Is it optimal to combine forecast with a simple average?

F. Chan [a] **and L. Pauwels** [b]

[a]*School of Economics and Finance, Curtin University, Australia*
[b]*Business Analytics, University of Sydney, Australia*
Email: L.Pauwels@econ.usyd.edu.au

Abstract: This paper proposes a unified framework to study the theoretical properties of forecast combination. By setting up the forecast combination problem as a panel data model, the paper obtains the necessary and sufficient conditions for optimal weight as well as the necessary and sufficient conditions for the simple average to be the optimal weight under Mean Squared Forecast Errors (MSFE). These conditions are consistent with existing results in the literature but the derivations are much simpler due to the proposed framework. In addition to existing results, this paper also establishes two useful theoretical results. First, it derives the necessary and sufficient conditions for a single model to outperform simple average of forecasts. As argued in the paper, it is unlikely that any individual model would satisfy these conditions in practice and therefore, it explains the empirical observation that simple average of forecasts often outperforms any single model. More importantly, it provided a theoretical explanation on the superiority of forecast combinations, at least in the MSFE sense. Second, the paper also shows that the MSFE of simple average of forecast decreases as the number of model increases. This implies that a single model is unlikely to be superior over simple average of forecasts if the number of models increases in the combination.

This paper shows that the proposed framework is also useful in studying the forecast combination puzzle. The paper verifies the existing view that the puzzle may be a result of estimation error in the optimal weight but more importantly, it identifies an additional cause of the puzzle. Specifically, the MSFE may be an inconsistent estimator of the forecast variance and thus, it may produce inconsistent results on the forecast performance of different models with different weighting schemes. A series of Monte Carlo experiments provided overwhelming support of this explanation. An important implication of this results is that selecting optimal model based on naïve comparison of MSFE values without further statistical test may produce inconsistent results.

Keywords: *Forecast combination, averaging, optimal weights, mean squared error*

1 INTRODUCTION

In the literature on forecasting, several studies have provided insightful discussions on forecast combination. However, theoretical results on forecast combination have been derived with different techniques or under different assumptions. Consequently, it is unclear from that body of literature if there are deeper connections underlying these different theoretical results.

This paper proposes a unified framework to analyse the theoretical properties of forecast combination. The framework simplifies the derivations of existing results and shows that they can be derived with ease. More importantly, the framework provides a solid foundation to study forecast combination and leads to further insight into the theoretical properties of forecast combination. Specifically, this paper provides the necessary and sufficient conditions on the optimal weight that minimises the Mean Squared Forecast Errors (MSFE). This result also yields the necessary and sufficient conditions for the simple average to be the optimal weight.

This paper also establishes the necessary and sufficient conditions that allows forecasts of a single model to outperform the simple average of forecasts in the MSFE sense. By defining the difference in MSFE as a simple bilinear form, it is trivial to demonstrate that these conditions rarely hold in practice and thus the result provides a simple theoretical justification on the superiority of forecasts from averaging models over single models. This confirms the findings from the seminal paper by Bates and Granger (1969). The result also suggests that the MSFE of any affine forecast combination decreases as the number of model increases under a mild set of assumptions.

Third, it provides a theoretical investigation on forecast combination puzzle, namely why forecast combination using simple average often outperforms complicated weighting schemes in the MSFE sense. This puzzle was first raised in Clemen (1989) and formally named “forecast combination puzzle” by Stock and Watson (2004). Recent studies, such as Smith and Wallis (2009) and Claeskens et al. (2014), have argued that the forecast combination puzzle is due to the presence of finite sample errors from estimating the optimal weights. Specifically, Smith and Wallis (2009) suggests that the simple average forecast combination beats complex weighting scheme with respect to MSFE partly because the weights have to be estimated, which contains finite sample errors. Claeskens et al. (2014) follows in the footsteps of Smith and Wallis (2009) and shows that when the weights need to be estimated the forecast combination is biased and the variance of the combination is larger than in the fixed-weights case such as the simple average. Other and earlier explanations pointing to the estimation error as the source of the problem include Clemen and Winkler (1986), which investigates parameter instability as the underlying motive for error, Hendry and Clements (2004) that considers discrete shift in the data generating process and forecasting models that are misspecifed, and also in Aiolfi and Timmermann (2006). Elliott (2011), on the other hand, investigates the hypothesis that the size of the gains from combination are outweighed by the estimation error. Furthermore, Elliott (2011) examines the sizes of the theoretical gains to optimal forecast combination and provide the conditions under which averaging and optimal combination are equivalent.

These results can be easily derived from the proposed unified framework. More importantly, this paper identifies another cause of the puzzle. Specifically, it shows that the calculation of MSFE is merely a variance estimator of the forecast errors, which may not be consistent. Subsequently, selecting optimal model based on naïve comparison of MSFE without further statistical testing will lead to biased results.

2 FRAMEWORK AND ASSUMPTIONS

This section introduces the framework and the assumptions to analyse the theoretical properties of forecast combination. Let f_{it} denotes an unbiased [1]. Their forecasts of a variable of interest y_t for model i, where $i = 0, \cdots, k$, at time t then

$$y_t = f_{it} + \nu_{it} \qquad i = 0, \cdots, k, \tag{1}$$

where ν_{it} are the forecast errors. Without loss of generality, let f_{0t} be the “best” unbiased forecast of variable y_t, based on the forecast criterion $g(f_{it})$ such that $\mathbb{E}[g(\nu_{0t})] < \mathbb{E}[g(\nu_{it})] \quad \forall i = 1, \cdots, k$, where $\mathbb{E}(\cdot)$ is the expectation operator. Let $u_{it} = \nu_{0t} - \nu_{it}$ and rearranging equation (1) gives

$$f_{it} = y_t - \nu_{0t} + u_{it} \qquad i = 1, \cdots, k. \tag{2}$$

[1] The assumption of unbiasedness is made for mathematical convenience. Using the regression approach proposed in Granger and Ramanathan (1984), Hsiao and Wan (2014) presented a convenient way to accommodate biased forecasts in studying forecast combination. While it is straightforward to incorporate such approach here, it does not change the implications of the theoretical results but it increases the algebraic complexity substantially.

This framework decomposes the prediction errors ν_{it} into two parts. The first part ν_{0t} represents the prediction error from the best model and the second part, u_{it}, represents the difference in prediction errors between the best model and model i.

Following the standard practice, this paper focuses on the matrix version of equation (2). Let $\mathbf{Y} = (y_1, \cdots, y_T)'$, $\mathbf{f}_t = (f_{1t}, \cdots, f_{kt})'$, $\mathbf{F} = (\mathbf{f}_1, \cdots, \mathbf{f}_T)'$, $\mathbf{F}_0 = (f_{01}, \cdots, f_{0T})$ and $\mathbf{u} = (\mathbf{u}_1, \cdots, \mathbf{u}_T)'$ with $\mathbf{u}_t = (u_{1t}, \cdots, u_{kt})'$, $\boldsymbol{\nu}_t = (\boldsymbol{\nu}_{1t}, \cdots, \boldsymbol{\nu}_{kt})'$ with $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_T)'$ and $\boldsymbol{\nu}_0 = (\nu_{01}, \cdots, \boldsymbol{\nu}_{0T})'$. Equation (2) can be written in matrix form as

$$\mathbf{F} = (\mathbf{Y} - \boldsymbol{\nu}_0) \otimes \mathbf{i}' + \mathbf{u} \tag{3}$$

where $\mathbf{i}$ denotes a $k \times 1$ vector of ones and $\otimes$ denotes the Kronecker product. Forecasts for $t = 1, \cdots, T$ based on a linear combination of forecasts from the k models is therefore

$$\mathbf{Fa} = \mathbf{Yi}'\mathbf{a} - \boldsymbol{\nu}_0\mathbf{i}'\mathbf{a} + \mathbf{ua}. \tag{4}$$

If $\mathbf{a}$ is an affine combination, i.e. $\mathbf{i}'\mathbf{a} = 1$, then $\boldsymbol{\nu}_0 + \mathbf{ua}$ is a $T \times 1$ vector containing the forecast errors from the forecast combination. If $\mathbf{a}$ does not represent an affine combination, then $\mathbf{Fa}$ does not produce unbiased forecasts, since $\mathbb{E}(\mathbf{Fa}) = \mathbb{E}(\mathbf{Y})\,\mathbf{i}'\mathbf{a}$ under the standard assumptions that $\mathbb{E}(\nu_{0t}) = \mathbb{E}(u_{it}) = 0$ for all i and t. It is for this reason that only affine combination of forecast are considered.

This framework is flexible enough to produce simple and complex forecast combination models. For example, Bates and Granger (1969) presented a simple combination model for two competing forecasts: $f_{ct} = af_{1t} + (1-a)f_{2t}$ with forecast error $\nu_{it} = y_t - f_{it}$, $i = 1, 2$. This simply implies that $\mathbf{f}_t = (f_{1t}, f_{2t})'$ and $\mathbf{a} = (a, 1-a)'$ in (4) and typically $0 \leq a \leq 1$.

Unless otherwise stated, this paper assumes the following:

Assumption 1. $\nu_{0t} \sim \text{iid}\left(0, \sigma_\nu^2\right)$.

Assumption 2. $\mathbb{E}(\mathbf{u}_t) = \mathbf{0}$ and $\mathbb{E}(\mathbf{u}_t\mathbf{u}_t') = \boldsymbol{\Omega}$ for all t where $\boldsymbol{\Omega}$ is a bounded matrix.

Assumption 3. $\mathbb{E}\left[g(u_{it})\, f(\nu_{0t})\right] = \mathbb{E}\left[g(u_{it})\right]\mathbb{E}\left[f(\nu_{0t})\right]$ for all $i = 1, \cdots k$ and any functions g and f.

3 Optimal weights and averages

This section applies the framework introduced in the previous section to derive several theoretical results. This includes the necessary and sufficient conditions for optimal weight under MSFE as well as the necessary and sufficient conditions for a single model to outperform simple average of forecasts. While the former is well known in the literature, the latter is new.

The theoretical results presented in this section focus on forecast combination with optimal weights. The following discussion assumes that the forecast criterion is MSFE as commonly chosen in the forecast literature, which implies that $g : \mathbb{R} \to \mathbb{R}^+$ is a differentiable function. That is:

$$g(\nu_{it}) = T^{-1}\boldsymbol{\nu}_i'\boldsymbol{\nu}_i$$

where $\boldsymbol{\nu}_i = (\nu_{i1}, \cdots, \nu_{iT})'$. Thus, the MSFE of a forecast combination, $\hat{\sigma}_{\mathbf{a}}^2$, based on the weight vector, $\mathbf{a}$, and $\boldsymbol{\nu}$ is

$$\hat{\sigma}_{\mathbf{a}}^2 = g(\boldsymbol{\nu}\mathbf{a}) = T^{-1}\left(\boldsymbol{\nu}_0'\boldsymbol{\nu}_0 + \mathbf{a}'\mathbf{u}'\mathbf{ua}\right) \tag{5}$$

The last line follows from the restriction that $\mathbf{i}'\mathbf{a} = 1$. Note that $\mathbb{E}\left(\hat{\sigma}_{\mathbf{a}}^2\right) = \mathbb{E}\left[g(\boldsymbol{\nu}\mathbf{a})\right] = \sigma_\nu^2 + \mathbf{a}\boldsymbol{\Omega}\mathbf{a} = \sigma_{\mathbf{a}}^2$. Equation (5) also provides a natural and practical estimator for $\boldsymbol{\Omega}$. However, its consistency relies on $T^{-1}\mathbf{u}'\mathbf{u} - \boldsymbol{\Omega} = o_p(1)$, which may not be true depending on the memory structure in $\mathbf{u}$. As discussed before, $\mathbf{u}$ is likely to be serially correlated in the time series context and further assumption on $\mathbf{u}$ would then be required to ensure $\hat{\boldsymbol{\Omega}}$ is a consistent estimator for $\boldsymbol{\Omega}$.

It is straightforward to derive a set of optimal weights by minimising the forecast error variance as first introduced by Bates and Granger (1969). The $k \times 1$ vector of optimal weight $\mathbf{a}$ is the solution to the following optimisation problem:

$$\mathbf{a} = \arg\min_{\mathbf{x}} \sigma_{\mathbf{a}}^2 = \mathbf{x}'\boldsymbol{\Omega}\mathbf{x} + \sigma_\nu^2 \qquad \text{s.t.} \qquad \mathbf{i}'\mathbf{x} = 1. \tag{6}$$

This can be solved by analysing the associated Lagrangian function which implies:

$$\mathbf{\Omega a}\left(\mathbf{a}'\mathbf{\Omega a}\right)^{-1} = \mathbf{i}. \tag{7}$$

Given the convexity of the objective function and the linearity of the constraint, equation (7) provides the necessary and sufficient condition to derive the optimal weight vector, $\mathbf{a}$. Note that $\mathbf{\Omega}_\nu\mathbf{a}\left(\mathbf{a}'\mathbf{\Omega}_\nu\mathbf{a}\right)^{-1} = \mathbf{i}$ implies $\mathbf{\Omega a}\left(\mathbf{a}'\mathbf{\Omega a}\right)^{-1} = \mathbf{i}$, under the constraint $\mathbf{i}'\mathbf{a} = 1$. It is thus straightforward to show that the closed form solution for the optimal weight vector does indeed satisfy equation (7). This closed form solution is in fact derived in Elliott (2011, p.5), which generalise Bates and Granger (1969).

The first observation from the above optimisation is that $\mathbf{a}$ does not depend on σ^2_ν. An immediate consequence is that forecast combination under affine combination cannot perform better than the best model. This is obvious from the objective function since $\mathbf{\Omega}$ is positive semi-definite and therefore $\mathbf{x}'\mathbf{\Omega x} \geq 0$ for all $\mathbf{x}$. Thus, the role of the optimal weights is to minimise the additional variance due to the deviations from the best model. Interestingly, this observation is not restricted to affine combination of forecasts, it also applies to linear combination of forecasts in general. As shown in the following proposition, forecasts based on linear combinations of k competing models cannot outperform the best model in the MSFE sense.

Proposition 1. $\mathbf{x}'\mathbf{\Omega}_\nu\mathbf{x} \geq \sigma^2_\nu$ *for all* $\mathbf{x} \in \mathbb{R}^k$.

Proposition 1 also suggests that $\mathbf{\Omega}$ contain all the necessary information to analyse forecast combination problems with respect to MSFE and hence equation (7) is often more convenient than the closed form solution as stated in Elliott (2011).

Proposition 2. *The simple average is the optimal weight if and only if* $\mathbf{\Omega i} = k^{-1}\mathbf{i}\left(\mathbf{i}'\mathbf{\Omega i}\right)$.

The proof of Proposition 2 is trivial from equation (7). There are some interesting implications of this result. First, it is obvious that if $\mathbf{\Omega} = \sigma^2\mathbf{I}$ for some $\sigma^2 < \infty$ then $\mathbf{\Omega}$ satisfies the condition in Proposition 2 and therefore the simple average will be the optimal weight. This implies that all deviations from the best model are uncorrelated with each other while the forecasts errors share the same correlation between each model. This is due to the fact that the variance-covariance matrix of the forecast errors is $\mathbf{\Omega}_\nu = \sigma^2_\nu\mathbf{i}\mathbf{i}' + \mathbf{a}'\mathbf{\Omega a}$. This result is consistent with the one derived in Aiolfi and Timmermann (2006). Furthermore, Hsiao and Wan (2014) also provide a necessary and sufficient condition where the simple average is an optimal combination. This condition covers the possibility that some of the models may produce biased forecasts, which would require to estimate a scaling constant.

The second observation is that if the deviations from the best model are not correlated with each other but the variances of the deviations are different then the simple average average will not be the optimal weight but it is still likely to perform better than any single model. In order to formalise this claim, this paper proposes the following bilinear form:

$$dV\left(\mathbf{x},\mathbf{z};\mathbf{\Omega}\right) = \left(\mathbf{x}+\mathbf{z}\right)'\mathbf{\Omega}\left(\mathbf{x}-\mathbf{z}\right) = \mathbf{x}'\mathbf{\Omega x} - \mathbf{z}'\mathbf{\Omega z}. \tag{8}$$

where $\mathbf{x} = (x_1, ..., x_k)'$ and $\mathbf{z} = (z_1, ..., z_k)'$ are two affine forecast combinations such that dV represents the difference in forecast variance between the two affine combinations. The weights vector under simple averaging is $\mathbf{z} = \frac{1}{k}\mathbf{i}$, which means that the difference in forecast variance between any affine combination $\mathbf{x}$ and the simple average can be expressed as

$$dV\left(\mathbf{x}, k^{-1}\mathbf{i};\mathbf{\Omega}\right) = \left(\mathbf{x}+\frac{\mathbf{i}}{k}\right)'\mathbf{\Omega}\left(\mathbf{x}-\frac{\mathbf{i}}{k}\right).$$

Hence, the forecasting performance of any affine forecast combination relative to the simple average can be analysed by examining the sign of the bilinear form as defined in equation (8). Note that the relative efficiency depends solely on the variance-covariance matrix of the *random deviations* from the best model. This is a sequence of affine combination and has some important implications.

Recently, Smith and Wallis (2009) and Claeskens et al. (2014) have demonstrated that the reason for the poor performance of optimal weights relative to the simple average in applications is tied to the effect of estimation errors, at least in the MSFE case. This facts can be simply demonstrated by using the bilinear form. Let $\hat{\mathbf{a}}_T = \mathbf{a} + \boldsymbol{\varepsilon}_T$ be an estimator of $\mathbf{a}$ where $\boldsymbol{\varepsilon}_T$ denotes the estimation error of $\mathbf{a}$ from a finite sample of T observations. So that the bilinear form can be written as

$$dV = \left(\hat{\mathbf{a}}_T+\frac{\mathbf{i}}{k}\right)'\mathbf{\Omega}\left(\hat{\mathbf{a}}_T-\frac{\mathbf{i}}{k}\right) = dV_0 + \boldsymbol{\varepsilon}_T\mathbf{\Omega}\boldsymbol{\varepsilon}_T. \tag{9}$$

where $dV_0 = \left(\mathbf{a}+\frac{\mathbf{i}}{k}\right)'\boldsymbol{\Omega}\left(\mathbf{a}-\frac{\mathbf{i}}{k}\right) < 0$. Since $\boldsymbol{\Omega}$ is positive semi definite, $\boldsymbol{\varepsilon}_T'\boldsymbol{\Omega}\boldsymbol{\varepsilon}_T \geq 0$, and therefore dV can be greater than 0 if $\boldsymbol{\varepsilon}_T'\boldsymbol{\Omega}\boldsymbol{\varepsilon}_T > |dV_0|$. That is, the simple average can outperform the estimated optimal weight if the estimation error of the optimal weight is large. This is consistent with the result given in Claeskens et al. (2014). If $\boldsymbol{\varepsilon}_T = o_p(1)$, it implies that $\boldsymbol{\varepsilon}_T'\boldsymbol{\Omega}\boldsymbol{\varepsilon}_T = o_p(1)$ by the Continuous Mapping Theorem. Thus, dV based on estimated weight will converge in probability to dV_0. However if $\boldsymbol{\varepsilon}_T$ is not $o_p(1)$ or has a very slow rate of convergence, then dV may be severely biased in finite and small sample, respectively.

This also corroborates the findings of Smith and Wallis (2009) that the reason for the poor performance of optimal weights relative to averaging in finite sample is tied up with the estimation error generated when estimating the weights. The properties of forecast combinations with optimal weights are derived under the assumption that the combination weights are fixed and ignore that the weights have to be estimated. Claeskens et al. (2014) provides the theory that shows that when accounting for the optimal weights estimation, the forecast combination can be biased and its variance larger than assuming the weights are fixed.

The current framework provides an insight on the source of the estimator error. The computation of the optimal weight often involves $\boldsymbol{\Omega}$ which is usually not known in practice and therefore it must be estimated based on the forecast errors from individual models, namely, $\nu_{it} = \nu_{0t} + u_{it}$. Recalled a natural estimator for $\boldsymbol{\Omega}_\nu$ is $\hat{\boldsymbol{\Omega}}_\nu$ as defined in equation (5), which is consistent if $T^{-1}\boldsymbol{\nu}_0'\boldsymbol{\nu}_0 - \sigma^2_\nu = o_p(1)$ and $T^{-1}\mathbf{u}'\mathbf{u} - \boldsymbol{\Omega} = o_p(1)$. While the convergence of $T^{-1}\boldsymbol{\nu}_0'\boldsymbol{\nu}_0$ is ensured by Assumption (1), the convergence of $T^{-1}\mathbf{u}'\mathbf{u}$ requires further assumption due to possible serial correlation in u_{it}. Moreover, even if the appropriate conditions are satisfied, T is generally small and therefore, estimation errors are likely to be substantial in most practical situations.

4 Monte Carlo Simulation

This section conducts several Monte Carlo simulations to demonstrate the theoretical results presented in previous sections. The simulation setup relates closely to Smith and Wallis (2009) with the true data generating process (DGP) follows the autoregressive process of order 2:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \nu_{0t} \qquad \nu_{0t} \sim NID(0, \sigma^2_\nu). \tag{10}$$

Note that $\mathbb{E}(y_t) = 0$. Obviously, the best forecast is based on the true DGP which is $f_t = \phi_1 y_{t-1} + \phi_2 y_{t-2}$. Consider two competing models namely, a naïve forecast and an AR(1) process. Specifically,

$$f_{1t} = y_{t-1} \qquad \text{and} \qquad f_{2t} = \theta y_{t-1}. \tag{11}$$

The paper considers the following parameter values:$\phi_1 = 0.5$, $\phi_2 = -0.3$, $\sigma^2_\nu = 0.5$, and $\theta = -0.9525$, which yields a positive covariance between the two competing forecast models . Specifically, $\boldsymbol{\Omega}_\nu = \begin{pmatrix} 0.718 & 0.143 \\ 0.143 & 1.392 \end{pmatrix}$ respectively.

The numbers of in-sample observations (N) considered in this simulation are 10, 20, 50 and 100, and the numbers of out-of-sample forecast observations (T) are 5, 10, 20, 50, 100 and 500. Replication is set at 1000 in each case.

The Monte Carlo simulations consider the following hypotheses

$$H_0 : \mathbb{E}\left(\sigma^2_{\mathbf{i}}\right) - \mathbb{E}\left(\sigma^2_j\right) \leq 0 \qquad \text{and} \qquad H_1 : \mathbb{E}\left(\sigma^2_{\mathbf{i}}\right) - \mathbb{E}\left(\sigma^2_j\right) > 0, \quad j = \text{dgp},\ \mathbf{a},\ \hat{\mathbf{a}},\ \hat{\mathbf{a}}_{nc},$$

where $\sigma^2_{\mathbf{i}}$ is the variance of the simple average forecast and dgp is the variance of the sampling error of the data generating process (true model), $\mathbf{a}$ is the optimal weight, $\hat{\mathbf{a}}$ is the estimated optimal weight based on N in-sample observations and $\hat{\mathbf{a}}_{nc}$ denotes the estimated optimal weight with the restriction that the correlation between forecast is 0. The Diebold-Mariano test.

Tables 1 contains the relative forecast performances from different forecast combination weights based on different number of in-sample and out-of-sample forecasts. Table 1 contains the results in the case when the forecasts from the two models as defined in the equations in (11) are positively correlated. The entries in each column are the proportion of time when the simple average forecasts produced lower MSFE than the other weighting schemes.

As shown in column 3 (σ^2_{dgp}) of Table 1, the simple average appears to outperform the forecast from the true data generating process in the MSFE sense over 40% of the time when $T = 5$. The Diebold-Mariano test suggests that less than 4% of the cases are statistically significant. Note that the significance level of test is set at 5%, the result can potentially be explained by the type I error associated with the test. The implication of this finding is that naïve comparison of MSFE without statistical testing can produce inconsistent results. This is supported by the fact that as T increases the performance of the simple average decreases and the number of statistically significant cases decreases to 0. This confirms the theoretical results that the calculation of MSFE contains significant estimation error.

In Column 5 of Table 1, the simple average may appear to outperform the actual optimal weight when the number of out-of-sample forecast is small but the associated Diebold-Mariano test suggested that only a very small fraction of those cases are statistically significant. Note that the number of within sample observation is not relevant in this case, as the forecast combination is based on the actual theoretical values. Thus, there is no estimation error. As T increases, the simple average begins to perform worse than the actual optimal weight. This corresponds to the fact that as T increases, the MSFE is tending closer towards the forecast variance and therefore this result provides evidence for the hypothesis that the forecast combination puzzle is partly due to estimation error and partly due to sampling error when calculating the MSFE.

The same seem to apply to estimated optimal weight. Although, the simple average may often appears to be superior, especially when N is small, the Diebold-Mariano test once again reveals that only a small fraction of these cases are statistically significant (see Column 7 in Table 1). As N increases, the impact of estimation error in $\hat{\mathbf{a}}$ decreases and the simple average becomes less successful. As T increases, the sample error in estimating the forecast variance also decreases and the estimated optimal weight forecasting performance is often superior.

ACKNOWLEDGEMENTS

The first author wishes to acknowledge the financial support from Curtin University and the second author acknowledges the financial support from the University of Sydney Business School.

REFERENCES

Aiolfi, M. and Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, 135(1–2):31 – 53.

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

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2014). A simple theoretical explanation of the forecast combination puzzle. SSRN: http://dx.doi.org/10.2139/ssrn.2342841.

Clemen, R. and Winkler, R. (1986). Combining economic forecasts. *Journal of Business and Economic Statistics*, 4:39–46.

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

Elliott, G. (2011). Averaging and the optimal combination of forecasts. University of California, San Diego.

Granger, C. and Ramanathan, R. (1984). Improved methods of combining forecast accuracy. *Journal of Forecasting*, 19:197–204.

Hendry, D. and Clements, M. (2004). Pooling of forecasts. *Econometrics Journal*, 1:1–31.

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

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Table 1. Relative Forecast Performances of Different Forecast Combination Weights based on Positively Correlated Forecasts

In-sample (N)	Out-of-sample (T)	σ^2 (dgp)[1]	DM Test [2]	σ^2 $(\mathbf{a})$	DM Test	σ^2 $(\hat{\mathbf{a}})$	DM Test	σ^2 $(\mathbf{a}_{nc})$	DM Test
10	5	0.422	0.037	0.321	0.019	0.386	0.044	0.323	0.036
	10	0.388	0.03	0.259	0.009	0.356	0.055	0.273	0.045
	20	0.329	0.024	0.167	0.005	0.296	0.085	0.205	0.071
	50	0.219	0.015	0.040	0	0.250	0.174	0.185	0.166
	100	0.142	0.012	0.008	0	0.229	0.188	0.183	0.183
	500	0	0	0	0	0.148	0.141	0.139	0.139
20	5	0.413	0.037	0.359	0.016	0.366	0.032	0.284	0.024
	10	0.385	0.029	0.253	0.009	0.262	0.017	0.176	0.014
	20	0.302	0.015	0.132	0	0.191	0.027	0.101	0.023
	50	0.243	0.012	0.048	0	0.098	0.036	0.047	0.034
	100	0.15	0.005	0.007	0	0.060	0.037	0.037	0.037
	500	0	0	0	0	0.041	0.037	0.037	0.037
50	5	0.433	0.049	0.331	0.020	0.321	0.021	0.237	0.015
	10	0.366	0.022	0.258	0.010	0.247	0.008	0.154	0.003
	20	0.346	0.019	0.149	0.004	0.155	0.004	0.068	0.003
	50	0.242	0.02	0.044	0	0.048	0.002	0.006	0.002
	100	0.136	0.002	0.009	0	0.016	0.002	0.003	0.002
	500	0	0	0	0	0.001	0.001	0.001	0.001
100	5	0.426	0.053	0.368	0.022	0.372	0.022	0.286	0.017
	10	0.387	0.024	0.230	0.006	0.230	0.008	0.149	0
	20	0.323	0.021	0.147	0.001	0.146	0.001	0.071	0
	50	0.230	0.014	0.050	0	0.045	0	0.006	0
	100	0.171	0.006	0.005	0	0.011	0	0.001	0
	500	0	0	0	0	0	0	0	0

[1] The columns labeled with $\sigma^2(j)$ present the proportion of time when the simple average produces lower MSFE than the other weight schemes. $\sigma^2(j)$ denotes the MSFE from the forecast combination based on the weight vector j for $j =$ dgp, $\mathbf{a}$, $\hat{\mathbf{a}}$, $\hat{\mathbf{a}}_{nc}$ where dgp denotes the variance of the sampling error of the data generating process (true model), $\mathbf{a}$ denotes the optimal weight, $\hat{\mathbf{a}}$ denotes the estimated optimal weight based on N in-sample observations and $\hat{\mathbf{a}}_{nc}$ denotes the estimated optimal weight with the restriction that the correlation between forecast is 0.

[2] DM test denotes the Diebold-Mariano test.