0386	-0.0028
0.2	1.0430	1.0527	0.7420	-0.00012	0.5735	0.0839	0.3580	-0.0013	0.0390	-0.0005
0.3	1.0345	1.0542	0.6500	-0.00002	0.4642	0.1562	0.4015	-0.0014	0.0383	-0.0046
0.4	1.0228	1.0570	0.5764	-0.00032	0.3956	0.2328	0.3914	-0.0009	0.0379	-0.0002

References

Abadir, K. M. & Magnus, J. R. (2005). Matrix Algebra, New York: Cambridge University Press.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis, New York: John Wiley & Sons.

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

Elliott, G. (2011). Averaging and the optimal combination of forecasts. UCSD working paper, econweb.ucsd.edu/∼grelliott/AveragingOptimal.pdf.

Graefe, A., Armstrong, J. S., Jones Jr., R. J. & Cuzán, A. G. (2014). Combining forecasts: An application to elections. International Journal of Forecasting, 30, 43–54.

Hansen, B. E. (2008). Least-squares forecast averaging. Journal of Econometrics, 146, 342–350.

Hsiao, C. & Wan, S. K. (2014). Is there an optimal forecast combination? Journal of Econometrics, 178, 294–309.

Liang, H., Zou, G., Wan, A. T. K. & Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association, 106, 1053–1066.

Magnus, J. R. & De Luca, G. (2014). Weighted-average least squares (WALS): A review. Journal of Economic Surveys, doi:10.1111/joes.12094.

Smith, J. & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics, 71, 331–355.

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

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be