Averaging and the Optimal Combination of Forecasts*

Graham Elliott

University of California, San Diego

9500 Gilman Drive

LA JOLLA, CA, 92093-0508

September 27, 2011

Abstract

The optimal combination of forecasts, detailed in Bates and Granger (1969), has empirically often been overshadowed in practice by using the simple average instead. Explanations of why averaging might in practice work better than constructing the optimal combination have centered on estimation error and the effects variations of the data generating process have on this error. The flip side of this explanation is that the size of the gains must be small enough to be outweighed by the estimation error. This paper examines the sizes of the theoretical gains to optimal combination, providing bounds for the gains for restricted parameter spaces and also conditions under which averaging and optimal combination are equivalent. The paper also suggests a new method for selecting between models that appears to work well with SPF data.

*Corresponding author : Graham Elliott, Department of Economics, 9500 Gilman Drive, LA JOLLA, CA, 92109, phone (858) 534 4481, fax (858) 534 7040, email:grelliott@ucsd.edu. The author thanks participants at the Clive W. Granger Memorial Conference (Nottingham).

1 Introduction

In a seminal paper, Clive Granger along with Bates (Bates and Granger 1969) considered the combination of a pair of forecasts for the purpose of constructing a better forecast. As in portfolio optimization, the idea was to use the relative variances and covariances to construct a weighted average of the forecasts that minimized the mean square error of the combined forecast, improving our ability to construct forecasts. The idea readily extended to a set of m forecasts. Granger and Ramanathan (1984) later showed that this method, when restricting weights to sum to one, was numerically equivalent to weights constructed from a restricted least squares regression of the outcome variable on the forecasts.

In empirical application it has turned out that the optimal weights of Bates and Granger (1969), when replaced with sample estimates, often does not appear to be a better method for many sets of data than taking a simple average of the forecasters. Clemen (1989) surveys the literature up until the time of its publication and notes that over a very large number of papers averaging forecasts appears to be a more robust procedure in practice than optimal combination. More recent work has only reiterated this finding. Stock and Watson (2001) considered 49 methods for constructing forecasts of a large number of US macro series. Combining the methods proved to be the favored strategy, rather than using individual models. Averaging the models, along with taking the median of the forecasts, worked best.

This somewhat surprising result has spurred a very large literature, see Timmermann (2006) for a review. A large number of papers have made the reasonable suggestion that estimation error in the estimated weights is the underlying problem (see Smith and Wallis (2009) for a recent example). Approaches to the problem based on the idea that estimation error in the weights require two pieces to their argument — first, that gains from optimal combination are not too large and second that estimation error is large. By far the largest amount of attention has been paid to the second of these pieces, whilst this paper examines the first part.

Regarding the possibility that estimation error may be large, the literature has focussed on two aspects of this problem. First, researchers have suggested alternative data generating procedures that would make estimation error larger and hence large enough to be expected to outweigh theoretical gains. For example Clemen and Winkler (1986) consider parameter instability as a cause. Clements and Hendry (2004) consider situations where the data generating process has discrete shifts and forecasting models are misspecified. Both of these elements can increase estimation error.

A second strand of the literature has taken as given that estimation error is the problem, and focussed attention on providing methods that might mitigate the estimation error. One direction of attack was the realization that with estimation error other estimators of the optimal weights could be considered. Clemen and Winkler (1986) suggested priors based on exchangeability of the forecasts and a prior on the variance covariance which would result in the average forecast being optimal. Diebold and Pauly (1990) suggest two empirical Bayes shrinkage methods, where the estimates are a weighted average of the least squares estimates for the weights and the average weights vector. The shrinkage factor here is the same for each forecast, so the method does not allow for some weights to be shrunk differently from others. Chan, Stock and Watson (1999) suggest factor models. Many other methods, often ad hoc, have also been suggested as attempts to exploit the data for constructing weights. Bates and Granger (1969) suggested ignoring covariances between the forecasts, constructing weights as the variance of that forecast error as a proportion of the sum of all the variances of the forecast errors. Aiolfi and Timmermann (2006) suggest ranking the forecast error variances, then constructing the weights as the inverse of the rank of that forecast divided by the sum of the inverses of the ranks of each of the forecasts (so the best forecast has weight one divided by the sum of the inverse of the rank of the forecasts, the second has weight equal to one divided by twice the sum of the inverse of the ranks, etc.). This method places a very sharply declining weighting scheme even if variances are quite similar, and ignores the possibility that correlations can be employed to reduce forecast error). Swanson and Zheng (2001) suggest choosing estimating not only the model with all of the forecasts, but every combination of subsets of the forecasts, choosing the preferred set to combine using information criteria.

However it still must be the case that reasonable variance covariance matrices of forecast errors are such that the expected gain in practice is not as large as some of the possible parameterizations suggest. An implication of the empirical success of averaging is that it would appear that the types of variance covariances that result in very large gains from varying these weights are empirically unlikely and that the reasonable parameter space to consider is much more restricted than might otherwise be considered.

This paper seeks to improve our understanding of how large gains from optimal combination can be. We extend general results for which there is no gain. We also consider a more restricted but yet theoretically and empirically reasonable parameter space. Considering models where much of the error is common to all forecasts, we might expect that both the variances of different individuals forecast errors are not too dissimilar and also that their correlations are likely to be positive. An extensive examination of the Survey of Professional Forecasters (SPF) data suggests that this is true. Other empirical evidence for positively correlated forecast errors arises from the often noted fact that typically forecasters are all on one side of the actual (clustering). With these restrictions we provide results that allow us to better understand the bound on the size of the expected gains from optimal combination. The theoretical results of this paper could be considered complimentary to those that argue for a large estimation error explanation of the success of averaging, by providing results that shed light on the first piece of the puzzle mentioned above.

We also present, based on the theoretical findings, a simple yet robust method for forecast combination that both allows the combination method to include averaging when this method is likely to be a good approach and also to approximate the optimal weights when they are likely to be a much better approach than averaging. The methods are examined in Monte Carlo experiments and applied to the SPF data.

The next section reviews the forecast combination problem and discusses the aims of this paper and its place in the literature. We examine Survey of Professional Forecaster data to illustrate reasonable parameter spaces for the variance covariance matrix of forecast errors. Section 3 presents the main results of the paper, along with deriving a set of methods for practical forecast combination. The methods are analyzed both in Monte Carlo experiments and with application to the SPF data in Section 4. A final section concludes.

2 Combining Forecasts

An enormous literature was spawned by the seminal ideas of Bates and Granger (1969). This paper suggested that for situations when two forecasts were available, the covariance structure could be employed to optimally combine them into a single forecast that would have better properties when the forecaster has a mean square loss function. The idea readily extends to the combination of an m vector of h step ahead forecasts $f_{t,h}$ of an outcome y_{t+h}.

Define the m dimensional vector of forecast errors $e_{t,h} = \iota_m y_{t+h} - f_{t,h}$ where ι_k will denote a vector of ones of length k. Consider a combined forecast $\omega_0 + \omega' f_{t,h}$ where ω_0 is a constant and ω is an m vector of weights. The combined forecast can be written

$$\begin{aligned} y_{t+h} - \omega_0 - \omega' f_{t,h} &= y_{t+h}(\omega'\iota_m - 1) - \omega_0 + \omega'(\iota_m y_{t+h} - f_{t,h}) \\ &= (y_{t+h}(\omega'\iota_m - 1) + E[(e_{t,h})] - \omega_0) + \omega'(e_{t,h} - E(e_{t,h})) \end{aligned}$$

Then if $E[e_{t,h}] = \mu$ and $E[e_{t,h}e'_{t,h}] = \Sigma$ the MSE of the forecast using the combined forecast under the restriction that the weights sum to one is

$$\begin{aligned} E[y_{t+h} - \omega_0 - \omega' f_{t,h}]^2 &= E[(E[(e_{t,h})] - \omega_0) + \omega'(e_{t,h} - E(e_{t,h}))]^2 \\ &= (\mu - \omega_0)^2 + \omega'\Sigma\omega. \end{aligned}$$

Clearly minimizing MSE here involves setting[1] $\omega_0 = \mu$ and then choosing weights to minimize $\omega'\Sigma\omega$ subject to the constraint $\omega'\iota_m = 1$. The solution to this is to set $\omega^{opt} = (\iota'_m\Sigma^{-1}\iota_m)^{-1}\iota_m\Sigma^{-1}$. This is the multivariate generalization of Bates and Granger (1969). In this case MSE is equal to $(\iota'_m\Sigma^{-1}\iota_m)^{-1}$.

The problem could also be set up to minimize the mean square error employing the forecasts $f_{t,h}$ as data, in which case the weights could be constructed by considering the regression of y_{t+h} on a constant and $f_{t,h}$ subject to the restriction that the sum of the coefficients on the regressors $f_{t,h}$ sum to one. As shown by Granger and Ramanathan (1984), the population restricted least squares estimator combination weights are identical to the Bates and Granger (1969) optimal combination weights. Hence the result for ω^{opt} can be considered an alternative expression for the restricted least squares estimator. However the above form of the expression will be useful in section 3 below.

Whilst the result given above yields the optimal combination for the population problem, in practice we still need estimates from the data to make the methods operational. Typically a plug-in estimator such as estimating $\hat{\Sigma} = (T-k)^{-1}\sum_{t=1}^{T} e_{t,h}e'_{t,h}$ and $\hat{\mu} = T^{-1}\sum e_{t,h}$ are employed. The resulting plug-in estimated weights do not have any optimality properties in terms of minimizing MSE except in the limit. For example any consistent estimator for Σ or consistent estimator for the restricted least squares coefficients will result in a different

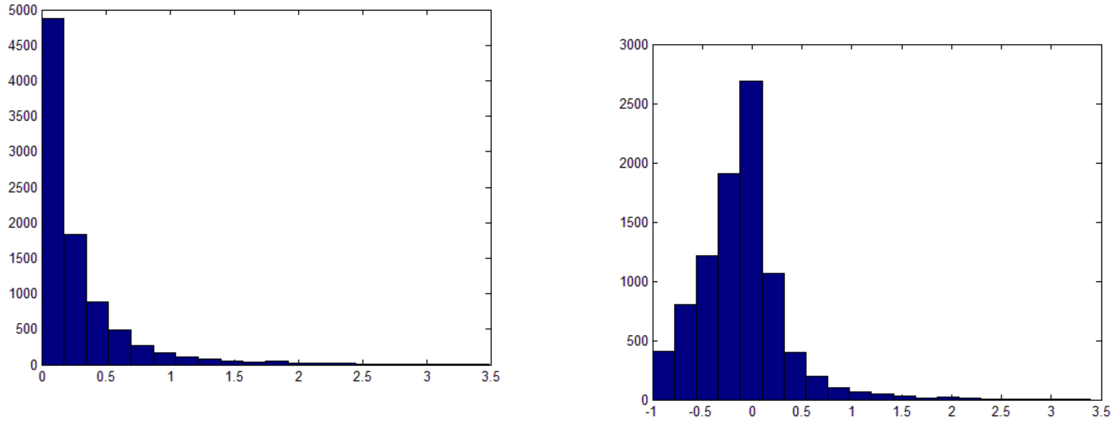
[1] In population the mean takes care of any bias in $\omega' f_{t,h}$. Since this paper is concerned with population results we will ignore the mean in the remainder of the paper.

construction method with different small sample properties but the same large sample justification. Hence shrinkage estimators (where the shrinkage disappears asymptotically) will also yield estimators that have the same asymptotic justification but may well improve over standard restricted least squares methods or standard estimators for $\hat{\Sigma}$.

Clemen (1989) summarizes the performance in practice up to 1989 but with results still relevant today. It is by now a general folklore result that simply averaging the forecasts (perhaps with still estimating the constant term to take into account biases) turns out to work better than using the above plug-in estimators for the optimal combination when we look at out of sample performance. The average forecast is $\iota_m' f_{t,h}$ and hence involves no estimation. MSE for this method of combination is given by $m^{-2}\iota_m'\Sigma\iota_m$. Thus whilst it is always true that $(\iota_m'\Sigma^{-1}\iota_m)^{-1} \leq m^{-2}\iota_m'\Sigma\iota_m$, for out of sample situations it is often the case that when $\hat{\Sigma}_{in}$ uses an 'in sample' set of observations for estimation of the combining weights that the out of sample loss is $(\iota_m'\hat{\Sigma}_{in}^{-1}\iota_m)^{-2}\iota_m'\hat{\Sigma}_{in}^{-1}\Sigma_{out}\hat{\Sigma}_{in}^{-1}\iota_m \geq m^{-2}\iota_m'\Sigma_{out}\iota_m$ when Σ_{out} is the variance covariance of the out of sample forecast errors.

This can be demonstrated from examining forecast data from the Survey of Professional Forecasters database. This is a set of quarterly forecasts that have been provided by a revolving set of forecasters. Forecasts are provided for a variety of macroeconomic variables, for the purposes of this paper we restrict attention to forecasts of real GDP, using forecasts from the first quarter of 1980. The panel of forecasts is an extremely unbalanced panel, with forecasters dropping in and out of the sample and often missing some quarters. Forecasters active early in the sample disappear for good, others active now were not participants earlier in the sample. For both panels in Figure 1 I have, for each triplet of forecasters in the survey, constructed a series of forecast errors for all of the periods for which each of the three forecasters provided forecasts, dropping all triplets of forecasters for which we have less than 15 observations. Of the 256 forecasters that have provided forecasts over the sample, this leaves 9015 groups of three forecasters. For each of these, I have set the 'in sample' period to be the first two thirds of each sample, leaving one third of the observations for the out of sample evaluation. Define the percentage loss of averaging over the optimal combination using a plug-in estimator as described above to be the estimated relative loss. Figure 1 reports a histogram of estimated average loss over all triplets in the SPF sample. The first panel shows results for in sample estimated weights. As must be true, averaging performs worse than the optimal combination, often by quite a large percentage[2]. The second panel repeats this exercise with the same estimates for the weights but now evaluated on the out of sample data, the result is dramatically different. For more than half of the triplets, loss from averaging outperforms the optimal combination. This gain is often quite large.

Figure 1: Percentage Loss from Averaging with respect to that from Optimal Combination.



Note: First panel shows in sample results, the second panel shows out of sample results.

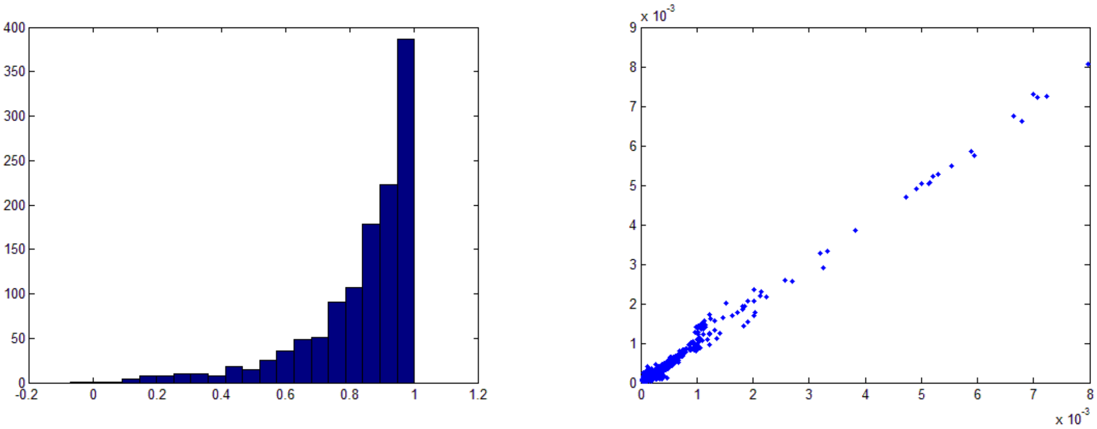
An issue to be examined to understand for the argument that it is the problem with estimation error that is the reason that averaging tends to work better in practice than estimating optimal weights is that, given the relationship with restricted least squares, we expect this error asymptotically to be of the order of $(m-1)/T$ percent. In out of sample calculations it will possibly be larger. For instability of the weights, this error could also be larger. But for modest m, this can only be the explanation if the population relative loss from averaging over optimal combination $m^{-2}\iota_m' \Sigma \iota_m (\iota_m' \Sigma^{-1} \iota_m)^{-1} - 1$ is quite large. The next section presents various results on the size of this expression. We first show a number of results for which this relative loss is zero. We then turn to examining the potential size of this relative loss under restrictions that are motivated by stylized facts for forecasts.

Before examining the theoretical results, we first present these stylized facts in forecast combination. A first point is that forecasts often tend to track each other, which is to say that they are closer to each other than they are to the outcome in most forecasting

[2] For the purposes of readability of the graph, I have dropped a few combinations of forecasters where the relative gains were larger than 3.5 (33 in the first panel, 9 in the second).

situations. Such a result arises naturally since the unforecastable shock is common to all forecast errors. It is also likely that different forecasters are observing similar information. However this leads to positively correlated forecast errors, which results in implications for reasonable specifications for Σ. This can be demonstrated in the SPF data by examining the correlations of all possible pairs of forecasters in the sample.

Figure 2: Correlations and Variances of Pairs of Forecast Errors



Note: The first panel shows the correlations between each pair of forecasters. The second panel is a scatterplot of the variances of each of the forecasters in each pair.

The first panel of Figure 2 shows the correlations between all pairs of forecasters over the sample. The data was constructed using the same method as for Figure 1 however considering pairs rather than triplets. Clearly, the majority of the forecast errors are strongly positively correlated. Indeed the mode is close to one, whereas there are few forecast errors that have correlations less than 0.4. Only a single forecast error of the 1229 pairs of forecasts has an estimated negative correlation.

The second panel of Figure 2 presents a scatterplot of the variances of the forecast errors for each of the forecasters in each pair. By and large these lie very close to the 45^o line, indicating that the variance of forecast errors is relatively similar across forecasters. This too makes sense — forecasters who are very poor are likely to not continue sending in forecasts to the SPF (or more generally poor forecasters are likely to be weeded out of any sample, poor forecast methods are also likely to be ignored if they are substantially inferior to other methods). Hence we might expect that there is very little difference in the MSE of different forecasters.

3 Relative Gains from Optimal Weights

This section provides theoretical results on specifications of Σ that either mean that relative loss is small or zero and alternatively results where relative loss is large. For restricted spaces of models we are able to bound the difference in loss between the best case optimal weights loss and the loss that would arise if we averaged over the forecasts. We first present some general results for which the average of the forecasts is the optimal combination, then present results for specific (but empirically reasonable) subcases to understand istuations where averaging will be a poor approach.

3.1 Equivalences between averaging and optimal weights

A first result on optimal results being equivalent to averaging over the forecasts is to note that if the unit vector lies in the eigen space of Σ then averaging is optimal. This follows as when $\Sigma\iota_m = \lambda\iota_m$ for a scalar λ then $\Sigma^{-1}\iota_m = \lambda^{-1}\iota_m$ and

$$\begin{aligned}\omega^{opt} &= (\iota_m'\Sigma^{-1}\iota_m)^{-1}\Sigma^{-1}\iota_m \\ &= (\lambda^{-1}\iota_m'\iota_m)^{-1}\lambda^{-1}\iota_m \\ &= m^{-1}\iota_m.\end{aligned}$$

This gives some indication as to the types of Σ that will result in equal weights being optimal — they are situations in which the row sums of Σ are equal to each other. Special cases have been noted in the literature — for example when Σ has all variances equal to each other and all covariances equal to a constant (Capistran and Timmermann (2007), Hsiao and Wan (2010)). In the case of $m > 4$ the 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;

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

(the second of these is a permutation of the first where the third and fourth forecasters are swapped with each other). However the set of data generating processes (specifications of

Σ) for which averaging is optimal is quite large. Averaging is optimal if and only if the row sums of Σ are equal for each row, thus for a great many specifications of Σ. A point to take from this result is that there are very many points in the space generated by Σ for which losses are equivalent for both optimal and averaging. One expects by continuity of the loss function as a function of the elements of Σ that gains from optimal combination near these points are small and easily outweighed by estimation error.

The result that the averaging vector lies in the eigen space of Σ for the optimal combination weights to equal the averaging weights is a necessary condition — the only optimal combination that is also an eigen vector of Σ is the averaging weights. To see this, note that we require that $\Sigma\omega^{opt} = \lambda\omega^{opt}$ for λ scalar for the optimal vector to be an eigen vector. However $\Sigma\omega^{opt} = (\iota_m'\Sigma^{-1}\iota_m)^{-1}\Sigma\Sigma^{-1}\iota_m = (\iota_m'\Sigma^{-1}\iota_m)^{-1}\iota_m$ and $\lambda\omega^{opt} = \lambda(\iota_m\Sigma^{-1}\iota_m)^{-1}\Sigma^{-1}\iota_m$ so on equating and dividing both sides by $\lambda(\iota_m\Sigma^{-1}\iota_m)^{-1}$ we have $\lambda^{-1}\iota_m = \Sigma^{-1}\iota_m$, so for ω^{opt} to be an eigen vector ι_m must be an eigen vector. From the result above this means averaging must be optimal.

Results showing that optimal weights and averaging lead to similar losses under MSE are also available for forecast combination situations where m is large and there is a common component to the forecast errors, which is likely in practice. Consider a model where there is a common component to the forecast error as well as an idiosyncratic one, so $e_{t,h} = \iota_m\varepsilon_{t+h}+v_{t+h}$ where ε_{t+h} is a univariate forecast error common to all forecasters and v_{t+h} is the $mx1$ vector of idiosyncratic forecast errors. Then $\Sigma = \sigma_\varepsilon^2\iota_m\iota_m' + \sigma_\varepsilon^2\tilde{\Sigma}$ where σ_e^2 is the variance of the common component of the forecast error. If the idiosyncratic component is such that the largest eigen vector of its variance covariance matrix is bounded above, then for large m loss from optimal combination and loss from averaging is the same.

Proposition 1 *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*

$$\begin{aligned} (i) \quad \lim_{m\to\infty} m^{-2}\iota_m'\Sigma\iota_m &= \sigma_\varepsilon^2 \\ (ii) \quad \lim_{m\to\infty} (\iota_m'\Sigma^{-1}\iota_m)^{-1} &= \sigma_\varepsilon^2 \end{aligned}$$

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

This result suggests that when there is idiosyncratic error where there is no common factor in this error, then with enough forecasts to combine averaging is likely to do as well as the optimal combination in population[3]. Hence when estimation error is added to the problem, we might well expect estimated optimal combinations to perform worse than simple averaging.

3.2 Limits to gains from Optimal Combination

For fixed m, it is still the case that there exist models (choices of Σ) for which the gains to optimal combination can be theoretically large. For example two forecasters with perfectly negatively correlated forecast errors can be combined to obtain a combination forecast that is perfect, with no estimation error. Thus for placing reasonable bounds on the size of the gains from optimal combination over averaging it will be useful to restrict the space of Σ to regions that are theoretically and empirically reasonable.

One restriction that follows from the good empirical performance of the averaging method is to constrain Σ such that weights are nonnegative. This could be thought of as a loose prior that places weight only on models for which forecasters are not so differently informed that the optimal combination would result in offsetting weights. Alternatively we can directly think of this as a loose prior that says that the weights can deviate from equal weights but not so far that any can be negative. Secondly, motivated by the empirical results of the previous section that the variances of forecast errors across forecasters are very similar and that the correlations between forecast errors are rarely negative we first restrict the diagonal elements of Σ to be equivalent to each other so that Σ is a scaled correlation matrix and secondly restrict the correlations to be nonnegative. Theoretical considerations suggested that this would be the likely outcome for the forecast combination problem in most situations.

Since the MSE loss from averaging is given by $m^{-2}\iota_m'\Sigma\iota_m$ and the MSE loss from optimal combination is $(\iota_m'\Sigma^{-1}\iota_m)^{-1}$, the relative (percentage) gain from optimal weights over averaging is given by[4] $m^{-2}(\iota_m'\Sigma\iota_m)(\omega^{opt\prime}\Sigma\omega^{opt})^{-1} - 1$. Since scale multiples of Σ result in

[3] This result is similar in spirit to the Grenander and Rosenblatt (1957) result of the equivalence of GLS and OLS under conditions on the variance covariance structure in time series models. Here optimal weights is GLS for the mean and average weights is OLS for the mean, however we have no obvious structure for the variance covariance matrix since the model is spatial (a time series assumption would be that more distant observations are less related). I thank Peter Phillips for pointing this out.

[4] Ordinarily $\omega^{opt\prime}\Sigma\omega^{opt} = (\iota_m'\Sigma^{-1}\iota_m)^{-1}$, however this notation requires Σ to be nonsingular. We write in the more general form here to allow for specifications of Σ such that subblocks Σ_a are nonsingular, in which

equivalent values for this expression, without loss of generality we consider only correlation matrices for Σ for the remainder of this section. Subject to the above constraints, we are interested in characterizing both the specifications of Σ and the optimal weights that would arise from these specifications that make this percentage gain as large as possible.

The general Lagrangian to be solved is

$$\mathcal{L} = (\iota_m' \Sigma \iota_m)(\omega^{opt\prime} \Sigma \omega^{opt})^{-1} + \sum_{i,j=1, i \neq j}^{m} \lambda_{ij} r_{ij} + \sum_{i,j=1, i \neq j}^{m} \gamma_{ij}(1 - r_{ij}) + \sum_{i=1}^{m} \lambda_i u_i' \Sigma^{-1} \iota_m \qquad (2)$$

where $\lambda's$ and $\gamma's$ are such that they are nonnegative, $\lambda_{ij} r_{ij} = \gamma_{ij}(1 - r_{ij}) = \lambda_i u_i' \Sigma^{-1} \iota_m = 0$ for any $(i, j) = 1, ..., m$, $i > j$ and the restrictions hold. Here u_i is an $mx1$ vector of zeros with a value of unity at the i^{th} row.

The optimization problem is difficult because, as can be seen from the results of the previous subsection, the objective function is non convex and hence all solutions of the Lagrangian that arises from the constrained optimization must be checked to ensure a global solution. Despite this, we can make a few general points before presenting explicit results.

First, the optimal solution must be a corner solution. This can be seen through ignoring the restrictions and from the derivative of $\mathcal{L}$ with no restrictions with respect to element r_{ij} of Σ. From this derivative we have that

$$\begin{aligned} 0 &= 2(\iota_m' \Sigma^{-1} \iota_m) - (\iota_m' \Sigma \iota_m)(\iota_m' \Sigma^{-1}(u_i u_j' + u_j u_i') \Sigma^{-1} \iota_m) \\ &= 2(\iota_m' \Sigma^{-1} \iota_m) - 2(\iota_m' \Sigma \iota_m)(\iota_m \Sigma^{-1} \iota_m)^2 \omega_i \omega_j \\ &= 1 - (\iota_m' \Sigma \iota_m)(\iota_m \Sigma^{-1} \iota_m) \omega_i \omega_j . \end{aligned}$$

where ω_k is the k^{th} optimal forecast weight (k^{th} element of ω). Since there are numerous results of this type for each (i, j), then we have

$$(\iota_m' \Sigma^{-1} \iota_m) \omega_i \omega_j = (\iota_m \Sigma^{-1} \iota_m) \omega_i \omega_k$$

for $j, k = 1, ..., m$, not i and so $\omega_j = \omega_k$. Hence solving the first order conditions of the unconstrained problem results in a minimum, not a maximum. Any maximum will be a corner solution.

A second result that can be shown is that specifications of Σ such that, in block form,

$$\Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}' & \Sigma_b \end{pmatrix}$$

case some of the weights are zero and $\omega^{opt\prime} \Sigma \omega^{opt} = (\iota_a' \Sigma_a^{-1} \iota_a)^{-1}$.

for some integer $a < m$, that when not all weights are nonzero a local maximum occurs with $\Sigma_a = I_a$ so the first a forecast errors are uncorrelated, and all of the remaining forecasts are perfectly correlated with each other. In this case the optimal weights are to average over the first a forecasts and give zero weight to the remaining ones. This result holds for all $1 < a < (m+2)/2$. This result is presented in lemma 1 of the appendix.

This result provides some understanding of the types of specifications for Σ that will result in averaging being a poor forecast combination method relative to optimal combination. As we have seen with the SPF data, typically correlations between forecast errors are large and positive. In situations where there are a few lesser correlated forecasts, it is likely to be better to concentrate weights on these forecasts and ignore to a great extent the highly positively correlated forecasts.

Table 1: Relative Loss at each (m,a) pair

	m=3	4	5	6	7	8	9	10
a=2	0.1111	0.2500	0.3600	0.4444	0.5102	0.5625	0.6049	0.6400
3		0.1250	0.3200	0.5000	0.6531	0.7813	0.8889	0.9800
4			0.1200	0.3333	0.5510	0.7500	0.9259	1.0800
5				0.1111	0.3265	0.5625	0.7901	1.0000
6					0.1020	0.3125	0.5556	0.8000
7						0.0938	0.2963	0.5400
8							0.0864	0.2800
9								0.0800

Notes: Relative loss at each local maxima is given by

$$(a-1) - m^{-1}(2a - 2a^2) + m^{-2}(a^3 - a^2)$$

We can also evaluate the relative loss at these local maxima. Although it is possible that relative loss could be larger than n these values (the local maxima may not be global maxima) this loss gives some idea of the size of possible losses from using averaging. Table 1 gives the size of the relative loss for the losses at each local maxima for each m (we also include results for other possible values of a, which is useful in the next section). From the table we see that the relative largest loss is quite modest for small m. As m becomes larger, the size of the potential loss from averaging over optimal combination increases. However such gains require more and more uncorrelated forecasts to be achieved. In practice many of these potential gains will outweigh estimation error, although for small m estimation error could be large enough to outweigh the potential gains for all but the most extreme specifications for Σ. These numbers give some indication as to the outcomes in bad case scenarios, however to argue that they are bounds requires tedious checking of all of the other potential solutions. For the $m = 3$ and $m = 4$ cases this is feasible, as noted in the following proposition.

Proposition 2 *The relative loss* $m^{-2}(\iota_m'\Sigma\iota_m)(\omega^{opt\prime}\Sigma\omega^{opt})^{-1} - 1$ *where* Σ *is a nonnegative correlation matrix with elements constrained so that the optimal weights* ω^{opt} *are nonnegative is maximized for* $m = 3$ *and* $m = 4$ *when* $a = 2$. *At this point*

(a) $\Sigma_a = I_2$, $\omega_1 = \omega_2 = 0.5$*, all other weights are zero*

(b) $\Sigma_b = \iota_{m-2}\iota_{m-2}'$ *so the last* $(m-2)$ *forecast errors are perfectly correlated.*

(c) The maximal relative loss is $1/9$ *for* $m = 3$ *and* 0.25 *for* $m = 4$.

These results put an upper bound on the gain from using optimal weights over averaging for both the $m = 3$ and $m = 4$ cases. As previously noted, if this value is small then it would be unsurprising that estimation error in constructing the optimal weights could outweigh the potential gains from using optimal weights. When $m = 3$, the maximal gain is equal to just 11%. For iid forecast errors with a distribution near enough to normality, the relative loss due to estimation error would be expected to be of the order $(m-1)/T$ where T is the number of observations. So for small sample sizes estimation error could outweigh the 11% loss from averaging. However most designs for Σ will have a lower loss than 11% from averaging, and the estimation error of this size is a best case scenario and should be thought of as a lower bound on estimation error (fatter tails in forecast errors would for example drive the size of the estimation error up, as would models with Σ time varying so we are estimating an average of the weights). At $m = 4$ the maximal gain can be 25%. Again, for most specifications for Σ this loss will be lower.

4 Best Subset Averaging Procedure

The theoretical results presented in the previous section and the numerical results presented in Table 1 suggest that averaging over subsets of the forecasts can capture many of the possible specifications for Σ where averaging over all the forecasts might be costly. The theoretical results suggest that such a model captures the best case scenario for estimating the weights as opposed to simply averaging. This suggests a method that might lie in between the empirically successful approach of averaging and the theoretically compelling approach of attempting to obtain the optimal weighting combination. The 'in between' method, which we label best subset averaging, would be to examine the losses that arise on an estimation sample from averaging every possible subset of the forecasts (ignoring subsets that include just one forecast). The best of these is then chosen as the forecast combination on new outcomes of the forecasts to be combined. There is a sense in which this method captures the best of both worlds. By never having to actually estimate the weights the method is fast and does not — for each subset considered — result in the estimation error from estimating the weights. On the other hand as we have noted the models can pick up many models for which simple averaging is likely to be a poor method. In contrast to Ridge, common shrinkage and simple Bayesian approaches, this method does not involve shrinking all of the weights towards the averaging vector but allows for models to be selected that are 'far' from this prior.

The method itself is straightforward to program and quick to run. In the case of $m = 4$, the method requires consideration of the average over all four forecasts, the four possible combinations of three of the four forecasts, and six possible combinations of two of the forecasts. Hence it requires taking only 11 averages and a single estimation of the variance covariance matrix to construct the estimator. The method could also be varied to favor one of the averages (presumably the total average), by selecting not the smallest but the smallest so long as it below the total average by some pre-set percentage. If not, the total average would be taken. In the results below we consider the method that chooses the smallest regardless of which model it is.

We present some Monte Carlo results to illustrate the properties of the best subset averaging procedure. For the design we have four possible forecasts, with the forecast error being mean zero normal with variance covariance matrix Σ. All variances are set to unity, and we consider variations in the correlations designed to capture various possibilities for the optimal vector. Table 2 gives the values for the correlations along with the optimal weights using the known value for Σ. In DGP1 we consider a situation where the correlation matrix is of the form in (1), so averaging is optimal. The next two specifications of Σ (DGP2 and

DGP3) are models for which relative loss from averaging over optimal combination is half of its maximal amount defined in the previous section (the maximal amount for $m = 4$ is 25%). We include two designs to capture the situations where even though the relative loss is not at its maximum, it is possible at this relative loss for the weights to be not too far from averaging a subset of the forecasts as well as those for which this is less true. DGP2 captures the situation where the optimal vector is not too far from averaging the first two forecasts. DGP3 has weights that are far from averaging any subset, with the first forecast garnering most of the weight and the second somewhat less with all having some positive weight. Finally, in DGP4 we present the case (close to) the models identified in Proposition 2 (we make the correlation between the last two forecasts close to but not equal to one). This last case is one where we expect the optimal and best subset averaging procedures to do well relative to averaging since the relative loss from averaging is near its maximum.

Table 2: Monte Carlo designs

	r_{12}	r_{13}	r_{23}	r_{14}	r_{24}	r_{34}	Weights
DGP1	0.4	0.4	0.1	0.1	0.4	0.4	(0.25,0.25,0.25,0.25)
DGP2	0.1	0.6	0.4	0.4	0.6	0.8	(0.44,0.44,0.06,0.06)
DGP3	0.066	0.2	0.7	0.4	0.5	0.8	(0.45,0.38,0.12,0.05)
DGP4	0	0.1	0.9	0.4	0.6	0.8	(0.5,0.5,0,0)

For each of the designs we generate 80 observations, using 40 observations for estimating weights and the remaining 40 for evaluating the loss. Results are reported in Table 3. The first column presents loss from estimating the optimal combination weights in the first sample and applying these weights (without any updating) to the second sample. The second column reports the loss from averaging the forecasts over the evaluation sample relative to the loss from the first column of results (estimating the optimal weights out of sample). A value greater than unity represents a larger loss from averaging. The third column reports the out of sample loss from the best averaging procedure relative to the result of the first column. The last column reports the loss from estimating the optimal combination weights in sample rather than out of sample, and is included for comparison to the first column of results. Each number reported is the average over 10000 replications.

For DGP1, where the averaging method is optimal, unsuprisingly averaging is by far the

Table 3: Monte Carlo results

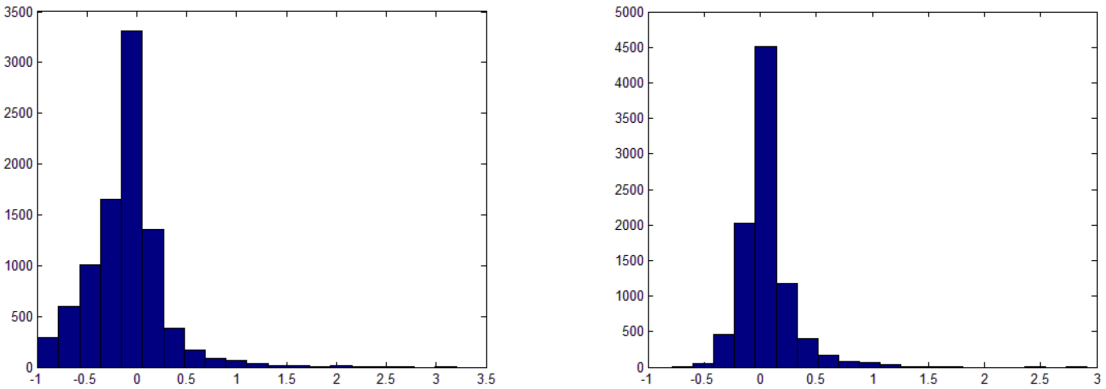
	Optimal Combination Relative Loss	Averaging Forecasts	Best Subset Averaging	In Sample Optimal Combination
DGP1	0.4287	0.9302	0.9990	0.5021
DGP2	0.4926	1.0501	0.9751	0.5769
DGP3	0.4689	1.0508	0.9868	0.5492
DGP4	0.4526	1.1237	0.9632	0.5302

Note: The first and last columns are average MSE, the second and third columns are average MSE relative to the first column.

best performer. Notice that the improvement over estimating the weighting vector is about 7.5%, precisely what we would expect from an asymptotic approximation to the effect of estimation error (which is $(m-1)/T = 3/40 = 0.075$). The best subset averaging method does not with probability one choose averaging over all of the forecasts, however still does as well as estimation. For the remaining DGP's, we expect that estimating weights will do better than averaging since the models have been chosen so that the relative loss from averaging is not insignificant (it is 12.5% for DGP2 and DGP3 and 20% for DGP4). In all cases averaging does indeed become the least successful method. DGP2 was chosen so that the weights are close to one of the models considered by the best subset averaging methods, and indeed this method performs well. It is here that we can see how there can be gains by searching over a smaller subset of models rather than estimating the weights — there is a considerable gain for the best subset averaging procedure over estimating the optimal weights. When the optimal weights are however quite far from any of the models considered by the best subset averaging procedure, there are costs. For DGP3 the best subset averaging method still performs better than estimated weights, however is only slightly better. For the final DGP, in which averaging over the first two forecasts is optimal and averaging over all of them is far from optimal, the best subset averaging procedure does very well in capturing the possibility that averaging is not a good approach but does so in a way that is far better than directly estimating the weights. Overall, the argument for the best subset averaging procedure is that it works well when averaging is a better approach than estimating the optimal combination weights (which appears to be the most empirically relevant situation) and also does well for a host of possible situations for which averaging is likely to turn out to be a poor estimation technique, doing either very well for the multitude of models for which the optimal weights are not so far from averaging over a subset and still much better than averaging when it is not.

We can also examine the method with the SPF data. We repeat the exercise of section 2 but now compare the best subset averaging procedure with both the estimated optimal weights (in panel 1) and the average weights (in panel 2). In both cases we can see that the best subset averaging procedure does well in practice. The first panel shows the relative loss for the best subset average procedure over estimating the optimal weights. In 64.5% of the triplets of forecasters, we have that the best subset averaging procedure provides a better out of sample loss (relative loss is negative). This compares with the similar value of 64% for averaging over estimating the weights (pictured in the second panel of Figure 1 in Section 2). The second panel shows the relative loss of using the best subset averaging procedure with respect to averaging over all three forecasts. The large peak at zero shows that these methods are the same for a large number of forecast triplets. However there are many cases where the best subset averaging does better than simple averaging (47% of the triplets).

Figure 3: Relative loss from the Best Subset Averaging Procedure



Note: The first column gives the loss from the best subset averaging procedure relative to estimated optimal weights. The second panel repeats this procedure relative to the loss from average weights.

We also include some additional results from the same empirical analysis that generated Figure 3. In Table 4 we show, for each of the three considered methods, the proportion of triplets for which that method was the best performer out of sample and the proportion for which it was the worst out of sample performer. The results are quite striking — both the estimated optimal weights and the averaging procedure were often the best performer, more so that the best subset averaging procedure. However the estimated optimal combination weights were half the time the worst performer, whilst the best subset averaging procedure was the worst performer in only one out of every ten triplets.

Table 4: Proportion of Triplets method is best or worst

	Optimal Combination	Averaging	Best Average
Best	0.28	0.45	0.27
Worst	0.59	0.29	0.12

Overall, the best subset averaging procedure does appear to have desirable properties. The currently considered best approach is averaging, which is simple and often works well. The best subset averaging method is just as simple, straightforward to apply even in unbalanced panels of forecasts, and appears to be robust in practice. In the data we see from Figure 3 panel 2 that they are often very similar in MSE. However in Monte Carlo results we see situations where it can indeed help and provide a robust method either in situations where averaging is good and where the optimal weights depart from the average weights. The Monte Carlo results suggest that it will not do too poorly in either situation — the results of Table 4 back this up with data where the best subset average is rarely the worst performer of the three.

5 Conclusion

The optimal weights of Bates and Granger (1969) do not appear to work well in practice. This finding has been reiterated in many studies with data across many decades. This paper attempts to shed some light on the possibility that the problem is in estimation error. For estimation error to be the reason why estimating optimal weights does not work out of sample, it must be that the difference in the population losses between the optimal weights and the loss from averaging are small enough that estimation error overwhelms the potential gain. To this end, this paper examines situations where gains might be small and characterizes potential bounds on the size of such gains.

We first extended results on specifications of the variance covariance matrix of forecast errors that lead to averaging being optimal. It is known that when this matrix is a scale multiple of a correlation matrix with all correlations equal to each other that this result holds. This is a special case of a more general result. We show that if and only if the unit vector is a scalar multiple of an eigen vector of this variance covariance matrix then using average weights is optimal. This means that whenever the rows of this variance covariance matrix sum to the same value, averaging will be optimal. Obviously when these rows are close to equivalent it will be near optimal. This is a large space of possible specifications for the variance covariance matrix, and hence we might expect that for a very large set of specifications of this matrix that potential, gains from optimal weights are small.

We also derive the worst case scenario for averaging when combining three or four forecasts under the empirically relevant constraints that the variance covariance matrix is a correlation matrix with nonnegative correlations and when weights are constrained to be nonnegative. We show that the maximal relative loss is 11% for combining three forecasts and 25% for combining four forecasts. The worst case scenario for averaging is when two of the forecast errors are uncorrelated and the remaining ones are correlated with the first two in such a way that the optimal combination is to take the average of the first two forecasts only. From the results with the SPF data, it seems unlikely that we might have two uncorrelated forecast errors in practice. If we allow the first two to be correlated, the (maximal) gain of optimal weights over averaging declines precipitously. For example if the correlations remain such that the optimal combination is to average over the first two forecasts, the relative loss as a function of the correlation of the first two forecast errors (r_{12}) is $(2/9) * (5+4r_{12})/(1+r_{12}) - 1$ which is under 6% at $r_{12} = 0.3$ and under 5% at $r_{12} = 0.4$ (from the SPF data, these are values for which most of the estimated correlations were larger). Hence we expect that with these restrictions it might be common that estimation error is larger than the potential gain.

We also examined results for when we have a larger number of forecasts to combine. Conditions were derived for which when there are many forecasts averaging and optimal combination might be expected to be similar. We also found models for which averaging might work poorly, and showed that for large numbers of forecasts the gains from optimal combination could be quite large even with the restrictions mentioned in the previous paragraph. However these situations could be considered somewhat unlikely, although when they happen averaging might be a poor approach.

Finally, we presented what we called the best subset averaging method. Since the largest gains occur (under the restrictions on the specification mentioned above) when averaging over a subset of the forecasts is optimal, we suggest a procedure that examines averages of all possible subsets of the forecasts (which for even a moderately large number of forecasts is a small set of averages) and chooses the model with the smallest MSE. This method has the nice property that it can mimic both the averaging procedure when it is good and the optimal combination when it is much better than averaging across all forecasts. It is shown to be a robust method in Monte Carlo and when applied to forecasting real GDP using the SPF data.

6 Proofs

Proof. (Proposition 1)

For any symmetric non-negative definite matrix Φ we have that there exist matrices C and Λ such that $C\Lambda C' = \Phi$ where Λ has zeros in the off diagonals and the eigen values of Φ 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

$$\Phi = C\Lambda C' = \sum_{i=1}^{m} \lambda_i c_i c_i'.$$

Consider first the term from the sample average, i.e

$$E[(y_{T+h} - f_T^a)^2] = \sigma_\varepsilon^2 + \sigma_\varepsilon^2 m^{-2} \iota_m' \tilde{\Sigma} \iota_m.$$

Examining the second term, we have

$$\begin{aligned} m^{-2}\iota_m'\tilde{\Sigma}\iota_m &= m^{-2}\iota_m' \left(\sum_{i=1}^{m} \lambda_i c_i c_i' \right) \iota_m \\ &\leq m^{-2}\lambda_{\max}\iota_m' \left(\sum_{i=1}^{m} c_i c_i' \right) \iota_m \\ &= m^{-2}\lambda_{\max}\iota_m'\iota_m \\ &= \frac{\lambda_{\max}}{m} \\ &\rightarrow 0. \end{aligned}$$

Note that we used $I_m = CC' = \sum_{i=1}^{m} c_i c_i'$ in the third line.

Now consider the result from the optimally combined forecast method.

$$\begin{aligned}(\iota_m' \Sigma^{-1} \iota_m)^{-1} &= (\iota_m' \left[\sigma_\varepsilon^2 \iota_m \iota_m' + \sigma_\varepsilon^2 \tilde{\Sigma}\right]^{-1} \iota_m)^{-1} \\ &= \sigma_\varepsilon^2 (\iota_m' \left[\iota_m \iota_m' + \tilde{\Sigma}\right]^{-1} \iota_m)^{-1}.\end{aligned}$$

To solve this, note that for G nonsingular and H rank 1 that $(G+H)^{-1} = G^{-1} - (1 + tr(HG^{-1}))^{-1}(G^{-1}HG^{-1})$ (see Miller (1981) for example). For our problem $\iota_m \iota_m'$ is rank one, $tr\left(\iota_m \iota_m' \tilde{\Sigma}^{-1}\right) = \iota_m' \tilde{\Sigma}^{-1} \iota_m$ and hence

$$(\iota_m \iota_m' + \tilde{\Sigma})^{-1} = \tilde{\Sigma}^{-1} - (1 + \iota_m' \tilde{\Sigma}^{-1} \iota_m)^{-1} \tilde{\Sigma}^{-1} \iota_m \iota_m' \tilde{\Sigma}^{-1}$$

and

$$\begin{aligned}(\iota_m' \Sigma^{-1} \iota_m)^{-1} &= \sigma_\varepsilon^2 (\iota_m' \left[\iota_m \iota_m' + \tilde{\Sigma}\right]^{-1} \iota_m)^{-1} \\ &= \sigma_\varepsilon^2 \left(\iota_m' \tilde{\Sigma}^{-1} \iota_m - (1 + \iota_m' \tilde{\Sigma}^{-1} \iota_m)^{-1} (\iota_m' \tilde{\Sigma}^{-1} \iota_m)^2\right)^{-1} \\ &= \sigma_\varepsilon^2 \left(\frac{\iota_m' \tilde{\Sigma}^{-1} \iota_m}{1 + \iota_m' \tilde{\Sigma}^{-1} \iota_m}\right)^{-1} \\ &= \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \left(\iota_m' \tilde{\Sigma}^{-1} \iota_m\right)^{-1}.\end{aligned}$$

Hence for the result we require that $\lim_{m\to\infty} \left(\iota_m' \tilde{\Sigma}^{-1} \iota_m\right)^{-1} = 0$. This follows as

$$\begin{aligned}\iota_m' \tilde{\Sigma}^{-1} \iota_m &= \iota_m' \left(\sum_{i=1}^{m} \lambda_i^{-1} c_i c_i'\right) \iota_m \\ &\geq \lambda_{\max}^{-1} \iota_m' \left(\sum_{i=1}^{m} c_i c_i'\right) \iota_m \\ &= \frac{m}{\lambda_{\max}}\end{aligned}$$

and so $\left(\iota_m' \tilde{\Sigma}^{-1} \iota_m\right)^{-1} \leq m^{-1} \lambda_{\max} \to 0$ if the largest eigen value is bounded. ■

Lemma 1 *For the problem in (2) then there exist local maxima such that*

(a) for any integer $a \leq (m+2)/2$ we have $\Sigma_a = I_a$ and $\Sigma_b = \iota_{m-a} \iota_{m-a}'$

(b) locally optimal weights are $1/a$ for the first a forecasts and zero for the remaining forecasts.

(c) locally optimal relative loss is $(a-1) - m^{-1}(2a - 2a^2) + m^{-2}(a^3 - a^2)$.

Proof. We rewrite the Lagrangian

$$\mathcal{L} = (\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a)+\sum_{i,j=1,i\neq j}^{m}\lambda_{ij}r_{ij}+\sum_{i,j=1,i\neq j}^{m}\gamma_{ij}(1-r_{ij})+\sum_{i=1}^{a}\lambda_i u_i'\Sigma_a^{-1}\iota_a+\lambda_r'(\iota_{m-a}-\Sigma_{ab}'\Sigma_a^{-1}\iota_a)$$

to allow for singular Σ. The Lagrange multipliers λ_{ij} refer to the nonnegativity constraint on r_{ij}, the γ_{ij} refer to the constraints that $r_{ij} \leq 1$, λ_i refer to the nonnegativity of the weights (for which the numerators are $u_i'\Sigma_a^{-1}\iota_a$, the denominators are positive and nonzero) and λ_r is a vector of constraints on the condition required so that the weights on the remaining forecasts are zero.

For the result we require that at $\Sigma_a = I_a$ that the equality constraint holds (λ_r nonzero) and the remaining multipliers are positive.

The first order conditions for $i = 1, ...a, j = i+1, ..., a$ can be written as

$$\begin{aligned} 0 &= 2(\iota_a'\Sigma_a^{-1}\iota_a) - 2(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a)^2\omega_i\omega_j + \lambda_{ij} - \gamma_{ij} \\ &\quad -(\iota_a'\Sigma_a^{-1}\iota_a)\sum_{k=1}^{a}\lambda_k\left(u_k'\Sigma_a^{-1}u_i\omega_j + u_k'\Sigma_a^{-1}u_j\omega_i\right) \\ &\quad +(\iota_a'\Sigma_a^{-1}\iota_a)\lambda_r'\left(\Sigma_{ab}'\Sigma_a^{-1}u_i\omega_j + \Sigma_{ab}'\Sigma_a^{-1}u_j\omega_i\right). \end{aligned}$$

For $i = 1, ..., a$, $j = a+1, ..., m$ we have

$$0 = 2(\iota_a'\Sigma_a^{-1}\iota_a) + \lambda_{ij} - \gamma_{ij} - (\iota_a'\Sigma_a^{-1}\iota_a)\omega_i\lambda_r' u_j^{m-a}$$

where u_j^{m-a} is an $(m-a)x1$ vector of zeros with a value of unity in the j^{th} column.

For $i = a+1$,...,m and $j = i+1, ..., m$ we have that

$$0 = 2(\iota_a'\Sigma_a^{-1}\iota_a) + \lambda_{ij} - \gamma_{ij}. \tag{3}$$

The last constraint can only be satisfied if $r_{ij} = 1$ (so $\gamma_{ij} = 2(\iota_a'\Sigma_a^{-1}\iota_a) > 0$). This result makes sense since $(\iota_m'\Sigma\iota_m)(\iota_m'\Sigma^{-1}\iota_m) = (\iota_a'\Sigma_a\iota_a + \iota_b'\Sigma_b\iota_b + 2\iota_a'\Sigma_{ab}\iota_b)(\iota_a'\Sigma_a^{-1}\iota_a)$ which is strictly increasing in r_{ij} in Σ_b. Hence the optimal solution when not all weights are nonzero is when all of the remaining forecast errors are perfectly correlated. At $\Sigma_a = I_a$ the weights ω_i are all equal to $1/a$ and $(\iota_a'\Sigma_a^{-1}\iota_a) = a$. Since $u_i'\Sigma_a^{-1}\iota_a = 1 > 0$ then $\lambda_i = 0$ and so these terms drop. At $r_{ij} = 0$ for $i = 1, ..a, j = i+a, ..., a$ then $\gamma_{ij} = 0$ for these terms. We can now rewrite the first two sets of FOC as

$$\begin{aligned} 0 &= 2a - 2(\iota_m'\Sigma\iota_m) + \lambda_{ij} + \lambda_r'\left(\Sigma_{ab}'u_i + \Sigma_{ab}'u_j\right) \\ 0 &= 2a + \lambda_{ij} - \gamma_{ij} - \lambda_r' u_j^{m-a}. \end{aligned}$$

Consider an internal solution for r_{ij}, i=1,..,a and $j = a+1, ..., m$. Then each element of λ_r is equal to $2a \neq 0$ so the constraint holds and $\lambda_r = 2a\iota_b$. It follows from the perfect correlation between the last $(m-a)$ forecasts that $\Sigma_{ab} = r_a \otimes \iota_b'$ where r_a is any row of Σ_{ab}. So the first FOC becomes

$$0 = 2a - 2(\iota_m'\Sigma\iota_m) + \lambda_{ij} + 2a(m-a)\left(r_{i,a+1} + r_{j,a+1}\right).$$

So we have a local max so long as for all $(i = 1, ..., a; j = i+1, ..., a)$ that

$$\lambda_{ij} = 2(\iota_m'\Sigma\iota_m) - 2a - 2a(m-a)\left(r_{i,a+1} + r_{j,a+1}\right) \geq 0.$$

For this solution

$$\begin{aligned}(\iota_m'\Sigma\iota_m) &= a + (m-a)^2 + 2(m-a)r_a'\iota_{ma} \\ &= a + (m-a)^2 + 2(m-a)\end{aligned}$$

and so we require

$$2(\iota_m'\Sigma\iota_m) - 2a - 2a(m-a)\left(r_{i,a+1} + r_{j,a+1}\right) = 2(m-a)[m - a + 2 - a\left(r_{i,a+1} + r_{j,a+1}\right) \geq 0$$

so $(m+2) \geq a(1 + r_{i,a+1} + r_{j,a+1})$. Since $r_a'\iota_{ma} = 1$ then $a(1 + r_{i,a+1} + r_{j,a+1}) \leq 2$ so the result holds for all a such that $a \leq (m+2)/2$.

Relative loss at these local optima is equal to

$$\begin{aligned}m^{-2}(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a) - 1 &= m^{-2}\left(a + (m-a)^2 + 2(m-a)\right)a - 1 \qquad (4)\\ &= a - 1 + m^{-1}(2a - 2a^2) + m^{-2}(a^3 - a^2).\end{aligned}$$

■

Proof. Proposition 2.

There are a large number of possible corner solutions, since there are a large number of correlations. Further, just as shown above that there are many sets of correlations that lead to the averaging weights being optimal there are many combinations that lead to the weights and loss being the same (flat spots on the surface to be optimized). We reduce the problem slightly without loss of generality by setting r$_{12}$ to be the smallest correlation. One of the possible cases for maxima can be ruled out for the general case, so we begin with this.

For any model with some weights zero, we have the situation of Lemma 2. For any m and a we cannot have a solution where $r_a = 0$ (so Σ_{ab} is a matrix of zeros) as $r_a' \Sigma_a^{-1} \iota_a = 1$ for the constraint that the remaining $(m - a)$ weights are zero, so cannot hold at this point. We have also shown earlier that the unconstrained problem results in a minima, not a maximum. Remaining cases need to be examined one by one for both the $m = 3$ and $m = 4$ cases.

$m = 3$ case:

For the $m = 3$ we first consider the $a = 2$ case. Here $\Sigma_b = 1$ and so we only have elements in $\Sigma_{ab} = r_a$ and Σ_a which contains only the element r_{12}. The solution is most easily found directly since there are only three correlations to examine. From lemma 2 we know that the stated solution is a local maximum. At this point we have that relative loss from (4) is $10/9 - 1 \approx 0.11\%$. We need to show that other local maxima yield smaller relative losses. Setting $a = 2$ we have that the constraint $1 = \Sigma_{ab}' \Sigma_a^{-1} \iota_a$ simplifies to $r_{13} + r_{23} = 1 + r_{12}$ for Σ_a nonsingular. For $r_{12} < 1$ we have that $\omega_1 = \omega_2$ (since the rows of Σ sum to the same number). At this point $(\iota_m' \Sigma \iota_m) = 2(1 + r_{12} + 1 + 2(r_{13} + r_{23})) = 4(1 + r_{12}) + 1$ and so the function to be optimized does not depend on r_{13} or r_{23}. Substituting in for $(\iota_m' \Sigma \iota_m)(\iota_a' \Sigma_a^{-1} \iota_a)$ results in an optimand tha has no internal solution for r_{12}, the function is maximized at the boundary point $r_{12} = 0$ which is the claimed restricted maximum. For $r_{12} = 1$, the first two forecasts are perfectly correlated and so $r_{13} = r_{23}$, the relative loss is $[(5 + 4r_{13})/9] - 1$ which is no larger than zero.

For the case of $a = 3$ (so $\Sigma_a = \Sigma$) we have already ruled out the unrestricted case. In general FOC for this problem are of the form

$$0 = 2(\iota_m' \Sigma^{-1} \iota_m) - 2(\iota_m' \Sigma \iota_m)(\iota_m' \Sigma^{-1} \iota_m)^2 \omega_i \omega_j + \lambda_{ij} - \gamma_{ij}$$

for each $(i, j) = (1, 2), (1, 3)$ and $(2, 3)$. For the model where a single correlation is zero and the remainder are interior (say $r_{12} = 0$ with r_{13} and r_{23} interior) we have that $\omega_1 = \omega_2$ and hence by direct calculation the constraint $r_{13} + r_{23} = 1 + r_{12}$ holds. Hence in this case we are in the situation of the previous paragraph, where all results have been ruled out apart from the claimed optimal solution. For $r_{12} = r_{13} = 0$ we have that from the FOC for r_{23} (interior) that $\omega_3 = 1/(\omega_2(\iota_m' \Sigma \iota_m)(\iota_m' \Sigma^{-1} \iota_m))$ and from the FOC for r_{12} then $\lambda_{12} = 2(\iota_m' \Sigma^{-1} \iota_m)(\omega_1 - 1)$. For a nonnegative result this requires that $\omega_1 = 1$ and the remaining forecasts have zero weight. For this case relative loss is below the claimed maxima for any r_{23}. This also rules out the case of $r_{12} = r_{23} = 0$ by rearranging the forecasts. For r_{12} and r_{13} interior and the r_{23}

on its bound we have that $\omega_2 = \omega_3$ and hence we would have the case where there are only weights on the last two forecasts (which is the same as the case of the previous paragraph after rearranging the forecasts).

Finally, we need to rule out cases where some are interior and some correlations are unity. If all three correlations are perfectly correlated, then the optimal solution is to take one of them, averaging over all of them leads to a relative loss of zero (since the average is equal to any one of the forecasts). If two are perfectly correlated, then we are at the point of combining two forecasts which means even weights on the two forecasts. Averaging then leads to a relative loss that is at most zero.

$m = 4$ case:

For the $m = 4$ problem, there are a much larger set of possibilities to rule out. We need to rule out all of the other possibilities when $a = 2$ and all of the possibilities when $a = 3$. First note that relative loss for the claimed optimum is from plugging into (4) equal to 0.25.

For the problem when $a = 2$, we have that $\omega_1 = \omega_2$ (this is true for any correlation) and so $\omega^{opt\prime}\Sigma\omega^{opt} = (\iota_a'\Sigma_a\iota_a)/4 = (1 + r_{12})/2$. From lemma 2 we have that $r_{34} = 1$ at the optimal solution (this follows from FOC of the form in (3). From lemma 2 we have that $\Sigma_{ab}' = \iota_2 \otimes r_a'$ where r_a has two elements (the first is $r_{13} = r_{23}$ and the second is $r_{14} = r_{24}$). Hence the restriction for the last two weights zero is $r_a'\Sigma_a^{-1}\iota_a = 1$ which from the $m = 3$ case we have seen is $r_a'\iota_2 = 1 + r_{12}$. At this solution

$$\begin{aligned}
(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a) &= \left(\frac{2}{1 + r_{12}}\right)(2(1 + r_{12}) + 4 + 4r_a'\iota_a) \\
&= 4 + \frac{2(4 + 4(1 + r_{12}))}{1 + r_{12}} \\
&= 4 + 8 + 8/(1 + r_{12}).
\end{aligned}$$

Here the optimal choice is $r_{12} = 0$ and hence $(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a) = 20$ and is at the stated optimal solution..

For the problem at $a = 3$, if $r_a = \iota_a$ then the restriction $\iota_a'\Sigma_a^{-1}\iota_a = 1$ hence loss is unity. Then $(\iota_a'\Sigma_a\iota_a + 1 + 2r_a'\iota_a) < 12$ and so this cannot be a global minimum. Also, $\Sigma_a = I_3$ results in a loss of $3(3 + 1 + 2r_a'\iota_a) = 3(4 + 2) = 18 < 20$ so this cannot be a global maximum. Similarly, if Σ_a were singular due to all of the first three forecast errors being perfectly correlated then the weights are anything that sums to one, and optimal loss $\omega_a'\Sigma\omega_a = 1$ so loss is at most $1(1 + 1 + 2\iota_a'r_a) \leq 8 < 20$ so again this is not a global max. What remains to be shown is that there are no internal solutions or partial internal solutions, which is more difficult.

The FOC are of the form (where the tilda refers to that variable multipled by $(\iota_a'\Sigma_a^{-1}\iota_a)^{-1}$)

$$\begin{aligned} 0 &= 1 - w_i w_j L + \tilde{\lambda}_{ij} - \tilde{\gamma}_{ij} - \tilde{\lambda}_r(r_a'\Sigma_a^{-1}u_i w_j + r_a'\Sigma_a^{-1}u_j w_i) \qquad r_{ij} \text{ in } \Sigma_a \\ 0 &= 1 + \tilde{\lambda}_{ij} - \tilde{\gamma}_{ij} + \tilde{\lambda}_r w_i \qquad r_{ij} \text{ i=1,2,3}, j = 4. \end{aligned} \tag{5}$$

If all of the r_{i4} are interior, then $\tilde{\lambda}_r = -1/w_i$ for $i = 1, 2, 3$ and hence the weights are equal. For this to be true, we know that $r_{12} = r_{13} = r_{23} = r$ and hence

$$\Sigma_a^{-1} = \frac{1-r}{1-3r^2+2r^3}\begin{pmatrix} 1+r & -r & -r \\ -r & 1+r & -r \\ -r & -r & 1+r \end{pmatrix}.$$

Since $r_a'\Sigma_a^{-1}\iota_a = 1 = \left(\frac{(1-r)^2}{1-3r^2+2r^3}\right) r_a'\iota_a$ then we have the result for $r_a'\iota_{ma}$. Further $(\iota_a'\Sigma_a^{-1}\iota_a) = 3(1-r)^2/(1-3r^2+2r^3)$, so relative loss is

$$\begin{aligned} L &= \frac{3(1-r)^2}{1-3r^2+2r^3}\left[3 + 6r + 1 + 2\left(\frac{(1-r)^2}{1-3r^2+2r^3}\right)^{-1}\right] \\ &= \frac{3(4-2r-8r^2+6r^3)}{1-3r^2+2r^3} + 6. \end{aligned}$$

At $r = 0$, this is 18<20 and hence not a solution. In the limit as r approaches one (by iterating l'hopital) we have a limit of 16<20 and hence not a solution. Could also just show that this is decreasing in r.

A further possibility is that one of the elements of r_{i4} are zero. Setting the first to zero, with the remaining two interior results in the FOC

$$\begin{aligned} 0 &= 1 + \tilde{\lambda}_r w_i \qquad \text{i=2,3}, j = 4. \\ 0 &= 1 + \tilde{\lambda}_{ij} + \tilde{\lambda}_r w_i \qquad (i,j) = (1,4). \end{aligned}$$

For this solution $\tilde{\lambda}_r = -1/\omega_2 = -1/\omega_3$ so $\omega_2 = \omega_3$. Further, $\tilde{\lambda}_{14} = (\omega_1/\omega_2) - 1 \geq 0$ so $\omega_1 \geq \omega_2$. For equalty of the weights $u_2'\Sigma_a^{-1}\iota_a = u_3'\Sigma_a^{-1}\iota_a$ which results in the expression

$$(r_{13} - r_{12})(1 + r_{23} - r_{12} - r_{13}) = 0.$$

At the solution $(1 + r_{23} - r_{12} - r_{13}) = 0$ then $w_1 = 0$ and since $\omega_1 \geq \omega_2 = \omega_3$ this cannot be a solution for the FOC. For the solution $r_{12} = r_{13}$,the problem is slightly simplified. From

$r_a'\Sigma_a^{-1}\iota_a = 1$ we have that $r_{24} = (u_2'\Sigma_a^{-1}\iota_a)^{-1} - r_{34}$ so $r_a'\iota_a = (u_2'\Sigma_a^{-1}\iota_a)^{-1}$. The problem can now be stated as

$$\max_{r_{12},r_{23}} \left(\frac{3 + r_{23} - 4r_{12}}{1 + r_{23} - 2r_{12}^2}\right)(4 + 4r_{12} + 2r_{23}) + 2\frac{(1 + r_{23} - 4r_{12})}{1 - r_{12}}.$$

With the constraint that $r_{12} \leq r_{23}$ this is maximized as $r_{23} \to 1$ and $r_{12} \to 0$, which is the case where the second and third forecast errors are perfectly correlated and is the maximal solution given above where we give weight equal to one half on the first forecast and a quarter to each of the identical second and third forecasts.

The next possibility to check is that only one of the elements of r_a is nonzero. In this case, we have that

$$\begin{aligned}(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a) &= (\iota_a'\Sigma_a\iota_a)(\iota_a'\Sigma_a^{-1}\iota_a) + (1 + 2r_{34})(\iota_a'\Sigma_a^{-1}\iota_a)\\ &\leq 10 + 3(\iota_a'\Sigma_a^{-1}\iota_a)\\ &\leq 10 + 3 * 3\\ &= 19\end{aligned}$$

which is less than the worst case for the solution above.

Finally, with r_a elements interior consider some elements of Σ_a as corner solutions. With two zero correlations, set $r_{12} = r_{13} = 0$. In this case by direct calculation we have

$$(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a) = \left[\frac{1 + 2(1 + r_{23})^2}{(3 + r_{23})^2}\right](4 + 2r_{23} + 2(r_{14} + r_{24} + r_{34})).$$

The restriction $r_a'\Sigma_a^{-1}\iota_a = 1$ yields $r_{14} = 1 - (r_{24} + r_{34})/(1 - r_{23})$. Substituting this into the above expression yields

$$(\iota_m'\Sigma\iota_m)(\iota_a'\Sigma_a^{-1}\iota_a) = \left[\frac{1 + 2(1 + r_{23})^2}{(3 + r_{23})^2}\right]\left(6 + 2r_{23} - 2(r_{24} + r_{34})\left(\frac{r_{23}}{1 - r_{23}}\right)\right).$$

The FOC for the derivative with respect to either r_{24} or r_{34} yields

$$0 = -2\left[\frac{1 + 2(1 + r_{23})^2}{(3 + r_{23})^2}\right]\left(\frac{r_{23}}{1 - r_{23}}\right).$$

For $0 < r_{23} < 1$ there is no solution so this cannot be a maximum. As $r_{23} \to 1$ both the second and third forecast errors become perfectly correlated and hence the loss is identical to the result where we combine the first two forecasts (where here we can have any weights on the second and third such that they sum to one half, with one half weight on the first).

With a single zero correlation (set $r_{12} = 0$) we follow the same strategy as above. The restriction $r_a' \Sigma_a^{-1} \iota_a = 1$ which yields

$$r_{34} = \frac{r_{14}(1 - r_{23})(1 + r_{23} - r_{13}) + r_{24}(1 - r_{13})(1 + r_{13} - r_{23})}{(1 - r_{13} - r_{23})}$$

where $r_{13} + r_{23} < 1$ (if these are equal we have the optimal result of weights of one half on the first and second forecasts as above). Substituting this into $(\iota_m' \Sigma \iota_m)(\iota_a' \Sigma_a^{-1} \iota_a)$ with direct calculation for $(\iota_a' \Sigma_a^{-1} \iota_a)$ yields an expression for $(\iota_m' \Sigma \iota_m)(\iota_a' \Sigma_a^{-1} \iota_a)$ that depends on $\{r_{13}, r_{23}, r_{14}, r_{24}\}$. From the first order condition with respect to r_{14} and calculation results in an expression involving both r_{13} and r_{23}, which cannot be satisfied for a real solution ruling out this possibility as an optimal solution. ■

References

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

Bates, J., and C. Granger (1969): “The Combination of Forecasts,” *Operations Research Quarterly*, 20, 451–468.

Capistran., C., and A.Timmermann (2009): “Forecast Combination with Entry and Exit of Experts,” *Journal of Business and Economic Statistics*, 27, 428–440.

Clemen, R. (1989): “Combining Forecasts: A Review and Annotated Bibliography,” *International Journal of Forecasting*, 5, 559–581.

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

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

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

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

Hsiao, C., and S. Wan (2010): “Is there an optimal forecast combination,” .

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

Stock, J., and M. Watson (2001): “A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series,” in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive Granger*, ed. by R. Engle, and H. White, pp. 1–44. Oxford University Press, Oxford.

Swanson, N. R., and T. Zheng (2001): “Choosing among competing econometric forecasts: A regression-based forecast combination using model selection,” *Journal of Forecasting*, 20, 425–440.

Timmermann, A. (2006): “Forecast Combinations,” in *Handbook of Forecasting Volume 1*, ed. by G. E. et. al, pp. 135–196. Elsevier, Amsterdam.

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