Combining Forecasts - On Why Averaging beats Optimal Linear Weights[*]

Graham Elliott[†] Jie Liao [‡]

May 5, 2025

Abstract

A continuing puzzle in constructing a point forecast by combining individual forecasts is that simple averaging often beats estimating optimal weights (the forecast combination puzzle). Most researchers have focused on the size of estimation error other difficulties in forecasting weights, despite this estimation procedure being a simple least squares regression. For this explanation to hold, gains from using optimal weights must be small. This paper focuses on this complementary part of the argument - we ask how big can the gains from optimal combination be in empirically and theoretically reasonable situations. Under these restrictions we show that gains can indeed be small, and that for gains to be large the best approach to forecast combination is to discard some of the forecasts and average over the remaining ones.

JEL codes: C22, C53, E37.

Keywords: Forecast Combination Puzzle, Forecast Averaging.

[*] This paper is a major revision of an earlier circulated paper "Averaging and the optimal combination of forecasts". A replication package is available at https://econweb.ucsd.edu/ gelliott/

[†] Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093, email: grelliott@ucsd.edu

[‡] Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093, email: j4liao@ucsd.edu

1 Introduction

We consider the situation, as in Bates and Granger (1969), where for the construction of a forecast of an outcome we have m point forecasts available and will employ a linear combination of them into a single number to forecast the outcome. Combining makes sense (Yang (2004)), and though Bates and Granger (1969) derived optimal weights for this forecast combination problem, even in their paper (and a great deal of subsequent work) it has been noticed that using the estimated optimal weights has often been outperformed by taking simple averages of the forecasts in constructing the forecast combination. It is a surprising result given that the estimation of the optimal weights (when they sum to one) simply requires restricted OLS and the estimation of $m-1$ weights (Granger and Ramanathan (1984). Linear regression is a workhorse model in statistics that generally provides very good results. This has become known as the forecast combination puzzle[1].

Indeed, it is rare in practice to favor optimal combination over simple averaging. Consensus Economics, a company that collects and sells combined forecasts, uses simple averages[2]. Recently in macroeconomic research there has been a renewed focus on overreaction in expectations based on combined forecasts, where simple averages are taken (Coibion and Gorodnichenko (2015), Bordalo et al. (2020)).

Most econometric explanations revolve around the idea that the puzzle is explained via estimation error - optimal forecast combinations require that the combination weights be estimated whereas under simple averaging there is no estimation error in the construction of the weights[3]. These problems are obviously exacerbated when the number of forecasts and hence weights to estimate is large relative to the sample size (Chan et al. (1999), Stock and Watson (2004)). Methods to resolve this difficulty for moderate m attempt to use relatively ad hoc methods that still allow the data to determine the weights but with hopefully less estimation error.

Other approaches suggest that a reason for regression results to be poor in practice is that the estimation error is large due to the lack of stationarity of the data used to construct the forecast combinations. For stable enough data, regression will accurately estimate weights, however if the optimal weights themselves are not stable over time then regression estimates the average optimal weights rather than the optimal weights at each forecast point. Methods to address this problem typically involve either simply using averages (Kang (1986)) or instead using rolling averages of data (which reduce the sample size, also increasing estimation error) or modeling the instability directly (which also can increase estimation error as a more complicated model is required)[4].

Our approach in this paper is to look at the other side of this argument. The use of

[1] Review articles include Timmermann (2006), Wang et al. (2020), Clemen (1989). The literature is too vast to mention all the papers on this topic

[2] See Consensus Economics which reports means of forecasts, which appears to be the simple average

[3] See Smith and Wallis (2009), Claeskens et al. (2016)

[4] For example Deutsch et al. (1994) models the instability.

estimated optimal weights results in the need to estimate $m-1$ parameters via restricted OLS. For estimation error to be the reason for which the forecast combination puzzle then a counterpart to the explanation must be that the gains from using the optimal weights - i.e. the potential value of optimal weights over averaging - must be small relative to the size of the estimation error. If this difference were large the estimation error would be an unlikely explanation for this puzzle. In this paper we characterize situations where gains are small, and show that they are empirically and theoretically relevant.

We consider the model

$$\begin{pmatrix} y_{t+1} \\ f_{t,t+1} \end{pmatrix} \sim \left[\begin{pmatrix} \mu_{t,t+1} \\ \mu_{t,t+1} \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \tilde{\Sigma} \end{pmatrix} \right] \tag{1}$$

The model assumes that the forecasters are getting the conditional mean of y_{t+1} correct (usually referred to in this literature as unbiased forecasts). However there is noise in each of the m individual forecasts $dim(\tilde{\Sigma}) = m$) as no forecaster knows the correct model. This is a situation where averaging or weighted averaging can possibly improve the accuracy of providing a single forecast by reducing the variance of the forecast error.

By definition of $\mu_{t,t+1}$ as the conditional mean (based on some joint information set) of y_{t+1} we have that the unforecastable component of the outcome is uncorrelated with the noise around the conditional mean in the forecasts. This is because if there were some correlation between the errors of the forecasts and the outcome to be forecast we would then use that to construct better estimates of $\mu_{t,t+1}$ changing the meaning of this conditional mean. Note also that in this model we do not assume stationarity of the outcomes or forecasts, but assume that the variance covariance matrix of the forecast errors is constant across time. We agree that there might be heteroskedasticity in practice, however we intend to show the results in the case that is most favorable to linear regression and least favorable to simple averaging.

Defining ι to be a vector of ones (usually $mx1$ unless otherwise indicated) the vector of forecast errors[5] $e_{t,t+1} = y_{t+1}\iota - f_{t,t+1}$ has a variance covariance matrix that follows from above equal to

$$\tilde{\Omega} = \sigma_\epsilon^2 \iota\iota' + \tilde{\Sigma}.$$

We are interested in a linear weighted average of the individual forecasts for use as a combined forecast, i.e. our point forecast is $\omega' f_{t,t+1}$. The resulting MSE loss from such a combined forecast is equal to

$$L = (1 - \omega'\iota_m)^2 \mu_{t,t+1}^2 + \sigma_\epsilon^2 + \omega'\tilde{\Sigma}\omega.$$

For much of what follows all comparisons between methods for which the weights sum to one regards only the variance component (as the bias squared component is zero) and so loss is $\sigma_\epsilon^2 + \omega'\tilde{\Sigma}\omega$. This does not mean that there is not a bias-variance trade-off in general,

[5] The restriction to one step ahead forecasts here is without loss of generality, and simplifies the exposition of the results.

but following this literature we focus on this problem where it is a variance minimization problem only.

For averaging, we have the weights $\omega^{ave} = m^{-1}\iota$ which results in a loss of $\sigma_\epsilon^2 + m^{-2}\iota_m'\tilde{\Sigma}\iota_m$. For the optimal combination weights, we have the weights $\omega^{opt\prime} = (\iota_m'\tilde{\Sigma}^{-1}\iota_m)^{-1}\iota_m'\tilde{\Sigma}^{-1}$ which results in a loss of $\sigma_\epsilon^2 + (\iota_m'\tilde{\Sigma}^{-1}\iota_m)^{-1}$. Note that optimal weights (provided they sum to one) based on $\tilde{\Omega}$ are identical[6] to those based on $\tilde{\Sigma}$.

Let the space of possible variance covariance matrices of the forecast error (in population) be denoted as $M_{\tilde{\Omega}}$ or $M_{\tilde{\Sigma}}$ for $\tilde{\Omega}$ and $\tilde{\Sigma}$ respectively. This paper aims to characterize the space of such matrices and restrictions on these spaces to understand how large possible gains to optimal combination can be. The general result is that gains from optimal combination are often small for empirically relevant problems, and as such it is highly likely that estimation error is too large to exploit these small gains. In the best case scenario for OLS estimation of the weights, the difference in expected loss from averaging over using estimated weights equals

$$\left(m^{-2}\iota'\tilde{\Sigma}\iota - (\iota'\tilde{\Sigma}^{-1}\iota)^{-1}\right) - \left(\sigma_\epsilon^2 + (\iota'\tilde{\Sigma}^{-1}\iota)^{-1}\right)\left(\frac{m-1}{T}\right) \tag{2}$$

when there are T observations available to estimate the weights. Both terms are nonnegative, the first term is the loss from averaging over optimal weights and the second is the term that accounts for estimation error. So we are concerned with examining the size of the first term, to show that it is often small.

For the figures later it is useful to consider relative loss from averaging

$$\left(\frac{m^{-2}\iota'\tilde{\Sigma}\iota - (\iota'\tilde{\Sigma}^{-1}\iota)^{-1}}{\sigma_\epsilon^2 + (\iota'\tilde{\Sigma}^{-1}\iota)^{-1}}\right) \tag{3}$$

which can be directly compared to $\frac{m-1}{T}$. Notice that the relative loss depends on the forecastability of the outcome (i.e. the magnitude of σ_ϵ^2, this issue is examined in Elliott (2016)).

Without restrictions on $\tilde{\Sigma}$ gains from optimal combination can be very large. For example if two forecasters provided forecasts that were perfectly negatively correlated around $\mu_{t,t+1}$ then the optimal combination between just these two forecasts would reveal $\mu_{t,t+1}$. This situation is obviously unlikely in practice. One primary takeaway from this paper is that under restrictions on $\tilde{\Sigma}$ motivated from looking at a long history of the Survey of Professional Forecasters data for real GDP forecasts is that gains from optimal combination can be very small. This gives credence for the explanations that the problem is estimation error because this argument only works if the potential gains are not so large as to overcome the need for estimation.

A second result is that under these restrictions we are able to show that the optimal combination under the worst case scenario for averaging is to discard some forecasts and average over the remaining forecasts. This too is a simple procedure that lays the ground for simple estimation methods, we suggest such a method in Section 5 below.

[6] See for example Elliott (2016)

Finally, we also show that for large numbers of forecasts to combine, restrictions on the covariance matrix of forecast errors around the conditional mean suggest that for many models averaging and optimal forecasts will give similar results in the absence of estimation error. Thus estimating optimal weights for large m is likely to result in a forecast combination that performs poorly.

2 Some Stylized Facts Using SPF data

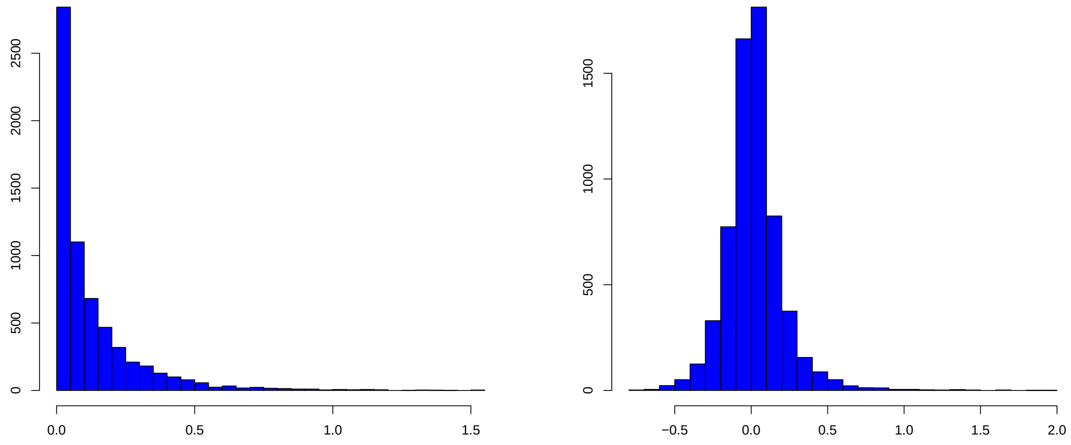
In this section, we use forecast data from the Survey of Professional Dataset(SPF). This dataset provides quarterly data for a range of macroeconomic variables from the fourth quarter of 1968 until the fourth quarter of 2019. For each period, different forecasters provide predictions for over the following four quarters. We treat this dataset as an "experiment" to investigate empirical facts relating to relative loss and more importantly motivate restrictions on $\tilde{\Sigma}$. We focus specifically on real GDP as the variable of interest. In the sample, each forecaster in each period is considered as an independent observation. By collecting m forecasters which have provided prediction for the same (not necessarily contiguous) t periods, we obtain for each group a dataset that includes forecasts and outcomes only from these m forecaster during these t periods. Once all possible such datasets are obtained, we aggregate them to construct a new dataset, where each smaller dataset serve as generating a single number in Figure 1 below. So each observation in Figure 1 is an estimate of MSE for a group of forecasters that provide forecasts for the same set of periods.

We first examine the in sample and out of sample performance of simple averaging versus optimal weights in forecast combination by examining the relative loss from averaging. In sample, by the properties of least squares the loss from averaging weights is always greater than the loss from optimal weights. Out of sample, estimation error increases MSE rather than decreases MSE, so the distribution can include positive or negative relative losses.

To see the empirical properties of relative loss (an estimate of Equation 3) in sample and out of sample, we collect all possible combinations of 3 forecasters who have provided forecasts for more than 40 same periods in the survey, resulting in total of 6648 triplets. For each triplet, we compute the MSE using both averaging and estimated optimal weights, and then derive the relative loss. This process is performed for both in sample and out of sample cases. As a result, we have two datasets for both types of relative loss, each derived from the computed losses across all triplets.

Figure 1 presents the histograms of these relative losses. The first plot shows in sample results for which averaging always performs worse than the estimated optimal weights, which is necessarily true due to the property of optimal weights minimizing in sample loss. The second plot, which shows the relative out of sample performance, exhibits a greater variability and wider spread. Approximately half of the triplets suggest that the averaging weight performs better than the optimal weights in out of sample.

It is often the case in forecasting that forecasters predicting similar outcome tend to align



(a) In Sample

(b) Out of Sample

Figure 1: Relative loss of averaging relative to Optimal combination

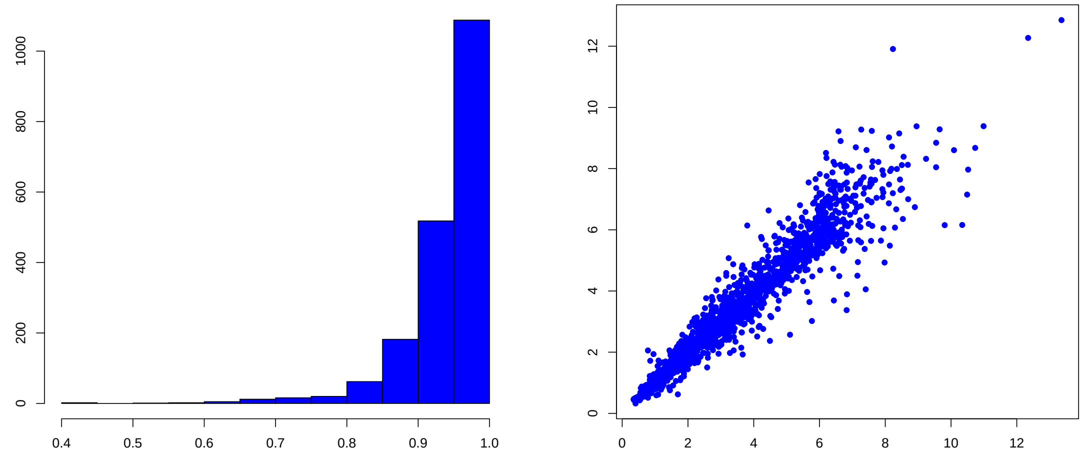
closely with each other. Indeed, it is standard advice in forecast combination to trim out forecasters who do poorly (Armstrong (2001),Timmermann (2006),Jose and Winkler (2008)) which would leave forecasters in the combination pool with similar variances. Different forecasters tend to carry similar information when predicting the same outcome, which leads to the highly positive correlations among forecasters. Empirically, the alignment and positive correlation is often reflected in the similar variances and a high positive correlation between forecast errors in $\tilde{\Omega}$ across forecasters for the same events, where $\tilde{\Omega} = \sigma_\epsilon^2 \iota\iota' + \tilde{\Sigma}$. However, the similarity may arise from the unforecastable shock ϵ_t that affects all forecasters equally.

To focus on the forecasters behavior excluding the influence of common shocks, we wish to separate $\tilde{\Sigma}$ from $\tilde{\Omega}$. Instead of directly calculating the variance and correlation of forecast error $y_{t+1} - f_{t,t+1}$ for each forecaster which only gives $\tilde{\Omega}$, we aim to obtain $\tilde{\Sigma}$ by analyzing $\mu_{t,t+1} - f_{t,t+1}$. However, since $\mu_{t,t+1}$ is unobservable, we need a model to estimate $\mu_{t,t+1}$. To achieve this, we estimate μ_t using an AR(1) model over the entire sample. These estimates enable us to construct an estimate of the variances and covariances of $\mu_{t,t+1} - f_{t,t+1}$ across all pairs of forecasters, which provides estimates of the elements in $\tilde{\Sigma}$ instead of $\tilde{\Omega}$.

We select all possible pairs of forecasters which provided prediction for at least 40 overlapping periods of real GDP in the SPF. For each pair, we compute both the variance and covariance in $\tilde{\Sigma}$ using the described approach. Figure 2 visualizes the variance and correlation for all pairs of forecasters in the sample, which provides empirical support for the restrictions imposed on $\tilde{\Sigma}$.

The left plot in Figure 2 shows the correlations for all pairs of forecasters predicting same periods of real GDP[7]. Clearly, all of the pairs exhibit the positive correlations in our

[7] Figure 2 using the actual forecast errors results in more extreme results, correlations are larger and



(a) Correlation

(b) Variances

Figure 2: Variance and Correlation between each pair of forecasters

sample. In addition, most of pairs centered around high correlations, demonstrating that forecasters tend to rely on similar information when making predictions. The right plot shows the scatter plot of the variance for all pairs of forecasters. Most points align closely along the 45^o line, suggesting that the forecasters are really similar in variance. Both plots provide evidence for assuming that $\tilde{\Sigma}$ is a positive definite matrix with identical diagonal and positive off-diagonal entries.

The empirical results from the SPF data inform possible constraints on the form of $\tilde{\Sigma}$. The variances of the forecast errors tend to lie on the 45 degree line, suggesting that the diagonals are all similar. We denote this value σ_f^2 and scale $\tilde{\Sigma} = \sigma_f^2 \Sigma$ where now Σ is a correlation matrix with ones on the diagonal and correlations elsewhere. We denote the correlations r_{ij}. The empirical results also suggest that forecast errors made by the forecasters around the true conditional mean $\mu_{t,t+1}$ are positively correlated. Below in Sections 3.2 and 3.3 we restrict ourselves to the space of correlation matrices where Σ is a non-negative correlation matrix.

3 Theoretical Results for a Fixed Number of Forecasts

Gains from combining forecasts using estimated weights require that the estimation error be smaller than the relative gains in using average weights, which require no estimation. If the relative gains are small, it would not be surprising that estimation error results in combined forecasts that are outperformed by simple averaging. It is also interesting to understand the

variances closer to the 45 degree line.

conditions under which the gains are large. We define $M_{\tilde{\Sigma}}$ to be the space of positive definite mxm matrices and M_{Σ} to be the spaces of positive definite matrices under the restrictions of Proposition 2 below.

3.1 The Space of $\tilde{\Sigma}$ when Averaging is Optimal.

Define the space $M^a_{\tilde{\Sigma}}$ to be the subspace of $M_{\tilde{\Sigma}}$ for which the optimal combination weights are equivalent to the simple average of the forecasts (hence the superscript a). This space can be large for even moderate m. Elliott and Timmermann (2016) (p314-5) show that if the unit vector lies in the eigen space of $\tilde{\Sigma}$, then average forecasts are optimal. This shows that for all $\tilde{\Sigma} \in M_{\tilde{\Sigma}}$ that if the row sums of $\tilde{\Sigma}$ are equal, then $\tilde{\Sigma} \in M^a_{\tilde{\Sigma}}$. This can be a large space even for moderate m.

Special cases have been noted in the literature — for example when $\tilde{\Sigma}$ has all variances equal to each other and all covariances equal to a constant (Capistran and Timmermann (2009), Hsiao and Wan (2014)). In the case of $m > 4$ the eigen value result is richer than the previous example even when the covariances are all equal. For example the following two matrices in the $m = 4$ case yield equal weights with nonequal covariances;

$$\tilde{\Sigma} = \begin{pmatrix} 1 & r_1 & r_2 & r_1 \\ r_1 & 1 & r_1 & r_2 \\ r_2 & r_1 & 1 & r_1 \\ r_1 & r_2 & r_1 & 1 \end{pmatrix} \text{ or } = \begin{pmatrix} 1 & r_1 & r_1 & r_2 \\ r_1 & 1 & r_2 & r_1 \\ r_1 & r_2 & 1 & r_1 \\ r_2 & r_1 & r_1 & 1 \end{pmatrix} \tag{4}$$

(the second of these is a permutation of the first where the third and fourth forecasters are swapped with each other). A point to take from this result is that the subset $M^a_{\tilde{\Sigma}}$ in $M_{\tilde{\Sigma}}$ is very large and spread out over $M_{\tilde{\Sigma}}$ in a noncontiguous way so there are very many points in $M_{\tilde{\Sigma}}$ for which losses are equivalent for both optimal and averaging.

The result that the averaging vector lies in the eigen space of $\tilde{\Sigma}$ for the optimal combination weights to equal the averaging weights is a necessary condition.

Proposition 1. *The only optimal combination that is also an eigen vector of $\tilde{\Sigma}$ is the averaging weights.*

Proofs of Propositions are in the appendix.

One way to think of this is that the eigen vector as a variance reduction approach is implied by replacing the requirement that the weights sum to one with the requirement that the sum of the squared weights sum to one (See Hsiao and Wan (2014), who argue for use of the estimated eigen vector for constructing a weighted combination forecast). Such a requirement generally results in biased combined forecasts, the only eigen vector that does not lead to biased combined forecasts is the equal weights as an eigen vector as in the proposition.

These results suggest that there is a large space of parameterizations of $\tilde{\Sigma}$ that will result in averaging weights being optimal. It is also important to realize that the set of such matrices is not a closed space, but are instead lots of different non contiguous points in $M_{\tilde{\Sigma}}$. Parameterizations not too far from this space will also result in the loss of using averaging over the known optimal weights will be small.

3.2 The Space of $\tilde{\Sigma}$ when Averaging is Worst.

Whilst the subspace $M^a_{\tilde{\Sigma}}$ where estimating weights does not help is large, there are possible sets of forecasts for which the optimal weights differ considerably from the average weights. Such situations are the best case scenario for successfully estimating the optimal weights from the data. The best gains from optimal combination require either one forecaster or subgroup be much better than others or that forecast errors be negatively correlated with each other. As we have discussed in the previous section, motivated by the SPF data, in many practical situations neither of these are true. So in this section we examine potential gains for a restricted space of variance covariance matrices of the forecast errors.

We restrict $\tilde{\Sigma} = \sigma_f^2 \Sigma$ such that Σ is a correlation matrix with nonnegative correlations $r_{ij} \geq 0$. Essentially we are restricting the problem to be one where the variances of the forecast errors are the same across forecasters, and their forecast errors are positively correlated. These restrictions were motivated by the empirical data in the previous section. With the additional restriction that the optimal weights are nonnegative ($\Sigma^{-1}\iota \geq 0$), we can define the space of matrices $M^o_{\Sigma} \in M_{\Sigma}$ such that the optimal weights result in the largest deviation of loss from averaging minus the loss from using these optimal weights, i.e. we find specifications of Σ such that we maximize

$$m^{-2}(\iota'\Sigma\iota) - (\iota'\Sigma^{-1}\iota)^{-1}. \tag{5}$$

These are the situations that are most advantageous to estimating weights rather than using a simple averaging approach. To do this consider blocking Σ into two blocks of dimension m_1 and $m - m_1$ respectively, so

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{11} \in R^{m_1 \times m_1}$ etc. Further, define ι_1 as an $m_1 x1$ vector of ones and ι_2 as an $(m - m_1)x1$ vector of ones.

Proposition 2. *Given m_1 and Σ_{22} positive definite such that its minimum row sum exceeds $\frac{m-2m_1}{m_1}$ then the set of solutions to maximizing (5) subject to both $\Sigma^{-1}\iota \geq 0$ and $r_{ij} \geq 0$ for $i, j = 1, ..., m$ where $j > i$ requires Σ to satisfy*
(i) $\Sigma_{11} = I_{m1}$
(ii) $\Sigma_{12} = m_1^{-1}\iota_1\iota_2'$

It follows directly from the results of Proposition 2 that the additional loss from using simple averaging over the optimal weights in the worst case scenario is equal to

$$m^{-2}(\iota'\Sigma\iota) - (\iota'\Sigma^{-1}\iota)^{-1} = \frac{1}{m^2}(2m - m_1 + \iota_2'\Sigma_{22}\iota_2) - \frac{1}{m_1}. \tag{6}$$

The proof of Proposition 2 shows that for the restrictions on Σ that deliver non negative weights, having fixed both m_1 and Σ_{22}, that the worst case scenario for using simple averaging over the Bates-Granger optimal weights is a situation where the first m_1 forecast errors are uncorrelated and the correlation between the first m_1 forecast errors and the remaining forecast errors takes a very specific form.

It follows directly from this restriction that the optimal weights in this worst case scenario for simple averaging is to take the simple average over the first m_1 forecasts (see Lemma 1 in the appendix), i.e.

$$w^{opt} = \begin{pmatrix} (1/m_1)\iota_1 \\ 0 \end{pmatrix}$$

We fix both m_1 and Σ_{22} in the theorem, and now consider how varying these makes choosing to average a subset of the forecasts leads to a greater advantage over averaging over all the forecasts. Since under averaging, given the restrictions on Σ_{11} and Σ_{12} identified in the theorem for the local maxima, we have that loss is $m^{-2}(2m - m_1 + \iota_2'\Sigma_{22}\iota_2)$, that loss is increasing in $m - m_1$ with the latter $m - m_1$ forecast errors as correlated as possible whilst retaining the positive definiteness of Σ. At the same time loss under the optimal weights is not dependent on Σ_{22}.

To characterize the form of Σ_{22}, we use two steps. First, it follows directly that the optimand is increasing in $\iota_2'\Sigma_{22}\iota_2$. Hence it follows directly that the worst case scenario for averaging is when the correlations in Σ_{22} are as large as possible. Second, we require that the minimal eigen value of Σ is nonnegative, we wish to make the correlations as large as possible whilst keeping the smallest eigen value nonnegative so that Σ remains positive semidefinite.

For the second step, consider Σ_{22} such that all of the correlations are equivalent and equal to ρ_{22}. From Cadima et. al (2010) we can characterize the eigen values for this matrix. We have that $(m_1 - 1)$ of the eigen values are equal to one, $(m - m_1 - 1)$ of the eigen values are equal to $1 - \rho_{22}$. The remaining two eigen values are given by the formula

$$\frac{1}{2}\left(2 + \rho_{22}(m - m_1 - 1) \pm \sqrt{4(\frac{m - m_1}{m_1}) + \rho_{22}^2(m - m_1 - 1)^2}\right).$$

From these results we can examine the range of ρ_{22} for which Σ is positive definite given the solution of the optimization problem. We require that

$$\frac{1}{2}\left(2 + \rho_{22}(m - m_1 - 1) - \sqrt{4(\frac{m - m_1}{m_1}) + \rho_{22}^2(m - m_1 - 1)^2}\right) > 0$$

for some $0 < \rho_{22} < 1$. Rearranging this and solving for a bound on ρ_{22} the inequality requires that

$$\rho_{22} > \frac{m - 2m_1}{m_1(m - m_1 - 1)}.$$

For a solution we need that this holds for some $0 < \rho_{22} < 1$. Thus

$$\rho_{22} \in \left(\frac{m - 2m_1}{m_1(m - m_1 - 1)}, 1 \right).$$

For the case where $m_1 \geq m/2$ then the lower eigen vector does not depend on ρ_{22} so is valid for the entire range, i.e. $0 < \rho_{22} < 1$. For $m_1 = 1$ there is no ρ_{22} for which this holds, as for all m we would need $\rho_{22} \geq 1$.

These results suggest that the worst case scenario for averaging over using optimal weights is when Σ_{22} has off diagonal values that are equal to each other and as close to one as possible. This means that the 'dropped' forecasters are basically giving the same forecast as each other. Note that this result also has the effect that for smaller m_1, so $m - m_1$ is larger, the additional losses from averaging over all the forecasters will be larger. In the numerical work that follows we choose $\rho_{22} = 0.99$.

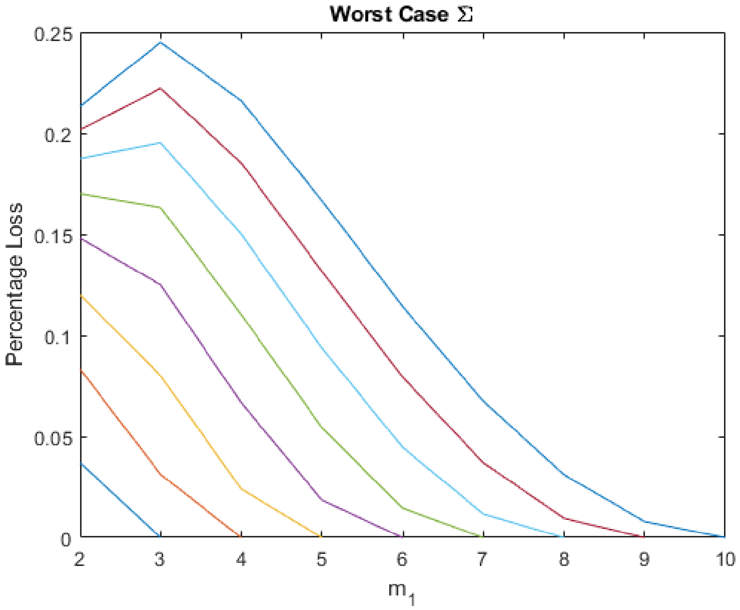


Figure 3: Relative Losses for Σ following from the results of Proposition 2 for $m = 3$ to $m = 10$. Curves further to the right are for larger m. We set $\sigma_\epsilon^2 = \sigma_f^2 = 1$.

The numerical results in Figure 3 show values from Equation 6 divided by the optimal loss (here $\sigma_\epsilon^2 + 1/m_1$ with σ_ϵ^2 and $\sigma_f^2 = 1$) to give relative losses from using averaging over the most extreme parameterization of Σ. Results are shown for each possible m_1 for $m = 3$ to $m = 10$, with successive curves to the right being for larger m. We see that the relative loss can be large for these extreme situations where the optimal weighting is to drop $m - m_1$ of the forecasts and average over the rest when this is large. But even for this extremal case there are many potential m_1 for any m for which the gains are likely small relative to the estimation error which is of order $(m - 1)/T$. For example with $T = 100$ and $m = 6$ we have that for $m_1 > 4$ the gain is smaller than expected estimation error.

In some sense these results are artificial - with large m we can have situations where there $m - m_1$ is large but this means that the dimensions of Σ_{22} is large and can be populated with large correlations making the difference between averaging and optimal weights large. However what this means in practice is that one has very many almost identical forecasts that are 'padding' the set of forecasts. Practically one might ignore these forecasts if they are almost identical. This suggests that reasonable worst case Σ would have m_1 closer to m and hence less of an available difference in the losses.

The result does remain that there are possible empirically relevant cases where averaging is not a particularly good approach, and we would be better off choosing a subset of forecasts to average over.

3.3 Gains from Optimal Combination over the Space of Σ.

The results of Proposition 2 show the worst case scenarios for using the averages of the forecasts to construct combinations relative to using the optimal combination in population. However the precise nature of Σ for this worst case is quite extreme - a subset of uncorrelated forecast errors as well as another subset of highly correlated forecast errors that have a very precise correlation with the first set (the restriction on Σ_{12} in Proposition 2). Obviously it is more likely that the variance covariance matrix results in a set of optimal weights and a difference in losses that lies between no gain and these upper bounds on the gains. This section considers the distribution of such gains whilst retaining the conditions of Proposition 2.

So what these calculations do not do is show where the 'mass' of relative gains from optimal combination lie over possible correlations between forecasts. For most such parameterizations, the gain in population from using optimal combinations lies well below the worst case scenario.

In Figure 4 we show the distribution of relative gains from using optimal weights over averaging for randomly drawn Σ that satisfy the assumptions of Proposition 2 for $m = 4$ to $m = 8$ (higher peaks are larger m). We scale Equation 2 by $\sigma_\epsilon^2 + (\iota_m' \Sigma \iota_m)^{-1}$ (setting $\sigma_f^2 = \sigma_\epsilon^2 = 1$) so that we report the percentage loss from using average weights over optimal weights. These values on the x-axis can be compared directly to $(m-1)/T$.

The point to note is that in comparison with the results in Figure 3, the mass of draws of Σ in M_Σ have gains from using optimal weights that are very far from the potentially worst case scenario gains derived in the previous section. For all of these cases the vast majority of losses from using averaging over optimal weights are below 5%, which is a small gain when one must estimate the weights unless sample sizes are quite large. This indicates that we might very much expect that estimation error vastly outweighs the potential gains from optimal combinations for variance covariance matrices that lie in M_Σ.

These gains can be compared directly to the expected loss from estimation, which in the best case scenario would be to compare these results to $(m-1)/T$ where we have T observations for estimating the weights. It is notable that as m gets larger, the distributions

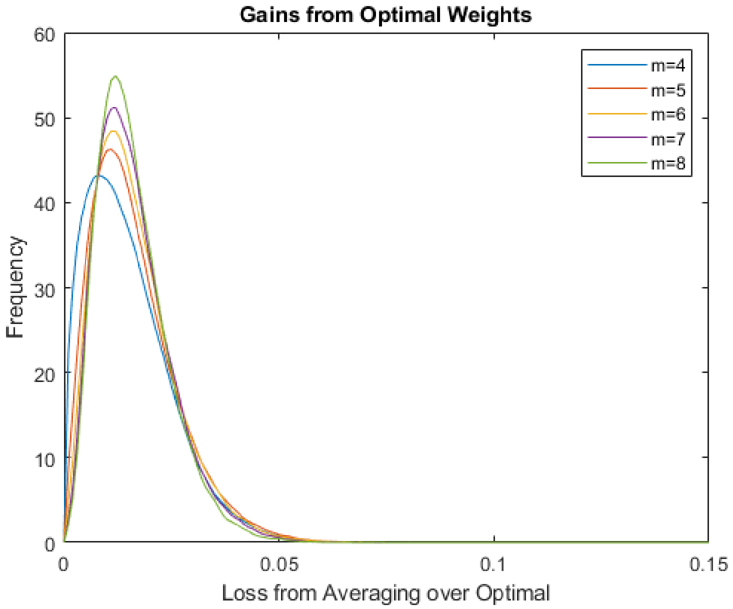


Figure 4: Relative Losses across Parameterizations of Σ for $m = 4, 5, 6, 7$. Curves further to the right are for larger m.

of gains is not moving much to the right, so for larger m we have a larger effect from estimation error without there being much expected gain from estimation.

For example consider the case where $m = 7$, as shown in Figure 5. Here the vertical lines show the expected losses from estimation for $T = 200$ and $T = 100$, both relatively large sample sizes for forecast combination exercises. For a sample size of $T = 100$ nearly all of the gains from optimal combination are smaller than the expected sampling error. At $T = 200$ still 90% of them are.

4 Results when the Number of Forecasts is Large.

When the number of forecasts becomes larger, we might expect that averaging outperforms estimated optimal combinations because the number of weights needed for estimation becomes larger, making the estimation error component larger. Whilst this is true, there are also implications for the space of $\tilde{\Sigma}$ for which optimal weights have relatively smaller or larger gains which can make gains from optimal combination more difficult to obtain even when we do not factor in estimation error.

The following result shows that for a general $\tilde{\Sigma}$ (without the restrictions of Proposition 2) then if the largest eigen value of $\tilde{\Sigma}$ is bounded as m becomes large, then averaging and optimal weights yield the same loss.

In the fixed m case we saw that there was a wide set of possible $\tilde{\Sigma}$ for which average weights are optimal, here we see that the space of such matrices when m is large is itself

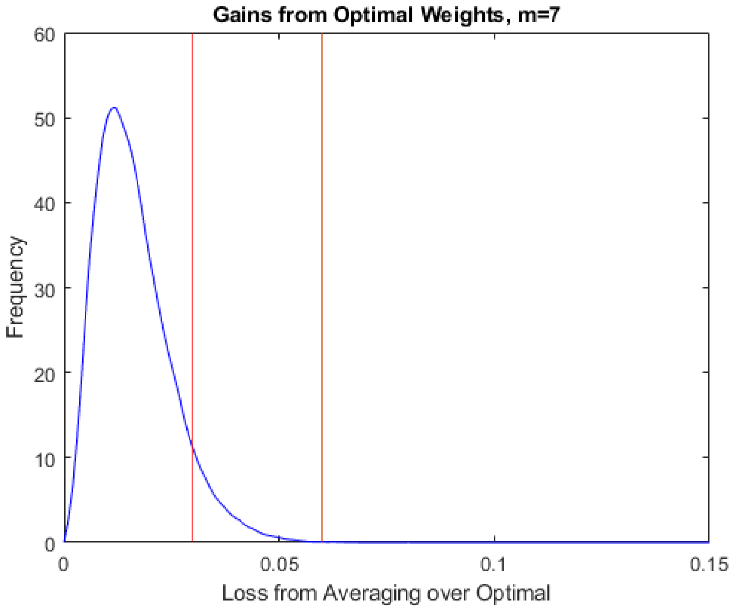


Figure 5: Relative Losses across Parameterizations of Σ for $m = 7$. The orange vertical line shows the expected loss from estimation with $T = 100$ and the red line with $T = 200$.

large. Thus even without estimation error we might expect that averaging performs as well as using optimal combinations. Such a result also suggests that the approach of Consensus Economics is justified theoretically.

Proposition 3. *Let λ_i, i=1,...,m be the eigen values of $\tilde{\Sigma}$, and assume that $\lambda_{\max} = \max_{i=1,\ldots,m} \lambda_i < K$ for some finite K. Then*

$$\begin{array}{rrcl} (i) & m^{-2}\iota_m'\tilde{\Sigma}\iota_m & = & o(m^{-1}) \\ (ii) & (\iota_m'\tilde{\Sigma}^{-1}\iota_m)^{-1} & = & o(m^{-1}) \end{array}$$

and so MSE loss from both methods is the same for large m.

Bounding the largest eigen value of the variance covariance matrix of forecast errors essentially says that no single eigen vector can completely explain the variance of the forecast error. As a result addition linear combinations of the forecasts will have some ability to explain the forecast error. As the number of forecasts increases, we expect the weights from both averaging and optimal weights to give similar losses.

An example is a Toeplitz structure for the variance covariance matrix. Here

$$\Sigma = \begin{pmatrix} 1 & \rho & 0 & \ldots & 0 \\ \rho & 1 & \rho & 0 & \ldots \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & \ldots & 0 & \rho & 1 \end{pmatrix}$$

We do not have equal weights for this case, although many of the weights ($m-2$ of them) are close to even weights. The row sums are not the same because of the first and last rows. The eigen values are bounded over m, between zero and $1+2\rho$ so this matrix satisfies the assumptions of Proposition 3. The loss from averaging is

$$\frac{1}{m^2}\iota'\Sigma\iota = \frac{m+2\rho(m-1)}{m^2}$$

which asymptotes to zero at rate m as $m \to \infty$.

For the specification of Σ in Proposition 2, we know that for all m the optimal forecast combination is to average over a subset of the forecasts providing a gain over averaging. Since this remains true for any m, it must be true as m gets large. Thus we expect that the largest eigen value for that special case must be diverging. This is indeed the case. The worst case (for averaging) matrix Σ is now

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} I_{m_1} & \frac{1}{m_1}\iota_1\iota_2' \\ \frac{1}{m_1}\iota_2\iota_1' & \Sigma_{22} \end{pmatrix}$$

with off diagonals of Σ_{22} fixed for a large correlation, write as ρ_{22}.

From Cadima et al. (2010) we have that the largest eigen value of Σ is given by

$$\frac{1}{2}\left(2+\rho_{22}(m-m_1-1)+\sqrt{4(\frac{m-m_1}{m_1})+\rho_{22}^2(m-m_1-1)^2}\right)$$

For a fixed m_1 the largest eigen value diverges as $m \to \infty$. As expected (since loss for the optimal case is $\frac{1}{m_1}$ for each m) this violates the conditions where a bounded eigen value means that the two approaches yield the same loss for large m.

For the worst case scenario from Proposition 2 we can evaluate the relative loss from averaging as m gets large. Consider allowing both $m \to \infty$ and $m_1 \to \infty$ such that $\frac{m_1}{m} \to c$. Relative loss normalized by m is then (from Equation 6)

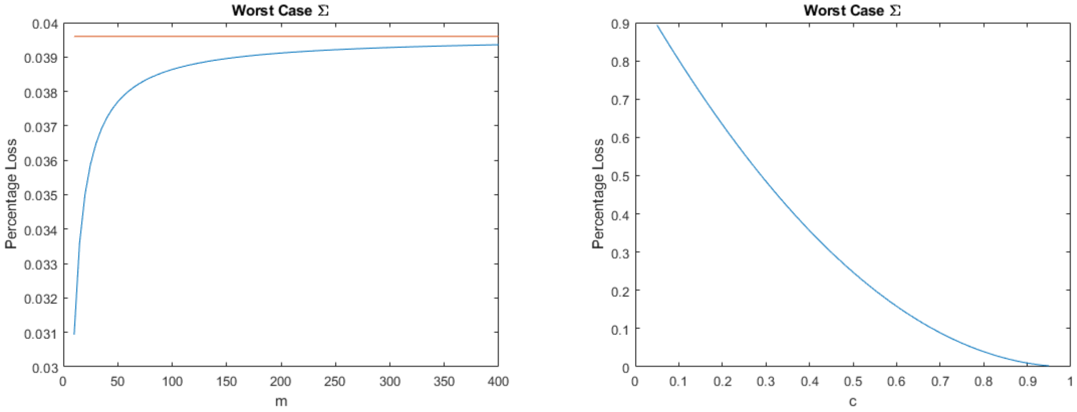
$$\begin{aligned}\left(\frac{m^{-2}(\iota'\Sigma\iota)-(\iota'\Sigma^{-1}\iota)^{-1}}{\sigma_\epsilon^2+(\iota'\Sigma^{-1}\iota)^{-1}}\right) &= \frac{\frac{1}{m^2}(2m-m_1+(m-m_1)+((m-m_1)^2-(m-m_1))\rho_{22})-1/m_1}{\sigma_\epsilon^2+\frac{1}{m_1}} \\ &\to \sigma_\epsilon^{-2}\rho_{22}(1-c)^2\end{aligned}$$

as $m \to \infty$.

A few observations follow from this result. First, the worst case relative loss from using averaging (under the conditions of Proposition 2) is for any m_1 growing with rate m, which is the same rate as which estimation error grows with m. Hence they are of the same order, we would then expect that trade-off's between these exist for all m, even m large. Second, the term in the numerator that dominates is due to the correlations in Σ_{22}, which means there are a fraction of the forecasters that are nearly perfectly correlated. As in the fixed m case we would argue that this is unlikely, and that a larger m_1 relative to m is more reasonable (so c is relatively large).

We can examine this numerically. In Figure 6 we examine the relative loss as in the above equation. In the left panel we set $c = 4/5$ and vary m, in the right panel we set $m = 200$ and vary c. In the left panel we see that over all m, the limit relative loss is quite low. As in the equation above, as the number of forecasts gets large but the relative number of forecasts we average over in the worst case also gets large, there is not a great difference asymptotically (in m) between averaging and optimal weights. Here the maximal gain is about 4% which for large m would be very small relative to estimation error.

In the right hand panel of Figure 6 we can see the effect of varying c. For c much smaller, the gain can be much larger (at $c = 1/5$ it is near 60%, which still might not be enough to offset estimation error if m is large relative to the number of observations). Overall we might then expect that averaging is likely to work much better for larger m, given that these gains are upper bounds given the empirically motivated restrictions on Σ.



(a) c=4/5

(b) m=200.

Figure 6: Relative loss when $\frac{m_1}{m} \to c$.

Hence again for large m we might expect that averaging performs better than estimating weights.

5 How well do methods work?

In this section we propose a new method for estimating the combination weights, and compare in a number of Monte Carlo experiments how this method and standard methods perform.

5.1 A subset Averaging Approach

The additional method we suggest follows from the results of Section 3.2. For the worst case scenario we have shown that the optimal weights are to average over a subset of the forecasts. This suggests considering looking for the best subset to average over. By refining the set of models we examine in the estimation procedure, the hope is that estimation error is smaller for this procedure as it does not look at models that are likely to only provide small gains.

Our subset average weights are constructed by looking for each $m_1 = 3, ..., m$ every possible permutation of m_1 forecasts. The subset with the smallest in sample MSE is chosen to be the set of forecasts to average over. Even for values of m large from the perspective of actual applications, this method is very fast. This is because only sample averages need be estimated, no matrices are inverted, and so the search procedure is quick. After computing all of the sample averages, the models are simply ordered and the best one chosen. We include the full sample average over all of the forecasts in the procedure, so it is possible that the full average is chosen.

This method introduces estimation error into the construction of the weights because of the search across models will due to sampling choose the model that is not best in population on occasion. However by searching over a sparse set of models it is expected that the sampling error would likely be smaller than many other approaches, and that a subset average close enough to the population optimal weights will be chosen with a high probability.

5.2 Other Methods

As we noted in the introduction, even in Bates and Granger (1969) it was understood that estimating weights did not necessarily result in better performance. The expected loss for restricted OLS when we include estimation error is $(\sigma_\epsilon^2 + \sigma_f^2 \omega' \Sigma \omega)(1 + \frac{m-1}{T})$ when there are T observations available for estimating the weights. This can be greater than the expected loss under averaging because of the additional $\frac{m-1}{T}$ term. The literature broadly has suggested a number of methods that reduce the need for estimation and hence result hopefully in better performance. We examine restricted OLS and simple averaging. The two additional methods we include are weights based on the inverse of the individual in sample MSE's and weights based on ranks of these MSE's. For our Monte Carlo designs, because the population value of the individual MSE's are equivalent, the estimated weights based on inverse MSE's will converge to the average weights, so for large enough sample sizes the methods will perform similarly. Weights based on ranks however will be somewhat random and this method cannot replicate the average weights (it is designed for situations where there are more and less dominant forecasters), so would be expected to perform poorly for these simulations.

We examine in the Monte Carlo results the performance of a number of combination methods in addition to the one proposed in this paper,

- Restricted OLS so that weights sum to one, without imposing non negative weights;
- Average Weights;
- Shrinkage Weights (the first method with shrinkage towards the second listed method) (Diebold and Pauly (1990),Stock and Watson (2004)), denoted Shrink;

- Weights based on the inverse of the individual MSE's (denoted MSE);
- Weights based in the rankings of the individual MSE's (Aolfi and Timmermann (2006)), denoted Rank.

5.3 Monte Carlo Results

The Monte Carlo results of this section explore the results of Sections 3 and 4. For the fixed m relatively small we consider three designs for the variance covariance matrix of the forecasts (around their conditional mean). The first accords to Section 3.2 results, where $\tilde{\Sigma} = \sigma_f^2 \Sigma$ and Σ is of the form that accords with the results of Proposition 2 so the optimal weights are the average of a subset of the forecasters. The second design chooses Σ randomly (with correlations uniformly distributed on $[0, 1]$) such that Σ is positive definite and is of the form for the conditions of Proposition 2 (so has unit variances and is a correlation matrix with nonzero elements, and the optimal weights are nonnegative). The third design is the same as the second design but we drop the binding constraint that the weights are nonnegative, which allows for larger differences between the expected loss using optimal weights over averaging. This third design explores the importance of this constraint numerically.

For each of the tables the Monte Carlo results reported are constructed using $T = 100$ to estimate the weights for any of the estimation methods. All results are numerically equivalent for any parameterization of $\mu_{t,t+1}$. Reported are the losses averaged over Monte Carlo draws where for each estimated set of weights we draw 1000 out of sample values for evaluation and we average over 10000 estimates of the weights. Panels are for different values of σ_f^2. A larger value for this parameter can be interpreted as the outcome being relatively more forecastable, and so the differences between methods is on average stronger.

For each of the combination methods, the out of sample loss depends on the sample size, σ_ϵ^2, σ_f^2 and Σ. We fix the estimation sample size at T=100, set $\sigma_\epsilon^2 = 1$ as a normalization, and vary σ_f^2. Recall that a smaller σ_f^2 means that y_{t+1} is less forecastable, which mutes the relative differences between methods (Elliott (2016)). Results reported are from using the estimated weights to predict 1000 times for each estimated set of weights. For averaging there is no estimation error, and the estimated MSE is close to the population MSE for each m. For optimal weights there is estimation error, on average it results in estimated losses that are larger than simple averaging.

For the optimal combination via restricted OLS, we report the asymptotic ($T \to \infty$ with m fixed) loss $\sigma_\epsilon^2 + \sigma_f^2/(\iota_m' \Sigma^{-1} \iota_m)$ in the first column and the expected loss from this method with estimation error in the second column. The expected loss under averaging $\sigma_\epsilon^2 + \frac{\sigma_f^2}{m^2}(\iota_m' \Sigma \iota_m)$ is in the fourth column. The first column numbers are always at least as good if not better than the fourth, given the optimality of the weights in the first column. However adjusted for estimation error (the second column) typically this ranking is reversed.

Results in Table 1 examine the procedures when $\Sigma \in M_\Sigma^o$ with a scaling coefficient σ_f^2 as noted in the each panel of the table. We set $\sigma_\epsilon = 1$ as a normalization. We have different

m	m_1	Optimal Weights			Average Weights		Subset	Shrink	MSE	Rank
		Pop	E(Est)	Est	Pop	Est				
$\sigma_f^2 = 0.75$										
3.000	2.000	1.375	1.403	1.404	1.417	1.417	1.390	1.416	1.419	1.454
4.000	3.000	1.250	1.288	1.288	1.281	1.281	1.281	1.279	1.282	1.328
5.000	2.000	1.375	1.430	1.431	1.492	1.491	1.395	1.482	1.491	1.507
5.000	4.000	1.188	1.235	1.237	1.210	1.210	1.233	1.209	1.212	1.264
6.000	2.000	1.375	1.444	1.448	1.517	1.516	1.400	1.502	1.517	1.527
6.000	5.000	1.150	1.208	1.211	1.167	1.167	1.204	1.165	1.168	1.223
7.000	2.000	1.375	1.458	1.464	1.536	1.536	1.402	1.516	1.536	1.538
7.000	5.000	1.150	1.219	1.224	1.196	1.195	1.206	1.190	1.196	1.245
7.000	6.000	1.125	1.193	1.197	1.138	1.137	1.185	1.136	1.139	1.195
8.000	2.000	1.375	1.471	1.480	1.551	1.552	1.405	1.527	1.552	1.549
8.000	5.000	1.150	1.230	1.237	1.227	1.228	1.208	1.217	1.228	1.269
8.000	7.000	1.107	1.185	1.191	1.117	1.117	1.171	1.116	1.118	1.175
$\sigma_f^2 = 1$										
3.000	2.000	1.500	1.530	1.530	1.556	1.555	1.513	1.553	1.557	1.603
4.000	3.000	1.333	1.373	1.375	1.375	1.375	1.365	1.373	1.377	1.439
5.000	2.000	1.500	1.560	1.563	1.656	1.655	1.522	1.642	1.655	1.678
5.000	4.000	1.250	1.300	1.301	1.280	1.278	1.295	1.276	1.280	1.349
6.000	2.000	1.500	1.575	1.578	1.689	1.687	1.525	1.668	1.687	1.699
6.000	5.000	1.200	1.260	1.265	1.222	1.223	1.258	1.221	1.224	1.298
7.000	2.000	1.500	1.590	1.596	1.714	1.715	1.528	1.689	1.715	1.717
7.000	4.000	1.250	1.325	1.330	1.376	1.375	1.302	1.360	1.376	1.422
7.000	5.000	1.200	1.272	1.278	1.261	1.261	1.260	1.254	1.263	1.329
7.000	6.000	1.167	1.237	1.242	1.184	1.183	1.232	1.181	1.185	1.260
8.000	2.000	1.500	1.605	1.614	1.734	1.735	1.530	1.702	1.735	1.731
8.000	5.000	1.200	1.284	1.293	1.303	1.304	1.263	1.289	1.305	1.358
8.000	7.000	1.143	1.223	1.230	1.156	1.156	1.214	1.155	1.158	1.235
$\sigma_f^2 = 1.25$										
3.000	2.000	1.625	1.657	1.659	1.694	1.694	1.639	1.691	1.697	1.754
4.000	3.000	1.417	1.459	1.462	1.469	1.469	1.448	1.466	1.472	1.548
5.000	2.000	1.625	1.690	1.693	1.820	1.820	1.648	1.804	1.820	1.849
5.000	4.000	1.312	1.365	1.367	1.350	1.350	1.359	1.347	1.352	1.439
6.000	2.000	1.625	1.706	1.710	1.861	1.861	1.651	1.837	1.860	1.875
6.000	5.000	1.250	1.312	1.316	1.278	1.277	1.309	1.274	1.280	1.372
7.000	2.000	1.625	1.723	1.729	1.893	1.893	1.654	1.860	1.893	1.899
7.000	5.000	1.250	1.325	1.330	1.327	1.326	1.312	1.317	1.328	1.411
7.000	6.000	1.208	1.281	1.286	1.230	1.230	1.278	1.227	1.232	1.326
8.000	2.000	1.625	1.739	1.749	1.918	1.918	1.656	1.876	1.918	1.914
8.000	5.000	1.250	1.338	1.345	1.379	1.380	1.316	1.361	1.381	1.448
8.000	7.000	1.179	1.261	1.268	1.195	1.195	1.254	1.193	1.197	1.293

Table 1: Results for Pop are $\sigma_\epsilon^2 + \sigma_f^2 \omega' \tilde{\Sigma} \omega$ for Optimal and Average weights, For E(Est) this is augmented by $1 + (m-1)/T$. Other columns are averages over using $T = 100$ to construct weight estimates and the average MSE is reported.

values for Σ for each m, m_1 pair, some pairs are omitted for table readability. The optimal weights here then are to average over the first m_1 forecasts and ignore the remaining forecasts. Comparing the population MSE (Pop) from combination for optimal weights vs average weights shows the size of the gains from ignoring the remaining $m - m_1$ forecasts. They are larger for larger σ_f^2, as the forecastability of the outcome becomes better. When restricted OLS is used to estimate the weights (without knowledge of their form, just imposing that they sum to one) then the expected estimation error increases the expected MSE, which is given in the E(Est) column. This may be larger or smaller than using average weights - the basic point that estimation error can outweigh the gains from optimal combination over averaging.

Comparing estimated to average weights, it is still the case that averaging can be better despite this being the largest deviation between optimal and average weights under the conditions of Proposition 2. The difference in the average weights to optimal weights for forecasts that have nonzero weights is $(m_1 - m)/(m_1 m)$. For models where m_1 is close to m, this will be smaller. The results in the Table (comparing the out of sample MSE's for Average vs. Optimal) show that in such cases averaging still outperforms estimating the optimal weights. As we noted in discussing this result, values with m_1 much smaller than m are unlikely models where a large number of the forecasters have essentially the same forecast. When m_1 is small relative to m the difference between the optimal weights and averaging weights is much larger, as are the population MSE's, and here estimating the weights can outperform averaging, even substantially.

This is the best case scenario for the subset approach to estimation, which has the same population expected MSE but the search procedure introduces estimation error. For a larger Σ_f^2 and a larger $m - m_1$ the subset approach outperforms other methods. Not only does it have the advantage that it is 'looking in the right direction', these are situations where it is relatively better to get the model correct and the best weights are often zero which is a corner solution for other estimators.

The MSE method is basically equivalent to the averaging approach because the variances are all equal here in the Monte Carlo design, so in population the MSE approach will produce weights that are consistent for the average weights. Hence the only difference is estimation error in computing the individual MSE's. In contrast the rank method cannot produce even weights by design, and is best suited to situations where some forecasting methods outperform others consistently, which is not the design of this Monte Carlo. Hence the method is generally the poorest for these results. This intuition for the results extends to the following tables as well.

Next consider the problem in Section 3.3 where the restrictions on the space for Σ are as in Proposition 2 and drawn randomly rather than having the worst case solution occur. Table 2 gives the results for the average out of sample MSE.

The results in Table 2 show the results from drawing random correlation matrices Σ. Since we are drawing different Σ for each Monte Carlo round, the optimal weights for each MC draw are different. The theoretical values for Pop and E(Est) for the optimal weights

Random Draws from Σ

m	Optimal Weights			Average Weights		Subset	Shrink	MSE	Rank
	Pop	E(Est)	Est	Pop	Est				
$\sigma_f^2 = 0.5$									
4.000	1.256	1.294	1.295	1.267	1.267	1.285	1.267	1.268	1.292
5.000	1.232	1.281	1.283	1.243	1.243	1.267	1.243	1.244	1.271
6.000	1.215	1.276	1.280	1.227	1.227	1.255	1.227	1.227	1.256
7.000	1.202	1.274	1.280	1.214	1.213	1.245	1.213	1.214	1.244
8.000	1.193	1.276	1.283	1.204	1.203	1.237	1.203	1.203	1.235
$\sigma_f^2 = 0.75$									
4.000	1.384	1.426	1.427	1.401	1.401	1.419	1.401	1.402	1.438
5.000	1.347	1.401	1.404	1.365	1.365	1.390	1.365	1.366	1.407
6.000	1.322	1.388	1.393	1.340	1.340	1.371	1.340	1.341	1.384
7.000	1.303	1.381	1.387	1.320	1.320	1.355	1.320	1.321	1.366
8.000	1.193	1.276	1.283	1.204	1.203	1.237	1.203	1.203	1.235
$\sigma_f^2 = 1$									
4.000	1.512	1.557	1.559	1.535	1.534	1.553	1.534	1.536	1.585
5.000	1.463	1.522	1.525	1.487	1.487	1.513	1.486	1.488	1.542
6.000	1.430	1.501	1.506	1.453	1.454	1.486	1.453	1.455	1.512
7.000	1.404	1.489	1.495	1.427	1.427	1.464	1.426	1.428	1.488
8.000	1.385	1.482	1.490	1.407	1.407	1.448	1.406	1.408	1.469
$\sigma_f^2 = 1.25$									
4.000	1.640	1.689	1.691	1.668	1.668	1.687	1.668	1.670	1.731
5.000	1.579	1.642	1.645	1.609	1.608	1.635	1.608	1.610	1.677
6.000	1.537	1.614	1.620	1.567	1.567	1.600	1.566	1.569	1.640
7.000	1.505	1.596	1.603	1.534	1.534	1.573	1.533	1.536	1.610
8.000	1.481	1.585	1.594	1.509	1.508	1.553	1.507	1.510	1.587

Table 2: for Pop are $\sigma_\epsilon^2 + \sigma_f^2 \omega' \tilde{\Sigma} \omega$ for Optimal and Average weights, For E(Est) this is augmented by $1 + (m-1)/T$. Other columns are averages over using $T = 100$ to construct weight estimates and the average MSE is reported.

and Pop for the average weights then are different for each MC draw, the number reported for loss is the average across the draws. The remaining columns are averages of the out of sample performance as discussed above.

With random draws of Σ the optimal weights often result in losses that are not that much better than averaging, as we discussed in Section 3.3. This can be seen by comparing the columns for Pop for both the optimal and averaged weights. Further, when we (on average) include estimation error (compare E(Est) with Pop for averaging) then averaging has a smaller loss. This is indicating that for the majority of Σ that satisfy the empirical restrictions motivated by the SPF data, that the gains from using optimal weights over averaging is outweighed by estimation error. The out of sample MC results show that the theoretical results are indicative of what happens in practice. As above the MSE method approximates averaging because of the MC design, whereas the rank method is not suited to this MC design. Subset averaging, which includes averaging over the whole set of forecasts as a possible outcome, often does better than simply estimating the weights but is inferior to averaging. However the additional losses from estimation here are small and suggest it might be a worthwhile estimation approach.

We can using the same experiment as in Table 2 examine the outcomes from in and out of sample estimation versus averaging, as we did in the empirical section in Figure 1. For all of the rows in Table 2, the MSE for the averaging method outperforms the estimation of weights in a range of just above 50% ranging up to about 54%. In Figure 7 we have chosen $m = 4$ and $\sigma_f^2 = 1$ (which is one of the rows in the table) although for other values for m results are similar. The same shapes as we saw from the data are apparent here as well for models drawn from Σ as in the table. The main exception is that for the out of sample results we see that the distribution is a little more skewed towards using estimated weights - here $T = 100$ so estimates are more precise than we might expect in the data.

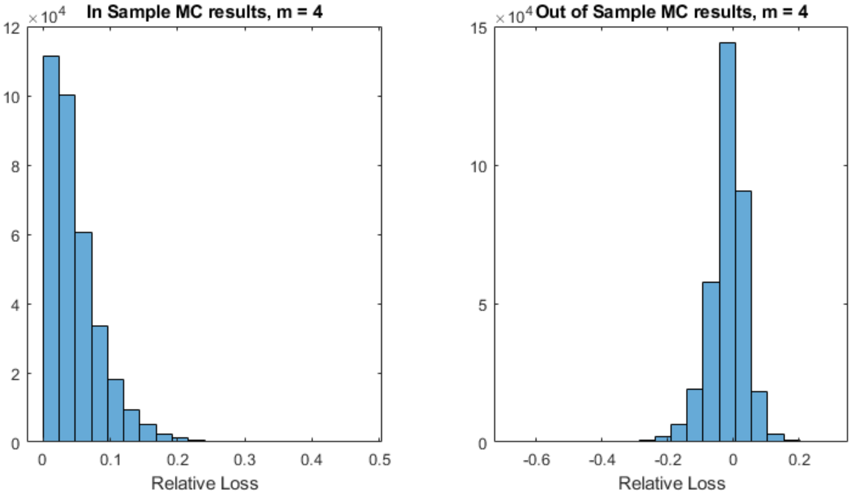


Figure 7: Relative Losses for in and out of sample evaluation

Random Draws from Σ without positive weights

m	Optimal Weights			Average Weights		Subset	Shrink	MSE	Rank
	Pop	E(Est)	Est	Pop	Est				
$\sigma_f^2 = 0.5$									
4.000	1.237	1.274	1.275	1.289	1.289	1.288	1.288	1.290	1.312
5.000	1.202	1.250	1.253	1.266	1.266	1.270	1.265	1.267	1.291
6.000	1.176	1.235	1.239	1.249	1.249	1.256	1.247	1.249	1.275
7.000	1.157	1.226	1.231	1.235	1.235	1.246	1.232	1.235	1.262
$\sigma_f^2 = 0.75$									
4.000	1.355	1.396	1.397	1.434	1.434	1.425	1.433	1.435	1.468
5.000	1.303	1.355	1.358	1.399	1.399	1.394	1.397	1.400	1.436
6.000	1.265	1.328	1.332	1.373	1.373	1.373	1.370	1.374	1.413
7.000	1.235	1.309	1.315	1.352	1.352	1.356	1.348	1.353	1.394
$\sigma_f^2 = 1$									
4.000	1.474	1.518	1.520	1.578	1.578	1.561	1.577	1.580	1.623
5.000	1.404	1.460	1.463	1.532	1.532	1.519	1.530	1.533	1.582
6.000	1.353	1.420	1.425	1.497	1.497	1.488	1.493	1.498	1.550
7.000	1.313	1.392	1.398	1.469	1.469	1.465	1.464	1.470	1.525
$\sigma_f^2 = 1.25$									
4.000	1.592	1.640	1.642	1.723	1.723	1.697	1.721	1.725	1.779
5.000	1.505	1.565	1.568	1.665	1.665	1.642	1.662	1.667	1.727
6.000	1.441	1.513	1.518	1.621	1.621	1.603	1.617	1.623	1.688
7.000	1.392	1.475	1.482	1.586	1.586	1.573	1.580	1.588	1.656

Table 3: for Pop are $\sigma_\epsilon^2 + \sigma_f^2 \omega' \tilde{\Sigma} \omega$ for Optimal and Average weights, For E(Est) this is augmented by $1 + (m-1)/T$. Other columns are averages over using $T = 100$ to construct weight estimates and the average MSE is reported.

For the third design, Table 3 shows the results where Σ is a correlation matrix but we relax the requirement that the optimal combination weights are all nonnegative. Since this was the binding constraint in Proposition 2, we expect that the differences between the losses for the optimal and averaging weights to be larger. And this is clearly the case, comparing Pop for the two methods. Now even with estimation error, it is often the case that estimating the weights will be superior to averaging.

There are a number of takeaways from these results. First, there is a large space for $\tilde{\Sigma}$, in combination with large enough sample sizes, for which the estimation error argument for why averaging outperforms restricted OLS cannot be the correct argument. As we have argued in this paper, empirically reasonable assumptions on $\tilde{\Sigma}$ flip this result. It is also the case that sample sizes matter and can flip the result back towards averaging. It is often very difficult to obtain a coherent long time series of forecasts for which to estimated the weights.

A direct implication of our results is that under these empirically relevant assumptions on $\tilde{\Sigma}$, the best gains are going to be equal to (or perhaps close to) weights that are an average on a subset of the forecasts. The subset method we introduce in this section works well for the Monte Carlo results presented. When the optimal method does result in averaging over a subset, the method works better than either averaging or estimating the weights with restricted OLS. For the restricted versions of Σ, it is still performing quite well.

6 Conclusion

The forecast combination puzzle is that simple averaging of forecasts typically beats estimating optimal weights via restricted OLS. Optimal combination weights in population must always lead (weakly) to a smaller loss than averaging by definition, however estimation error can reverse this ranking. Such a reversal can really only happen if the gains from optimality are small relative to the estimation error.

It is clearly the case that there exist possible correlation structures in the forecasts such that gains from optimality can be large. For example two forecasters who have forecast errors around the true conditional mean that are negatively correlated allows major gains from optimal combination. One forecaster that is much better than the others can be upweighted, resulting in large gains from optimal combination. In this paper, motivated by empirical results using the Survey of Professional Forecasters and their forecasts of real GNP, we show that the types of reasonable restrictions on the variance covariance of the forecast errors are such that for the most part, gains are small. When the number of forecasts to be combined is large, the space of possible covariance matrices that deliver large gains from optimal combinations can be shown to be small under some restrictions on the eigen values of the covariance matrix.

We characterize within these restrictions the worst case for averaging, and find that the result is to average over a subset of the forecasts. This motivates a method for estimating weights that appears in Monte Carlo to work well and reduce estimation error through searching over a smaller number of models.

References

Aolfi, M. and Timmermann, A. (2006). Persistence of forecasting performance and combination strategies. *Journal of Econometrics*, 135:31–53.

Armstrong, J. S. (2001). *Combining Forecasts*, pages 417–439. Springer US, Boston, MA.

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

Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2020). Overreaction in macroeconomic expectations. *American Economic Review*, 110(9):2748–82.

Cadima, J., Calheiros, F., and Preto, I. (2010). The eigenstructure of block-structured correlation matrices and its implications for principal component analysis. *Journal of Applied Statistics*, 37:577–589.

Capistran, C. and Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business and Economic Statistics*, 27:428–440.

Chan, Y., Stock, J., and M.W.Watson (1999). A dynamic factor model framework for forecast cobmbination. *Spanish Economic Review*, 1:91–122.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–581.

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–78.

Deutsch, M., Granger, C. W., and Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, 10(1):47–57.

Diebold, F. and Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, 6:503–508.

Elliott, G. (2016). Forecast combination when outcomes are difficult to predict. *Empirical Economics*, 53:7–20.

Elliott, G. and Timmermann, A. (2016). *Economic Forecasting.* Princeton University Press, Princeton and Oxford.

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

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

Jose, V. R. R. and Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24(1):163–169.

Kang, H. (1986). Unstable weights in the combination of forecasts. *Management Science*, 32(6):683–695.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71:302–355.

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

Timmermann, A. (2006). Forecast combinations. In et. al, G. E., editor, *Handbook of Forecasting Volume 1*, pages 135–196. Elsevier, Amsterdam.

Wang, X., Hyndman, R., Lee, F., , and Kang, Y. (2020). Forecast combinations: an over 50-year review. *manuscript*.

Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(1):176–222.

7 Appendix - Proofs of Results

Proof of Proposition 1.

Proof. To see this, we want to show that if $w^{opt} \neq m^{-1}\iota_m$, then w^{opt} is not an eigenvector of Σ. Suppose, to the contrary, that there is λ such that $\lambda w^{opt} = \Sigma w^{opt}$. We have $\Sigma w^{opt} = (\iota_m'\Sigma^{-1}\iota_m)^{-1}\Sigma\Sigma^{-1}\iota_m$ and $\lambda w^{opt} = \lambda(\iota_m\Sigma^{-1}\iota_m)^{-1}\Sigma^{-1}\iota_m$. Equating these two equations and dividing both sides by $\lambda(\iota_m\Sigma^{-1}\iota_m)^{-1}$, we have $\lambda^{-1}\iota_m = \Sigma^{-1}\iota_m$. Then $w^{opt} = (\iota_m'\Sigma^{-1}\iota_m)^{-1}\Sigma^{-1}\iota_m = (\iota_m'\lambda^{-1}\iota_m)^{-1}\lambda^{-1}\iota_m = m^{-1}\iota_m$, a contradiction to $w^{opt} \neq m^{-1}\iota_m$. □

Results and proof of Proposition 2.

Consider the following optimization problem

$$\max_{r_{ij}\in R} m^{-2}(\iota_m'\Sigma\iota_m) - (\iota_m'\Sigma^{-1}\iota_m)^{-1}$$

subject to

$$u_k'\Sigma^{-1}\iota_m \geq 0, \; k = 1, ..., m.$$

$$r_{ij} \geq 0, \; i = 1, \dots, m, \; j = i+1, \dots, m$$

where u_k is a $m \times 1$ vector of zeros with one in the k^{th} row. The space R is such that Σ is positive definite.

Then fix m_1, we can block the matrix Σ such that it equal to

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{11} \in R^{m_1 \times m_1}$.

Lemma 1. *For Σ with $\Sigma_{11} = I_{m1}$ and $\Sigma_{12} = m_1^{-1}\iota_1\iota_2'$, we have $\Sigma^{-1}\iota_m = \begin{pmatrix} \iota_1 \\ 0 \end{pmatrix}$*

Proof. From the usual inverse formula for block diagonal matrices we have

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\Sigma_{12}'\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1} \\ -\Sigma_{2.1}^{-1}\Sigma_{12}'\Sigma_{11}^{-1} & \Sigma_{2.1}^{-1} \end{pmatrix}, \text{ where } \Sigma_{2.1} = \Sigma_{22} - \Sigma_{12}'\Sigma_{12}.$$

So we have

$$\begin{aligned} \Sigma^{-1}\iota_m &= \begin{pmatrix} \Sigma_{11}^{-1}\iota_1 + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\Sigma_{12}'\Sigma_{11}^{-1}\iota_1 - \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\iota_2 \\ -\Sigma_{2.1}^{-1}\Sigma_{12}'\Sigma_{11}^{-1}\iota_1 + \Sigma_{2.1}^{-1}\iota_2 \end{pmatrix} \\ &= \begin{pmatrix} \iota_1 + m_1^{-1}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\iota_2\iota_1'\iota_1 - \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\iota_2 \\ -m_1^{-1}\Sigma_{2.1}^{-1}\iota_2\iota_1'\iota_1 + \Sigma_{2.1}^{-1}\iota_2 \end{pmatrix} \\ &= \begin{pmatrix} \iota_1 + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\iota_2 - \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{2.1}^{-1}\iota_2 \\ -\Sigma_{2.1}^{-1}\iota_2 + \Sigma_{2.1}^{-1}\iota_2 \end{pmatrix} \\ &= \begin{pmatrix} \iota_1 \\ 0 \end{pmatrix}. \end{aligned}$$

It follows directly that $(\iota_m'\Sigma^{-1}\iota_m) = m_1$ and so the corresponding optimal weights are

$$w^{opt} = \begin{pmatrix} (1/m_1)\iota_1 \\ 0 \end{pmatrix}$$

□

Proof of Proposition 2.

Proof. The Lagrangian function of the optimization problem is

$$L = m^{-2}\iota'\Sigma\iota - (\iota'\Sigma^{-1}\iota)^{-1} + \sum_{k=1}^{m}\lambda_k u_k\Sigma^{-1}\iota + \sum_{i=1}^{m_1}\sum_{j=i+1}^{m}\lambda_{ij}r_{ij}$$

where λ_k and λ_{ij} are the Lagrangian multipliers associated with constraints.
By optimality condition, if for Σ with r_{ij} satisfying the following:

1. $\frac{\partial L}{\partial r_{ij}} = 0,\ \frac{\partial L}{\partial \lambda_k} \geq 0,\ \frac{\partial L}{\partial \lambda_{ij}} \geq 0$
2. $\lambda_k,\ \lambda_{ij} \geq 0$
3. $\lambda_k u_k \Sigma^{-1}\iota_m = 0,\ \lambda_{ij}r_{ij} = 0$

then Σ is the optimal solution.
To show our proposed Σ satisfying these conditions, we firstly compute the partial derivatives of the Lagrangian function:

$$\frac{\partial L}{\partial r_{ij}} = m^{-2}\iota'\frac{\partial \Sigma}{\partial r_{ij}}\iota - (\iota'\Sigma^{-1}\iota)^{-2}(\iota'\Sigma^{-1}\frac{\partial \Sigma}{\partial r_{ij}}\Sigma^{-1}\iota) - \sum_{k=1}^{m}\lambda_k u_k\Sigma^{-1}\frac{\partial \Sigma}{\partial r_{ij}}\Sigma^{-1}\iota + \lambda_{ij}$$

$$
\begin{aligned}
&= 2m^{-2} - ((\iota'\Sigma^{-1}\iota)^{-1}\iota'\Sigma^{-1})(u_i u_j' + u_j u_i')((\iota'\Sigma^{-1}\iota)^{-1}\Sigma^{-1}\iota) \\
&\quad - \sum_{k=1}^{m} \lambda_k u_k \Sigma^{-1}(u_i u_j' + u_j u_i')\Sigma^{-1}\iota + \lambda_{ij} \\
&= 2m^{-2} - w^{opt\prime}(u_i u_j' + u_j u_i')w^{opt} - \sum_{k=1}^{m} \lambda_k u_k \Sigma^{-1}(u_i u_j' + u_j u_i')\Sigma^{-1}\iota + \lambda_{ij} \\
&= 2m^{-2} - 2w_i^{opt} w_j^{opt} - \sum_{k=1}^{m} \lambda_k u_k \Sigma^{-1}(u_i u_j' + u_j u_i')\Sigma^{-1}\iota + \lambda_{ij} \\
\frac{\partial L}{\partial \lambda_k} &= u_k \Sigma^{-1}\iota \\
\frac{\partial L}{\partial \lambda_{ij}} &= r_{ij}
\end{aligned}
$$

Note that $w^{opt} = ((1/m_1)\iota_1 \;\; 0)'$ by Lemma 1. Then, by optimality condition, we let $\lambda_k = 0$ as $u_k \Sigma^{-1}\iota_m > 0, \;$ for $k = 1, \ldots, m_1$.
Now, consider r_{ij} in (1,2) block, where $i = 1, \ldots, m_1, \; j = m_1 + 1, \ldots, m$.
For the same reason, we let $\lambda_{ij} = 0$ as $r_{ij} > 0$ in this block.
Then we need to determine if there exist $\lambda_k > 0$ for $k = m_1+1, \ldots, m$ such that the following equality holds:

$$
\begin{aligned}
\frac{\partial L}{\partial r_{ij}} &= 2m^{-2} - 2 \cdot m_1^{-1} \cdot 0 - \sum_{k=m_1+1}^{m} \lambda_k u_k \Sigma^{-1}(u_i u_j' + u_j u_i') \begin{pmatrix} \iota_1 \\ 0 \end{pmatrix} \\
&= 2m^{-2} - 2 \cdot m_1^{-1} \cdot 0 - \sum_{k=m_1+1}^{m} \lambda_k u_k \Sigma^{-1} u_j \\
&= 2m^{-2} - \sum_{k=m_1+1}^{m} \lambda_k \Sigma_{kj}^{-1} \\
&= 0, \forall i = 1, \ldots, m_1, \; j = m_1 + 1, \ldots, m, \text{ where } \Sigma_{kj}^{-1} \text{ is the (k,j) entry of } \Sigma^{-1}
\end{aligned}
$$

First note that for each j we have the same set of simultaneous equations over $j = m_1 + 1$ to m. We can write the set of equations as

$$
\begin{pmatrix}
\lambda_{m1+1}\Sigma_{m1+1,m1+1} + ... + \lambda_m \Sigma_{m,m1+1} \\
\lambda_{m1+1}\Sigma_{m1+1,m1+2} + ... + \lambda_m \Sigma_{m,m1+2} \\
... \\
\lambda_{m1+1}\Sigma_{m1+1,m} + ... + \lambda_m \Sigma_{m,m}
\end{pmatrix} = \Sigma_{2.1}^{-1}\lambda
$$

So

$$2m^{-2}\iota_2 - \Sigma_{2.1}^{-1}\lambda = 0$$

where $\lambda = (\lambda_{m_1+1}, \lambda_{m_1+2}, ..., \lambda_m)'$. This solves to

$$\lambda = \frac{2}{m^2}\Sigma_{2.1}\iota_2.$$

For a subspace of Σ_{22} satisfying that the minimum row sum of $\Sigma_{2.1} > 0$, it follows that $\lambda_k \geq 0$ for $k = m_1 + 1, \ldots, m$, and the optimality conditions hold in (1,2) block.
Now, consider r_{ij} in (1,1) block where $i = 1, \ldots, m_1,\ j = i + 1, \ldots, m_1$.
Let l_i be a $m_1 \times 1$ vector of zeros with one in i^{th} row.
We can substitute λ_k that we obtained from (1,2) block, and then we need to determine whether $\lambda_{ij} \geq 0$ for i, j in (1,1) block such that the following equality holds:

$$\begin{aligned}
\frac{\partial L}{\partial r_{ij}} &= 2m^{-2} - 2 \cdot m_1^{-1} \cdot m_1^{-1} - \sum_{k=m_1+1}^{m} \lambda_k u_k \Sigma^{-1}(u_i u_j' + u_j u_i') \begin{pmatrix} \iota_1 \\ 0 \end{pmatrix} + \lambda_{ij} \\
&= 2m^{-2} - 2m_1^{-2} - \sum_{k=m_1+1}^{m} \lambda_k (\Sigma_{ki}^{-1} + \Sigma_{kj}^{-1}) + \lambda_{ij} \\
&= 2m^{-2} - 2m_1^{-2} + \lambda' \Sigma_{2.1}^{-1} \Sigma_{12}' (l_i + l_j) + \lambda_{ij} \\
&= 2m^{-2} - 2m_1^{-2} + 2m^{-2} \iota_2' \Sigma_{2.1} \Sigma_{2.1}^{-1} \Sigma_{12}' (l_i + l_j) + \lambda_{ij} \\
&= 2m^{-2} - 2m_1^{-2} + 2m_1^{-1} m^{-2} \iota_2' \iota_2 \iota_1' (l_i + l_j) + \lambda_{ij} \\
&= 2m^{-2} - 2m_1^{-2} + 4m_2 m_1^{-1} m^{-2} + \lambda_{ij} \\
&= 0
\end{aligned}$$

Solving the equation, we get

$$\begin{aligned}
\lambda_{ij} &= 2m_1^{-2} - 2m^{-2} - 4m_2 m_1^{-1} m^{-2} \\
&= 2m_1^{-2} - 2m^{-2} - 4(m - m_1) m_1^{-1} m^{-2} \\
&= 2m_1^{-2} m^{-2} (m^2 - m_1^2 - 2m_1(m - m_1)) \\
&= 2m_1^{-2} m^{-2} (m - m_1)^2 \geq 0,
\end{aligned}$$

for all $i \in \{1, \ldots, m_1\}, j \in \{i + 1, \ldots, m_1\}$
Thus, for Σ_{22} in the specified space, we have that the optimality condition holds. □

Proof of Proposition 3.

Proof. For any symmetric non-negative definite matrix Σ we have that there exist matrices C and Λ such that $C \Lambda C' = \Sigma$ where Λ has zeros in the off diagonals and the eigen values of Σ for diagonal elements. The matrix C has columns c_i equal to the eigen vectors associated with the eigen values which are orthonormal so $C'C = CC' = I$. Note that

$$\begin{aligned}
\frac{1}{m^2} \iota_m' C \Lambda C' \iota_m &= \frac{1}{m^2} \sum_{i=1}^{m} \lambda_i \iota_m' c_i c_i' \iota_m \\
&\leq \frac{\lambda_{max}}{m^2} \iota_m' C C' \iota_m \\
&= \frac{\lambda_{max}}{m}
\end{aligned}$$

which goes to zero for λ_{max} bounded above. For the inverse of optimal loss ($\iota_m' \Sigma^{-1} \iota_m$) we have that this is equal to

$$\iota_m' C \Lambda^{-1} C' \iota_m \geq \frac{1}{\lambda_{max}} \iota_m' C C' \iota_m = \frac{m}{\lambda_{max}}$$

and so the inverse of this (optimal loss) is converging to zero. More elegantly we could just say that the first result is sufficient since if average weights are optimal the result holds by direct equality, if not then loss from averaging is greater than optimal loss so since it is going to zero, so is the optimal result.

□