TI 2014-127/III
Tinbergen Institute Discussion Paper

The Forecast Combination Puzzle: A Simple Theoretical Explanation

Gerda Claeskens[1]
Jan Magnus[2]
Andrey Vasnev[3]
Wendun Wang[4]

[1] KU Leuven, Belgium;
[2] Faculty of Economics and Business Administration, VU University Amsterdam, and Tinbergen Institute, the Netherlands;
[3] University of Sydney, Australia;
[4] Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute, the Netherlands.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

The forecast combination puzzle: A simple theoretical explanation

September 26, 2014

Gerda Claeskens
KU Leuven, Belgium

Jan R. Magnus
VU University, Amsterdam and Tinbergen Institute, The Netherlands

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

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

Abstract:
This papers 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 an application in the context of model selection.

Key words: forecast combination, optimal weights, model selection

JEL Codes: C53, C52

Corresponding author:
Andrey L. Vasnev
The University of Sydney Business School Merewether Building (H04)
Sydney, NSW 2006
Australia
E-mail: andrey.vasnev@sydney.edu.au

1 Introduction

When several forecasts of the same event are available, it is natural to try and find a (linear) combination of the forecasts which is 'best' in some sense. Empirical evidence and extensive simulations show that a simple average often performs best — better than a theoretically derived optimum. This finding is known as the 'forecast combination puzzle'. Its history is elegantly summarized in Section 4 of Graefe et al. (2014). A particularly rigorous attempt to explain the puzzle, using simulations and an empirical example, was undertaken by Smith and Wallis (2009) who showed that the reason lies in the estimation error.

One important fact has, however, been overlooked in all (or almost all) previous research, namely the fact that the optimal weight derivation and its estimation are separated. This is the case in Bates and Granger (1969) and it remains the case in later contributions, important and insightful as they may be, such as Hansen (2008), Elliott (2011), Liang et al. (2011), and Hsiao and Wan (2014). This separation is quite common in econometrics, although its dangers 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).

To better understand the problem, let us consider the well-known situation of feasible generalized least squares. We have a linear regression model under the simplest assumptions, except that the errors are not white noise but follow a first-order autoregressive process with parameter r. We typically estimate r, say by $\hat{r}$, and then assume that r is fixed at $\hat{r}$, thus ignoring the added noise caused by the estimation of r. The resulting generalized least squares estimator is then assumed to be normally distributed, even though this is clearly not the case. But asymptotically all is well, assuming that $\hat{r}$ is a consistent estimator of r.

In the forecast puzzle the role of r is played by the weights w, but now there is no easy 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 acknowledge explicitly that optimal weights should be derived by taking the estimation step into account. In order to highlight our main findings we 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 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 and $\rho = \operatorname{corr}(y_1, y_2)$ is the correlation.

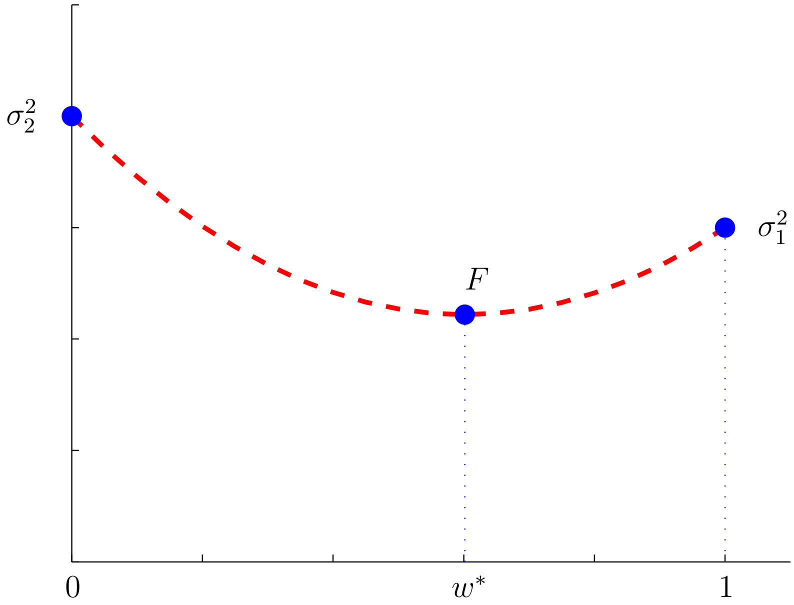

Figure 1: Variance of forecast combination, two dimensions: fixed weights

The variance is a quadratic function of w, as plotted in Figure 1. If $w = 0$ we obtain σ_2^2; if $w = 1$ we obtain σ_1^2. The optimum F is reached when $w = w^*$, 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

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

In another special case where w is independent of (y_1, y_2), the combination is unbiased ($\mathrm{E}\, y_c = \mu$) and

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

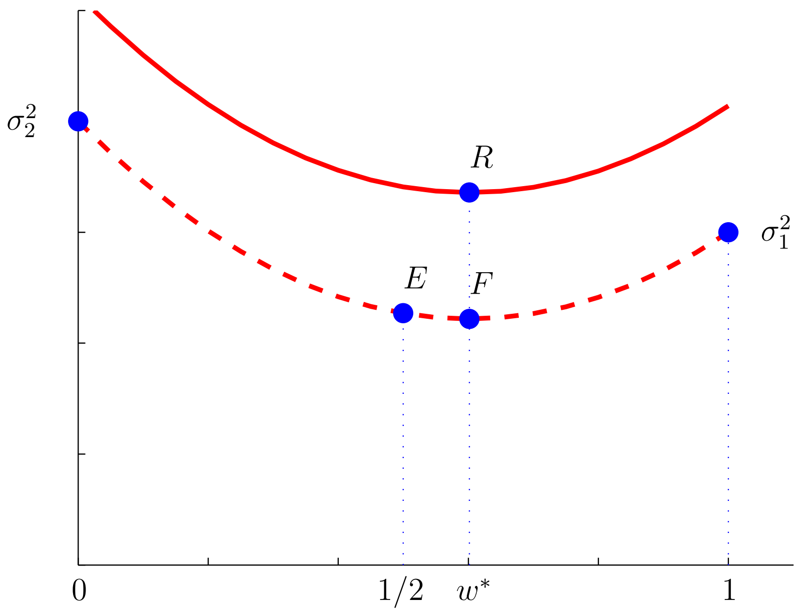

Figure 2: Variance of forecast combination, two dimensions: random weights under normality

In either case the variance is shifted upwards, as shown in Figure 2. The dashed curve is the same as in Figure 1, showing not only the optimum F but also the equal-weights point E. The solid line gives the variance as a function of $\mathrm{E}\, w$ and the optimum is reached at the same point w^* as before, but leading to a higher variance of the forecast combination. For comparison we add the equal-weights point $w = 1/2$ (point E), which is not optimal but its variance is smaller than in the estimated-weights case (point R). This figure provides the essence of our answer to the forecast combination puzzle.

In general, when the weights are estimated, the combined forecast will be biased, as in (3), and its variance is now given by

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

There are additional terms now that shift and distort the fixed-weights curve of Figure 1, and this is illustrated in Figure 3. 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 story would be the same.

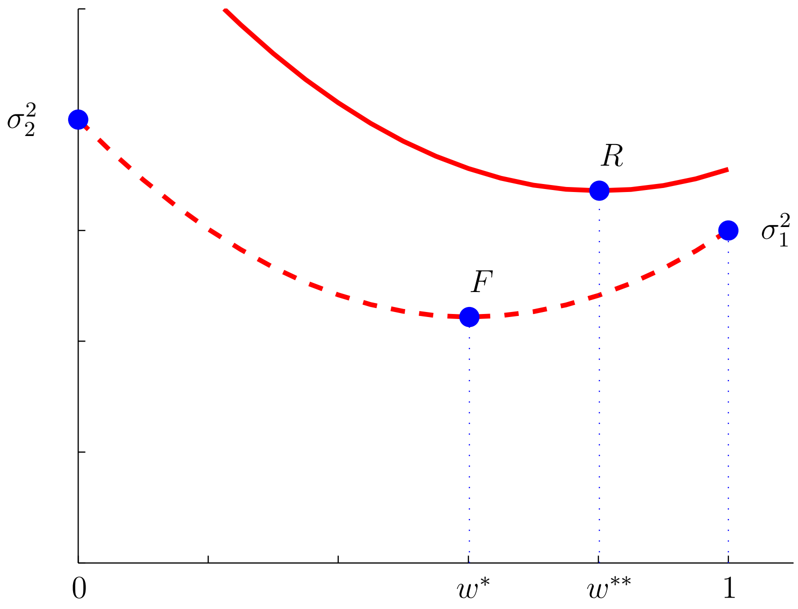

Figure 3: Variance of forecast combination, two dimensions: random weights, general case

These three graphs provide the essence of the paper. The underlying formulae will be derived in m rather than in two dimensions, but the story remains the same. 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. Section 5 provides the connection between forecast combination and forecast/model selection. Our explanation of the puzzle is summarized in Section 6 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

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

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,

$$1 = (e_j'\imath)^2 = (a_j'b)^2 \leq (a_j'a_j)(b'b) = (e_j'\Sigma_{yy}e_j)(\imath'\Sigma_{yy}^{-1}\imath) = \sigma_j^2/\operatorname{var}(y_c^*). \tag{11}$$

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

$$\operatorname{corr}(y_i, y_j) < \frac{\min(\sigma_i, \sigma_j)}{\max(\sigma_i, \sigma_j)}. \tag{12}$$

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

$$w^\dagger = \frac{V^{-1}\imath}{\imath' V^{-1}\imath}, \tag{13}$$

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

$$\operatorname{var}(y_c^\dagger) = \frac{\imath' V^{-1}\Sigma_{yy}V^{-1}\imath}{(\imath' V^{-1}\imath)^2}. \tag{14}$$

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

$$\frac{\operatorname{var}(y_c^\dagger)}{\operatorname{var}(y_c^*)} = \frac{x'Px}{x'x} \cdot \frac{x'P^{-1}x}{x'x} \tag{15}$$

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

$$1 \leq \frac{\operatorname{var}(y_c^\dagger)}{\operatorname{var}(y_c^*)} \leq \frac{(\lambda_1 + \lambda_m)^2}{4\lambda_1\lambda_m}, \tag{16}$$

where λ_1 and λ_m denote the largest and smallest eigenvalue of P, respectively. This provides an estimate of the 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 $\operatorname{E} y_j = \mu$, 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 $\Delta y_j = y_j - \operatorname{E} y_j$ 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

$$\operatorname{var}\begin{pmatrix} y \\ w \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yw} \\ \Sigma_{wy} & \Sigma_{ww} \end{pmatrix}, \tag{17}$$

we have

$$y_c = w'y = \mu + w'\Delta y, \tag{18}$$

and hence

$$\operatorname{E} y_c = \mu + \operatorname{E}(w'\Delta y) = \mu + \operatorname{tr}\Sigma_{wy}, \tag{19}$$

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

$$\operatorname{var}(y_c) = \operatorname{var}(w'\Delta y), \qquad \operatorname{MSE}(y_c) = \operatorname{var}(w'\Delta y) + (\operatorname{tr}\Sigma_{wy})^2. \tag{20}$$

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

$$w'\Delta y = (\operatorname{E} w)'(\Delta y) + (\Delta w)'(\Delta y), \tag{21}$$

so that

$$\operatorname{var}(w'\Delta y) = (\operatorname{E} w)'\Sigma_{yy}(\operatorname{E} w) + 2(\operatorname{E} w)'\operatorname{E}[(\Delta y)(\Delta y)'(\Delta w)] + \operatorname{var}[(\Delta w)'(\Delta y)]. \tag{22}$$

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*

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

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

and

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

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

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

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. In Section 5 (Proposition 5.1) we do discuss the case of biased forecasts, but then in terms of conditional expectations.

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 $\mathrm{MSE}(y_c)$ 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

$$w^{**} = \left(\frac{1 + \imath' \Sigma_{yy}^{-1} d}{\imath' \Sigma_{yy}^{-1} \imath} \right) \Sigma_{yy}^{-1} \imath - \Sigma_{yy}^{-1} d. \tag{23}$$

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 under joint symmetry (so that third-order moments vanish). With $\mathrm{E}\, w$ chosen optimally as w^{**}, the variance of y_c is given by

$$\mathrm{var}(y_c) = \frac{1 + 2\imath' \Sigma_{yy}^{-1} d - [(\imath' \Sigma_{yy}^{-1} \imath)(d' \Sigma_{yy}^{-1} d) - (\imath' \Sigma_{yy}^{-1} d)^2]}{\imath' \Sigma_{yy}^{-1} \imath} + \delta - (\mathrm{tr}\, \Sigma_{wy})^2. \tag{24}$$

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.

Symmetry. If the joint distribution of (y, w) is symmetric, then the mean and variance of the forecast combination $y_c = w'y$ are given by

$$\mathrm{E}\, y_c = \mu + \mathrm{tr}\, \Sigma_{wy} \tag{25}$$

and

$$\mathrm{var}(y_c) = (\mathrm{E}\, w)' \Sigma_{yy} (\mathrm{E}\, w) + \delta - (\mathrm{tr}\, \Sigma_{wy})^2. \tag{26}$$

This follows from the fact that the third-order moments $d = \mathrm{E}\, [(\Delta y)(\Delta y)'(\Delta w)]$ vanish under symmetry, so that $w^* = w^{**}$ and

$$\mathrm{MSE}(y_c) = (\mathrm{E}\, w)' \Sigma_{yy} (\mathrm{E}\, w) + \delta \tag{27}$$

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 the fourth-order term δ 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

$$\operatorname{E} y_c = \mu + \operatorname{tr} \Sigma_{wy} \tag{28}$$

and

$$\operatorname{var}(y_c) = (\operatorname{E} w)'\Sigma_{yy}(\operatorname{E} w) + \operatorname{tr}(\Sigma_{ww}\Sigma_{yy}) + \operatorname{tr}(\Sigma_{wy}\Sigma_{yw}). \tag{29}$$

This follows from the fact that multivariate normality implies symmetry, so that $d = 0$, and also, using Anderson (1958, p. 39),

$$\begin{aligned}\delta_{ij} \equiv \operatorname{E}[(\Delta w_i)(\Delta y_i)(\Delta w_j)(\Delta y_j)] &= \operatorname{cov}(w_i, y_i)\operatorname{cov}(w_j, y_j) \\ &\quad + \operatorname{cov}(w_i, w_j)\operatorname{cov}(y_i, y_j) + \operatorname{cov}(w_i, y_j)\operatorname{cov}(y_i, w_j),\end{aligned} \tag{30}$$

so that

$$\delta = \sum_{ij} \delta_{ij} = (\operatorname{tr} \Sigma_{wy})^2 + \operatorname{tr}(\Sigma_{ww}\Sigma_{yy}) + \operatorname{tr}(\Sigma_{wy}\Sigma_{yw}). \tag{31}$$

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,

$$\operatorname{E} y_c = \mu, \tag{32}$$

and its variance and mean squared error are given by

$$\operatorname{var}(y_c) = \operatorname{MSE}(y_c) = (\operatorname{E} w)'\Sigma_{yy}(\operatorname{E} w) + \operatorname{tr}(\Sigma_{ww}\Sigma_{yy}). \tag{33}$$

5 Model-selection weights

The random-weights framework provides a natural connection between forecast combination and model selection. In this case we consider m models, each model leading to a forecast. We then select *one* of the m models — the one we like best, based on some criterion. Next we take the forecast of the selected model, ignoring all other forecasts.

This procedure can be interpreted in terms of forecast combinations, be it of a rather special nature: all weights are zero except one, which is one. Of course, the weights are random, not fixed, and are defined via a zero-one indicator that takes the value one for the forecast of the selected model and zero for all other forecasts. This case is of particular interest to study post-selection forecasts.

It is well-known that ignoring the uncertainty involved with the selection procedure leads to incorrect inference (Kabaila, 1995; Pötscher, 1991; Hjort and Claeskens, 2003; Danilov and Magnus, 2004). For our purpose, the precise type of model selection is irrelevant, and our results hold for any model-selection criterion, for example AIC (Akaike, 1973), BIC (Schwarz, 1978), or Mallows' C_p (Mallows, 1973). The proposition below also covers 'smooth' weights: values proportional to the value of an information criterion assigned to the model (Burnham and Anderson, 2002), but we shall not consider smooth weights explicitly.

Before we can state our result for model-selection based forecasting, we have to extend Proposition 3.1 to the case of biased forecasts, because model selection typically occurs within a framework where all (or almost all) forecasts will be biased, unless all models are overspecified. Thus we define $\mathrm{E}\, y_j = \mu + \theta_j$ so that

$$\mathrm{E}\, y = \mu\imath + \theta \tag{34}$$

and

$$\Delta y = y - \mathrm{E}\, y = y - \mu\imath - \theta. \tag{35}$$

The forecast combination is then given by

$$y_c = w'y = \mu + w'\Delta y + w'\theta. \tag{36}$$

We could extend Proposition 3.1 straightforwardly by separating $\mathrm{E}\, w$ and Δw as before, but the formulae become cumbersome. Instead, it will prove useful to define the following two conditional expectations:

$$a_w = \mathrm{E}(\Delta y \mid w), \qquad A_w = \mathrm{E}((\Delta y)(\Delta y)' \mid w). \tag{37}$$

Given these definitions, we obtain the following result.

Proposition 5.1. *When the forecasts are biased with bias $\theta_j = \mathrm{E}\, y_j - \mu$, the mean, variance, and mean squared error of the forecast combination $y_c = w'y$ can be expressed as*

$$\mathrm{E}\, y_c = \mu + \mathrm{E}(a_w' w) + (\mathrm{E}\, w)'\theta,$$

$$\mathrm{var}(y_c) = \mathrm{E}(w' A_w w) - (\mathrm{E}\, a_w' w)^2 + 2\left[\mathrm{E}(a_w' w w') - (\mathrm{E}\, a_w' w)(\mathrm{E}\, w)'\right]\theta + \theta' \Sigma_{ww} \theta,$$

and

$$\mathrm{MSE}(y_c) = \mathrm{E}(w' A_w w) + 2\,\mathrm{E}(a_w' w w')\theta + \theta' \left[\Sigma_{ww} + (\mathrm{E}\, w)(\mathrm{E}\, w)'\right]\theta.$$

In the special case where all forecasts are unbiased, we have $\theta = 0$ and we obtain $\mathrm{E}\, y_c = \mu + \mathrm{E}(a_w' w)$ and

$$\mathrm{var}(y_c) = \mathrm{E}(w' A_w w) - (\mathrm{E}\, a_w' w)^2, \quad \mathrm{MSE}(y_c) = \mathrm{E}(w' A_w w). \tag{38}$$

This is the counterpart to Proposition 3.1, dealing with the same situation but expressed differently.

Proposition 5.1 can be applied directly to the case of model selection. In model selection we have

$$\Pr(w = e_j) = p_j \quad (j = 1, \dots, m), \qquad \sum_j p_j = 1, \tag{39}$$

where we recall that e_j denotes the m-dimensional vector with one in its j-th position and zeros elsewhere. The first two moments of the weights w are then given by

$$\mathrm{E}\, w = p, \qquad \mathrm{E}\, ww' = \mathrm{diag}(p), \qquad \Sigma_{ww} = \mathrm{diag}(p) - pp', \tag{40}$$

and the first two moments of the selected forecast are as follows.

Proposition 5.2. *When the weights are chosen through model selection, so that $\Pr(w = e_j) = p_j$ $(j = 1, \dots, m)$, the mean, variance, and mean squared error of the selected forecast y_c are given by*

$$\mathrm{E}\, y_c = \mu + \bar\theta_1, \qquad \mathrm{var}(y_c) = \bar v + \bar\theta_2 - \bar\theta_1^2, \qquad \mathrm{MSE}(y_c) = \bar v + \bar\theta_2,$$

where

$$\bar\theta_1 = \sum_j p_j(\theta_j + \eta_j), \qquad \bar\theta_2 = \sum_j p_j(\theta_j + \eta_j)^2, \qquad \bar v = \sum_j p_j(v_j - \eta_j^2),$$

and

$$\eta_j = \mathrm{E}(\Delta y_j \mid w = e_j), \qquad v_j = \mathrm{E}((\Delta y_j)^2 \mid w = e_j).$$

We see this immediately by letting

$$a_j = \mathrm{E}(\Delta y \mid w = e_j), \qquad A_j = \mathrm{E}((\Delta y)(\Delta y)' \mid w = e_j). \tag{41}$$

Then, in the notation of Proposition 5.1,

$$\mathrm{E}(a_w' w) = \sum_j p_j (a_j' e_j) = \sum_j p_j \eta_j, \tag{42}$$

$$\mathrm{E}(w' A_w w) = \sum_j p_j (e_j' A_j e_j) = \sum_j p_j v_j, \tag{43}$$

and

$$\mathrm{E}(a_w' w w') = \sum_j p_j (a_j' e_j e_j') = \sum_j (p_j \eta_j) e_j', \tag{44}$$

and the result follows from Proposition 5.1.

We note that $\theta_j + \eta_j$ is the conditional bias of the j-th model's forecast conditional on the j-th model, and that $v_j - \eta_j^2$ is the conditional variance of the j-th model's forecast conditional on the j-th model. The bias $\bar{\theta}_1$ is thus written as the weighted sum of the conditional biases in each of the models, each time conditioning on the specific model used. The variance is more complicated, because the unconditional variance is not simply the sum of the weighted conditional variances; hence we express the variance as MSE $-(\text{bias})^2$. But the mean squared error can be interpreted as the sum of the weighted conditional MSE-values of the j-th model's forecast conditional on the j-th model. Even though only a single model is selected, in the mean squared error all models' conditional MSE-values are combined. This highlights the important (but often ignored) fact that conditional inference, thus ignoring model selection, is incorrect and can be harmful.

6 Explanation of the puzzle

In their 'simple explanation of the forecast puzzle' Smith and Wallis (2009) offer three main conclusions in terms of 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 1–3 of Section 1. 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 (16), 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.

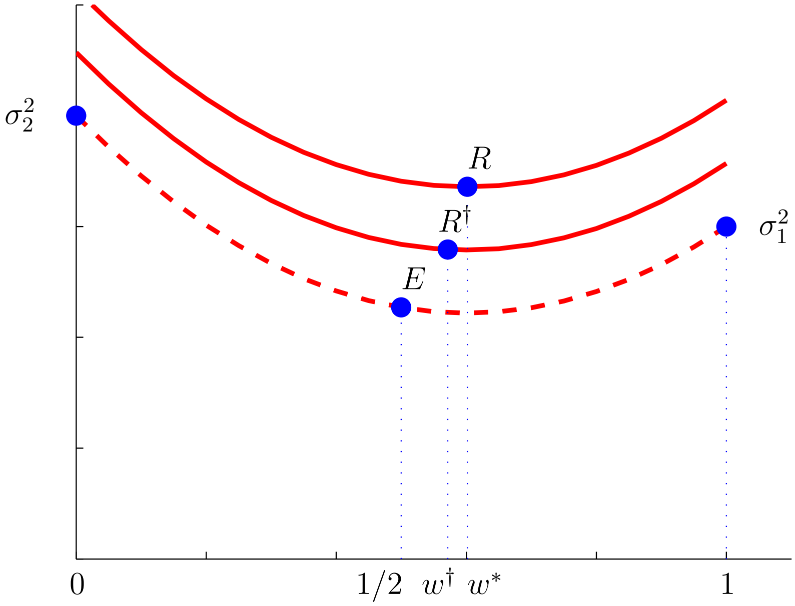

Figure 4: Variance of forecast combination, two dimensions: random weights under normality with and without covariances

Figure 4 provides a stylized illustration in two dimensions. The figure is identical to Figure 2, 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 (13). 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.

7 Concluding remarks

In analyzing the properties of a combined forecast, this paper follows an integrated approach where the estimation of the weight is explicitly accounted for from the start. Weight estimation always increases the variance of the combination. In some situations this increase may be small, but in the case where the optimal weight is estimated the increase is substantial and this explains the forecast combination puzzle. The special case of model selection is naturally accommodated in the integrated approach.

We analyzed the bias, variance, and mean squared error of the combined forecast, but other functions of the moments can be similarly analyzed, for example the absolute percentage error, mean absolute deviation, or directional accuracy.

Acknowledgements

The authors are grateful 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.

References

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csáki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiadó, Budapest.

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

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

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer Verlag.

Danilov, D. and J. R. Magnus (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics 122*, 27–46.

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

Graefe, A., J. S. Armstrong, R. J. Jones Jr., and A. G. Cuzáne (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.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–899.

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

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory 11*, 537–549.

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

Magnus, J. R. and G. De Luca (2014). Weighted-average least squares: A review. *Journal of Economic Surveys*, to appear.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics 15*, 661–675.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory 7*, 163–185.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

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

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