Some Theoretical Results on Forecast Combinations

Felix Chan
School of Economics and Finance, Curtin University
Laurent Pauwels
University of Sydney

Preliminary – Do not Quote
This draft: March 2015

Abstract: Since Bates and Granger (OR, 1969), both the theoretical and applied forecasting literature have embraced the use of forecast combinations. Despite the growing number of studies employing forecast combination, there are still some questions that remains unresolved. For example, why do combination of forecast models outperform any given forecast model? Or why a simple average with equal weight often outperforms complicated weighting schemes? This paper introduces a simple and general framework to analyse the problem of forecast combination under different forecast criteria, including, but not limited to, Mean Squared Error (MSE) and Mean Absolute Deviation (MAD). This framework provides the sufficient conditions that shows why combining forecast with simple average forecast combination often out-performs individual forecast models both in terms of MSE and MAD. Furthermore, it is possible to obtain various results on the performance of forecast combination when forecasts are correlated.

Keywords:*Forecast combination, panel data model, averaging, optimal weight, linear programming.*
JEL:C53, C33

1 Introduction

Since Bates and Granger (1969), both the theoretical and applied forecasting literature have embraced the use of forecast combinations. Despite the growing number of studies employing forecast combination, there is still a question that remains unresolved, namely why a simple average with equal weight often outperforms complicated weighting schemes in a mean-squared forecast errors (MSFEs) sense. This was coined by Clemen (1989) and named "forecast combination puzzle" by Stock and Watson (2004). There have several recent attempts to explaining the forecast combination puzzle, for example by Smith and Wallis (2009), Elliott (2011), Vasnev et al. (2014).

Smith and Wallis (2009) propose that simple average forecast combination beat complex weighting scheme partly because the weights have to be estimated, which tend to be unreliable. Smith and Wallis (2009) finds that if the optimal weights are close to equality, then simple average forecast combination is more accurate in the MSFE sense. Their argument is based on obtaining weights that minimises the MSFE and rely on the fact that the general weighting schemes nest the simple average, which can be extended to the forecast evaluation of competing (nested) models as in Clark and West (2007). They also show that the MSFE adjustment term that explains the discrepancy between the sample average and the weighted alternative can be calculated. The only advantage of using estimated weights over the simple average is when the forecast errors differ substantially, which does not seem to happen much practically as observed by Smith and Wallis (2009). Vasnev et al. (2014) follows in the footsteps of Smith and Wallis (2009) and show 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 investigate parameter instability as the underlying motive for error, and also Hendry and Clements (2004) that considers discrete shift in the data generating process and forecasting models that are mis-specifed. 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.

This paper proposes a general framework which can be used to derive existing results on forecast combination and to investigate further theoretical issues on forecast combination not considered before. First, the paper derives the conditions under which the simple average is the optimal weight. Second, it provides a theoretical investigation on forecast combination puzzle, especially in light of the recent findings that estimation error is at the source of the puzzle. Further insights are gained from the proposed approach. The main contribution of this paper lies in the formulation of the problem, or the framework, which permits to simply and generally expose the forecast combination puzzle. This framework is closely related to the model studied by Hsiao and Wan (2014) which imposed a multi-factor structure on the forecast error ν_{it} such that $\nu_{it} = \alpha_i + \mathbf{b}_i'\boldsymbol{\lambda}_t + \boldsymbol{\varepsilon}_{it}$ where $\boldsymbol{\lambda}_t$ represents a vector of common factors. From this multi-factor structure, Hsiao and Wan (2014) develop several eigenvector approaches to combining forecasts. They also provide the conditions under which their new approaches yield identical results to the regression approach of determining the optimal weight vector, as suggested in Granger and Ramanathan (1984). Hsiao and Wan (2014) also provide a necessary and sufficient condition where the simple average is an optimal combination, but require estimating a scaling constant in case models produce biased forecast. This paper contributes to the literature by showing that many of the results in Hsiao and Wan (2014) can be derived without the multi-factor structure and therefore provides more general results than Hsiao and Wan (2014). Finally, a set of conditions for which the simple average outperforms optimal weighting schemes based on MAD is provided. This has not been covered before by this literature.

The paper is organised as follows: Section 2 revisits forecast combination in the context of the proposed framework in the case when forecasts are evaluated using MSFE as a criterion. Moreover, it proposes a simple bilinear form to evaluate relative efficiency of two forecast combinations. Section 3 derives several theoretical results using the proposed framework including the conditions for which the simple average is the optimal weight. Section 4 discusses forecast combination and simple averaging in the context of the MAD evaluation criterion and concluding remarks can be found in

Section 5.

2 Model and assumptions

This section introduces the framework to analyse the theoretical properties of forecast combination. Let f_{it} denotes an unbiased 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

$$\mathbb{E}\left[g\left(\nu_{0t}\right)\right] < \mathbb{E}\left[g\left(\nu_{it}\right)\right] \quad \forall i = 1, \cdots, k,$$

where $\mathbb{E}(\cdot)$ is the expectation operator. Let $u_{it} = \nu_{0t} - \nu_{it}$ and rearranging equation (1) gives

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

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, 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 i. $\nu_{0t} \sim \text{iid}(0, \sigma_\nu^2)$.

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

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

Remark 1. The existence of second moment applies only to the forecast errors and not on the variable, y_t. Thus, the analysis in this paper applies equally to the case where $y_t \sim I(1)$ under Assumptions (i) – (iii).

Remark 2. Note that u_{it} is not required to be independently and identically distributed. This is particularly important in the time series context since u_{it} represents the deviations from the best model which is likely to be serially correlated. For example, let $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \nu_{0t}$ then $f_{0t} = \phi_1 y_{t-1} + \phi_2 y_{t-2}$ and let $f_{1t} = \phi_1 y_{t-1}$, then $u_{1t} = \phi_2 y_{t-2}$ which is clearly serially correlated.

Remark 3. Since $u_{it} = \nu_{0t} - \nu_{it}$, Assumptions (i) and (ii) are sufficient to characterise ν_{it} for $i = 1, \cdots, k$, and also imply that the variance-covariance matrix of the forecast errors $\boldsymbol{\Omega}_\nu = \sigma_\nu^2 \mathbf{ii}' + \boldsymbol{\Omega}$ is a bounded matrix.

Remark 4. The independence assumption as stated in Assumption (iii) may seem restrictive but it is appropriate in the context of forecast combination problems. This is due to the fact that most forecasts for time t are based on the information up to time $t-1$. More formally, let y_t be a stochastic process adapted the filtration $\mathfrak{F}_t$, then equations (1) and (2) imply $u_{it} \in \mathfrak{F}_{t-1}$ but $\mathbb{E}\left(\nu_{0t}|\mathfrak{F}_{t-1}\right) = \mathbb{E}\left(\nu_{0t}\right)$ under Assumption (i). Therefore, ν_{0t} and u_{it} are independent for all i and t.

3 Optimal weights and averages

3.1 Deriving optimal weights

This section presents some theoretical results concerning forecast combination with a specific focus on the optimal weight. The 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:

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

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

$$\begin{aligned}
\hat{\sigma}^2_{\mathbf{a}} =& g\left(\boldsymbol{\nu}\mathbf{a}\right) && (5)\\
=& T^{-1}\mathbf{a}'\boldsymbol{\nu}'\boldsymbol{\nu}\mathbf{a} && (6)\\
=& T^{-1}\left(\boldsymbol{\nu}_0'\boldsymbol{\nu}_0 + \mathbf{a}'\mathbf{u}'\mathbf{u}\mathbf{a}\right) && (7)
\end{aligned}$$

The last line follows from the restriction that $\mathbf{i}'\mathbf{a} = 1$. Note that $\mathbb{E}\left(\hat{\sigma}^2_{\mathbf{a}}\right) = \mathbb{E}\left[g\left(\boldsymbol{\nu}\mathbf{a}\right)\right] = \sigma^2_\nu + \mathbf{a}\boldsymbol{\Omega}\mathbf{a} = \sigma^2_{\mathbf{a}}$. Equation (7) 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

$$\mathbf{a} = \arg\min_{\mathbf{x}} \sigma^2_{\mathbf{a}} = \mathbf{x}'\mathbf{\Omega}\mathbf{x} + \sigma^2_\nu \tag{8}$$

$$\text{s.t.} \qquad \mathbf{i}'\mathbf{x} = 1. \tag{9}$$

This can be solved by analysing the associated Lagrangian function:

$$L = \mathbf{x}'\mathbf{\Omega}\mathbf{x} + \sigma^2_\nu + \lambda\left(1 - \mathbf{i}'\mathbf{x}\right)$$

with the first order condition being:

$$L_{\mathbf{x}}|_{\mathbf{x}=\mathbf{a},\lambda=\lambda^*} = 2\mathbf{a}'\mathbf{\Omega} - \lambda^*\mathbf{i}' = 0$$
$$L_{\lambda}|_{\mathbf{x}=\mathbf{a},\lambda=\lambda^*} = 1 - \mathbf{i}'\mathbf{a} = 0.$$

Post-multiply $L_{\mathbf{x}}$ by $\mathbf{a}$ and using L_λ yields:

$$\lambda^* = 2\mathbf{a}'\mathbf{\Omega}\mathbf{a}$$

which implies:

$$\mathbf{\Omega}\mathbf{a}\left(\mathbf{a}'\mathbf{\Omega}\mathbf{a}\right)^{-1} = \mathbf{i}. \tag{10}$$

Given the convexity of the objective function and the linearity of the constraint, equation (10) 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 $\mathbf{\Omega}\mathbf{a}\left(\mathbf{a}'\mathbf{\Omega}\mathbf{a}\right)^{-1} = \mathbf{i}$, 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 (10). This closed form solution is in fact derived in Elliott (2011), which generalise Bates and Granger (1969). Again in the simple combination proposed by Bates and Granger (1969), $\mathbf{x} = (x, 1-x)'$ and $\mathbf{\Omega}_\nu = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ as before. Then minimising the forecast error variance would yield

$a = \dfrac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$ as the optimal value, as implied in (10).

The first observation from the above optimisation is that $\mathbf{a}$ does not depend on σ_ν^2. 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 $\mathbf{x}'\mathbf{\Omega}\mathbf{x} \geq 0$ 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. $\mathbf{x}'\mathbf{\Omega}_\nu\mathbf{x} \geq \sigma_\nu^2$ *for all* $\mathbf{x} \in \mathbb{R}^k$.

Proof. See Appendix. □

Proposition 1 also suggests that $\mathbf{\Omega}$ contain all the necessary information to analyse forecast combination problems with respect to MSFE and hence equation (10) is often more convenient than the closed form solution as stated in Elliott (2011).

3.2 Is the simple average optimal?

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

The proof of Proposition 2 is trivial from equation (10). 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 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, $\mathbf{\Omega}_\nu = \sigma_\nu^2\mathbf{i}\mathbf{i}' + \mathbf{a}'\mathbf{\Omega}\mathbf{a}$. This result is consistent with those derived in 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 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:

$$\begin{aligned} dV\left(\mathbf{x},\mathbf{z};\mathbf{\Omega}\right) =& \left(\mathbf{x}+\mathbf{z}\right)'\mathbf{\Omega}\left(\mathbf{x}-\mathbf{z}\right) \\ =& \mathbf{x}'\mathbf{\Omega}\mathbf{x}-\mathbf{z}'\mathbf{\Omega}\mathbf{z}. \end{aligned} \tag{11}$$

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

$$\mathbf{z} = \frac{1}{k}\mathbf{i} \tag{12}$$

which means that the difference in forecast variance between any affine combination $\mathbf{x}$ and a 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 forecast 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 (11). Note that the relative efficiency depends solely on variance-covariance matrix of the *random deviations* from the best model. This is a sequence of affine combination and has some important implications. Specifically, equation (10) along with the bilinear form as defined in equation (11) provide an unified framework to analyse various problems arise from forecast combinations. The following corollaries provide some examples on the advantage of this framework.

Corollary 1. *If* $\mathbf{\Omega} = diag\left(\sigma_1^2, \cdots, \sigma_k^2\right)$, *then*

1. *The simple average is not the optimal weight if* $\sigma_i^2 \neq \sigma_j^2$ *for some* $i \neq j$.
2. *Let* $\bar{\sigma}^2 = k^{-1}\sum_{i=1}^{k}\sigma_i^2$, *model* j *will outperform a simple average forecast if and*

only if

$$\sigma_j^2 < \frac{\bar{\sigma}^2}{k}. \tag{13}$$

Proof. See Appendix. □

Corollary 1 shows that a single model j can outperform simple average forecast if the forecast variance of model j is less than the average forecast variance by a factor of k^{-1}. Moreover, under the assumption that $\bar{\sigma} < \infty$,

$$\frac{\bar{\sigma}^2}{k} = o(1).$$

This has two implications. First, if $\sigma_j^2 > 0$ then the simple average will eventually outperform model j as the number of models increases. This result is quite important as it helps to explain the superiority of simple average forecasts even when simple average is not the optimal weight. Second, if $\sigma_j^2 = 0$, that is, model j is the true model then the equality will hold as $k \to \infty$. In fact, this implication is not surprising as the inequality in the second part of Corollary 1 can be written as

$$\sigma_j^2 \leq V\left(\sum_{i=1}^{k} \frac{u_{it}^2}{k}\right),$$

where $V(x)$ denotes the variance of x. The expression above states that the variance of forecast errors from model j must be less than or equal to the variance of forecast error from the simple average.

The relationship in (13) is also presented in Smith and Wallis (2009) equation (6); they find that the simple average would outperform the weighted average systematically when they are theoretically equivalent. This difference in MSFE is also defined as the MSFE "adjustment" in Clark and West (2006).

The third observation from Corollary 2 is that the sum of each row in $\mathbf{\Omega}$ must be the same for simple average to be the optimal weight. That is, if $\mathbf{\Omega} = \{\sigma_{ij}\}$ with $\sigma_{ii} = \sigma_i^2$ then $\sum_{j=1}^{k} \sigma_{ij}^2 = \sigma$ for each $i = 1, .., k$ and some $\sigma < \infty$. The significance of this observation is that the sum of the variance and covariances of each k model must sum to the same constant. Along with the restriction that $\mathbf{\Omega}$ must be a symmetric matrix,

there are $\frac{k^2+k-2}{2}$ restrictions in $\mathbf{\Omega}$. Although the simple average may not be the optimal weight, the following proposition shows that it is likely that it will perform better than a single model.

Corollary 2. *Let $\mathbf{\Omega} = \{\sigma_{ij}\}$ with $\sigma_{ii} = \sigma_i^2$ for $i, j = 1, \cdots, k$ then the expected difference in MSFE between the r^{th} model and the simple average is*

$$dV\left(\mathbf{e}_r, k^{-1}\mathbf{i}; \mathbf{\Omega}\right) = \sigma_r^2 - \frac{\bar{\sigma}^2}{k} - \frac{2}{k}\bar{\sigma}_r - \frac{2}{k}\bar{\sigma}_{ij}. \tag{14}$$

where $\bar{\sigma}^2 = k^{-1}\sum_{i=1}^{k}\sigma_i^2$, $\bar{\sigma}_r = k^{-1}\sum_{j=1,j\neq r}^{k}\sigma_{rj}$ and $\bar{\sigma}_{ij} = k^{-1}\sum_{i=1,i\neq r}^{k}\sum_{j>i,j\neq r}^{k}\sigma_{ij}$.

Proof. See Appendix. □

The result in Corollary 2 is quite insightful in explaining the performance of a simple average relative to single model in general. Firstly, it is clear that

$$\lim_{k\to\infty} dV = \sigma_r \geq 0 \qquad \forall r = 1, ..., k,$$

under Assumption (ii). Therefore, simple average will in general outperform a single model when k is large. This is consistent with result in Corollary 1. Moreover, equation (14) reduces to $\sigma_j^2 - k^{-1}\bar{\sigma}^2$ if the deviations are not correlated with each other and further reduces to the result in Corollary 1 when $\sigma_{rj} = 0$ for all $r = 1, \cdots, k$.

In order to gain further insight on the implication of equation (14), consider once again two competing forecasts f_{1t} and f_{2t} with $\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$, and without loss of generality, assume $\sigma_1^2 < \sigma_2^2$. The expected difference in MSFE between model 1 and the simple average is $\frac{3\sigma_1^2 - \sigma_2^2}{4} - \frac{\sigma_{12}}{2}$, which will be negative if and only

$$\sigma_1^2 < \frac{\sigma_2^2 + 2\sigma_{12}}{3}. \tag{15}$$

That is, when $k = 2$, the model with the lowest forecast variance will outperform a simple average if and only if the above inequality hold. If $\sigma_{12} = 0$, the inequality reduces to the case under Corollary 1. An interesting observation is that the inequality

is more likely to hold when $\sigma_{12} > 0$ than when $\sigma_{12} < 0$. This justifies the conventional belief that forecast combination between diversified forecasts tends to perform better. The inequality also holds when $\sigma_{12} \geq \sigma_2^2$, but this cannot be the case since $\sigma_1 < \sigma_2$ by assumption and $\sigma_{12} \leq \sigma_1\sigma_2$ by the Hölder's Inequality. Therefore equation (15) gives the maximum bound of σ_1^2 in order for an individual model to outperform the simple average when $k = 2$.

3.3 Estimation error

Recently, Smith and Wallis (2009) and Vasnev et al. (2014) have demonstrated that the reason for the poor performance of optimal weights relative to a 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

$$dV = \left(\hat{\mathbf{a}}_T + \frac{\mathbf{i}}{k}\right)' \boldsymbol{\Omega} \left(\hat{\mathbf{a}}_T - \frac{\mathbf{i}}{k}\right) \tag{16}$$

$$= \left(\mathbf{a} + \frac{\mathbf{i}}{k}\right)' \boldsymbol{\Omega} \left(\mathbf{a} - \frac{\mathbf{i}}{k}\right) + \boldsymbol{\varepsilon}_T' \boldsymbol{\Omega} \boldsymbol{\varepsilon}_T \tag{17}$$

$$= dV_0 + \boldsymbol{\varepsilon}_T \boldsymbol{\Omega} \boldsymbol{\varepsilon}_T. \tag{18}$$

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 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 Vasnev et al. (2014). If $\boldsymbol{\varepsilon}_T = o_p(1)$, $\boldsymbol{\varepsilon}_T = o_p(1)$ 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 taking a simple average 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. Vasnev 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 estimator error. The computation of the optimal weight often involves $\mathbf{\Omega}$ 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 $\mathbf{\Omega}_\nu$ is $\hat{\mathbf{\Omega}}_\nu$ as defined in equation (7), which is consistent if

$$T^{-1}\boldsymbol{\nu}_0'\boldsymbol{\nu}_0 - \sigma_\nu^2 = o_p(1) \qquad \text{and} \qquad T^{-1}\mathbf{u}'\mathbf{u} - \mathbf{\Omega} = o_p(1).$$

While the convergence of $T^{-1}\boldsymbol{\nu}_0'\boldsymbol{\nu}_0$ is ensured by Assumption (i), 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.

Another insight relates to the relative performance between the optimal weight, estimated optimal weight and simple average. If we consider the fact that forecast combination is itself a forecasting model, then forecasts based on optimal weight and simple average can be considered as two competing forecast models. Let $\mathbf{\Sigma}_\nu$ be the variance-covariance matrix of the forecast errors from combining the optimal weight and the simple average models, that is

$$\mathbf{\Sigma}_\nu = \sigma_\nu^2 \mathbf{i}\mathbf{i}' + \mathbf{\Sigma} \tag{19}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{a}'\mathbf{\Omega}\mathbf{a} & k^{-1}\mathbf{a}'\mathbf{\Omega}\mathbf{i} \\ k^{-1}\mathbf{a}'\mathbf{\Omega}\mathbf{i} & k^{-2}\mathbf{i}'\mathbf{\Omega}\mathbf{i} \end{pmatrix}. \tag{20}$$

It is possible to choose an optimal weight vector $\mathbf{b} = (b, 1-b)$ to minimise $\mathbf{x}'\mathbf{\Sigma}_\nu\mathbf{x}$. Note that b has the interpretation of the number of times one should use the forecast from

optimal weight in order to minimise the forecast errors in the MSFE sense. It is clear from this interpretation that the only time when forecasts from the optimal weight will always outperform forecast from simple average is when $b = 1$, which occurs if and only if $\mathbf{a}'\mathbf{\Omega}\mathbf{a} = k^{-1}\mathbf{a}'\mathbf{\Omega}\mathbf{i}$. This means that even when the optimal weight can be obtained without any estimation error, there will be times when simple average outperforms the optimal weight. Furthermore, it is possible to derive an expression relating b with the estimator error as shown in Proposition 3.

Proposition 3. *Let* $\tilde{\mathbf{\Sigma}}_\nu = \mathbf{\Sigma}_\nu + \mathbf{\Lambda}$ *where*

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & 0 \end{pmatrix}$$

denotes the deviation from $\mathbf{\Sigma}_\nu$ *due to the estimation error of* $\mathbf{a}$*. Let* $\mathbf{b}$ *be the vector that minimises* $\mathbf{x}'\mathbf{\Sigma}_\nu\mathbf{x}$ *such that* $\mathbf{i}'\mathbf{b} = 1$*. Then*

$$db = -\frac{1}{\Delta} d\lambda_{11} + \left(\frac{2b-1}{\Delta}\right) d\lambda_{12} \tag{21}$$

where $\Delta = \mathbf{a}'\mathbf{\Omega}\mathbf{a} + k^{-2}\mathbf{i}'\mathbf{\Omega}\mathbf{i} - 2k^{-1}\mathbf{a}'\mathbf{\Omega}\mathbf{i}$*. Furthermore,* $db > 0$ *if and only if* $d\lambda_{11} \leq (2b-1)\, d\lambda_{12}$*.*

Proof. See Appendix. □

The last line in Proposition 3 implies that the estimated weight may perform better in the presence of increasing estimation error if and only if $d\lambda_{11} \leq (2b-1)\, d\lambda_{12}$ but this cannot happen due to Hölder's Inequality. Therefore, the presence of estimation error will always improve the performance of the simple average relative to the estimated optimal weight.

Given this result, it would appear important to examine if the simple average does in fact produce significantly better results than the estimated optimal weight. The simple test, inspired by the bilinear form, will be introduced in the next section.

4 Continuous and non-differentiable forecast evaluation criteria

The results from previous section explicitly assumed that the forecast criteria are differentiable functions. However, there are many forecast criteria that do not satisfy differentiability. One of such criteria is Mean Absolute Deviation, which is defined to be

$$MAD = T^{-1} \sum_{t=1}^{T} |u_t|. \tag{22}$$

Recall that the forecast errors from the forecast combination can be expressed as :

$$ua = Ya - F$$

and therefore the MAD can be expressed as

$$MAD = T^{-1} \sum_{t=1}^{T} |u_t' a|. \tag{23}$$

Define $w_t = |u_t' a| I_t$ and $v_t = |u_t' a| (1 - I_t)$ where $I_t = I(u_t' a \geq 0)$ is an indication function with $I(A) = 1$ if A is true and 0 otherwise. The problem of selecting a in order to minimise MAD as defined in equation (23) can be written as a linear programming problem:

$$a = \arg\max_{x \in X} \frac{\mathbf{i}'}{T}\mathbf{w} + \frac{\mathbf{i}'}{T}\mathbf{v} \tag{24}$$

$$\text{subject to} \qquad -\mathbf{w} + \mathbf{v} + u'x = \mathbf{0} \tag{25}$$

$$\mathbf{i}'x = 1 \tag{26}$$

where $\mathbf{w} = (w_1, \cdots, w_T)'$ and $\mathbf{v} = (v_1, \cdots, v_T)'$. Let $\mathbf{b} = \left(\mathbf{0}'_{T\times 1}, 1\right)'$ and

$$A = \begin{pmatrix} -\mathbf{I}_T & \mathbf{I}_T & u \\ \mathbf{0}_{1\times T} & \mathbf{0}_{1\times T} & \mathbf{i}_T \end{pmatrix}$$

then the linear programming problem can be written in the standard form:

$$\begin{aligned} \mathbf{a} = \arg\max_{x \in X} z = \mathbf{c}'\mathbf{x} \\ \text{subject to} \qquad \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

where $\mathbf{c} = T^{-1}\left(\mathbf{i}'_{2T}, \mathbf{0}'_k\right)'$ and $\mathbf{x} = (\mathbf{w}', \mathbf{v}', x')'$. By the Fundamental Theorem of Linear Programming, it is clear that if there is a optimal feasible solution, there is an optimal basic feasible solution. Lets' consider a feasible solution where only one model was selected. Without loss of generality, let assume the first model was selected, that is $x = \mathbf{e}_1$. Note that if $w_t > 0$ then $v_t = 0$ and vice versa by definition. Therefore, there will be exactly T non-zero elements from $(\mathbf{w}', \mathbf{v}')'$ to form a basic feasible solution. Moreover, if $w_t > 0$ then $w_t = u_{1t}$ and if $v_t > 0$ then $v_t = |u_{1t}|$. Since the index order in $\mathbf{x}$ is arbitrary, it is always possible to rewrite the basic feasible solution in the following form:

$$\left[\mathbf{B}, \mathbf{D}\right] \begin{bmatrix} \mathbf{x}_B \\ \mathbf{x}_D \end{bmatrix} = \begin{bmatrix} \mathbf{0}_T \\ 1 \end{bmatrix}$$

where $\mathbf{x}_B = (\mathbf{w}'_B, \mathbf{v}'_B, 1)'$ consists of all the non-zero elements in $\mathbf{x}$,

$$\mathbf{B} = \begin{bmatrix} K^{\alpha}\mathbf{I} & \mathbf{u}_{1t} \\ \mathbf{0}'_T & 1 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} K^{\beta}\mathbf{I} & \mathbf{u}_{-1t} \\ \mathbf{0}'_T & \mathbf{0}'_{k-1} \end{bmatrix}$$

with $K^{\alpha}\mathbf{I} = \text{diag}\left((-1)^{1-I_{1t}}\right)$, with $I_{it} = I\left(u_{it} < 0\right)$ and $I\left(A\right)$ is an indication function such that it equals to 1 if A is true and 0 otherwise.

The objective function evaluated at this basic feasible solution is $z = \mathbf{c}'_B\mathbf{x}_B + \mathbf{c}'_D\mathbf{x}_D = c'_B\mathbf{x}_B$, where $\mathbf{x}_D$ corresponds to all the zero elements in $\mathbf{x}$ at this particular basic feasible solution. Note that $B\mathbf{x}_B + D\mathbf{x}_D = \mathbf{b}$ where $\mathbf{b} = (\mathbf{0}'_T, 1)'$ and therefore it is straightforward to show that the changes in the objective function when changing from

one basic feasible solution to another is

$$\delta = \left(\mathbf{c}'_D - \mathbf{c}'_B\mathbf{B}^{-1}\mathbf{D}\right)\mathbf{x}_D. \tag{27}$$

Obviously, the change of basic feasible solution is only appropriate if $\delta < 0$ and therefore, a necessary condition for forecast combination to outperform single model forecast is that at least one element in $\mathbf{c}'_D - \mathbf{c}'_B\mathbf{B}^{-1}\mathbf{D}$ is less than 0.

Proposition 4. *If $\exists j = 1, ..., k$ with $j \neq i$ such that $T^{-1}\sum_{t=1}^{T}(-1)^{I_{it}}(u_{jt} - u_{it}) < 0$, then the weighted average of the forecasts from models i and j will outperform forecast from model i alone based on mean absolute errors.*

Note that the condition implies that $\delta_j < 0$ if u_{it} and u_{jt} have the opposite signs or if $|u_{jt}| < |u_{it}|$ if $I_{it} = I_{jt}$ on average. This make intuitive sense as either condition will reduce mean absolute error when model j is included in the forecasts. It also supports the claims that combining models with contradictory forecasts will reduce forecast errors on average.

Proposition 5. *Under the Fundamental Theorem of Linear Programming, if there is an optimal solution then there must be a optimal basic feasible solution. This means for k different models, a necessary condition for simple average to be optimal is that the forecast errors from the forecast combinations are 0 for at least k periods.*

5 Conclusion

This paper established several theoretical results concerning forecast combinations. By setting up the forecast combination problem as a panel data model, the paper was able to provide the necessary and sufficient condition for optimal weight as well as the necessary and sufficient condition for simple average to be the optimal weight under Mean Squared Forecast Errors (MSFE). It also provided theoretical justifications on the superior forecast performance of simple average or individual models in the MSFE sense. The paper also provided a theoretical exposition on the relative performance of simple average and estimated optimal weight. The results show that the performance

of simple average can often outperform the estimated optimal weight in the presence of estimation error. This theoretical justification is consistent with the empirical observation that the simple average often has superior performance over estimated optimal weight.

The paper also investigated the forecast combination problem under Mean Absolute Deviation (MAD). By applying the Fundamental Theorem of Linear Programming, the paper was able to establish the necessary and sufficient condition for the simple average to outperform a single model in the MAD sense. This result is new and the method adopted in the paper might suggest a feasible way to analyse the forecast combination problems for non-differentiable forecast criteria further.

A Proofs

Proof of Proposition 1: Recall that

$$\begin{aligned}\mathbf{x}'\mathbf{\Omega}_\nu\mathbf{x} =& \mathbf{x}'\left(\sigma_\nu^2\mathbf{i}\mathbf{i}' + \mathbf{\Omega}\right)\mathbf{x} \\ =& \sigma_\nu^2\mathbf{x}'\mathbf{i}\mathbf{i}'\mathbf{x} + \mathbf{x}'\mathbf{\Omega}\mathbf{x}.\end{aligned}$$

Since σ_ν^2 denotes the forecast variance from the best model and if there exists $\mathbf{x}$ such that $\mathbf{x}'\mathbf{i} \neq 1$ and $\mathbf{x}'\mathbf{\Omega}_\nu\mathbf{x} - \sigma_\nu^2 < 0$ then

$$\sigma_\nu^2\left(1 - \mathbf{x}'\mathbf{i}\mathbf{i}'\mathbf{x}\right) > \mathbf{x}'\mathbf{\Omega}\mathbf{x} \qquad \mathbf{i}'\mathbf{a} \neq 1. \tag{28}$$

The inequality cannot hold if $\mathbf{i}'\mathbf{x} > 1$ since it implies $\sigma_\nu^2 < 0$. Thus the only remaining case is $\mathbf{x}'\mathbf{i} < 1$. Let's consider the following minimisation problem:

$$\mathbf{a} = \arg\min_{\mathbf{x}} \mathbf{x}'\mathbf{\Omega}_\nu\mathbf{x} \text{ s.t. } \mathbf{x}'\mathbf{i} - 1 \leq 0.$$

The associated Lagrangian is $L = \mathbf{x}'\mathbf{\Omega}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{i} - 1)$. Evaluate at the optimal point, $\mathbf{a}$, the Karush-Kuhn-Tucker conditions give:

$$\begin{aligned} \mathbf{\Omega}\mathbf{a} - \lambda\mathbf{i} &= 0 \\ \lambda &\geq 0 \\ \lambda(\mathbf{a}'\mathbf{i} - 1) &= 0. \end{aligned}$$

The last line implies that either $\lambda = 0$ or $\mathbf{a}'\mathbf{i} = 1$. Consider the case $\lambda = 0$, this implies $\mathbf{\Omega}\mathbf{a} = 0$. Since $\mathbf{\Omega}$ has full rank, this implies $\mathbf{a} = 0$ which cannot be the case. Hence, $\mathbf{a}'\mathbf{i} = 1$ which implies the optimal weight must be an affine combination. Hence, the forecast variance from a linear combination of k competing forecasts cannot be less than the forecast variance from the best model. This completes the proof. ■

Proof of Proposition 2: Replace $\mathbf{a}$ by $k^{-1}\mathbf{i}$ yields the result. This completes the proof. ■

Proof of Corollary 1: Using equation (10), it is straightforward to show that

$$\frac{a_j}{a_i} = \frac{\sigma_i^2}{\sigma_j^2} \qquad \forall i, j = 1, ..., k.$$

Note that $a = \mathbf{i}/k$ if and only if $\sigma_i^2 = \sigma_j^2$ for all i and j which reduces to Case I. This completes the first part of the proposition. To show the second part, it is sufficient to derive the condition such that $dV(\mathbf{e}_j, k^{-1}\mathbf{i}; \mathbf{\Omega}) \leq 0$:

$$\begin{aligned} dV &= \left(\mathbf{e}_j - k^{-1}\mathbf{i}\right)' \mathbf{\Omega} \left(\mathbf{e}_j - k^{-1}\mathbf{i}\right) \\ &= -\frac{\sigma_1^2}{k^2} - \cdots - \left(\frac{1}{k^2} - 1\right)\sigma_j^2 - \cdots - \frac{\sigma_k^2}{k^2} \qquad (29) \\ &= \sigma_j^2 - \frac{1}{k^2}\sum_{i=1}^{k}\sigma_i^2. \qquad (30) \end{aligned}$$

The last line suggests that, $dV \leq 0$, that is, a single model j can perform at least as

well as a simple average when

$$\sigma_j^2 \left(\sum_{i=1}^{k} \sigma_i^2 \right)^{-1} \leq \frac{1}{k^2}. \tag{31}$$

Notice that by defining $\bar{\sigma}^2 = k^{-1} \sum_{i=1}^{k} \sigma_i^2$ and rewriting equation (31) as

$$\sigma_j^2 \leq \frac{\bar{\sigma}^2}{k},$$

This completes the proof. ■

Proof of Corollary 2: The performance of the simple average relative to a single model is gauged by the relative efficiency measure dV as defined in equation (11). This gives:

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

Without loss of generality, set $r = 1$ which means that

$$dV\left(\mathbf{e}_1, k^{-1}\mathbf{i}; \mathbf{\Omega}\right) = \frac{1}{k^2} \left(k+1, \mathbf{i}'_{k-1}\right) \mathbf{\Omega} \left(k-1, \mathbf{i}'_{k-1}\right)'.$$

Let $\mathbf{x} = \left(k+1, \mathbf{i}'_{k-1}\right)$ and $\mathbf{z} = \left(k-1, \mathbf{i}'_{k-1}\right)$ and re-write dV in terms of its elements

gives:

$$\begin{aligned}
dV\left(\mathbf{e}_1, k^{-1}\mathbf{i}; \mathbf{\Omega}\right) =& k^{-2}\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}x_i z_j \\
=& k^{-2}\left(\sum_{j=1}^{k}\sigma_{1j}x_1 z_j + \sum_{i=2}^{k}\sum_{j=1}^{k}\sigma_{ij}x_i z_j\right) \\
=& k^{-2}\left(\sigma_{11}x_1 z_1 + \sum_{j=2}^{k}\sigma_{1j}x_1 z_j + \sum_{i=2}^{k}\sigma_{i1}x_i z_1 + \sum_{i=2}^{k}\sum_{j=2}^{k}\sigma_{ij}x_i z_j\right) \\
=& \left(\frac{k^2-1}{k^2}\right)\sigma_{11} - \sum_{j=2}^{k}\frac{k+1}{k^2}\sigma_{1j} + \sum_{i=2}^{k}\frac{k-1}{k^2}\sigma_{i1} - \sum_{i=2}^{k}\sum_{j=2}^{k}\frac{1}{k^2}\sigma_{ij} \\
=& \sigma_{11} - \frac{1}{k^2}\sigma_{11} - \sum_{j=2}^{k}\frac{2}{k^2}\sigma_{1j} - \sum_{i=2}^{k}\sum_{j=2}^{k}\frac{1}{k^2}\sigma_{ij} \\
=& \sigma_{11} - \sum_{i=1}^{k}\frac{1}{k^2}\sigma_{ii} - \sum_{j=2}^{k}\frac{2}{k^2}\sigma_{1j} - 2\sum_{i=2}^{k}\sum_{j>i}^{k}\frac{1}{k^2}\sigma_{ij} \\
=& \sigma_{11} - \frac{\bar{\sigma}}{k} - \frac{2}{k}\bar{\sigma}_1 - \frac{2}{k}\bar{\sigma}_{ij}.
\end{aligned}$$

This completes the proof. ■

Proof of Proposition 3: Let $\mathbf{b}_1 = (b_1, 1-b_1)$ that minimises $\mathbf{x}'\tilde{\mathbf{\Sigma}}_\nu\mathbf{x}$, then

$$b_1 = \frac{\mathbf{a}'\mathbf{\Omega}\mathbf{a} - k^{-1}\mathbf{a}'\mathbf{\Omega}\mathbf{i} - \lambda_{12}}{\Delta + \lambda_{11} - 2\lambda_{12}}. \tag{32}$$

Totally differentiate b_1 with respective to λ_{11} and λ_{12} and evaluate the derivatives at $(\lambda_{11}, \lambda_{12}) = (0, 0)$ gives the first part of the proposition. For the second part, simply set $db > 0$ then rearrange will yield the result. This completes the proof. ■

Proof of Proposition 4: It is sufficient to verify that $T^{-1}\sum_{t=1}^{T}(-1)^{I_{it}}(u_{jt} - u_{it}) < 0$ will ensure that the j element in the relative cost vector, $\mathbf{c}'_D - \mathbf{c}'_B\mathbf{B}^{-1}\mathbf{D}$, is less than zero. Without loss of generality, let $i = 1$, define $\mathbf{x}_B = (\mathbf{w}'_B, \mathbf{v}'_B, 1)'$ and $\mathbf{x}_D\left(\mathbf{w}'_D, \mathbf{v}'_D, \mathbf{0}'_{k-1}\right)'$ such that $\mathbf{w}_B$ is a vector consists of elements in $\mathbf{w}$ corresponding to the positive elements in $u_{1t} > 0$. Similarly, $\mathbf{v}_B$ is a vector consists of elements in $\mathbf{v}$ corresponding to negative

elements in $u_{1t} < 0$. Therefore $\mathbf{x}_B$ forms a basic feasible solution equivalent to selecting the first model with the objective function equivalent to the mean absolute error in model 1. Define also $\mathbf{u}_{-1} = (\mathbf{u}_2, ..., \mathbf{u}_k)$ as a $T \times (k-1)$ matrix consists of all forecast errors from the remaining models.

Given this, forecast combination will improve mean absolute error, if one of the elements in $\mathbf{x}_D$ moves in to the basic feasible solution. However, this would also mean that one of the element in $\mathbf{w}_B$ and $\mathbf{v}_B$ will move out from the basic feasible solution. This means that the forecast errors at least one time period will be zero when combining forecasts.

Note that

$$\begin{aligned}
&\mathbf{c}'_D - \mathbf{c}'_B\mathbf{B}^{-1}\mathbf{D} \\
&= \begin{bmatrix} \frac{\mathbf{i}'_T}{T} & \mathbf{0}'_{k-1} \end{bmatrix} - \begin{bmatrix} \frac{\mathbf{i}'_T}{T} & 0 \end{bmatrix} \begin{bmatrix} K^\alpha \mathbf{I} & \mathbf{u}_1 \\ \mathbf{0}'_T & 1 \end{bmatrix}^{-1} \begin{bmatrix} k^\beta \mathbf{I} & \mathbf{u}_{-1} \\ \mathbf{0}'_T & \mathbf{i}'_{k-1} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\mathbf{i}'_T}{T} & \mathbf{0}'_{k-1} \end{bmatrix} - \begin{bmatrix} \frac{\mathbf{i}'_T}{T} & 0 \end{bmatrix} \begin{bmatrix} K^\alpha \mathbf{I} & K^\beta \mathbf{I}\mathbf{u}_1 \\ \mathbf{0}'_T & 1 \end{bmatrix} \begin{bmatrix} k^\beta \mathbf{I} & \mathbf{u}_{-1} \\ \mathbf{0}'_T & \mathbf{i}'_{k-1} \end{bmatrix} \\
&= \begin{bmatrix} 2\frac{\mathbf{i}'_T}{T} & -\frac{\mathbf{i}'_T}{T} K^\alpha \mathbf{I} \left(\mathbf{u}_{-1} - \mathbf{u}_1 \mathbf{i}'_{k-1} \right) \end{bmatrix}
\end{aligned}$$

The last $k-1$ elements in the vector can be written as

$$T^{-1} \sum_{t=1}^{T} (-1)^{I_{1t}} (u_{jt} - u_{1t}),$$

and therefore, forecast combination can reduce mean absolute error if $\exists j \neq i$ such that

$$\delta_j = T^{-1} \sum_{t=1}^{T} (-1)^{I_{it}} (u_{jt} - u_{it}) < 0.$$

This completes the proof. ■

REFERENCES

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

Clark, T. and West, K. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1-2):155–186. cited By (since 1996)81.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291 – 311. 50th Anniversary Econometric Institute.

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.

Timmermann, A. (2006). *Forecast Combinations*. Elsevier, Amsterdam.

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