Too similar to combine? On negative weights in forecast combination

Peter Radchenko[a], Andrey L. Vasnev[a] and Wendun Wang[b]

[a] *University of Sydney, New South Wales, Australia*
[b] *Erasmus University Rotterdam and Tinbergen Institute, the Netherlands*

April 1, 2021

Abstract:
This paper provides the first thorough investigation of the negative weights that can emerge when combining forecasts. The usual practice in the literature is to consider only convex combinations and to ignore or trim negative weights, i.e., set them to zero. This default strategy has its merits, but it is not optimal. We study the problem from a variety of different angles, and the main conclusion is that negative weights emerge when highly correlated forecasts with similar variances are combined. In this situation, the estimated weights have large variances, and trimming reduces the variance of the weights and improves the combined forecast. The threshold of zero is arbitrary and can be improved. We propose an optimal trimming threshold, i.e., an additional tuning parameter to improve forecasting performance. The effects of optimal trimming are demonstrated in simulations. In the empirical example using the European Central Bank Survey of Professional Forecasters, we find that the new strategy performs exceptionally well and can deliver improvements of more than 10% for inflation, up to 20% for GDP growth, and more than 20% for unemployment forecasts relative to the equal-weight benchmark.

Key words: forecast combination, optimal weights, negative weight, trimming, shrinkage

JEL Codes: C53, C58

Corresponding author:
Andrey Vasnev
The University of Sydney Business School
Room 4087, Abercrombie Building (H70)
Sydney, NSW 2006
Australia
E-mail: andrey.vasnev@sydney.edu.au

1 Introduction

The benefits of forecast combination are well known and well documented; see Timmermann (2006) for an exhaustive literature review and Elliott and Timmermann (2016) for a recent and detailed treatment. The recent M4 competition by Makridakis et al. (2018) shows that the overwhelming majority of the most accurate methods are combinations. It is quite likely that combinations will become default forecasting methods and benchmarks in the forecasting literature that follows the M4 competition.

The issue of negative weights (or weights outside the $[0, 1]$ interval[1]) that can emerge when combining forecasts is largely ignored in the forecast combination literature. Negative weights are usually set to zero either after estimation (see, e.g., Smith and Wallis, 2009) or using an optimization constraint (see, e.g., Post et al., 2019). Throughout this paper, we will refer to those approaches as *trimming*. While it often results in better empirical performance of the combined forecast, its theoretical properties have not been studied. In which situations can trimming improve the combined forecast? Is it always optimal to trim weights to zero? If not, how do we determine the optimal trimming threshold?

This paper offers the first comprehensive study of the negative weights and trimming. First, we study the theoretical conditions for negative weights to emerge in the unconditional framework of Bates and Granger (1969) and in the conditional framework of Gibbs and Vasnev (2018). In the unconditional framework, the negative weights are driven by high positive correlations. In the conditional framework, the same effect can also be observed if several forecasts conditionally under- or overestimate the true value. This interpretation is the most natural explanation for the negative weights to appear in practice. Another important observation is that the region where negative values are theoretically optimal is unstable, i.e., a small estimation error of the underlying parameters can result in significant changes in the weights and the forecast.

We then investigate the effect of estimation and trimming using the framework of Claeskens et al. (2016) and find the usual tradeoff between variance and bias. The positive effect of trimming comes from the reduction in the variance of the estimated weights, i.e., stabilization, but the threshold of zero is arbitrary and can be improved. We investigate five different versions of trimming: three two-step methods where the optimal weights are estimated and then trimmed, and two one-step methods where the trimming constraint is imposed in the estimation. Our fifth version is based on the portfolio selection method with gross-exposure constraints by Fan et al. (2012). While this method has been used in portfolio

[1]Cases with negative weights and weights larger than 1 are symmetric, so without loss of generality, we concentrate on the case of negative weights.

management (see, e.g., Ledoit and Wolf, 2017; Fan et al., 2016), we are the first to apply it to forecast combination and to derive the asymptotic distribution of the estimated weights, which is of interest in its own right.

Finally, as we find in a simulation study that the default strategy of trimming weights at zero is rarely optimal, we propose an optimal trimming threshold, i.e., an additional tuning parameter to deliver better forecasting performance. The new optimal trimming delivers solid improvements in our empirical application using the European Central Bank Survey of Professional Forecasters. In many cases, the optimal threshold is different from zero, and even a small relaxation can result in improvements above 10% relative to the traditional approach of trimming at zero. Using the optimal threshold can deliver improvements of more than 10% for inflation, up to 20% for GDP growth, and more than 20% for unemployment forecasts relative to the equal-weight benchmark (which is often used in practice).

The remainder of this paper is organized as follows. In Section 2, we investigate the issue of negative weights in the classical unconditional framework of Bates and Granger (1969), assuming that the weights are fixed. In Section 3, we extend our analysis to the conditional framework of Gibbs and Vasnev (2018). In Section 4, we obtain additional insights from the regression framework of Granger and Ramanathan (1984). In Section 5, we analyse the properties of the forecast combination when the weights are estimated and the estimation is explicitly taken into account as in Claeskens et al. (2016). Our suggestion for the optimal threshold is given in Section 6. Section 7 provides a simulation study similar to Smith and Wallis (2009). Section 8 presents an empirical illustration similar to Matsypura et al. (2018), and some concluding remarks are offered in Section 9. All proofs are provided in Appendix A, additional empirical results are in Appendix B, and additional simulation results are in Appendix C. The code for all methods proposed in this paper is freely available online[2].

2 Classical framework

This section investigates what drives the weights to be negative in the *classical* framework of Bates and Granger (1969). We first analyze combinations with two candidate forecasts in Section 2.1 to shed light on the intuition in Section 2.2 and then consider the general case with multiple forecasts in Section 2.3. Despite the possibility of negative weights, we maintain the restriction that all weights sum to one in this section.

[2] https://bit.ly/2PkTTKn

2.1 Negative weights in the case of two forecasts

We consider a linear combination of two forecasts y_1 and y_2 of an event μ:

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

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

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

where σ_1^2 and σ_2^2 are the variances of y_1 and y_2, respectively and $\rho = \mathrm{corr}(y_1, y_2)$ denotes the correlation. The optimal weight that minimizes $\mathrm{var}(y_c)$ is

$$w^* = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{2}$$

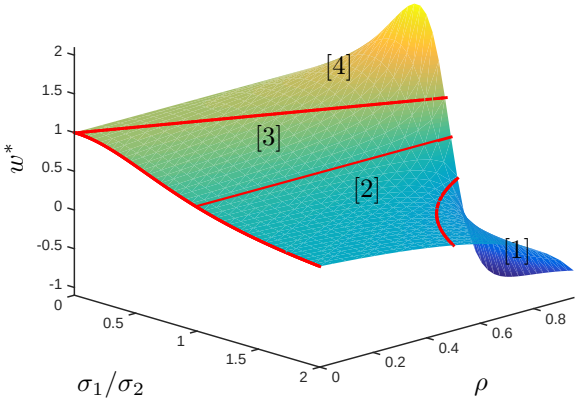
The theoretical results in this subsection are well known and derived in Bates and Granger (1969). We simply adopt them for our notation and present them visually in Figure 1.

For visual analysis, it is practical to study how w^* depends on ρ (the correlation between the two forecasts) and σ_1/σ_2 (the ratio of their variances). To this end, we rewrite w^* as

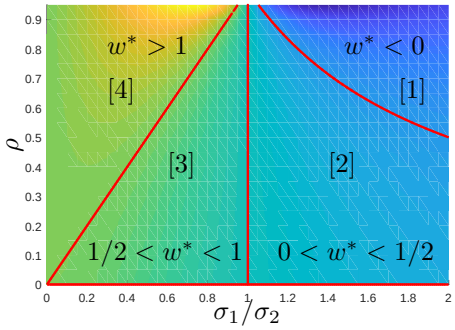
$$w^* = \frac{1 - \rho(\sigma_1/\sigma_2)}{(\sigma_1/\sigma_2)^2 + 1 - 2\rho(\sigma_1/\sigma_2)}. \tag{3}$$

Figure 1(a) depicts the three-dimensional plot of the optimal weight w^* as a function of ρ and σ_1/σ_2, and the view of this surface from the top is presented in Figure 1(b). Area [1] indicates the region where w^* is negative, and area [4] indicates the region where w^* is larger than 1 (implying that the weight of the second forecast $(1 - w^*)$ is negative). The nonnegative weights are in area [2], where $0 < w^* < 1/2$, and in area [3], where $1/2 < w^* < 1$. The lines on the surface indicate border cases. The line dividing [1] and [2] corresponds to $w^* = 0$, the line dividing [2] and [3] to $w^* = 1/2$ and $\sigma_1 = \sigma_2$, and the line dividing [3] and [4] to $w^* = 1$.

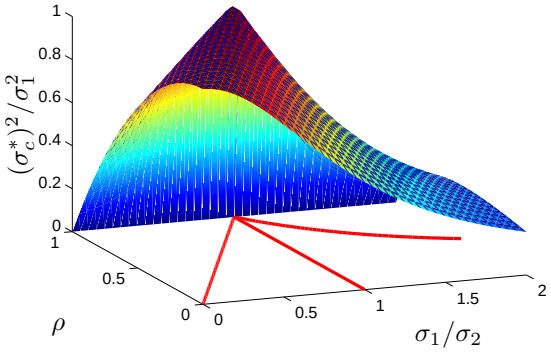
Figures 1(a) and 1(b) show how ρ and σ_1/σ_2 determine the values of optimal weights and reveal several interesting situations. First, negative optimal weights emerge when the correlation between the two forecasts is positive and large, i.e., $w^* < 0$ if $\rho > \sigma_2/\sigma_1$, or alternatively, $1 - w^* < 0$ if $\rho > \sigma_1/\sigma_2$. In contrast, the optimal weights are always positive (between 0 and 1) when the correlation ρ is zero (or negative). Second, when the correlation equals the variance ratio, one of the forecasts is not selected into the combination, i.e., $w^* = 0$ if $\rho = \sigma_2/\sigma_1$, and $w^* = 1$ if $\rho = \sigma_1/\sigma_2$. If the variance of one of the forecasts approaches zero, then the combination simply selects that forecast, e.g., $w^* \to 1$ when $\sigma_1/\sigma_2 \to 0$.



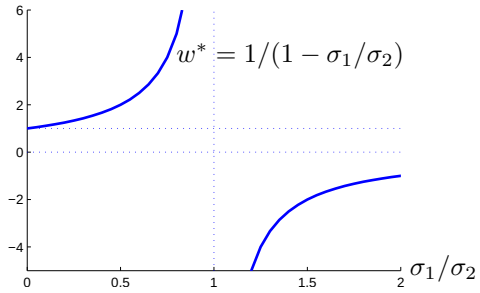
(a) Optimal weight w^* as a function of ρ and σ_1/σ_2.



(b) Optimal weight w^* as a function of ρ and σ_1/σ_2 (top view).



(c) Optimal variance $(\sigma_c^*)^2$ as a function of ρ and σ_1/σ_2.



(d) Optimal weight w^* as a function of σ_1/σ_2 when $\rho = 1$.

Figure 1: Optimal weight w^* and optimal variance $(\sigma_c^*)^2$ functions.

Finally, a discontinuity is observed when $\sigma_1 = \sigma_2$ and $\rho = 1$. Figure 1(d) depicts the optimal weight $w^* = 1/(1 - \sigma_1/\sigma_2)$ for perfectly correlated forecasts ($\rho = 1$). The optimal weight is always outside the $[0, 1]$ interval. When we go from situation $\sigma_1 < \sigma_2$ to situation $\sigma_1 > \sigma_2$, the weight flips from positive to negative. The area close to the discontinuity, i.e., when ρ is close to 1 and the variance σ_1 and σ_2 are close to each other, is very unstable, with minor changes in the parameters resulting in large changes in w^*; see Winkler and Clemen (1992) for the sensitivity analysis of the weight function given by (3).

The variance of the optimal forecast can be analysed in a similar way. Figure 1(c) shows the scaled version of the variance surface as a function of ρ and σ_1/σ_2:

$$(\sigma_c^*)^2/\sigma_1^2 = \frac{1 - \rho^2}{(\sigma_1/\sigma_2)^2 + 1 - 2\rho(\sigma_1/\sigma_2)}.$$

This theoretical surface is well behaved (and the combination always has superior properties) even in the regions when the optimal weight w^* is outside the $[0, 1]$ interval. However, this behaviour is not evidenced in the empirical and simulation literature, as the surface is distorted when w^* has to be estimated, which will be covered in Section 5.

2.2 Intuition behind negative weights

We understand from Section 2.1 that negative weights emerge when the correlation between the forecasts is high. We now explore why negative weights are useful in this situation. In the two-forecast framework, a negative weight implies that the combined forecast lies outside the range bounded by the two forecasts, which can be particularly useful when both forecasts over- or underestimate the event. This intuition can be best illustrated in the following special cases.

First, we consider the case of perfectly correlated forecasts, i.e., $\rho = 1$, and we assume $\sigma_1 > \sigma_2$ without loss of generality. Then, the optimal weight that minimizes $\text{var}(y_c)$ is always negative and given by

$$w^\dagger = -\frac{\sigma_2}{\sigma_1 - \sigma_2}.$$

In this situation, y_1 and y_2 essentially contain the same information, but y_1 is less reliable and farther away from μ due to its higher variance. In other words, both forecasts over- or underestimate μ, providing enough evidence to choose $y_c < y_2$ if $y_2 < y_1$ (overestimation case) or to choose $y_c > y_2$ if $y_2 > y_1$ (underestimation case).[3] The purpose of the negative weight is to obtain a combined forecast y_c

[3]Sometimes it is better to ignore y_1. For example, if $y_2 = \mu + \sigma_2\epsilon_2$ and $y_1 = y_2 + \epsilon_1$ (with uncorrelated ϵ_1 and ϵ_2, i.e., there is no new information in y_1 but only additional noise), then the optimal weight $w^* = 0$ and $y_c = y_2$, which is a sufficient statistic for μ.

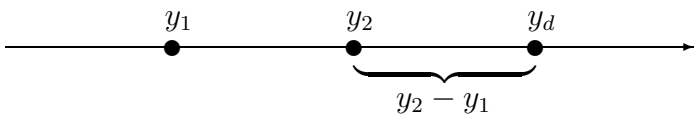
that lies outside the range bounded by y_1 and y_2.

Next, we consider the case in which the two forecasts are positively but not perfectly correlated, i.e., $0 < \rho < 1$, and when the optimal weight is moderately negative, i.e., $-1 < w^* < 0$. Assume that $y_1 < y_2$ without loss of generality. The negative optimal weight implies that $1 > \rho > \sigma_2/\sigma_1$, further suggesting that $\sigma_2 << \sigma_1$. Due to its large variance and high correlation with y_2, y_1 provides limited extra information and appears less reliable than y_2. Thus, both forecasts underestimate μ, and a good forecast should be greater than y_2. On the other hand, given that $-1 < w^* < 0$, y_1 is still useful because it indicates how much we need to correct y_2 and contributes to defining the upper bound of the combined forecast. In this case, the benefit of negative weights can be illustrated by rearranging $y_c = w^* y_1 + (1 - w^*) y_2$ as

$$y_c = \bar{w}^* y_d + (1 - \bar{w}^*) y_2,$$

where $\bar{w}^* = -w^* \in [0, 1]$ and $y_d = y_2 + (y_2 - y_1)$. Here, y_d can be interpreted as an upper bound for y_c given the information of y_1 and y_2. Now, y_c can be regarded as a combination of y_2 and y_d with weights in the [0,1] interval, as demonstrated in Figure 2.[4]

Figure 2: Rearrangement of combined forecasts with negative weights



Alternatively, one can also rewrite y_c as $y_c = y_2 + \bar{w}^*(y_2 - y_1)$, which explicitly shows that the combination should not be located between y_1 and y_2 because y_1 underestimates μ more than y_2, but the difference $y_2 - y_1$ provides information on the magnitude of correction. The geometry of this rearrangement is transparent in the special case when one forecast is a linear function of the other and presented in Figure 3. Further examples can be found in Magnus and De Luca (2016), in their Section 5, including a practical problem of inflation forecast. However, in the main part of their analysis, Magnus and De Luca (2016) submit to the general practice of keeping weights between zero and one.

[4] y_1 can always play a corrective role as long as the weight is finitely negative. For example, if $-2 < w^* < -1$, then the combination y_c comprises between $y_2 + (y_2 - y_1)$ and $y_2 + 2(y_2 - y_1)$ with weights in the [0,1] interval.

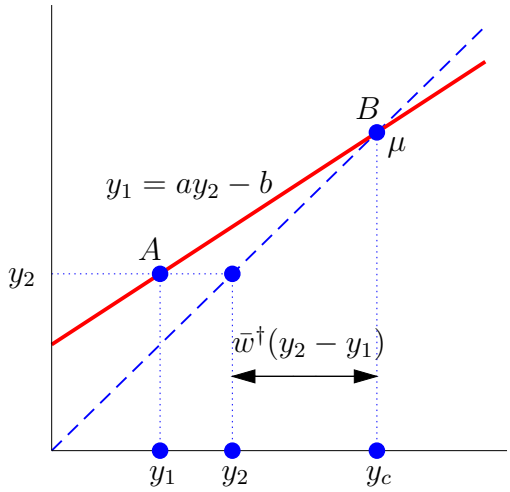


Figure 3: Two linearly dependent forecasts y_1 and y_2 are represented by point A on the solid line $y_1 = ay_2 - b$ with $a > 1$. They both underestimate μ, which is represented by point B. The dashed line represents the 45-degree line that allows us to project all points on the horizontal axis. In this case, $y_c = \mu = \frac{b}{a-1}$, the optimal weight $w^\dagger = -\frac{\sigma_2}{\sigma_1 - \sigma_2} = -\frac{1}{a-1}$, and $\bar{w}^\dagger = \frac{1}{a-1}$.

2.3 Negative weights in the case of multiple forecasts

We now consider the multivariate case when n forecasts are available to combine. Denote the vector of forecasts $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and the vector of fixed weights $\boldsymbol{w} = (w_1, \ldots, w_n)'$ that sum to one. If the original forecasts are unbiased and have variance Σ, then the linear combination

$$y_c = \boldsymbol{w}'\boldsymbol{y}$$

is unbiased and has variance $\boldsymbol{w}'\Sigma\boldsymbol{w}$, which has a minimum at

$$\boldsymbol{w}^* = \frac{\Sigma^{-1}\boldsymbol{\imath}}{\boldsymbol{\imath}'\Sigma^{-1}\boldsymbol{\imath}},$$

where $\boldsymbol{\imath}$ is the vector of ones.

Without loss of generality, we assume that all negative weights are collected at the beginning of $\boldsymbol{w}^*$ if there are any, so that we can partition $\boldsymbol{w}^*$ as

$$\boldsymbol{w}^* = \begin{pmatrix} \boldsymbol{w}^*_- \\ \boldsymbol{w}^*_+ \end{pmatrix},$$

where $\boldsymbol{w}^*_+$ and $\boldsymbol{w}^*_-$ are the vectors that contain all positive and (possibly) negative weights, respectively. Accordingly, we can partition the covariance matrix Σ and

unit vector $\boldsymbol{\imath}$ as

$$\Sigma = \begin{pmatrix} \Sigma_{--} & \Sigma_{-+} \\ \Sigma_{+-} & \Sigma_{++} \end{pmatrix} \text{ and } \boldsymbol{\imath} = \begin{pmatrix} \boldsymbol{\imath}_- \\ \boldsymbol{\imath}_+ \end{pmatrix}.$$

Using the inversion formula for the block matrix Σ, we can derive

$$\boldsymbol{w}_-^* = \frac{1}{\boldsymbol{\imath}'\Sigma^{-1}\boldsymbol{\imath}} \left(E^{-1}\boldsymbol{\imath}_- - E^{-1}\Sigma_{-+}\Sigma_{--}^{-1}\boldsymbol{\imath}_+ \right), \tag{4}$$

where $E = \Sigma_{--} - \Sigma_{-+}\Sigma_{++}^{-1}\Sigma_{+-}$ is the Schur complement.

We can see that the matrix of cross covariances Σ_{-+} plays a critical role here. First, if $\Sigma_{-+} = 0$, then $\boldsymbol{w}_-^*$ cannot be negative, which is similar to the two-dimensional case with uncorrelated forecasts. Second, if all elements of Σ_{-+} are negative, then $\boldsymbol{w}_-^*$ cannot be negative, as all elements of $E^{-1}\Sigma_{-+}\Sigma_{--}^{-1}\boldsymbol{\imath}_+$ in (4) are negative. The general condition when all elements of $\boldsymbol{w}_-^*$ are negative is given by the following proposition.

Proposition 2.1. *The vector of optimal weights contains negative elements* $\boldsymbol{w}_-^* < 0$ *if and only if* $E^{-1}\boldsymbol{\imath}_- < E^{-1}\Sigma_{-+}\Sigma_{--}^{-1}\boldsymbol{\imath}_+$ *elementwise.*

This condition is less intuitive than its two-dimensional version, but even here, we can observe that large positive elements of Σ_{-+} will make negative weights more likely to emerge.

The proposition would be clearer if we considered a special case in which there is only one negative weight, i.e., $\boldsymbol{w}_-^* = w_1^* < 0$ and $\boldsymbol{w}_+^* = (w_2^*, \ldots, w_n^*)' > 0$ elementwise. In this case,

$$\Sigma = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{+-}' \\ \boldsymbol{\sigma}_{+-} & \Sigma_{++} \end{pmatrix}$$

and $E = \sigma_{11} - \boldsymbol{\sigma}_{+-}'\Sigma_{++}^{-1}\boldsymbol{\sigma}_{+-}$ is a scalar, where $\boldsymbol{\sigma}_{+-}'$ is the covariance between the y_1 and the remaining forecasts $(y_2, \ldots, y_n)$, so Equation (4) can be simplified

$$w_1^* = \frac{1 - \boldsymbol{\sigma}_{+-}'\Sigma_{++}^{-1}\boldsymbol{\imath}_+}{1 - 2\boldsymbol{\sigma}_{+-}'\Sigma_{++}^{-1}\boldsymbol{\imath}_+ + E\boldsymbol{\imath}_+'\Sigma_{++}^{-1}\boldsymbol{\imath}_+ + (\boldsymbol{\sigma}_{+-}'\Sigma_{2++}^{-1}\boldsymbol{\imath}_+)^2}. \tag{5}$$

Following Proposition 2.1, we can obtain the condition when the weight w_1^* is negative as

$$w_1^* < 0 \Longleftrightarrow 1 < \boldsymbol{\sigma}_{+-}'\Sigma_{++}^{-1}\boldsymbol{\imath}_+.$$

Here, it is obvious that large positive elements of $\boldsymbol{\sigma}_{+-}$ will drive the negative weight. If Σ_{++} is diagonal, that is, the forecasts $y_2, \ldots, y_n$ are uncorrelated with each other but correlated with y_1 then

$$\boldsymbol{\sigma}_{+-}'\Sigma_{++}^{-1}\boldsymbol{\imath}_+ = \sum_{j=2}^{n} \sigma_{1j}/\sigma_{jj} = \sum_{j=2}^{n} \rho_{1j}\sigma_{11}^{1/2}\sigma_{jj}^{1/2}/\sigma_{jj} = \sigma_{11}^{1/2}\sum_{j=2}^{n} \rho_{1j}/\sigma_{jj}^{1/2},$$

which can easily exceed 1, especially if n is large.[5]

An alternative necessary and sufficient condition[6] for the existence of negative weights can be derived using the adjoint matrix. Let E_j be the $n \times (n-1)$ matrix obtained from the identity matrix I_n by deleting the jth column. Then $E_i'\Sigma E_j$ is the $(n-1) \times (n-1)$ matrix obtained from Σ by deleting row i and column j.

Proposition 2.2. *All elements of the vector of optimal weights $\boldsymbol{w}^*$ are nonnegative if and only if*

$$(-1)^j \sum_{i=1}^{n} (-1)^i |E_i'\Sigma E_j| \geq 0 \tag{6}$$

for all j.

Proof. See Appendix A. □

In the special case $n = 2$, the condition reduces to $\sigma_{12} \leq \min(\sigma_{11}, \sigma_{22})$, a well-known result. In the special case $n = 3$, the condition can be written in terms of variances $\sigma_i^2 = \sigma_{ii}$ and correlations $r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ as

$$\begin{aligned}
(1 - r_{12}^2)\sigma_1\sigma_2 - (r_{23} - r_{12}r_{13})\sigma_1\sigma_3 - (r_{13} - r_{12}r_{23})\sigma_2\sigma_3 &\geq 0, \\
-(r_{23} - r_{12}r_{13})\sigma_1\sigma_2 + (1 - r_{13}^2)\sigma_1\sigma_3 - (r_{12} - r_{13}r_{23})\sigma_2\sigma_3 &\geq 0, \\
-(r_{13} - r_{12}r_{23})\sigma_1\sigma_2 - (r_{12} - r_{13}r_{23})\sigma_1\sigma_3 + (1 - r_{23}^2)\sigma_2\sigma_3 &\geq 0.
\end{aligned}$$

In general, the above analysis shows that negative weights emerge when candidate forecasts are highly and positively correlated. Typically, this occurs when the forecasts simultaneously over- or underestimate the true value. An important advantage of negative weights is to correct joint over- or underestimation and allow the combination to lie outside the range of candidate forecasts in this situation.

3 Conditionally optimal weight framework

Thus far, we have analysed the situations in which negative weights can emerge in the classical unconditional framework of Bates and Granger (1969). In some cases, an additional information set is available and can be used to predict the forecasting errors that further influence the optimal weights. The information set can include, for example, past forecasting errors or variables that were used or

[5] Note that, as in all other cases, if the first forecast is uncorrelated with all other forecasts, i.e., $\boldsymbol{\sigma}_{+-} = \mathbf{0}$, then $w_1^* = 1/(1 + \sigma_{11}\boldsymbol{\imath}_+'\Sigma_{++}^{-1}\boldsymbol{\imath}_+)$, which is always inside the $[0, 1]$ interval.

[6] The authors are extremely grateful to Jan Magnus for suggesting this alternative.

not used in constructing forecasts. In this section, we investigate what makes the weights to be negative in the conditional framework of Gibbs and Vasnev (2018), where this additional information is used to predict forecasting errors.

For the purpose of this presentation, we stay in the framework of combining two forecasts, but we now need to explicitly introduce time in our notation. Assume that we need to forecast the future realization of μ_{T+1} and there are two individual forecasts $y_{1,T}$ and $y_{2,T}$ that are based on the information available at time T. Similar to the previous section, we use a linear combination $y_{c,T} = wy_{1,T} + (1 - w)y_{2,T}$. Assume further that an information set I_T is available that can be used to predict the forecasting errors $e_{1,T+1} = \mu_{T+1} - y_{1,T}$ and $e_{2,T+1} = \mu_{T+1} - y_{2,T}$. The information set I_T contains variables available at time T that might or might not be used to construct the original forecasts $y_{1,T}$ and $y_{2,T}$. For example, if the previous forecasting errors are used, then $I_T = \{e_{1,T}, e_{2,T}\}$.

The errors can now be decomposed as

$$e_{1,T+1} = b_{1,T} + \xi_{1,T+1}$$

and

$$e_{2,T+1} = b_{2,T} + \xi_{2,T+1},$$

where $b_{1,T} = \mathrm{E}(e_{1,T+1}|I_T)$, $b_{2,T} = \mathrm{E}(e_{2,T+1}|I_T)$ and $\mathrm{E}(\xi_{1,T+1}|I_T) = \mathrm{E}(\xi_{2,T+1}|I_T) = 0$. In our example with the previous forecasting errors, $b_{1,T} = \phi_1 e_{1,T}$ and $b_{2,T} = \phi_2 e_{2,T}$ capture the first-order correlation in the forecasting errors using AR(1) models, while $\sigma^2_{\xi_1} = \mathrm{var}(\xi_{1,T+1}|I_T)$ and $\sigma^2_{\xi_2} = \mathrm{var}(\xi_{2,T+1}|I_T)$ are the conditional variances.

Conditional on I_T, the error of the combined forecast $e_{c,T+1} = \mu_{T+1} - y_{c,T} = we_{1,T+1} + (1 - w)e_{2,T+1}$ will have a nonzero expectation:

$$\mathrm{E}(e_{c,T+1}|I_T) = wb_{1,T} + (1 - w)b_{2,T}$$

and its variance

$$\mathrm{var}(e_{c,T+1}|I_T) = w^2\sigma^2_{\xi_1} + (1 - w)^2\sigma^2_{\xi_2} + 2w(1 - w)\rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2},$$

where $\rho_{\xi_1,\xi_2} = \mathrm{corr}(\xi_{1,T+1}, \xi_{2,T+1}|I_T)$, and we use the fact that $\mathrm{var}(b_{1,T}|I_T) = \mathrm{var}(b_{2,T}|I_T) = 0$.

In this situation, one should consider minimizing the conditional mean squared forecasting error (MSFE) to balance both the bias and the variance components:

$$\mathrm{MSFE}(w|I_T) = (wb_{1,T} + (1-w)b_{2,T})^2 + w^2\sigma^2_{\xi_1} + (1-w)^2\sigma^2_{\xi_2} + 2w(1-w)\rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2}$$

and the optimal solution in this case is

$$w^*(I_T) = \frac{\sigma^2_{\xi_2} - \rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2} + b^2_{2,T} - b_{1,T}b_{2,T}}{\sigma^2_{\xi_1} + b^2_{1,T} + \sigma^2_{\xi_2} + b^2_{2,T} - 2\rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2} - 2b_{1,T}b_{2,T}}. \tag{7}$$

We will refer to $w^*(I_T)$ as *the conditionally optimal weight*, and I_T explicitly indicates the information set used for conditioning.

The conditional weight formula (7) resembles the unconditional formula (2) but contains new terms: $b_{2,T}^2$ and $-b_{1,T}b_{2,T}$ in the numerator and $b_{1,T}^2$, $b_{2,T}^2$ and $-2b_{1,T}b_{2,T}$ in the denominator. The denominator cannot be negative, and the numerator has two parts. The first classical part $\sigma_{\xi_2}^2 - \rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2}$ can become negative if $\rho_{\xi_1,\xi_2} > \sigma_{\xi_2}/\sigma_{\xi_1}$. This case is covered in detail in Section 2 and means that the individual forecasts are highly correlated. The second conditional part $b_{2,T}^2 - b_{1,T}b_{2,T}$ is negative if $b_{1,T}b_{2,T} > b_{2,T}^2 > 0$. This happens when the additional information suggests that $b_{1,T}$ and $b_{2,T}$ are of the same sign, and it is an indication that both forecasts are expected to over-/underestimate μ_{T+1}. In this situation, $w^*(I_T)$ is more likely to be negative than in the classical framework.

An important message of the conditional analysis is that negative optimal weights are even more frequent than in the unconditional framework. If the additional information suggests that the candidate forecasts are conditionally biased in the same direction, then negative conditional optimal weights can occur even under weak correlation. For example, even when $\rho_{\xi_1,\xi_2} = 0$, the conditional optimal weight

$$w^*(I_T) = \frac{\sigma_{\xi_2}^2 + b_{2,T}^2 - b_{1,T}b_{2,T}}{\sigma_{\xi_1}^2 + b_{1,T}^2 + \sigma_{\xi_2}^2 + b_{2,T}^2 - 2b_{1,T}b_{2,T}}$$

can be negative if $b_{1,T}b_{2,T} > \sigma_{\xi_2}^2 + b_{2,T}^2$. This was not possible in the classical framework with zero correlation.

4 Regression framework

Granger and Ramanathan (1984) showed that the combination weights can also be estimated via regressions with or without restrictions. Thus, we investigate the possibility of negative weights in such a regression framework. Their most general model is a regression without constraints (Method C in their paper). This method attains the best theoretical minimum sum of squared forecast errors (though the empirical results in the subsequent literature are mixed), and it is useful for our demonstration.

The linear regression of the historical individual forecasts $y_{1,t}$ and $y_{2,t}$ on the known realizations of μ_{t+1} can be formulated as

$$\mu_{t+1} = w_0 + w_1 y_{1,t} + w_2 y_{2,t} + e_{c,t+1}, \tag{8}$$

and it may produce negative weights. For example, consider the OLS estimator of the first coefficient

$$\widehat{w}_1 = \frac{\sum (y_{2,t}^{dm})^2 \sum y_{1,t}^{dm}\mu_{t+1}^{dm} - \sum y_{1,t}^{dm}y_{2,t}^{dm} \sum y_{2,t}^{dm}\mu_{t+1}^{dm}}{\sum (y_{1,t}^{dm})^2 \sum (y_{2,t}^{dm})^2 - (\sum y_{1,t}^{dm}y_{2,t}^{dm})^2}, \tag{9}$$

where deviations from the mean $y_{1,t}^{dm} = y_{1,t} - \bar{y}_1$, $y_{2,t}^{dm} = y_{2,t} - \bar{y}_2$, and $\mu_{t+1}^{dm} = \mu_{t+1} - \bar{\mu}$ are used for compactness of the formula (with the averages $\bar{y}_1$, $\bar{y}_2$, and $\bar{\mu}$ computed using the historical forecasts and realizations). The denominator and $\sum (y_{2,t}^{dm})^2$ cannot be negative. $\sum y_{1,t}^{dm} \mu_{t+1}^{dm}$ and $\sum y_{2,t}^{dm} \mu_{t+1}^{dm}$ are likely to be positive because the individual forecasts are constructed or selected to predict the target variable. If $\sum y_{1,t}^{dm} y_{2,t}^{dm}$ is negative, i.e., the two forecasts are negatively correlated, then $\widehat{w}_1$ cannot be negative. However, if the two forecasts are positively correlated, then $\widehat{w}_1$ can be negative, and a negative weight is more likely to appear if the forecasts are highly correlated.

The fact that highly correlated forecasts are responsible for negative weights is the same as before, but the new angle provides us with additional insights. If $y_{1,t}$ and $y_{2,t}$ are highly correlated, then regression (8) suffers from imperfect multicollinearity[7], and the estimated coefficients will have a large variance

$$\text{var}(\widehat{w}_1) = \frac{1}{T} \left(\frac{1}{1 - \rho^2} \right) \frac{\text{var}(e_c)}{\text{var}(y_1)},$$

and be highly negatively correlated, $\text{corr}(\widehat{w}_1, \widehat{w}_2) = -\rho$. The large variance of the estimated weight when ρ is close to 1 was first discovered by Winkler and Clemen (1992). The large estimation error in the weights will affect the performance of the combination. This effect is studied in the next section.

5 Estimated weights and trimming

The previous analysis exposed a rich set of situations in which negative weights are the optimal choice, at least in theory when all parameters are known. However, the underlying parameters are typically unknown in practice and need to be estimated, causing estimation errors in the weights. In this section, we bring the estimation of the weights explicitly into the analysis, and we examine the effect of trimming, which is widely used in forecast combinations.

5.1 Forecast combination using estimated weights

To analyse the influence of estimating unknown weights on the properties of combination, we employ the framework of Claeskens et al. (2016) that allows for random weights. To make the argument as transparent as possible, we return to the unconditional framework covered in Section 2 and assume that the estimation of the weight is done independently from the individual forecasts (some comments

[7]Lichtendahl Jr. and Winkler (2020) suggest to exclude one of the forecasts from the combination due to redundancy.

about the general case are provided at the end of this subsection). We still linearly combine two forecasts of an event μ:

$$y_c = \widehat{w} y_1 + (1 - \widehat{w}) y_2,$$

but now $\widehat{w}$ is random because it is estimated from the data. Since the mean squared forecast error (MSFE) is one of the most common criteria to evaluate the forecasts, we mainly focus on studying the MSFE of combined forecasts. The bias-variance tradeoff regarding weight estimation and trimming also applies to other similar criteria, such as mean absolute forecast error (MAFE).

If $\widehat{w}$ is independent of y_1 and y_2, then the combination will remain unbiased, namely, $\mathrm{E}\, y_c = \mu$, and the variance is given by

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

Note that the independence assumption is critical to preserve unbiasedness. For the variance, the first three terms are similar to the terms in Equation (1) with the difference being that we now need to use $\mathrm{E}\, \widehat{w}$, while the last term $\operatorname{var}(\widehat{w}) \operatorname{var}(y_1 - y_2)$ is new. This additional term is critical in understanding the effect of the weight estimation. In the presence of estimation error for the underlying parameters and weights, this additional variance term is positive and sometimes can be large. A bias in $\widehat{w}$ does not bias y_c, but a positive $\operatorname{var}(\widehat{w})$ always results in an upward shift of the variance curve and produces a suboptimal combination (in terms of minimum variance). This situation is demonstrated by Fig. 1 in Claeskens et al. (2016).

In the general framework of Claeskens et al. (2016), when the estimated weights and forecasts are correlated, the formulae have additional terms, but the main lesson is the same. Weight estimation brings additional noise that inflates the variance of the combination; therefore, variance reduction techniques (e.g., trimming) can be beneficial. In addition, if $\widehat{w}$ is correlated with $y_1 - y_2$, the combination will be biased from the beginning, as demonstrated by Fig. 2 in Claeskens et al. (2016), so trimming will simply modify the existing bias.

5.2 Effect of trimming

Existing studies involving forecast combinations often restrict weights to be within the [0,1] interval, which can be done by trimming the negative weights to zero after estimation (see, e.g., Smith and Wallis, 2009) or using an optimization constraint (see, e.g., Post et al., 2019). This approach generally results in an upward bias of the weights while simultaneously reducing their variances. In this section, we study the simplified setting to isolate the effect of trimming negative weights on

the combination, i.e., we consider

$$w^{\mathrm{TR}} = \begin{cases} \widehat{w}, & \widehat{w} \geq 0 \\ 0, & \widehat{w} < 0 \end{cases} . \tag{10}$$

In this setting, trimming is equivalent to imposing a nonnegativity constraint during the estimation. More general situations are considered in Section 6.

For simplicity, we assume that the initial weight estimator is unbiased, i.e., $\mathrm{E}\,\widehat{w} = w^*$; then, the trimmed weight w^{TR} is upward biased because $\mathrm{E}\,w^{\mathrm{TR}} > \mathrm{E}\,\widehat{w} = w^*$ but at the same time has a lower variance than the original estimated weight, i.e., $\mathrm{var}(w^{\mathrm{TR}}) < \mathrm{var}(\widehat{w})$. Thus, the overall effect of trimming negative weights on combination is determined by the relative size of the bias and variance reduction, both caused by trimming. Such a bias-variance tradeoff further depends on the discrepancy between $\widehat{w}$ and the threshold 0 as well as the variance of the initial estimated weight.

Figure 4 shows the above argument in graphical terms similar to the figures of Claeskens et al. (2016). The original weight $\widehat{w}$ allows us to reach the lowest point R but on a higher curve (due to the variance caused by estimation), while w^{TR} allows us to reach R^{TR}, which is not the lowest point (due to the bias), but its curve is lower (due to the variance reduction). As a result, trimming will achieve better combined forecast performance, i.e., smaller MSFE. The relative position of the middle curve is controlled by the variance reduction, i.e., the greater the variance reduction is, the lower the curve. The bias affects the distance between R^{TR} and the optimal point on the middle curve, i.e., the greater the bias is, the less optimal the combination. The variance reduction and the bias can be controlled by using a trimming threshold different from 0.

The effect of trimming is similar to the effect of using equal weights. As analysed in Elliott (2011) and Claeskens et al. (2016), equal weights do not coincide with theoretical optimal weights in general, but the main benefit of using the fixed equal weights is the reduction of the estimation error, i.e., their variance is reduced to zero. Thus, the overall effect of using equal weights on the MSFE of combination depends on the difference between the equal and optimal weights as well as the variance reduction. In situations when the theoretical optimal weight is negative, the bias of the equal weights could be large, so using fixed weights is not beneficial. However, using trimming (especially with varying thresholds) allows us to simultaneously extract the benefits of the variance reduction and to control the bias.

Jagannathan and Ma (2003) also discuss the role of nonnegativity constraints from different perspectives in the portfolio allocation setup. They show that the nonnegativity constraint is equivalent to unconstrained optimization with a shrunk covariance matrix. This is due to the variance of the element whose nonnegativity constraint is binding as well as its covariance with other elements that are

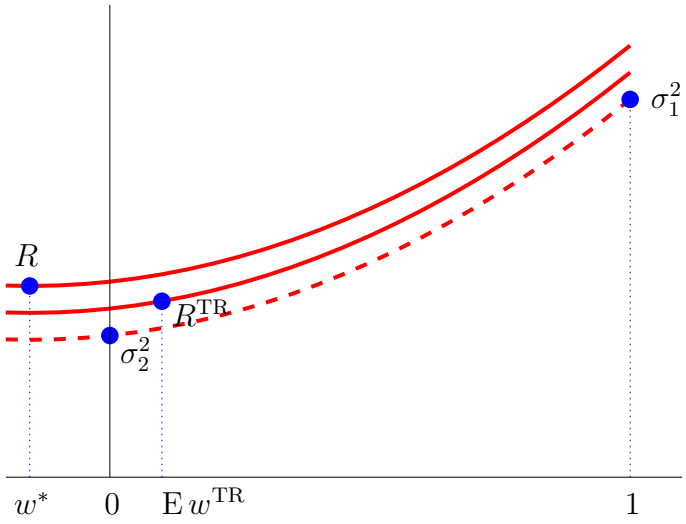


Figure 4: The effect of trimming when $w^* < 0$. The bottom dashed curve represents the scenario with fixed weights, and two dots correspond to the original forecasts y_1 and y_2. The top curve contains R, which represents the combination with the initial weight $\widehat{w}$. The middle curve contains R^{TR}, which represents the combination with the trimmed weight w^{TR}.

reduced. Shrinkage methods are generally beneficial in forecast combinations; see Roccazzella et al. (2020). Hence, imposing nonnegative constraints can be useful even when they are not true in the population.

6 Relaxing the nonnegativity constraint

Our previous analysis shows that, on the one hand, there are situations where negative weights are theoretically optimal. On the other hand, when weights are estimated, trimming (which does not allow the weights to be negative) can improve the performance of the combined forecast if the variance reduction by trimming dominates the bias. This suggests that a better way to balance the trade off between the bias and variance is to relax the nonnegativity assumption and use some negative threshold in trimming. The main advantage of using a negative threshold is to control the degree of bias while still enjoying variance reduction. We first consider trimming methods for a prespecified trimming threshold (where we consider two-step and one-step options) and then discuss how to choose the optimal threshold.

6.1 Trimming methods for a given threshold

There are two options for forcing the weights to be above a prespecified threshold $-c$ for some $c > 0$. First, we can estimate the optimal weights and then trim those that are below $-c$. We refer to this option as two-step trimming methods. Second, we can simultaneously estimate and trim the weights using constrained optimization. We refer to this option as one-step trimming methods. In this section, we consider the general case with n weights. We apply one common threshold across all weights, but there might be situations where one can benefit from different thresholds for different weights.

6.1.1 Two-step trimming methods

We consider three methods in this class. First, we consider a simple trimming method that truncates any weight less than $-c$. In other words, this method simply replaces zero in (10) with a negative parameter $-c$ for each element $\widehat{w}_i$ of the estimated vector of weights $\widehat{\boldsymbol{w}} = (\widehat{w}_1, ..., \widehat{w}_n)'$, i.e.,

$$w_i^{\text{TR1}} = \alpha_1 \times \begin{cases} \widehat{w}_i, & \widehat{w}_i > -c \\ -c, & \widehat{w}_i \leq -c \end{cases}, \tag{11}$$

where an additional scaling parameter α_1 is required to satisfy the constraint $(\boldsymbol{w}^{\text{TR1}})'\imath = 1$. The resulting weights are biased but with a lower variance $\text{var}(\boldsymbol{w}^{\text{TR1}}) < \text{var}(\widehat{\boldsymbol{w}})$. Thus, they can potentially provide the benefits of trimming while still maintaining the flexibility of being negative. Note that, in this case, the minimum weight is $-\alpha_1 c > -c$ due to rescaling if $c > 0$. The minimum trimmed weight is exactly $-c$ only when $c = 0$.

It is possible to strictly control the minimum weights and guarantee them to be precisely $-c$. In the second method, scaling is applied only to the untrimmed weights, i.e.,

$$w_i^{\text{TR2}} = \begin{cases} \alpha_2 \times \widehat{w}_i, & \widehat{w}_i > -c \\ -c, & \widehat{w}_i \leq -c \end{cases}, \tag{12}$$

where the scaling parameter α_2 is chosen to satisfy the constraint $(\boldsymbol{w}^{\text{TR2}})'\imath = 1$. Again, the variance of such trimmed weights will decrease. A disadvantage of the above methods (11) and (12) is that all weights below the threshold are treated in the same way. This is not always satisfactory since the magnitude of those weights provides additional information that can be included during the trimming process.

Thus, motivated, we propose the third trimming method as

$$w_i^{\text{TR3}} = \begin{cases} \alpha_3 \times \widehat{w}_i, & \widehat{w}_i > -c \\ \frac{-c}{\min \widehat{w}_i} \times \widehat{w}_i, & \widehat{w}_i \leq -c \end{cases}. \tag{13}$$

The scaling of the weights below the threshold sets the smallest weight to be exactly $-c$, while all other weights below the threshold are mapped to the $(-c, 0)$ interval. This version preserves the ratio between the weights, i.e., if $\widehat{w}_i \leq \widehat{w}_j \leq -c$, then $\widehat{w}_i/\widehat{w}_j = w_i^{\text{TR3}}/w_j^{\text{TR3}}$ (and similarly for the weights above the threshold). In other words, if forecast i had a heavier weight than forecast j before trimming, it will remain so after trimming. Again, α_3 is chosen to satisfy the constraint $(\boldsymbol{w}^{\text{TR3}})'\imath = 1$.

6.1.2 One-step trimming methods

In contrast to the first class of methods that utilize a given weight $\widehat{\boldsymbol{w}}$, the second class of methods estimates and trims the weights in a single, joint step via constraint optimization. We consider two methods in this class, differing in their constraints.

First, we can directly impose the minimum weight restriction on the optimization problem for the weight estimation as

$$\begin{aligned} \boldsymbol{w}^{\text{TR4}} = \arg\min_{\boldsymbol{w}} \ & \boldsymbol{w}'\Sigma\boldsymbol{w} \\ & \boldsymbol{w}'\imath = 1 \\ & w_i \geq -c. \end{aligned} \tag{14}$$

Second, we propose to employ an L_1-norm constraint in the optimization problem, drawing insights from Fan et al. (2012). Instead of restricting the minimum weights using (14), we can restrict the L_1-norm of the weights $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{n} |w_i|$,

$$\begin{aligned} \boldsymbol{w}^{\text{TR5}} = \arg\min_{\boldsymbol{w}} \ & \boldsymbol{w}'\Sigma\boldsymbol{w} \\ & \boldsymbol{w}'\imath = 1 \\ & \|\boldsymbol{w}\|_1 \leq 1 + \tilde{c}. \end{aligned} \tag{15}$$

If $\tilde{c} = 0$, the resulting weights are nonnegative. If $\tilde{c} = +\infty$, then this constraint drops out, and the weights have no bounds. For $0 < \tilde{c} < +\infty$, the weights can be negative but always bounded. Note that $\tilde{c}$ itself is not the minimum weight that we are prepared to tolerate, but of course, they are associated with each other. For example, $\tilde{c} = 1$ implies that the weights cannot be less than -0.5.

Unlike constraint (14), this L_1-norm constraint (15) also imposes an upper bound on the weight. As shown by Jagannathan and Ma (2003), imposing upper-bound restrictions in constrained optimization (15) is equivalent to the same unconstrained optimization but with an inflated covariance matrix, $\widetilde{\Sigma}$, which contains a larger variance and covariance for forecasts whose upper-bound constraint is binding. Since a large weight is typically assigned to a forecast with low covariances with others, the occurrence of large weights, when the candidate forecasts

are actually highly correlated, may indicate large estimation errors. In this case, the adjustment induced by the upper-bound constraints reduces the estimation error, achieving similar effects as shrinkage, and thus improves the performance of the combination.

A range of useful properties of optimization problem (15) was explored in Fan et al. (2012); see their Theorems 1–3, including the effect of using an estimated covariance matrix $\widehat{\Sigma}$ instead of the unknown Σ. We complement their investigation with the study of the asymptotic distribution of the estimated weights.

Let $\widehat{\boldsymbol{w}}$ denote the solution to the sample version of problem (15), in which Σ is replaced with the sample covariance matrix $\widehat{\Sigma}$. Given a weight vector $\boldsymbol{w}$, we write $\boldsymbol{w}_{-1} = (w_2, ..., w_n)'$ and similarly define $\widehat{\boldsymbol{w}}_{-1}$ for the estimated weight and $\boldsymbol{w}^*_{-1}$ for the theoretically optimal solution. Since the weights in $\widehat{\boldsymbol{w}}$ are required to sum to one, we focus on deriving the asymptotic distribution of $\widehat{\boldsymbol{w}}_{-1}$.

Theorem 6.1. *Suppose that $\boldsymbol{w}^*$ is the unique solution to the optimization problem* (15) *and that regularity assumptions A1–A4 in Appendix A are satisfied. Let $\boldsymbol{Z} \sim \mathcal{N}(0, V)$ and $\tilde{\boldsymbol{Z}} \sim \mathcal{N}(0, SVS')$, where the positive definite matrices V and S are defined in Appendix A.*

If the L_1-norm constraint is not binding, i.e., $\|\boldsymbol{w}^*\|_1 < 1 + \tilde{c}$, then

$$\sqrt{T}(\widehat{\boldsymbol{w}}_{-1} - \boldsymbol{w}^*_{-1}) \xrightarrow{d} \boldsymbol{Z}.$$

Alternatively, if the L_1-norm constraint is binding and $\|\boldsymbol{w}^*\|_1 = 1 + \tilde{c}$, then

$$\sqrt{T}(\widehat{\boldsymbol{w}}_{-1} - \boldsymbol{w}^*_{-1}) \xrightarrow{d} S^{-1} Proj_{\mathcal{C}} \tilde{\boldsymbol{Z}},$$

where $Proj_{\mathcal{C}} \tilde{Z}$ denotes the projection of $\tilde{Z}$ onto a convex cone $\mathcal{C}$, which is the tangent cone of the constraint set at the point $\boldsymbol{w}^*_{-1}$, defined in Appendix A.

Proof. See Appendix A. □

The theorem provides a useful result that can be used, for example, to construct confidence intervals or to test whether the weights are significantly different from the equal weights.

6.1.3 Numerical example

To demonstrate the effect of the above five trimming methods, let us consider a numerical example when we have three forecasts with $\sigma_1^2 = 1$, $\sigma_2^2 = 3$, $\sigma_3^2 = 5$, and all correlations $\rho_{ij} = 0.9$. The optimal weight $\boldsymbol{w}^* = (1.6, -0.2, -0.4)'$ has two negative components. If $c = 0.1$, then $\boldsymbol{w}^{\text{TR1}} = (1.14, -0.07, -0.07)'$ with the minimum weight above the trimming threshold, $\boldsymbol{w}^{\text{TR2}} = (1.2, -0.1, -0.1)'$ with all trimmed

weights set to the trimming value. The option $\boldsymbol{w}^{\mathrm{TR3}} = (1.15, -0.05, -0.1)'$ discriminates between the weights below the threshold. The optimization problems deliver $\boldsymbol{w}^{\mathrm{TR4}} = (1.2, -0.1, -0.1)'$ and $\boldsymbol{w}^{\mathrm{TR5}} = (1.1, 0, -0.1)'$ (with $\tilde{c} = 1.3226$ chosen to match the minimum weights of the previous options). Figure 5 gives a graphical representation of this numerical example in the trilinear coordinates.

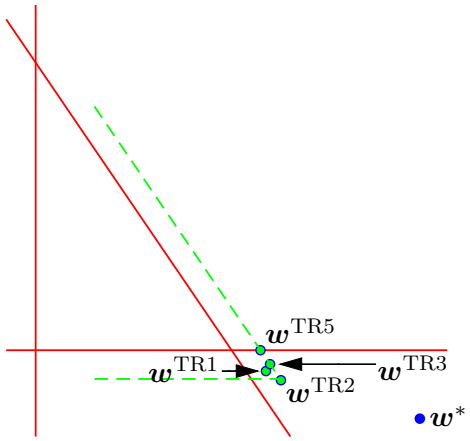


Figure 5: Numerical example in the trilinear coordinates. The solid vertical line corresponds to the weights with $w_1 = 0$, the solid horizontal line to $w_2 = 0$, and the solid crossing line to $w_3 = 0$. The dashed lines outline the border of the constraint weights set with $w_i > -c$. The trimmed versions $\boldsymbol{w}^{\mathrm{TR2}}$, $\boldsymbol{w}^{\mathrm{TR3}}$, and $\boldsymbol{w}^{\mathrm{TR5}}$ of the original weight $\boldsymbol{w}^{*}$ are on the border, and $\boldsymbol{w}^{\mathrm{TR1}}$ is inside.

6.2 Data-driven threshold

The parameter c (or $\tilde{c}$) can be specified by researchers, and then the choice of the parameters determines the bias-variance tradeoff induced by trimming. A natural next step is to let the data speak for itself and find a data-driven method to choose the trimming threshold.

One possibility is to select a threshold based on pseudo out-of-sample MSFE.[8] In this case, we divide the available data into two parts, $[1, \lfloor \tau T \rfloor] - 1]$ and $[\lfloor \tau T \rfloor, T]$, where $0 < \tau < 1$ and $\lfloor \cdot \rfloor$ take the closest integer. The first part ($\lfloor \tau T \rfloor - 1$ periods) is used to estimate the covariance matrix and the weight vector $\widehat{\boldsymbol{w}}$, and the second part ($\lfloor (1-\tau)T \rfloor$ periods) is used to compute the pseudo out-of-sample MSFE as

$$\mathrm{MSFE}(c, \tau) = \frac{1}{\lfloor (1-\tau)T \rfloor} \sum_{t=\lfloor \tau T \rfloor}^{T-1} (\boldsymbol{y}_t' \boldsymbol{w}^{\mathrm{TR}}(c) - \mu_{t+1})^2, \tag{16}$$

[8] If researchers evaluate the performance of forecasts via the MAFE, they can also select a threshold based on MAFE with the similar procedure.

where $\boldsymbol{y}_t' = \{y_{1,t}, \ldots, y_{n,t}\}'$ is the vector of individual forecasts of μ_{t+1} available at time t and $\boldsymbol{w}^{\mathrm{TR}}(c)$ is the vector of trimmed weights given the threshold parameter $-c$ using one of the methods from Section 6.1.

To increase the stability and reliability of the MSFE evaluation, we follow Fan and Yao (2003) and Fan et al. (2012) and consider the average MSFE (AMSFE) over a series τ_k as

$$\mathrm{AMSFE}(c) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{MSFE}(c, \tau_k), \tag{17}$$

where K is a prespecified number that increases with T. The optimal threshold is chosen by minimizing the AMSFE function,

$$c^* = \arg\min_{c} \mathrm{AMSFE}(c). \tag{18}$$

Finally, the optimal threshold can be used to construct the combination $\boldsymbol{y}_T'\boldsymbol{w}^{\mathrm{TR}}(c^*)$ to forecast μ_{T+1} out of sample.

The estimation of the optimal threshold will of course increase the variance of the weights (and the combination) relative to the case with the fixed threshold. Monte Carlo simulations in Section 7 and our empirical example in Section 8 demonstrate that the benefits of relaxing the nonnegativity constraint outweigh any costs associated with finding the optimal threshold.

7 Monte Carlo illustration

In this section, we demonstrate the effect of trimming via Monte Carlo simulations. We base our simulation design on Smith and Wallis (2009) with a small modification[9], so that the optimal theoretical weight can be negative. The modification allows us to simplify the data generating process. We draw a sequence of $T+1$ observations from a strictly stationary AR(1) process

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

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

$$\sigma_z^2 = \mathrm{var}(z_t) = \frac{1}{1-\phi_1^2}$$

[9]The data generating process is simplified from AR(2) to AR(1); the first forecast is the naïve 'no-change' forecast from Case 1, and the second forecast is based on a two-period lag from Case 2. This modification is sufficient to produce the negative optimal weights, while the original design of Smith and Wallis (2009) produces only positive theoretical optimal weights and sometimes negative estimated weights.

and the first two autocorrelation coefficients are

$$\rho_1 = \mathrm{corr}(z_t, z_{t-1}) = \phi_1, \qquad \rho_2 = \mathrm{corr}(z_t, z_{t-2}) = \phi_1^2.$$

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

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

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

Since the forecast z_{T+1} is random rather than fixed, we define $e_{1t} = z_t - z_{t-1}$ and $e_{2t} = z_t - \rho_2 z_{t-2}$, and consider the forecast errors

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

Their variances are

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

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

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

The optimal weight is negative if $\rho > \sigma_2/\sigma_1$, which is equivalent to $\phi_1 < 0$ in this case.

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

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

Our purpose is to better understand the uncertainty caused by the estimation of weights and the effect of trimming. We estimate the theoretically optimal weight w^* as

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

but we do *not* estimate the parameter ϕ_1; it is set to its true value. Any uncertainty shown in the simulations is therefore caused by weight estimation.

In the case of two forecasts, the different trimming versions introduced in Section 6.1 are equivalent, and we can simply consider

$$w^{\mathrm{TR}}(c) = \max(\widehat{w}, -c) = \begin{cases} \widehat{w}, & \widehat{w} > -c, \\ -c, & \widehat{w} \le -c \end{cases} \tag{19}$$

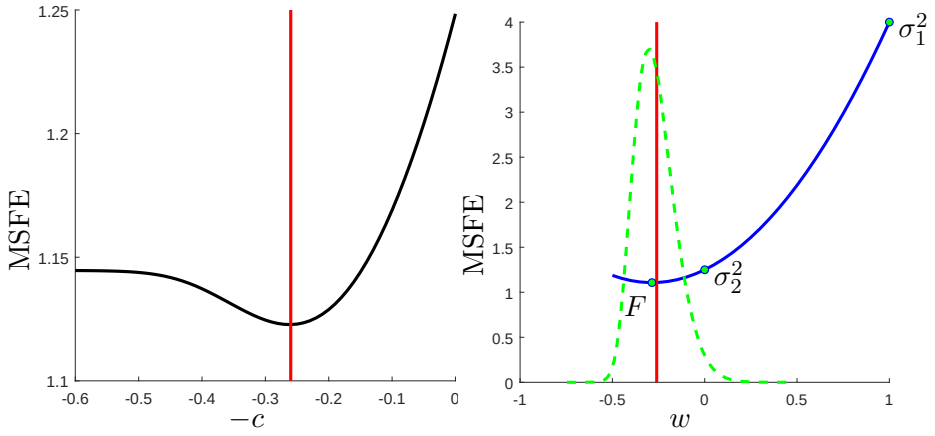


Figure 6: Optimal trimming for case $\phi_1 = -0.5$. The left subplot shows the MSFE as a function of the threshold $-c$, and the vertical line indicates the optimal threshold -0.26. The right subplot shows the same vertical line, the density of the estimated $\widehat{w}$ is given by the dashed line, and the solid line shows MSFE as a function of a fixed weight with point F indicating the theoretically optimal weight $w^* = -0.29$ and two other points indicating the original forecasts.

for a range of different thresholds between -1 and 0.

The experiment is repeated 1,000,000 times, which suffices to control the simulation error. For a given ϕ_1, each run produces a value of $\widehat{w}$ and w^{TR}, and the values of the forecast errors $e_1 = z_{T+1} - y_1$, $e_2 = z_{T+1} - y_2$, and $e_c = z_{T+1} - y_c$ for both weights. Finally, we compute the MSFE across all runs.

Figure 6 gives a detailed illustration for the case of $\phi_1 = -0.5$. It shows MSFE as a function of the threshold, the optimal threshold, the theoretically optimal weight and the density of the estimated weight (Figure C.1 in Appendix C also shows the variance and the bias components). We can see that the optimal trimming (-0.26) is very close to the theoretically optimal weight (-0.29), so the bias in the trimmed estimator is small. The estimated optimal weight has variance 0.0125, which is reduced to 0.0049 in the estimator with the optimal trimming. This variance reduction is the main explanation for the improved performance. In contrast, when trimming is done at zero, the variance is reduced to 0.0001, but the bias is 0.2868 and negates the benefits of the variance reduction.

The proximity of the theoretical optimal weight w^* and the optimal trimming threshold reveals another interesting fact: we should trim when $\widehat{w} < w^* < 0$, but not when $w^* < \widehat{w} < 0$. This finding is consistent with the optimal trimming of a normal random variable[10] and the following more general proposition.

[10] If $\widehat{w} \sim N(w^*, \sigma_w^2)$, then $\mathrm{E}(w^{\mathrm{TR}}(c) - w^*)^2$ is minimized for $-c = w^*$, i.e., it is optimal to trim only $\widehat{w} < w^*$.

Proposition 7.1. *Suppose that* $MSFE(w) = E(y_c - z_{T+1})^2$ *has a minimum at* w^* *and is non-decreasing for* $w \geq w^*$. *The minimum* w^* *is estimated with* $\widehat{w}$, *such that* $\mathrm{E}(\widehat{w}) = w^*$, *and its trimmed version* $w^{TR}(c)$ *constructed using threshold* $-c$. *Then, function* $E\big[MSFE\big(w^{TR}(c)\big)\big]$ *achieves its minimum when* $-c = w^*$.

Proof. See Appendix A. □

In empirical studies, neither the optimal weight nor the optimal threshold is available. However, finding a trimming threshold is arguably an easier task than estimating w^* as it does not involve estimation and inversion of the covariance matrix.

It is useful to compare MSFE of the trimming methods with the performance of the true optimal weights. The difference between the optimal trimming (-0.26) and the theoretically optimal weight (-0.29) is 10.3%, but the difference between the MSFE using the optimal trimming (1.123) and the MSFE using the theoretically optimal weight (1.107) is only 1.4%. This is due to the fact that we are very close to the minimum, so the first derivative is close to zero, and we only see second-order effects in the MSFE. Nevertheless, the difference is statistically significant. The optimal threshold minimizes the gap between MSFEs, but it does not bridge it.

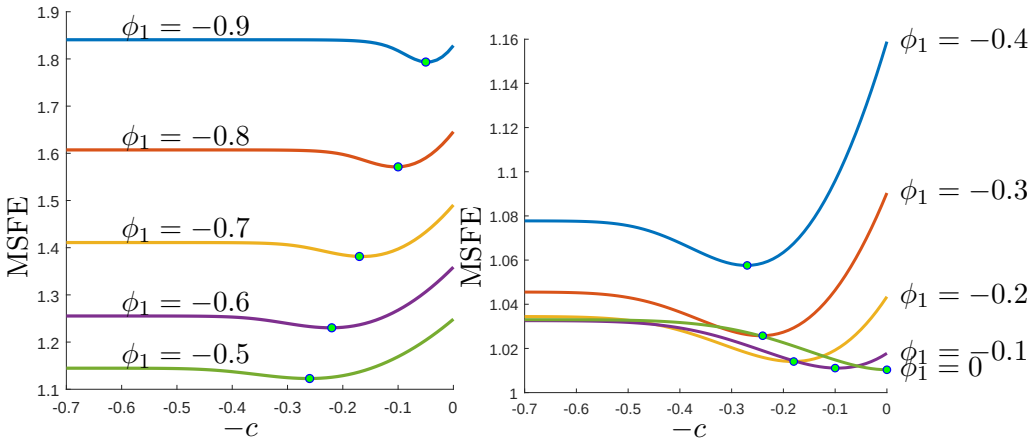


Figure 7: For each value of ϕ_1, the solid line shows the MSFE as a function of the threshold $-c$, and the point indicates the optimal trimming parameter that achieves the minimum on this curve. The left panel shows the curves for $\phi = -0.9, \ldots, -0.5$, while the right panel shows the curves for $\phi = -0.4, \ldots, 0.0$. The results are shown in two panels with different scales to maximize the visibility and clarity of the results.

Figure 7 shows how the MSFE curve and the optimal trimming change when different values of ϕ_1 are used. We can see that there are three possible scenarios:

1. Optimal trimming is zero ($\phi_1 = 0$).

2. Optimal trimming is not zero, but trimming at zero is better than no trimming ($\phi_1 = -0.1$, $\phi_1 = -0.9$).

3. Optimal trimming is not zero, but trimming at zero is worse than no trimming ($-0.8 \leq \phi_1 \leq -0.2$).

The second scenario is the most common situation in our empirical illustration, i.e., trimming at zero is beneficial relative to no trimming, but the optimal trimming delivers even better performance. Figure C.2 in Appendix C gives the simulation results for the case of 3 forecasts. Even though the curves are more complicated and contain multiple local minima, the main conclusion remains the same: we can improve the forecasting performance by choosing a threshold different from zero.

8 Empirical illustration

We illustrate our theoretical discussions using the European Central Bank (ECB) Survey of Professional Forecasters (SPF). The SPF provides quarterly forecasts on inflation (HICP), real GDP growth (RGDP) and the unemployment rate (UNEM) since 1999. In this paper, we employ the data from Q4 1999 to Q2 2018. We follow Matsypura et al. (2018) to focus on the one- and two-year-ahead forecast horizons and evaluate the performance of the combination based on the last 4 years (16 quarters) observations from Q2 2014 with expanding windows. There are approximately 100 forecasters that participate in the survey, although a number of forecasters do not respond at every period. Thus, the data constitute an unbalanced panel. To facilitate the computation of forecast combination and avoid outliers, we remove forecasters who do not respond for at least 24 quarters (6 years) following Matsypura et al. (2018). To compare the forecasts with the actual outcomes, we collect the ECB macroeconomic indices and use the final revision of the data whenever possible.

Our theoretical analysis in Sections 2–4 suggests that negative weights are more likely to appear when candidate forecasts are highly correlated or when the forecasts simultaneously over-/underestimate the true values. Thus, we first examine, for each variable, how often the true values fall outside the range of all expert forecasts. Figure 8 plots the time series of expert forecasts and the true value for the three variables. Of the three macroeconomic variables, RGDP seems the most difficult to forecast, with true values lying outside the forecast range in 83% of the periods for the 1-year-ahead forecast, and 81% for the 2-year-ahead forecast. HICP has the second-highest misforecasting percentage, with 43% and 51% of true values falling outside the range for the 1-year- and 2-year-ahead forecasts,

respectively. UNEM has the lowest percentage, while there are still 37% and 43% of periods when all forecasts are simultaneously over- or underestimated. These statistics suggest that the forecasts generally exhibit a high degree of similarity, and it further implies a high correlation between forecasts, leading to a large probability of the presence of negative weights.[11] From a different perspective, since the true values often lie outside the range of forecasts, a convex combination of candidate forecasts (using weights between 0 and 1) does not seem appropriate. In this case, negative weights that concavely combine the candidate forecasts are expected to produce forecasts close to the true values.

To further understand the behaviour of optimal weights, we examine the empirical distribution of the optimal weights, estimated from

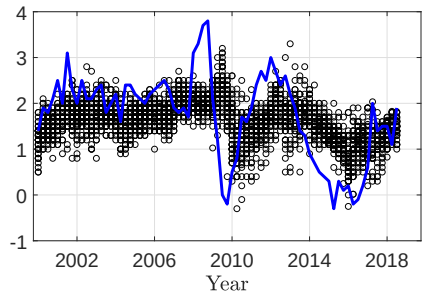
$$\boldsymbol{w}^* = \frac{\Sigma^{-1}\boldsymbol{\imath}}{\boldsymbol{\imath}'\Sigma^{-1}\boldsymbol{\imath}}.$$

To compute the optimal weights, we first need to estimate the covariance matrix Σ. Since there still exist missing observations after filtering out infrequent forecasters, we follow Matsypura et al. (2018) to compute the covariance matrix using the overlapping periods between each pair of forecasters. Specifically, let $\mathcal{T}_i \subseteq \{1, 2, ..., T\}$ be the set of periods in which a forecast is available for the ith forecaster. Then, the covariance matrix can be obtained by

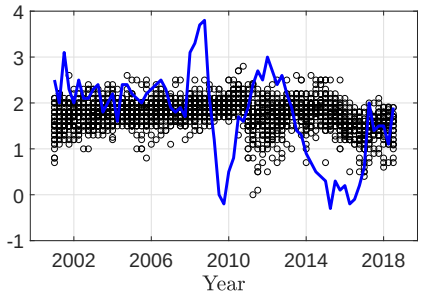
$$\hat{\sigma}_{ij} = \begin{cases} \frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}[y_{it} - E(y_{it})]^2 & \text{if } i = j \\ \frac{1}{|\mathcal{T}_i\cap\mathcal{T}_j|}\sum_{t\in\mathcal{T}_i\cap\mathcal{T}_j}[y_{it} - E(y_{it})][y_{jt} - E(y_{jt})] & \text{if } i \neq j. \end{cases}$$

When there is no overlap between two forecasters, we set the covariance to zero. To guarantee the invertability of the covariance matrix, we employ the nearPD function from the R package Matrix to obtain the nearest positive definite matrix. When the number of forecasters is larger in the training sample than in the evaluation sample, we use a part of the covariance matrix that involves only the forecasters who also respond in the evaluation sample. We report the quantiles of optimal weights across all forecasters for the entire forecasting period in Table 1. In all cases, the minimum estimated optimal weights are far less than zero, especially for the 2-year-ahead forecast of RGDP, in line with previous findings that the forecasts of RGDP are often highly correlated. The negativity extends to the 0.4 quantile for all three variables, indicating a large proportion of negative weights.

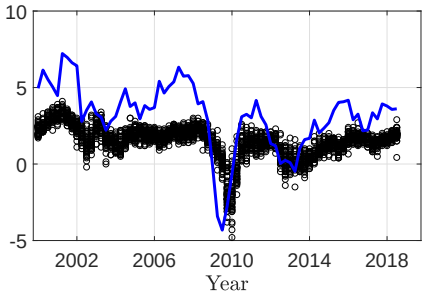
[11] This is confirmed by Matsypura et al. (2018), who showed that the partial correlations between forecaster errors are high, especially for 2-year-ahead RGDP forecasts.



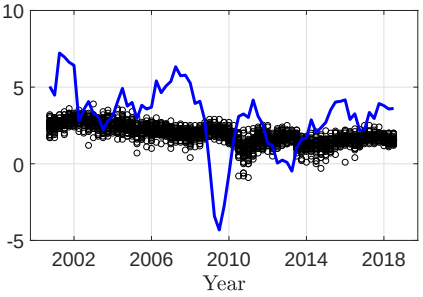
(a) HICP 1-year forecast



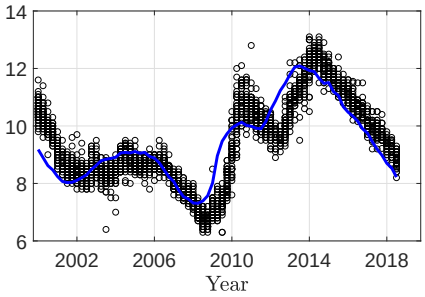
(b) HICP 2-year forecast



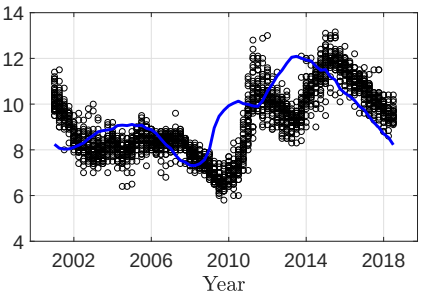
(c) RGDP 1-year forecast



(d) RGDP 2-year forecast



(e) UNEM 1-year forecast



(f) UNEM 2-year forecast

Figure 8: Time-series plots of the forecasts (∘) and true values (—).

8.1 Forecast combination using fixed trimming threshold

Given the high correlation between forecasts, the optimal weights are likely to be negative. We evaluate the performance of combined forecasts while allowing for negative weights and those using trimmed weights. The theory in Section 5 shows that trimming negative weights may lead to superior combined forecasts by decreasing estimation uncertainty. Thus we trim the negative weights using the five methods described in Section 6. We first range the threshold for trimming c from $-\infty$ (no trimming) to zero to examine how the performance varies across

Table 1: Quantiles of estimated optimal weights across all forecasters for the entire forecasting period

	1-year			2-year		
Quantile	HICP	RGDP	UNEM	HICP	RGDP	UNEM
0.0	−9.3223	−9.0726	−14.4107	−10.1253	−30.0940	−3.2778
0.1	−0.7105	−0.6153	−0.6554	−0.6201	−0.9821	−0.6978
0.2	−0.3827	−0.2678	−0.3571	−0.3437	−0.5043	−0.3319
0.3	−0.2142	−0.1335	−0.1996	−0.2317	−0.2952	−0.1744
0.4	−0.1108	−0.0326	−0.0970	−0.0889	−0.0983	−0.0632
0.5	−0.0039	0.0355	0.0149	0.0183	0.0683	0.0131
0.6	0.1256	0.1094	0.1163	0.1112	0.2223	0.1402
0.7	0.2563	0.1913	0.2343	0.2401	0.4001	0.2750
0.8	0.4365	0.3096	0.4415	0.4227	0.6264	0.5092
0.9	0.8556	0.6402	0.7600	0.7277	1.0094	0.7625
1.0	11.0675	9.1278	11.2622	7.8712	25.4851	2.7262

different threshold values. Note that the threshold imposed here is fixed and thus time-invariant. We will discuss the optimal (time-varying) threshold in the next section.

To evaluate the performance of various weighting schemes, we focus on the mean squared forecast error (MSFE). The results of using mean absolute forecast error (MAFE) are qualitatively similar and provided in Appendix B. Table 2 and 3 present the MSFE averaged across the whole testing period (16 quarters). We report the ratio of the MSFE of forecasts using trimmed weights over that using equal weights, and thus, a value smaller than 1 indicates better performance than the equal-weight combination. We also test the significance of the difference between the trimmed-weight and equal-weight combinations using the two-sided modified Diebold-Mariano (DM) test of Harvey et al. (1997). Several interesting findings deserve special attention. First, when no trimming is implemented ($c = -\infty$), the combined forecast using estimated optimal weights performs worse than the equal-weight combination (except for the case of 1-year ahead RGDP forecast), confirming the explanation of combination puzzle (Claeskens et al., 2016) that optimal weights typically do not outperform equal weights due to large estimation error. In most cases trimming negative weights to zero ($c = 0$) leads to better performance to those using equal weights (except for HICP), although the magnitude of improvement varies across different variables, horizons, and trimming methods.

Second, we find that in many cases, the MSFE first improves and then deteriorates when we vary the threshold for trimming from zero to $-\infty$. In particular,

Table 2: Relative MSFE of combined forecasts using five trimmed weights as a function of the trimming threshold c (two-step trimming)

		1-year forecast horizon			2-year forecast horizon		
	c	HICP	RGDP	UNEM	HICP	RGDP	UNEM
TR1	$-\infty$	8.959	0.874	1.065	3.814	5.339	4.625
	-5.0	4.175	0.878	0.910	2.320*	1.007	4.625
	-4.5	4.155	0.877	0.910	2.309*	1.015	4.625
	-4.0	4.139	0.876	0.910	2.296*	1.042	4.625
	-3.5	4.129	0.890	0.910	2.285*	1.039	4.625
	-3.5	4.049	0.904	0.907	2.276*	0.954	4.594
	-2.5	4.005	0.914	0.903	2.060**	0.938	4.397
	-2.0	3.196	0.870	0.814	1.742*	0.973	2.190*
	-1.5	2.197*	0.823*	0.644	1.552	0.938	1.808
	-1.0	1.655*	0.821**	0.555**	1.416	0.843	1.436
	-0.5	1.196	0.852**	0.594***	1.129	0.900*	0.735
	0.0	1.021	0.972***	0.914**	1.026	0.968	0.804***
TR2	$-\infty$	8.959	0.874	1.065	3.814	5.339	4.625
	-5.0	6.456*	0.901	0.903	2.803**	1.533	4.625
	-4.5	7.364	0.914	0.903	2.729**	1.399	4.625
	-4.0	6.720*	0.941	0.904	2.773**	1.292	4.625
	-3.5	7.185	0.937	1.227	2.685**	1.231	4.625
	-3.5	6.795*	0.931	2.624	2.595**	1.257	4.615
	-2.5	6.044*	0.925	2.571	2.473**	1.107	4.547
	-2.0	5.331*	0.911	2.017	2.302**	0.994	3.727
	-1.5	4.275*	0.873	1.606	2.124**	0.891	3.326
	-1.0	3.202*	0.849	0.852	1.898**	0.802	2.174
	-0.5	2.205**	0.871	0.547**	1.431	0.778	1.251
	0.0	1.021	0.972***	0.914**	1.026	0.968	0.804***
TR3	$-\infty$	8.959	0.874	1.065	3.814	5.339	4.625
	-5.0	5.681*	0.874	0.955	2.803**	1.002	4.625
	-4.5	5.371*	0.874	0.950	2.729**	0.990	4.625
	-4.0	5.130*	0.872	0.946	2.647**	0.984	4.625
	-3.5	4.890	0.876	0.936	2.581**	0.977	4.625
	-3.5	4.669	0.881	0.921	2.521**	0.933	4.615
	-2.5	4.475	0.887	0.914	2.396**	0.894	4.547
	-2.0	4.018	0.877	0.872	2.168**	0.878	3.537
	-1.5	3.220*	0.856	0.708	1.921**	0.865	2.657*
	-1.0	2.393*	0.841	0.568*	1.664*	0.810	1.921*
	-0.5	1.552	0.850*	0.490***	1.317	0.841	0.854
	0.0	1.021	0.972**	0.914**	1.026	0.968	0.804***

Notes: This table presents relative mean squared forecast error (MSFE) of different weight schemes, averaged across the whole testing period (16 quarters). We normalize all numbers by dividing by the MSFE of the equal-weight combination, and thus a value smaller than 1 indicates better performance than the equal-weight combination. ***, **, * indicate that the difference is significant at 1%, 5%, and 10%, respectively, based on two-sided modified DM tests.

Table 3: Relative MSFE of combined forecasts using five trimmed weights as a function of the trimming threshold c (one-step trimming)

		1-year forecast horizon			2-year forecast horizon		
	c	HICP	RGDP	UNEM	HICP	RGDP	UNEM
TR4	$-\infty$	8.959	0.874	1.065	3.814	5.339	4.625
	-5.0	5.719^{***}	0.979^{*}	1.701^{*}	4.216^{**}	2.342	7.290^{*}
	-4.5	5.169^{**}	0.934	4.950^{*}	3.944^{***}	1.586	9.406^{**}
	-4.0	3.473^{*}	1.532	4.011^{*}	4.547^{**}	1.141	8.259^{*}
	-3.5	4.350^{**}	1.163	4.337^{**}	3.992^{***}	1.920	7.705^{*}
	-3.5	3.133	1.089	2.140^{**}	3.605^{**}	1.234	6.812^{*}
	-2.5	3.503	1.183	1.702^{*}	2.345^{**}	1.338	8.115^{**}
	-2.0	3.764^{*}	0.979	1.328^{*}	2.321	1.445	5.489^{**}
	-1.5	2.161	0.957	1.575	2.112^{**}	1.184	3.339
	-1.0	1.810^{*}	0.706	0.780	1.947^{**}	1.030	1.786
	-0.5	2.005	0.710^{***}	0.484^{**}	1.578	0.713	0.906
	0.0	0.865^{*}	0.893^{***}	0.695^{***}	0.902	0.810^{***}	0.976
TR5	$-\infty$	8.959	0.874	1.065	3.814	5.339	4.625
	-5.0	1.739	0.937	0.549^{***}	1.266^{*}	0.568^{***}	1.796
	-4.5	1.594	0.751^{**}	0.567^{***}	1.342	0.546^{***}	1.666
	-4.0	1.596	0.776^{*}	0.472^{***}	1.237	0.545^{***}	1.455
	-3.5	1.544	0.778^{**}	0.450^{***}	1.201	0.540^{***}	1.321
	-3.5	1.430	0.774^{**}	0.296^{***}	1.112	0.536^{***}	1.101
	-2.5	1.238	0.776^{**}	0.276^{***}	1.032	0.538^{***}	0.926
	-2.0	1.134	0.787^{**}	0.236^{***}	1.024	0.566^{***}	0.748
	-1.5	1.027	0.831^{**}	0.198^{***}	0.993	0.603^{***}	0.589^{*}
	-1.0	0.937	0.872^{**}	0.199^{***}	0.970	0.654^{***}	0.474^{***}
	-0.5	0.896	0.888^{***}	0.303^{***}	0.961	0.741^{***}	0.530^{***}
	0.0	0.866^{*}	0.893^{***}	0.695^{***}	0.902	0.810^{***}	0.976

Notes: This table presents relative mean squared forecast error (MSFE) of different weight schemes, averaged across the whole testing period (16 quarters). We normalize all numbers by dividing by the MSFE of the equal-weight combination, and thus a value smaller than 1 indicates better performance than the equal-weight combination. ***, **, * indicate that the difference is significant at 1%, 5%, and 10%, respectively, based on two-sided modified DM tests.

the MSFE using trimmed weights decreases when we impose a stiffer trimming condition (larger thresholds). However, when the threshold is close to zero, further increasing the value of the threshold leads to a larger MSFE in many cases of RGDP and UNEM forecasting. This result implies that there may exist an optimal trimming threshold that is possibly nonzero (but close to zero), which we will examine in the next section. A certain fluctuation of performance as c increases is observed for the one-step trimming methods (TR4 and TR5) when c is highly negative because the constraints in (14) or (15) are not binding; thus, the minimum trimmed weights may fluctuate while still satisfying the constraints. When c is closer to zero, such that constraints are binding, we also find monotonic behaviour of combination for TR4 and TR5 as c increases.

Third, among the three variables, trimming (around zero) leads to the largest MSFE reduction with respect to the equal-weight combination for UNEM. For 1-year-ahead forecasting, the reduction can be more than 40% for two-step trimming methods and even 80% for TR5, and concerning 2-year-ahead forecasting, the reduction ranges from approximately 10% to approximately 40%. The largest MSFE reduction primarily mostly occurs when we set the trimming threshold $c = -0.5$ or -1. The good performance of trimmed weights for UNEM suggests that the error caused by estimating unknown weights is sizeable. With large estimation errors of optimal weights, trimming is particularly useful because it reduces such errors and thus improves forecasting efficiency. The large MSFE reduction for UNEM is also in line with our descriptive analysis above that negative optimal weights appear less often in UNEM forecasting than in HICP and RGDP, and they are also of smaller magnitude (less negative). In this case, trimming around zero does not sacrifice much bias while substantially reducing the variance.

Finally, and importantly, comparing across the five trimming methods, we find that TR5 has superior performance. For HICP, trimming at zero using TR5 leads to an MSFE ratio smaller than 1 in both the one-year and two-year horizons. In contrast, TR1, TR2, and TR3 all lead to an MSFE ratio larger than 1 when trimming at zero. For RGDP, although all trimming methods produce better combinations than equal weighting when the threshold is larger than -2, the improvement of TR5 is the most sizeable and significant. A similar comparison applies to UNEM, for which the forecasts with TR5 lead to the best combination with at most an 80% improvement over the equal-weight combination for the one-year horizon and 50% for the two-year horizon. Further examination reveals that TR5 imposes a stronger shrinkage effect on the weights, such that the trimmed weights are centred around zero. In other words, the minimum (or maximum) of the resulting weights is typically higher (or lower) than that of other methods, especially those produced by two-step trimming, and thus, the variance of estimated weights is smaller. Table B.4 in Appendix B uses Theorem 6.1 to show that the

proportion of cases where TR5 weights are statistically different from the equal weights is substantial.

Our final observation is that relatively small values of the trimming threshold c achieve the improvement. This might be due to a large number of forecasts present in the SPF. One might expect that the optimal threshold decreases when the number of forecasters increases.

8.2 Forecast combination using data-driven threshold

The previous section employs a fixed trimming threshold that is invariant over time and demonstrates that there may exist an "optimal" threshold that minimizes the MSFE in some cases. In practice, the trimming threshold is, of course, unknown, so we investigate the performance of data-driven thresholds in this section. We employ the method described in Section 6.2 to determine the threshold based on the pseudo out-of-sample AMSFE. We choose the threshold from the range between 0 to $c_{\min}$ with a step of 0.1 and divide the training sample (Q4 1999 to Q1 2014) into two subsamples, one for estimating the covariance matrix and the weight and the other for computing the out-of-sample AMSFE. We consider four partitions $\tau_k \in \{0.8, 0.85, 0.9, 0.95\}$. If multiple trimming thresholds lead to the same AMSFE, we take the largest threshold value among them. Note that this optimal trimming threshold is time-varying since the training set is expanding and the optimal weights are estimated at each time when the forecast combination is made. We consider two choices of $c_{\min}$, i.e., $c_{\min} = \{-5, -2\}$. A smaller value of $c_{\min}$ allows larger negative weights and simultaneously increases the weight variance. We will compare how the range of the trimming threshold affects forecasting performance.[12]

Table 4 presents the relative MSFE obtained from the trimmed-weight combination, where the trimming threshold is determined based on the pseudo out-of-sample AMSFE. Again, all numbers are relative to the MSFE of the equal-weight combination, and the significance of their difference is tested using the modified DM test. We highlight the best trimming method in bold for each variable. For HICP, TR4 and TR5 perform very similarly, and both produce better combinations than equal weights, but the difference is not statistically significant. For RGDP, all five methods produce significantly better combinations than the equal-weight combination, while TR5 leads to the greatest improvement of approximately 11% for the 1-year forecast and 22% for the 2-year forecast. For UNEM, again, all methods beat equal-weight combinations. The difference is particularly significant and sizeable for 1-year forecast, and TR5 leads to the largest improvement of

[12] We also consider $c_{\min} = -10$, and unreported results show that it leads to almost identical results as $c_{\min} = -5$.

Table 4: Relative MSFE of combined forecasts using trimmed weights with data-driven threshold based on pseudo out-of-sample evaluation

	1-year horizon			2-year horizon		
	HICP	RGDP	UNEM	HICP	RGDP	UNEM
			$c_{\min} = -2$			
TR1	1.243	0.915***	0.813***	1.037*	0.880***	0.832
TR2	1.166	0.934***	0.812***	1.040*	0.922***	0.881
TR3	1.161	0.895***	0.761***	1.072*	0.901***	**0.784***
TR4	**0.865**	0.896***	0.700***	**0.873**	0.810***	0.957
TR5	0.870	**0.887*****	**0.461*****	0.907	**0.781*****	0.866
			$c_{\min} = -5$			
TR1	1.243	0.900***	0.823***	1.037*	0.876***	0.832
TR2	1.166	0.931***	0.812***	1.040*	0.922***	0.881
TR3	1.161	0.912***	0.761***	1.072*	0.901***	**0.784***
TR4	**0.865**	0.893***	0.703***	**0.874**	0.810***	0.957
TR5	**0.865**	**0.884*****	**0.457*****	0.907	**0.781*****	0.866

Notes: This table presents relative mean squared forecast error (MSFE) of different weight schemes, averaged across the whole testing period (16 quarters). We normalize all numbers by dividing by the MSFE of the equal-weight combination, and thus a value smaller than 1 indicates better performance than the equal-weight combination. ***, **, * indicate that the difference is significant at 1%, 5%, and 10%, respectively, based on two-sided modified DM tests. The minimum number in each column is highlighted in bold.

approximately 54% in this case, followed by roughly 30% improvement by TR4. The order of improvement is much larger than the one observed in the previous studies (see, e.g., Matsypura et al., 2018), therefore, the practical significance of the results is an important contribution of this paper.

Compared with the *best* fixed threshold trimming, allowing for dynamic optimal thresholds does not improve forecasting performance except for 1-year UNEM forecasting. This is not surprising because estimating the dynamically optimal threshold introduces extra uncertainty and thus tends to inflate the final MSFE, especially for those variables that are difficult to forecast, such as HICP and RGDP. Nevertheless, the best trimming threshold is often unknown in practice, and a poor choice of the threshold can lead to much worse performance than equal-weight combination. The data-driven threshold approach performs almost as well as the *ex ante* best fixed threshold and robustly outperforms equal weighting. Hence, it offers a feasible way of estimating the unknown best threshold in practice.

Interestingly, allowing for a large range of searching sets of optimal thresholds ($c_{\min} = -5$) sometimes leads to a poorer combination than using a small range ($c_{\min} = -2$), e.g., for HICP forecasting. A possible explanation is that the pseudo out-of-sample evaluation leaves more room for estimation error to play a role. Typically, allowing for a wide range of searching sets introduces a larger estimation error. Unreported results (available upon request) show that the time-varying pattern of the optimal threshold is rather stable and that highly negative weights only occur at rare time points. Hence, allowing for a large searching set of trimming thresholds gains limited consistency but introduces more uncertainty.

9 Practical recommendation

We have analysed the negative weights that can emerge when combining forecasts and the effects of trimming from a variety of different angles. Our practical recommendation is to trim negative weights using an optimal trimming threshold (rather than zero). This strategy allows us to optimally balance the positive effects of variance reduction and the negative effects of bias. The code for all methods proposed in this paper is freely available online[13].

Acknowledgements

The authors are grateful to Jan Magnus and Gael Martin for stimulating conversations, and acknowledge Daniel Hsiao and Steven Vethman for excellent research assistance. This paper was presented at ISF2019 in Thessaloniki, at the Business

[13] https://bit.ly/2PkTTKn

Analytics research seminar in the University of Sydney in 2020, and at the virtual Econometric Society World Congress 2020. We thank the participants for their positive feedback.

References

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

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting 32*, 754–762.

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

Elliott, G. and A. Timmermann (2016). Forecast combinations. In *Economic Forecasting*, Chapter 14, pp. 310–344. Princeton University Press.

Fan, J., A. Furger, and D. Xiu (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics 34*, 489–503.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer-Verlag.

Fan, J., J. Zhang, and K. Yu (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association 107:498*, 592–606.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics 22*(4), 1993–2010.

Gibbs, C. G. and A. L. Vasnev (2018). Conditionally optimal weights and forward-looking approaches to combining forecasts. *working paper*.

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

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*, 281–291.

Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance 58*, 1651–1683.

Ledoit, O. and M. Wolf (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies 30*, 4349–4388.

Lichtendahl Jr., K. C. and R. L. Winkler (2020). Why do some combinations perform better than others? *International Journal of Forecasting 36*, 142–149.

Magnus, J. R. and G. De Luca (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys 30*, 117–148.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting 34*, 802–808.

Matsypura, D., R. Thompson, and A. L. Vasnev (2018). Optimal selection of expert forecasts with integer programming. *Omega 78*, 165–175.

Neudecker, H. and A. M. Wesselman (1990). The asymptotic variance matrix of the sample correlation matrix. *Linear Algebra and its Applications 127*, 589–599.

Post, T., S. Karabati, and S. Arvanitis (2019). Robust optimization of forecast combinations. *International Journal of Forecasting 35*, 910–926.

Roccazzella, F., P. Gambetti, and F. Vrins (2020). Optimal and robust combination of forecasts via contrained optimization and shrinkage. *working paper*.

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

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pp. 135–196. Elsevier, North-Holland.

Winkler, R. L. and R. T. Clemen (1992). Sensitivity of weights in combining forecasts. *Operations Research 40*, 609–614.

Appendix A: Assumptions and proofs

Proof of Proposition 2.2. Let $\Sigma = (\sigma_{ij})$ be a positive definite matrix of order $n \times n$. Let E_j be the $n \times (n-1)$ matrix obtained from I_n by deleting the jth column. Then $\Sigma_{ij} = E_i'\Sigma E_j$ is the $(n-1) \times (n-1)$ matrix obtained from Σ by deleting row i and column j. Now define

$$c_{ij} = (-1)^{i+j}|\Sigma_{ij}|.$$

The matrix $C = (c_{ij})$ is called the cofactor matrix and its transpose C' is the adjoint matrix. We have

$$\Sigma C' = C'\Sigma = |\Sigma| I_n$$

and hence

$$\Sigma^{-1} = \frac{1}{|\Sigma|} C'.$$

We wish to establish necessary and sufficient conditions such that all components of $\Sigma^{-1}\imath$ are nonnegative, which occurs if and only all components of $C'\imath$ are nonnegative.

The jth column of C is given by

$$Ce_j = (-1)^j \begin{pmatrix} (-1)^1|E_1'\Sigma E_j| \\ (-1)^2|E_2'\Sigma E_j| \\ \vdots \\ (-1)^n|E_n'\Sigma E_j| \end{pmatrix}$$

and hence the jth component of $C'\imath$ is given by

$$e_j'C'\imath = (-1)^j \sum_{i=1}^{n} (-1)^i |E_i'\Sigma E_j|.$$

Hence all components of $\Sigma^{-1}\imath$ are nonnegative if and only if

$$(-1)^j \sum_{i=1}^{n} (-1)^i |E_i'\Sigma E_j| \geq 0$$

for all j.

Assumptions for Theorem 6.1. We rewrite the criterion function in optimization problem (15) as a function of $\boldsymbol{w}_{-1}$:

$$\boldsymbol{w}'\Sigma\boldsymbol{w} = F(\boldsymbol{w}_{-1}),$$

where $\boldsymbol{w} = (w_1, \boldsymbol{w}_{-1}')' = (1 - \sum_{j=2}^{n} w_j, w_2, ..., w_n)'$. To derive the asymptotic distribution of $\widehat{\boldsymbol{w}}$, we impose the following assumptions:

A1: $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$ is a stationary ergodic m-dependent sequence;

A2: the moments of $\boldsymbol{y}_1$ exist, and are finite, up to the fourth order;

A3: the gradient vector $\nabla F(\boldsymbol{w}^*_{-1})$ is zero;

A4: the Hessian matrix $\nabla^2 F(\boldsymbol{w}^*_{-1})$ is non-singular.

The first two assumptions are needed to establish asymptotic normality for the sample covariance matrix. The last two assumptions specify mild regularity conditions needed for the asymptotics of $\widehat{\boldsymbol{w}}$.

Notation for Theorem 6.1. We write Σ in the block form

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{22} is an $(n-1)\times(n-1)$ matrix and Σ_{11} is a scalar. Writing $\mathbf{1}$ for the $(n-1)$-dimensional vector of ones, we define

$$\tilde{\Sigma} = \Sigma_{22} + \Sigma_{11}\mathbf{1}\mathbf{1}' - \mathbf{1}\Sigma_{12} - \Sigma_{21}\mathbf{1}'$$

and note that $\tilde{\Sigma}$ is an $(n-1)\times(n-1)$ positive semi-definite symmetric matrix. We show in the proof of Theorem 6.1 that $\nabla^2 F(\boldsymbol{w}^*_{-1}) = \tilde{\Sigma}$. We let $S = \tilde{\Sigma}^{1/2}$ and note that matrix S is positive definite by assumption A4. We define vectors $\boldsymbol{s} = \text{sign}(\boldsymbol{w}^*)$ and $\boldsymbol{z} = I(\boldsymbol{w}^* = \mathbf{0})$, where we apply functions $\text{sign}(\cdot)$ and $I(\cdot)$ element by element, and write $\boldsymbol{z} = (z_1, \boldsymbol{z}'_{-1})'$ and $\boldsymbol{s} = (s_1, \boldsymbol{s}'_{-1})'$, as before. We define

$$\mathcal{C} = \{S\boldsymbol{\delta},\ \text{s.t. } \boldsymbol{\delta} \in \mathbb{R}^{n-1},\ \boldsymbol{\delta}'\boldsymbol{s}_{-1} + |\boldsymbol{\delta}|'\boldsymbol{z}_{-1} - \mathbf{1}'\boldsymbol{\delta}s_1 + |\mathbf{1}'\boldsymbol{\delta}|z_1 \le 0\},$$

where $|\boldsymbol{\delta}|' = (|\delta_1|, ..., |\delta_{n-1}|)$, and note that $\mathcal{C}$ is a convex cone in $\mathbb{R}^{n-1}$.

Let $\boldsymbol{\mu} = E[\boldsymbol{y}_1]$. Given a matrix A, we write $vec(A)$ for the vector formed by stacking together the columns of A. Let W denote the variance for the limiting distribution of $T^{1/2}[vec(\widehat{\Sigma}) - vec(\Sigma)]$. By Theorem 1 and formula (3.11) in Neudecker and Wesselman (1990), we have

$$W = E[(\boldsymbol{y}_1 - \boldsymbol{\mu})(\boldsymbol{y}_1 - \boldsymbol{\mu})' \otimes (\boldsymbol{y}_1 - \boldsymbol{\mu})(\boldsymbol{y}_1 - \boldsymbol{\mu})'] - vec(\Sigma)vec(\Sigma)', \tag{A.1}$$

where $\otimes$ is the Kronecker product. We write I for the $(n-1)\times(n-1)$ identity matrix and define C as the following $(n-1)\times n^2$ matrix:

$$C = \begin{pmatrix} w_1^*\mathbf{1} & -w_1^* I & \cdots & w_n^*\mathbf{1} & -w_n^* I \end{pmatrix}.$$

Finally, we define

$$V = \tilde{\Sigma}^{-1} C W C' \tilde{\Sigma}^{-1},$$

which is an $(n-1)\times(n-1)$ matrix.

Proof of Theorem 6.1. Observe that

$$F(\boldsymbol{w}_{-1}) = \begin{pmatrix} 1-\mathbf{1}'\boldsymbol{w}_{-1} & \boldsymbol{w}_{-1}' \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} 1-\mathbf{1}'\boldsymbol{w}_{-1} \\ \boldsymbol{w}_{-1} \end{pmatrix},$$

and hence the Hessian matrix of F is given by

$$\nabla^2 F(\boldsymbol{w}_{-1}) = \Sigma_{22} + \Sigma_{11}\mathbf{1}\mathbf{1}' - \mathbf{1}\Sigma_{12} - \Sigma_{21}\mathbf{1}' = \tilde{\Sigma}.$$

We define $\boldsymbol{\tau} = \boldsymbol{w} - \boldsymbol{w}^*$, $\boldsymbol{\delta} = \boldsymbol{w}_{-1} - \boldsymbol{w}^*_{-1}$ and $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{w}}_{-1} - \boldsymbol{w}^*_{-1}$, to simplify the exposition. Writing a two-term Taylor expansion for $F(\boldsymbol{w}_{-1})$ at $\boldsymbol{w}^*_{-1}$, and noting assumption A3, we derive

$$F(\boldsymbol{w}_{-1}) - F(\boldsymbol{w}^*_{-1}) = \boldsymbol{\delta}'\tilde{\Sigma}\boldsymbol{\delta}. \tag{A.2}$$

Let $\widehat{\Sigma}$ denote the sample covariance matrix for $\boldsymbol{y}_1$ and define $\widehat{\Delta} = \widehat{\Sigma} - \Sigma$. Note that

$$\begin{aligned} \boldsymbol{w}'\widehat{\Delta}\boldsymbol{w} - \boldsymbol{w}^{*\prime}\widehat{\Delta}\boldsymbol{w}^* &= \left[\boldsymbol{w}'\widehat{\Delta}\boldsymbol{w} - \boldsymbol{w}'\widehat{\Delta}\boldsymbol{w}^*\right] + \left[\boldsymbol{w}'\widehat{\Delta}\boldsymbol{w}^* - \boldsymbol{w}^{*\prime}\widehat{\Delta}\boldsymbol{w}^*\right] \\ &= \boldsymbol{w}'\widehat{\Delta}\boldsymbol{\tau} + \boldsymbol{\tau}'\widehat{\Delta}\boldsymbol{w}^* \\ &= 2\boldsymbol{\tau}'\widehat{\Delta}\boldsymbol{w}^* + \boldsymbol{\tau}'\widehat{\Delta}\boldsymbol{\tau}. \end{aligned} \tag{A.3}$$

We define functions $R(\boldsymbol{w}) = \boldsymbol{w}'\Sigma\boldsymbol{w}$ and $\widehat{R}(\boldsymbol{w}) = \boldsymbol{w}'\widehat{\Sigma}\boldsymbol{w}$. Note that $\widehat{\boldsymbol{w}}$ minimizes $\widehat{R}(\boldsymbol{w})$ under the constraints in (15). Using formulas (A.2) and (A.3), we derive

$$\begin{aligned} \widehat{R}(\boldsymbol{w}) - \widehat{R}(\boldsymbol{w}^*) &= \left[\widehat{R}(\boldsymbol{w}) - R(\boldsymbol{w})\right] + \left[R(\boldsymbol{w}^*) - \widehat{R}(\boldsymbol{w}^*)\right] + \left[R(\boldsymbol{w}) - R(\boldsymbol{w}^*)\right] \\ &= \boldsymbol{w}'\widehat{\Delta}\boldsymbol{w} - \boldsymbol{w}^{*\prime}\widehat{\Delta}\boldsymbol{w}^* + \left[F(\boldsymbol{w}_{-1}) - F(\boldsymbol{w}^*_{-1})\right] \\ &= 2\boldsymbol{\tau}'\widehat{\Delta}\boldsymbol{w}^* + \boldsymbol{\tau}'\widehat{\Delta}\boldsymbol{\tau} + \boldsymbol{\delta}'\tilde{\Sigma}\boldsymbol{\delta}. \end{aligned}$$

Note that $\boldsymbol{\tau} = (-\mathbf{1}'\boldsymbol{\delta}, \boldsymbol{\delta})'$ and let $\boldsymbol{Z}_T = T^{1/2}\tilde{\Sigma}^{-1}(\mathbf{1} \quad -I)\widehat{\Delta}\boldsymbol{w}^*$, where vector $\mathbf{1}$ and matrix I are defined as before. It follows that $\boldsymbol{\tau}'\widehat{\Delta}\boldsymbol{w}^* = -T^{-1/2}\boldsymbol{\delta}'\tilde{\Sigma}\boldsymbol{Z}_T$. By the central limit theorem for the sample covariance matrix (Neudecker and Wesselman, 1990, Theorem 1), we have $\widehat{\Delta} = O_p(T^{-1/2})$ and $\boldsymbol{Z}_T = O_p(1)$. Consequently,

$$\widehat{R}(\boldsymbol{w}) - \widehat{R}(\boldsymbol{w}^*) = \boldsymbol{\delta}'\tilde{\Sigma}\boldsymbol{\delta} - 2T^{-1/2}\boldsymbol{\delta}'\tilde{\Sigma}\boldsymbol{Z}_T + O_p(T^{-1/2}\|\delta\|^2). \tag{A.4}$$

Because $\widehat{R}(\widehat{\boldsymbol{w}}) \leq \widehat{R}(\boldsymbol{w}^*)$, we then have

$$\widehat{\boldsymbol{\delta}}'\tilde{\Sigma}\widehat{\boldsymbol{\delta}} \leq O_p\Big(T^{-1/2}[\|\widehat{\boldsymbol{\delta}}\| + \|\widehat{\boldsymbol{\delta}}\|^2]\Big).$$

As we mentioned before, $\tilde{\Sigma}$ is a positive semi-definite matrix. Thus, by formula (A.2) and assumption A4, we have $\widehat{\boldsymbol{\delta}}'\tilde{\Sigma}\widehat{\boldsymbol{\delta}} \geq \kappa\|\widehat{\boldsymbol{\delta}}\|^2$, for some positive constant κ. Consequently,

$$\|\widehat{\boldsymbol{\delta}}\|^2 = O_p\Big(T^{-1/2}\|\widehat{\boldsymbol{\delta}}\| + T^{-1/2}\|\widehat{\boldsymbol{\delta}}\|^2]\Big),$$

which implies $\|\widehat{\boldsymbol{\delta}}\| = O_p(T^{-1/2})$ and establishes the $T^{-1/2}$ rate of convergence for $\widehat{\boldsymbol{w}}$.

Let $\mathcal{S}$ denote the constraint set for $\boldsymbol{w}_{-1}$ in problem (15), more specifically,

$$\mathcal{S} = \{\boldsymbol{w}_{-1} \in \mathbb{R}^{n-1},\ \text{s.t. } |1 - \mathbf{1}'\boldsymbol{w}_{-1}| + \|\boldsymbol{w}_{-1}\|_1 \leq 1 + \tilde{c}\}. \tag{A.5}$$

We derive the limiting distribution for $T^{1/2}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1})$ by applying Theorem 4.4 in Geyer (1994). An analysis of the proof of this theorem shows that for its conclusion to hold, the following conditions are sufficient: (i) $\widehat{\boldsymbol{\omega}}_{-1} = \boldsymbol{\omega}^*_{-1} + O_p(T^{-1/2})$; (ii) stochastic bound

$$T\Big[\widehat{R}(\boldsymbol{w}^* + T^{-1/2}\boldsymbol{v}_T) - \widehat{R}(\boldsymbol{w}^*)\Big] = \boldsymbol{v}_T'\tilde{\Sigma}\boldsymbol{v}_T - 2\boldsymbol{v}_T'\tilde{\Sigma}\boldsymbol{Z}_T + o_p(1) \tag{A.6}$$

holds for every $O_p(1)$ random vector sequence $\boldsymbol{v}_T$; (iii) constraint set $\mathcal{S}$ is Chernoff regular at $\boldsymbol{\omega}^*_{-1}$; and (iv) random sequence $\tilde{\Sigma}\boldsymbol{Z}_T$ converges in distribution to a mean zero Gaussian vctor. We have already established (i). Condition (ii) follows directly from approximation (A.4). Condition (iii) is only imposed to rule out pathological cases. It is satisfied in our setting, because $\mathcal{S}$ is formed via finitely many union and intersection operations, applied to a finite collection of closed half-spaces. By Theorem 1 and formula (3.11) in Neudecker and Wesselman (1990),

$$T^{1/2}[vec(\widehat{\Sigma}) - vec(\Sigma)] \xrightarrow{d} N(0, W),$$

which implies that $\boldsymbol{Z}_T$ converges in distribution to $\boldsymbol{Z}$, and hence condition (iv) holds.

To apply the result in Geyer (1994), we need to define the tangent cone of $\mathcal{S}$ at the point $\boldsymbol{w}^*_{-1}$. We denote this tangent cone by $\tilde{\mathcal{C}}$. A vector $\boldsymbol{\delta}$ lies in $\tilde{\mathcal{C}}$ if and only if there exists a sequence ϵ_n converging to 0 and a sequence $\boldsymbol{u}_n \in \mathcal{S}$ converging to $\boldsymbol{w}^*_{-1}$, such that $[\boldsymbol{u}_n - \boldsymbol{\omega}^*_{-1}]/\epsilon_n \to \boldsymbol{\delta}$. It follows from the definition of $\mathcal{S}$ in (A.5) that

$$\tilde{\mathcal{C}} = \{\boldsymbol{\delta} \in \mathbb{R}^{n-1},\ \text{s.t. } \boldsymbol{\delta}'\boldsymbol{s}_{-1} + |\boldsymbol{\delta}|'\boldsymbol{z}_{-1} - \mathbf{1}'\boldsymbol{\delta}s_1 + |\mathbf{1}'\boldsymbol{\delta}|z_1 \leq 0\}.$$

We apply the aforementioned result in Geyer (1994) to conclude that $T^{1/2}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1})$ converges in distribution to the minimizer of $\boldsymbol{v}'\tilde{\Sigma}\boldsymbol{v} - 2\boldsymbol{v}'\tilde{\Sigma}\boldsymbol{Z}$ over $\boldsymbol{v} \in \tilde{\mathcal{C}}$. Note that $\mathcal{C} = \tilde{\Sigma}^{1/2}\tilde{\mathcal{C}}$. Thus, we can write the solution to the aforementioned optimization problem as follows:

$$\min_{\boldsymbol{v}\in\tilde{\mathcal{C}}} \boldsymbol{v}'\tilde{\Sigma}\boldsymbol{v} - 2\boldsymbol{v}'\tilde{\Sigma}\boldsymbol{Z} = \min_{\boldsymbol{v}\in\tilde{\mathcal{C}}} \|\tilde{\Sigma}^{1/2}\boldsymbol{v} - \tilde{\Sigma}^{1/2}\boldsymbol{Z}\|^2 = \tilde{\Sigma}^{-1/2}\mathrm{Proj}_{\mathcal{C}}\tilde{\Sigma}^{1/2}\boldsymbol{Z},$$

which establishes the second convergence result in Theorem 6.1. The first convergence result in Theorem 6.1 follows from the observation that $\mathcal{C}$ spans the entire space $\mathbb{R}^{n-1}$ when $\boldsymbol{w}^*_{-1}$ is in the interior of the constraint set $\mathcal{S}$.

Proof of Proposition 7.1. MSFE is a function of the weight $w \in \mathbb{R}$ that appears in the combimation y_c To simplify the expressions, we write $H(c)$ for $E\big[\text{MSFE}\big(w^{\text{TR}}(c)\big)\big]$ and $G(w)$ for $\text{MSFE}(w)$. To complete the proof, we will demonstrate that $H(c) \geq H(-w^*)$ for all c.
First, we consider the case $-c < w^*$ and note that $w^{\text{TR}}(c) \neq w^{\text{TR}}(-w^*)$ implies $w^{\text{TR}}(-w^*) = w^*$. Thus, on the event $w^{\text{TR}}(c) \neq w^{\text{TR}}(-w^*)$ we have

$$G\big(w^{\text{TR}}(c)\big) \geq G\big(w^*\big) = G\big(w^{\text{TR}}(-w^*)\big).$$

Consequently, $G\big(w^{\text{TR}}(c)\big) \geq G\big(w^{\text{TR}}(-w^*)\big)$ with probability one, and hence $H(c) \geq H(-w^*)$.
Finally, we consider the case $-c > w^*$. We recall that $G(w^*)$ is non-decreasing for $w \geq w^*$, and note that

$$w^{\text{TR}}(c) \geq w^{\text{TR}}(-w^*) \geq w^*$$

with probability one. Consequently, we $G\big(w^{\text{TR}}(c)\big) \geq G\big(w^{\text{TR}}(-w^*)\big)$, and hence $H(c) \geq H(-w^*)$.

Appendix B: Additional empirical results

MAFE results

As shown in Tables B.1–B.3, our conclusion of the empirical analysis in Section 8 remains the same when we evaluate the forecasts using the MAFE. For the fixed thresholds (see Tables B.1 and B.2), the MAFE in most cases first improves and then deteriorates when we vary the threshold for trimming from zero to $-\infty$, and the one-step trimming methods (TR4 and TR5) generally outperform the two-step trimming (TR1–TR3). For the data-driven thresholds (see Table B.3), TR4 and TR5 again both produce better combinations than equal weights and two-step trimming methods.

Statistical difference between TR5 and equal weights

Table B.4 presents the proportion of times when the trimmed weights obtained by TR5 are significantly different from equal weights at 95% confidence level. More specifically, we employ the Wald statistic to test for the difference between the trimmed weights produced by TR5 and equal weights at each evaluation time point, and take the ratio of the number of times when the test rejects the null

Table B.1: Relative MAFE of combined forecasts using five trimmed weights as a function of the trimming threshold c (two-step trimming)

		1-year forecast horizon			2-year forecast horizon		
	c	HICP	RGDP	UNEM	HICP	RGDP	UNEM
TR1	$-\infty$	2.285	0.886	0.844	1.817	1.386	1.605
	-5.0	1.696	0.889	0.794	1.557	0.881	1.605
	-4.5	1.680	0.889	0.794	1.553	0.900	1.605
	-4.0	1.671	0.888	0.794	1.548	0.926	1.605
	-3.5	1.671	0.900	0.794	1.544	0.929	1.605
	-3.5	1.643	0.910	0.792	1.540	0.876	1.602
	-2.5	1.621	0.915	0.791	1.495	0.834	1.586
	-2.0	1.526	0.902	0.786	1.389	0.883	1.304
	-1.5	1.371	0.890	0.717	1.303	0.896	1.160
	-1.0	1.232	0.896	0.653	1.214	0.874	1.036
	-0.5	1.084	0.923	0.697	1.079	0.916	0.754
	0.0	1.016	0.987	0.945	1.022	0.976	0.907
TR2	$-\infty$	2.285	0.886	0.844	1.817	1.386	1.605
	-5.0	2.081	0.902	0.790	1.674	1.029	1.605
	-4.5	2.152	0.908	0.791	1.660	1.006	1.605
	-4.0	2.094	0.920	0.791	1.668	0.988	1.605
	-3.5	2.159	0.919	0.881	1.651	0.979	1.605
	-3.5	2.106	0.919	1.061	1.632	0.982	1.604
	-2.5	2.017	0.917	1.056	1.606	0.936	1.599
	-2.0	1.938	0.915	0.991	1.561	0.888	1.528
	-1.5	1.779	0.903	0.955	1.502	0.827	1.460
	-1.0	1.613	0.895	0.758	1.432	0.758	1.271
	-0.5	1.407	0.921	0.631	1.246	0.805	1.035
	0.0	1.016	0.987	0.945	1.022	0.976	0.907
TR3	$-\infty$	2.285	0.886	0.844	1.817	1.386	1.605
	-5.0	2.002	0.887	0.811	1.674	0.877	1.605
	-4.5	1.961	0.886	0.809	1.660	0.871	1.605
	-4.0	1.924	0.886	0.808	1.643	0.867	1.605
	-3.5	1.878	0.889	0.804	1.629	0.862	1.605
	-3.5	1.830	0.893	0.798	1.615	0.839	1.604
	-2.5	1.781	0.899	0.796	1.588	0.811	1.599
	-2.0	1.706	0.898	0.791	1.526	0.795	1.499
	-1.5	1.594	0.896	0.744	1.447	0.789	1.387
	-1.0	1.439	0.898	0.671	1.334	0.813	1.232
	-0.5	1.206	0.917	0.605	1.177	0.858	0.866
	0.0	1.016	0.987	0.945	1.022	0.976	0.907

Notes: This table presents relative mean absolute forecast error (MAFE) of different weight schemes, averaged across the whole testing period (16 quarters). We normalize all numbers by dividing by the MAFE of the equal-weight combination, and thus a value smaller than 1 indicates better performance than the equal-weight combination.

Table B.2: Relative MAFE of combined forecasts using five trimmed weights as a function of the trimming threshold c (one-step trimming)

		1-year forecast horizon			2-year forecast horizon		
	c	HICP	RGDP	UNEM	HICP	RGDP	UNEM
TR4	$-\infty$	2.285	0.886	0.844	1.817	1.386	1.605
	-5.0	2.374	1.210	1.762	1.731	1.588	2.652
	-4.5	2.326	1.296	1.458	2.146	1.016	2.624
	-4.0	2.084	0.860	1.121	1.594	1.039	2.307
	-3.5	2.017	1.113	1.816	1.568	1.097	2.235
	-3.5	2.366	0.965	1.612	1.595	0.978	1.820
	-2.5	1.321	1.025	1.198	1.624	1.115	1.947
	-2.0	1.435	0.953	1.154	1.484	0.888	1.952
	-1.5	1.299	0.799	0.761	1.305	0.891	1.636
	-1.0	1.454	0.745	0.846	1.459	0.836	1.178
	-0.5	1.282	0.783	0.570	1.331	0.769	0.875
	0.0	0.940	0.939	0.815	0.955	0.882	0.964
TR5	$-\infty$	2.285	0.886	0.844	1.817	1.386	1.605
	-5.0	1.244	0.874	0.798	1.279	0.710	1.186
	-4.5	1.176	0.914	0.662	1.163	0.710	1.164
	-4.0	1.198	0.847	0.597	1.101	0.709	1.049
	-3.5	1.165	0.864	0.524	1.084	0.705	0.975
	-3.5	1.105	0.898	0.490	1.071	0.699	0.879
	-2.5	1.055	0.890	0.445	1.037	0.702	0.826
	-2.0	1.014	0.899	0.422	1.045	0.724	0.736
	-1.5	0.968	0.925	0.402	1.019	0.735	0.632
	-1.0	0.951	0.946	0.390	0.985	0.765	0.572
	-0.5	0.958	0.947	0.485	0.979	0.828	0.673
	0.0	0.940	0.939	0.815	0.955	0.882	0.963

Notes: This table presents relative mean absolute forecast error (MAFE) of different weight schemes, averaged across the whole testing period (16 quarters). We normalize all numbers by dividing by the MAFE of the equal-weight combination, and thus a value smaller than 1 indicates better performance than the equal-weight combination.

Table B.3: Relative MAFE of combined forecasts using trimmed weights with data-driven threshold based on pseudo out-of-sample evaluation

	1-year horizon			2-year horizon		
	HICP	RGDP	UNEM	HICP	RGDP	UNEM
			$c_{\min} = -2$			
TR1	1.079	0.945	0.880	1.062	0.890	0.910
TR2	1.070	0.948	0.868	1.056	0.902	0.934
TR3	1.078	0.947	0.852	1.083	0.897	**0.870**
TR4	**0.940**	0.941	0.816	**0.925**	0.882	0.958
TR5	0.944	**0.938**	**0.663**	0.947	**0.857**	0.913
			$c_{\min} = -5$			
TR1	1.079	0.946	0.884	1.062	0.886	0.910
TR2	1.070	0.948	0.868	1.056	0.902	0.934
TR3	1.078	0.947	0.852	1.083	0.897	**0.870**
TR4	**0.940**	0.941	0.816	**0.934**	0.882	0.958
TR5	0.943	**0.939**	**0.663**	0.947	**0.857**	0.913

Notes: This table presents relative mean absolute forecast error (MAFE) of different weight schemes, averaged across the whole testing period (16 quarters). We normalize all numbers by dividing by the MAFE of the equal-weight combination, and thus a value smaller than 1 indicates better performance than the equal-weight combination.

hypothesis over the total number of evaluation periods. We report this proportion for a grid of fixed thresholds (upper panel of the table) and for a data-driven threshold (lower panel). We find that there are generally a larger proportion of times when the difference between two weights is significant if we use a highly negative threshold. Interestingly, for both fixed and data-driven thresholds, the trimmed weights more frequently differ from equal weights for UNEM forecasting than for HICP. This result is in line with a large improvement of TR5 forecast combination over the equal-weight combination, as shown in Tables 3 and 4 in the paper.

Table B.4: Proportion of times when the trimmed weights by TR5 are significantly different from equal weights at 95% confidence level

	1-year forecast horizon			2-year forecast horizon		
c	HICP	RGDP	UNEM	HICP	RGDP	UNEM
			Fixed threshold			
−5.0	0.19	0.38	0.44	0.44	0.44	0.31
−4.5	0.25	0.38	0.38	0.44	0.44	0.31
−4.0	0.19	0.38	0.38	0.44	0.50	0.38
−3.5	0.25	0.31	0.31	0.44	0.50	0.31
−3.5	0.19	0.31	0.31	0.38	0.44	0.38
−2.5	0.19	0.31	0.31	0.44	0.44	0.31
−2.0	0.25	0.31	0.31	0.38	0.31	0.31
−1.5	0.19	0.25	0.31	0.38	0.31	0.31
−1.0	0.13	0.31	0.31	0.38	0.31	0.31
−0.5	0.13	0.25	0.31	0.31	0.31	0.31
0.0	0.13	0.06	0.31	0.31	0.31	0.38
			Data-driven threshold			
	0.13	0.06	0.31	0.31	0.38	0.38

Notes: This table presents the proportion of times over the entire evaluation periods, at which the Wald statistics for testing the difference between trimmed weights using TR5 and equal weights are significantly different at 95% confidence level.

Appendix C: Additional simulation results

To investigate the trade-off between variance and bias, we consider those components in our simulation example (detailed in Figure 6). For the case of $\phi_1 = -0.5$, we decompose MSE $= \mathrm{E}(w^{\mathrm{TR}} - w^*)^2$ into the variance (which becomes smaller as we trim closer to zero) and the bias (which increases and reaches its maximum at zero), see Figure C.1. The decomposition reveals that the bias component is driving the sharp increase in MSE when the threshold is approaching zero. The shape of the MSE curve resembles the shape of the MSFE curve in Figure 6. This finding further strengthens our conclusions: the common practice of trimming at zero can be improved dramatically as even small deviations from zero will produce a sharp decrease in the bias. This is clearly visible in our empirical example as well.

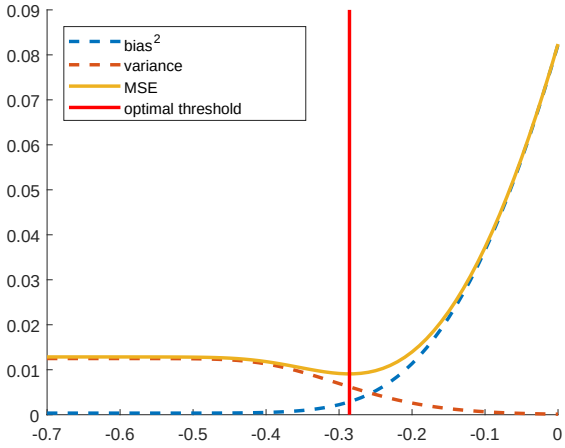


Figure C.1: The tradeoff between the bias and the variance in the case of $\phi_1 = -0.5$. The decreasing dashed curve shows the variance of w^{TR}, the increasing dashed curve shows the squared bias of w^{TR}, and the solid curve shows the MSE as a function of the threshold $-c$. The optimal threshold that minimizes MSE (as well as MSFE) is given by the vertical line.

We performed additional simulations for the case of 3 forecasts. The set up is the same as in Section 7, $y_1 = z_T$, $y_2 = \rho_2 z_{T-1}$, with the additional third forecast being $y_3 = z_{T-2}$. We only investigate TR5 version as it shows the best results in our empirical study. Since TR5 involves constraint optimization, it is more time-intensive. As a result, we decrease the number of simulation to be 10,000. Figure C.2 presents the results.

The MSFE curves have more complicated shapes than in the cases of two forecasts (see Figure 7). The left local minimum is due to the trimming of the first negative forecast. As we go from left to right, the MSFE curve decreases due to

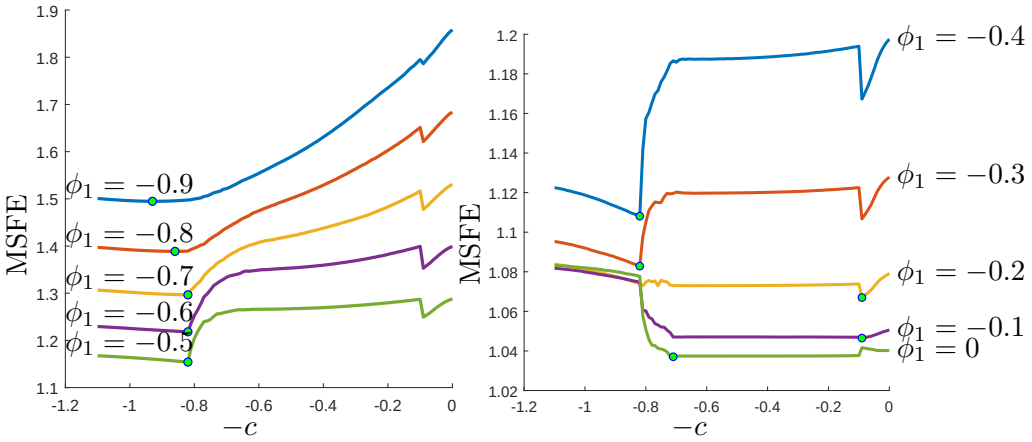


Figure C.2: The case of 3 forecasts and TR5 trimming. For each value of ϕ_1, the solid line shows the MSFE as a function of the threshold $-c$, and the point indicates the optimal trimming parameter that achieves the minimum on this curve. The left panel shows the curves for $\phi = -0.9, \ldots, -0.5$, while the right panel shows the curves for $\phi = -0.4, \ldots, 0.0$. The results are shown in two panels with different scales to maximize visibility and clarity of the results.

variance stabilization and sharply increases once we pass the first minimum due to the bias component that gives a heavy penalty once we go over the theoretically optimal value of the negative weight. The second local minimum is when we reach the second negative weight. Again, we see an initial dip due to the variance stabilization followed by the increase due to the increasing bias once we pass the optimal point. We still arrive at the same conclusion: the forecasting performance can be improved by choosing a threshold different from zero.