Linear Hybrid Shrinkage of Weights for Forecast Selection and Combination

Felix Schulz
KU Eichstätt-Ingolstadt
felix.schulz@ku.de

Thomas Setzer
KU Eichstätt-Ingolstadt
thomas.setzer@ku.de

Nathalie Balla
KU Eichstätt-Ingolstadt
nballa@ku.de

Abstract

Forecast combination is an established methodology to improve forecast accuracy. The primary questions in the current literature are how many and which forecasts to include (selection) and how to weight the selected forecasts (weighting). Although integrating both tasks seems appealing, we are only aware of a few data analytical models that integrate both tasks. We introduce Linear Hybrid Shrinkage (LHS), a novel method that uses information criteria from statistical learning theory to select forecasters and then shrinks the selection from their in-sample optimal weights linearly towards equality, while shrinking the non-selected forecasts towards zero. Simulation results show conditions (scenarios) where LHS leads to higher accuracy than LASSO-based Shrinkage, Linear Shrinkage of in-sample optimal weights, and a simple averaging of forecasts.

1. Introduction

Since the seminal papers of Reid [1] and Bates and Granger [2], improving accuracy of individual forecasts through their combination evolved to a predominating strategy in the forecasting literature (see, amongst others, [3, 4, 5, 6, 7]). Recent successes of forecast combination in real-world scenarios are observed, for example, in the area of short-term electricity demand forecasting [8]. By learning weighting schemes of forecasts an average accuracy improvement of 15.887% based on the Mean Absolute Percentage Error could be achieved. Thereby, the results of the combination based on French and Australian load data show better results than individual methods, especially on public holidays, which are considered to be particularly difficult to predict. The merits of forecast combination are also demonstrated in the area of oil price forecasting with a 13% reduction in the Mean Squared Prediction Error, showing that forecasting by combining several forecast methods based on oil price information is a viable alternative than relying on judgemental forecasts as a sole source in this realm [9]. Interestingly, both examples either use techniques for calculating optimal weights based on the in-sample error structure of forecasts or average the influence of forecasters for the combination, although further improvements might be achieved by introducing regularization like the shrinkage of optimal weights towards a simple average as shown in [10, 11, 12].

Besides the optimal weighting of forecasts, other studies have shown that a selection of forecasters to be included in forecast combination, i.e., how many and which forecasts to include (selection), can also lead to benefits. Mannes et al. [13] propose that either the whole crowd, a selected crowd, or the best member should be selected depending on the dispersion of expertise and the cancellation effects of errors of different forecasters. Further work has been published based on hard thresholding over filtering criteria from information theory to pick one single best forecast from the crowd [14].

While a fusion of both strategies into one holistic model seems promising, combined model selection and weighting has received scant attention so far as most articles either consider the weighting task or the group selection task in depth.

An exception is the recently published work by Diebold and Shin [15]. The authors propose and test combination methods that first select forecasts using the Least Absolute Shrinkage and Selection Operator (*LASSO*), relearn optimal weights and then shrink the weights of the selection to equal weights.

Inspired by this work, in this paper we propose Linear Hybrid Shrinkage (*LHS*) – a model that also selects a subgroup of forecasters and shrinks their weights. *LHS*, however, differs in the following aspects. First, instead of using the *LASSO* it uses information criteria like permutation based variable importance to rank the individual forecasters. Second, using the number of forecasters p' to be selected from p forecasters as a hyperparameter, *LHS*

URI: https://hdl.handle.net/10125/79599
978-0-9981331-5-7

shrinks the in-sample optimal weights of the top p' forecasters linearly towards their average weighting, while shrinking (out) the weights of the remaining $p-p'$ forecasters linearly towards zero.

One motivation for *LHS* stems from the behavior of *LASSO*-based Shrinkage. *LASSO* is a very popular shrinkage technique successfully used in a multitude of applications. A strength of *LASSO* is its tendency to select one forecast from a group of more strongly correlated forecasts and therewith its ability to shrink out several forecasters and reduce dimensionality. However, this strength can – to some extend – also turn into a weakness when multiple accurate and correlated forecasts and only a small or moderate set of training records are available. Indeed, in cases with limited data and many highly correlated and similarly performing individual forecasts, a situation common in practice, *LASSO* is therefore somewhat prone to overfitting as it often relies on fewer forecasters and might put everything "on one or few horses" although this might not be optimal in terms of out-of-sample accuracy.

A second motivation is based on findings in recent literature that recommend to select a forecaster subgroup and shrink their in-sample optimal weights to the simple average, while there is no analytical model to determine the best number of forecasters to be kept in the combination. *LHS* considers these findings by also shrinking a subgroup of forecasters towards their average weighting, while not removing remaining forecasters completely, but shrinking their weights towards zero to a controllable extent; thus, these are only shrunk out if cross-validation-based procedures have determined a maximal level of shrinkage.

The remainder of this paper is structured as follows. In Section 2, the foundations of (linear) forecast combination and shrinkage are summarized. Section 3 presents *LHS*, the method we introduce. The simulation-based experimental design used to study the behavior of *LHS* is described in Section 4. Section 5 presents and discusses the experimental results. The paper finishes with conclusions and an outlook to future research in Section 6.

2. Forecast Combination and Shrinkage

Let p out of $i \in 1,...,p$ forecasting models (henceforth: forecasters) generate not-perfectly collinear forecasts, $f = (f_1,\ldots,f_p)$. In a linear combination of forecasts, weights $w \in R^p$ with $\sum_{i=1}^{p} w_i = 1$ are assigned to form $f'w$, where f' is the transpose of the forecast vector. The task considered here is to find w that minimize a certain loss function with the combination, typically the Mean Squared Error (MSE). As shown in [2] for two forecasts and in [16] for its multivariate extension, assuming individually efficient, unbiased forecasts, i. e., with errors following a multivariate normal distribution with a mean of zero, the weights that minimize MSE (called optimal weights *OW*, w_o) can be derived with the error covariance matrix of forecasts Σ_e. With γ denoting a column vector of p ones, *OW* are defined in (1).

$$w_o = \frac{\Sigma_e^{-1}\gamma}{\gamma'\Sigma_e^{-1}\gamma} \tag{1}$$

Since the true Σ_e is usually not known, available training data on past forecast error $E \in R^{n\times p}$ for periods $t \in 1,...,n$ is used to estimate it as $\hat{\Sigma}_e = \frac{1}{n}E'E$. The estimated optimal weights $\hat{w_o}$ are shown in (2).

$$\hat{w_o} = \frac{\hat{\Sigma}_e^{-1}\gamma}{\gamma'\hat{\Sigma}_e^{-1}\gamma} \tag{2}$$

Estimated *OW*, $\hat{w_o}$ (with $\hat{w_o} \neq w_o$), are then applied to unseen forecasts (the evaluation or test data) which makes this combination scheme approach prone to overfitting. Therefore, shrinking *OW* for example towards equal weights (*EW*), w_e, can be expected to result in lower MSE on test data, in particular with higher levels of uncertainty due to less observed spread in forecast ability between forecasters and small amounts of training data [12].

One approach is to linearly shrink $\hat{w_o}$ towards w_e using a shrinkage parameter $\lambda \in [0,1]$, resulting in the weight vector $\hat{w}^{\lambda}$ defined in (3). With larger values of λ a greater degree of shrinkage towards w_e is applied, whereas $\lambda = 0$ corresponds to $\hat{w_o}$ and $\lambda = 1$ to w_e.

$$\hat{w}^{\lambda} = \frac{\lambda}{p}\gamma + (1-\lambda)\,\hat{w_o} \tag{3}$$

As an alternative to *Linear Shrinkage* as shown in (3), $\hat{w_o}$ can also be shrunk non-linearly using e. g. a *LASSO* or *Ridge* penalty on the weight deviations from equal weights as done in [15]. The authors coin the respective formulations Egalitarian LASSO (*eLASSO*) and Egalitarian Ridge (*eRidge*). With e_{it} indicating the error of forecast i on forecast event t, and λ as a hyperparameter that can be tuned using cross-validation, the *eRidge* regression formulation is shown in (4) and the *eLASSO* formulation is shown in (5).

$$\begin{aligned}\hat{w}_{eRidge} = \arg\min_{w}(&\sum_{t=1}^{n}(\sum_{i=1}^{p} w_i e_{it})^2 + \\ &\lambda\sum_{i=1}^{p}(w_i - \frac{1}{p})^2)\end{aligned} \tag{4}$$

$$\hat{w}_{eLASSO} = \arg\min_{w}(\sum_{t=1}^{n}(\sum_{i=1}^{p} w_i e_{it})^2 + \lambda \sum_{i=1}^{p} |w_i - \frac{1}{p}|) \tag{5}$$

A drawback of shrinking all forecasters towards equal weights is that poorly performing forecasts are retained and no forecast selection occurs. As optimal shrinkage levels increase with uncertainty (e. g. with decreasing number of training samples), weights of poor performing forecasts then approach weights of the individually best forecasts.

However, numerous researchers found that including poor performing models in a forecast combination can worsen forecast performance. Therefore, several papers have investigated how poorly performing forecast models can be removed from a combination with the result that forecast accuracy often improves by simply discarding models with the worst performance (e.g. [17, 18, 19]).

In this sense, Diebold and Shin [15] propose a data-analytical approach to first eliminate forecasts from the selection and then estimate *OW* and beneficial shrinkage parameters using cross-validation. The authors propose partially-egalitarian LASSO (*peLASSO*), which combines a *LASSO* penalty that selects and shrinks to zero, and a second penalty, which shrinks the remaining non-zero elements in $f(w)$ towards equality. *peLASSO* is shown in (6).

$$\hat{w}_{peLASSO} = \arg\min_{w}(\sum_{t=1}^{n}(\sum_{i=1}^{p} w_i e_{it})^2 + \lambda_1 \sum_{i=1}^{p} |w_i| + \lambda_2 \sum_{i=1}^{p} |w_i - \frac{1}{f(w)}|) \tag{6}$$

Due to the discontinuity of the objective function at $w_i = 0$, the function in (6) is implemented in two steps. In step 1 (selection to zero), forecasts are shrunk out using standard *LASSO* as shown in (7).

$$\hat{w}_{LASSO} = \arg\min_{w}(\sum_{t=1}^{n}(\sum_{i=1}^{p} w_i e_{it})^2 + \lambda \sum_{i=1}^{p} |w_i|) \tag{7}$$

In step 2 (shrinkage towards equality), the p' forecasts that survive step 1 are shrunk toward $\frac{1}{p'}$ using (4), named *peLASSO (eRidge)*, or (5), named *peLASSO (eLASSO)*, or directly setting the weights of the p' forecasts to $\frac{1}{p'}$, which is termed as *peLASSO (Avg.)*.

Inspired by the novel approach introduced by the authors, we now propose Linear Hybrid Shrinkage (*LHS*) – a model that also selects a subgroup of forecasters and shrinks their *OW*. *LHS* first uses information criteria like permutation-based variable importance to rank the individual forecasters. Second, using the number of forecasters p' to be selected as a hyperparameter, *LHS* simultaneously shrinks the in-sample *OW* of the top p' forecasters linearly towards their average weights, while shrinking (out) the weights of the remaining forecasters linearly towards zero.

3. Linear Hybrid Shrinkage

The aim of Linear Hybrid Shrinkage (*LHS*), the method we introduce, is to rank and select forecasts based on information criteria discussed later in this section, and to shrink the top p' forecasts from their in-sample *OW* to equality, while the remaining forecasters are shrunk towards zero (i.e., given high shrinkage, out of the selection).

To achieve both selection and weighting, *LHS* adjusts (3). Instead of shrinking all p forecasters towards $\frac{1}{p}$, only the top p'-ranked forecasters are shrunk towards their *EW*, $\frac{1}{p'}$. The remaining $p - p'$ forecasts are shrunk linearly towards zero.

As shown in Table 1, the decision which forecasters are shrunk towards *EW* can be represented by a vector v of size p, in which each element represents a forecaster and the selected forecasters (those that are shrunk towards *EW*) are set to 1, while the others (that are shrunk towards zero) are set to 0.

Top p'	v	$\frac{v}{\gamma' v}$
$p' = 1$	$(1, 0, 0)'$	$(1, 0, 0)'$
$p' = 2$	$(1, 1, 0)'$	$(\frac{1}{2}, \frac{1}{2}, 0)'$
$p' = 3$	$(1, 1, 1)'$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$

Table 1. Example of top p'-selection based on forecaster performance or importance

With γ as a column vector of p ones, the optimal weights w_o and the shrinkage parameter λ, *LHS* is formulated in (8).

$$\hat{w}_{p'}^{\lambda} = \frac{\lambda v}{\gamma' v} + (1 - \lambda)\, \hat{w}_o \tag{8}$$

As in (3), $\lambda = 0$ corresponds to $\hat{w}_o$. With increasing λ, a hyperparameter that will be tuned using cross-validation, $\hat{w_o}$ of p' forecasts will be increasingly shrunk towards $w_e = \frac{1}{p'}$ and the remaining ones to zero.

To determine the ranking of individual forecasters, i. e., to set their values in v, *LHS* uses two types of prior information which are now presented.

3.1. Selection Based on Forecaster Ability

As a first strategy we consider a common measure to assess forecast ability, i.e., forecaster performance. A forecaster's performance is thereby usually measured by an in-sample loss function such as the MSE, which corresponds to the error variance of an individual forecast if forecasts are efficient. Ranking the top p' forecasters, with $p' \in \{1, ..., p\}$ by their MSE leads to p possible shapes of v as shown in Table 1 for $p = 3$ forecasters, whereby, for reasons of brevity we assume forecasters are already ordered by their individual in-sample MSE in ascending fashion.

A potential disadvantage of using forecasters' individual performances (MSE) as ranking criterion is that a forecaster is evaluated in isolation and the contribution or importance of a forecaster in a combination model is not explicitly considered. Therefore, as a second ranking criterion, we propose a measure that quantifies the importance of a forecast in a combination.

3.2. Selection Based on Forecaster Importance

To take into account, e.g. interaction effects, mutual variance reduction effects or suppressor effects that might increase accuracy but cannot be accounted for in an isolated assessment of forecasters, we consider importance measures.

To estimate the importance of a forecaster for the accuracy of a combined prediction, several variable importance (VI) measures have been reported in the literature (for a more detailed review, see e.g. [20]). From the group of VI measures, we use the permutation-based VI, also called Model Reliance (MR), as proposed in [21].

The intuition of MR is to first learn a model, in our case in-sample *OW*, and then determine the MSE with the model on validation data. For each forecaster, its prediction values are then randomly permuted, the model is re-applied to the validation data, and again the MSE is measured, but this time using randomly permuted predictions from a forecaster i. The signal of this forecaster i is then considered as noise without predictive value, and the ratio of the MSE with noise and the MSE with the original data is considered as the importance I_i (or variable importance, VI_i) of a forecaster i as in (9).

$$I_i = \frac{\text{MSE}(\textit{OW}, \textit{i} \text{ under noise})}{\text{MSE}(\textit{OW}, \textit{i} \text{ without noise})} \tag{9}$$

The intuition is that removing the information of a forecaster's (predictions) leads to an increase in MSE. Consequently, the most important forecaster is considered as the one with the highest permutation-based VI value.

To determine $I_i \ \forall \ i$, we first split the available data in k folds, learn *OW* on $k - 1$ folds and evaluate their performance on the k-th fold by MSE to get the expected loss without noise. To introduce noise, we permutate the points in fold k for one forecaster i and again calculate the MSE in fold k given the *OW* learned. We shuffle the data points for this forecaster several times, average the results and repeat this process k times by shifting the folds for training and validation sets. Performing this procedure for each forecaster i, we receive a forecaster importance score I_i which is larger for forecasters for whom swapping the values reduces the accuracy of the combined forecast more.

As mentioned before, given p forecasters, there are p different shapes for v. To determine the selection of the final v, we consider v as a hyperparameter tuned via k-fold cross-validation as will be described in the next section.

4. Experimental Evaluation

We now describe the setup of our simulation-based experiments used to analyze the behavior of *LHS* in scenarios with different numbers of forecasts, training sample sizes, error variances and covariances. As commonly assumed, the simulation errors of individual forecasts generated in the simulations are efficient, i.e., time-invariant and following a multivariate normal distribution with zero means. Hence, the error variance of an individual forecast is the forecaster's MSE.

Subsequently, we provide and discuss the experimental outcome with *LHS* and compare the results to the results obtained with alternative weighting schemes, namely *OW*, *SA*, *Linear* and *eLASSO* shrinkage as well as shrinkage and selection via *peLASSO*.

4.1. Experimental Design

In our simulation-based experiments, we generate synthetic error samples according to a given covariance matrix of forecasters' errors (unknown to the models). The errors are drawn from a multivariate normal distribution with pairwise correlation ρ amongst forecasters' errors, with

No.	Individual Forecast i			
	1	2	3	4
1	1.00	1.26	1.58	2.00
2	1.00	1.44	2.08	3.00
3	1.00	1.58	2.52	4.00
4	1.00	1.71	2.92	5.00

Table 2. Example of simulated forecast variance vectors with $p = 4$

Treatment Combinations	
Forecaster	8, 10, 12, 14, 16
Correlation	0.4, 0.525, 0.65, 0.775, 0.9
Variance	2, 3, 4, 5
Train	30, 40, 50, 60, 70
Test	5000

Table 3. Overview of treatment combinations

$\rho \in \{0.4,\ 0.525,\ 0.65,\ 0.775,\ 0.9\}$. We assume error correlations of medium to high positive values, as usually assumed in scientific work on forecast combination.

For every scenario (a simulation with a set of treatments like ρ and other treatments introduced later in this section) a small empirical snapshot (n error observations) is drawn, that serves as training data. We consider $n \in \{30,\ 40,\ 50,\ 60,\ 70\}$. For all of the treatment combinations we run simulations with five different numbers of forecasters $p = 8,\ 10,\ 12,\ 14,\ 16$ and four different spreads of the individual forecasts' error variances. We let the variance of the p-th forecast σ_p^2 be 2, 3, 4 or 5, decreasing geometrically by $(\sigma_p^2)^{\frac{i-1}{p-1}}$ for the i-th forecast, with $i \in \{1, ...,\ p-1\}$. For instance, for $p = 4$ we obtain the four variance vectors shown in Table 2.

With five different numbers of forecasters, five training data sizes, four different variances as well as five correlations, we receive a total of 500 scenarios. Each scenario, i.e. treatment combination, is repeated ten times, with different errors drawn from multivariate normal distributions to increase robustness. Since the ultimate goal is to approximate the optimal weights behind the true underlying data generation process based on the given training patterns, a generous test set size of 5,000 data points is chosen. An overview of the treatment combinations can be found in Table 3.

Several benchmark methods are implemented and tested besides *LHS* to compare the performance of *LHS* and determine scenarios where it might be favorable to be used or dominated by alternative approaches.

As benchmark methods, *OW*, *SA*, *Linear Shrinkage* as well as non-linear shrinkage via *eLASSO* are used. The idea of crowd selection and subsequent averaging is reflected in the *peLASSO* method presented in [15], with the combination of *peLASSO (Avg.)* and *peLASSO (eRidge)* used for this purpose. For *LHS* and *Linear Shrinkage*, shrinkage is performed in 50 steps, whereby $\lambda = 0$ corresponds to a shrinkage level of 0% and $\lambda = 49$ to a maximum shrinkage level of 100%. *LASSO* learns max(λ) as the smallest value of lambda for which all coefficients are zero, i.e. EW for *eLASSO*, over the free available R package and respective function genlasso [22]. For the shrinkage in *peLASSO* over *eRidge*, a grid of λ ranging from zero to 3,000,000 is used to learn λ_{opt}, as depending on the scenario heavy penalization can be required according to [15].

Before we present and discuss the results obtained in our simulation, we will briefly illustrate how *LHS* works and how its hyperparameters are tuned.

4.2. Example LHS Weight Shrinkage Path and Associated MSE Curve

For illustration, we use an example simulation run with $p = 14$, $n = 60$, $\rho = 0.525$ and a high variance treatment with $\sigma_p^2 = 4$. As selection criterion, forecaster performance is taken which ranks the forecaster according to their MSE in the in-sample data. For the in-sample forecast performance of forecaster $i, ..., p$, we observe the following variance vector $MSE_i^{train} \in \{1.15,\ 1.04,\ 0.99,\ 0.99,\ 1.28,\ 1.76, 1.58,\ 2.20,\ 2.17,\ 2.62,\ 2.65,\ 2.54,\ 2.23,\ 3.01\}$ (Note that this is not corresponding to the true variance of the forecast, as we are only using a snapshot of the simulated data). After performing cross-validation, top $p' = 4$ was selected, leading to the following shape of $v = (1,\ 1,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)'$ and $\frac{v}{\gamma' v} = (0.25,\ 0.25,\ 0.25,\ 0.25,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)'$ accordingly.

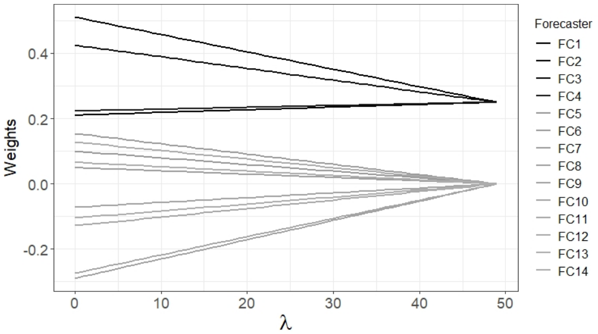

Figure 1. Example shrinkage path of weights over λ using selection over forecaster performance

Figure 1 shows the coefficients along the shrinkage path. With $\lambda = 0$, no shrinkage is performed and *OW* can be observed. With $\lambda = 49$, the maximum shrinkage level is reached and weights of non-selected forecasters (gray lines) are linearly shrunk to zero, while the weights of selected forecasters (black lines) are shrunk to $\frac{v}{\gamma' v}$ in a linear fashion.

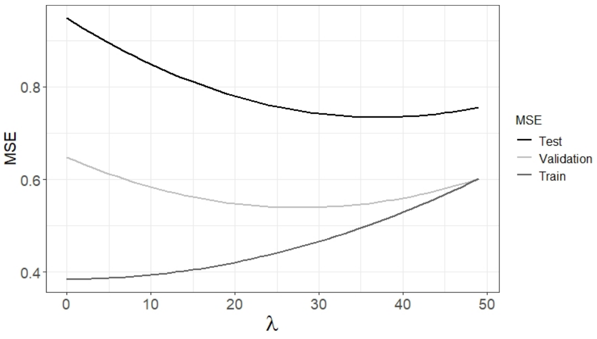

Figure 2. Example MSE development over λ

To analyze *LHS* performance, Figure 2 plots the development of MSE over λ. The dark gray line depicts the in-sample MSE, the gray line the MSE on validation data, and the black line the MSE on test data. As a shrinkage of zero (*OW*) is by design optimal on the training data but prone to overfitting and high MSE on the validation and test sample, the minimum MSE in-sample, but maximum MSE value out-of-sample can be observed at $\lambda = 0$. The curves show that the MSE monotonously increases in-sample, but show an U-shaped MSE curve on the validation and test set. On validation data, a medium optimal shrinkage level of 58% (or $\lambda = 29$) has been learned using cross-validation. Applying the learned model and shrinkage level to the test set a MSE of 0.747 is achieved. Inspecting the MSE test curve, it is noticeable that by cross-validation still a slightly overfitted model is learned. Yet the actual optimal lambda value on MSE test is 38, which would have yielded a test MSE of 0.734.

5. Experimental Results

We now present and discuss the experimental outcome. First, we provide aggregated results of the average test MSE of the considered combination approaches. Second, we drill-down the results to treatment combinations to analyze whether certain scenarios are dominated by certain weighting methods in terms of the ranges of treatment values. Third, we discuss the results and provide further analysis regarding optimal shrinkage levels and the number of forecasters included in a combination with the various approaches.

Method	MSE	λ_{opt}	Forecaster
OW	0.772		12
SA	1.317		12
Linear Shrinkage	0.727	17.241	12
eLASSO	0.725	2.820	12
LHS_P	0.714	31.528	3.006
LHS_VI	0.729	25.463	6.279
peLASSO (Avg.)	1.231	10.322	6.617
peLASSO (eRidge)	0.733	310,091	6.617

Table 4. Aggregated results over all scenarios

5.1. Aggregated Results

Aggregated experimental results are provided in Table 4. The table shows the average test MSE over all scenarios per combination method. Thereby, *LHS_P* denotes the *LHS* using forecaster performance as selection criterion, whereas *LHS_VI* uses the forecasters' variable importances. The methods *peLASSO (Avg.)* and *peLASSO (eRidge)* first use *LASSO* as shown in (7) for selection, whereby the first subsequently averages the survivors and the second learns new optimal weights for the survivors and another λ for optimal shrinkage towards equality via *eRidge*. Consequently, the lambda value shown for *peLASSO (Avg.)* applies to both *peLASSO* techniques.

The table shows that over all scenarios, all shrinkage approaches outperformed the *SA* as well as the *OW* approach with the exception of *peLASSO (Avg.)*. In detail, the *LHS*-based approaches, *Linear Shrinkage*, *eLASSO* as well as *peLASSO (eRidge)* lead to the lowest test MSE values between 0.714 and 0.733, compared to a mean test MSE with *SA* (*OW*) of 1.317 (0.772).

However, these aggregated results are limited regarding the insights one can gain from them, as MSE values are averaged over very different scenarios that entail very different MSE values. Therefore, we now drill-down the results and provide further insights for which scenarios which method performs best on average and can be recommended to be used.

5.2. Comparative Results with Different Treatment Value Combinations

As aforementioned, MSE results are only directly comparable for a particular variance–correlation treatment combination, as otherwise the minimum MSE that can be achieved differs strongly between different values of those parameters. Therefore, for each treatment combination (scenario), we determine the scenario-winner as the method that results in the lowest mean test MSE.

We then train a classification tree (using the open-source CART implementation rpart [23], available in the programming language R) in its default configuration and using Gini as splitting criterion, all treatments as predictor variables and the scenario-winner as the label of the target variable. The tree aims at finding subsets of parameter values (leaf nodes) that are pure, i.e., having a high concentration of scenarios with the same scenario-winner label. The labels correspond to the method labels in Table 4, with the exception of *peLASSO (eRidge)* and *Linear Shrinkage*, which are now called *Lin. Shrink.* and *peLAS(eRidge)* for shorter rendering. The resulting model trained with 500 scenarios is shown in Figure 3 as a binary tree, whereby the R package rpart.plot [24] was used for visualisation.

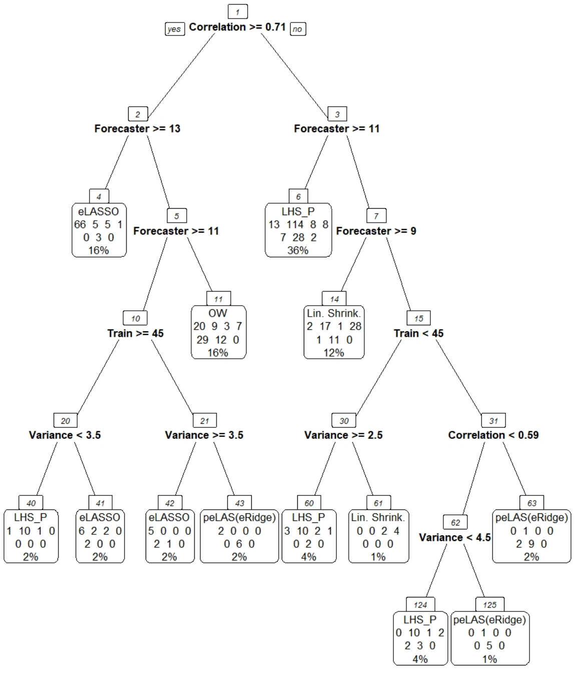

Figure 3. CART classification tree over all treatment combinations

The decision tree starts with the root node, indicated by the framed node number one. Below each node, the splitting criterion that leads to the maximum reduction in Gini impurity is shown. All observations that satisfy the condition follow the left branch, and observations that do not meet the condition are on the right branch.

Resulting leaf nodes are displayed with wider boxes containing the following information. First, the name of the most frequently chosen model, i.e., the scenario-winner in that node, is presented. Second, the absolute frequency per model in that node is reported. As *peLASSO (Avg.)* did not win a treatment combination, the numbers represent the following models within a leaf node from top left to (bottom) right: *eLASSO*, *LHS_P*, *LHS_VI*, *Linear Shrinkage*, *OW*, *peLASSO (eRidge)*, and *SA*. As a third information, the relative share of the 500 observations within a node is shown on the bottom of a box.

Summing all absolute frequencies for each model at the corresponding position within a leaf across all leaves, the most frequent scenario-winner is *LHS_P* with a total of 179 wins, followed by *eLASSO* with 118 wins, *peLASSO (eRidge)* with 80 wins, *Linear Shrinkage* with 51 wins, *OW* with 45 wins, and *LHS_VI* with 25 and *SA* with two wins. However, these overall figures largely depend on the scenarios considered (i.e., on the design of the experimental treatment space). The more interesting analysis is therefore the interpretation of the results with respect to the treatment combination where *LHS_P* is the majority label, and why.

A first observation is that all four variables *Forecaster*, *Correlation*, *Variance* and *Train* are selected in the tree. Thereby, the variables *Correlation* and *Forecasters* are used for the first splits, indicating a higher importance of these variables in finding treatment conditions with a high proportion of same scenario-winner labels.

Summing-up the percentages of observations, 80% of scenarios can be assigned to a dominant model by separating the correlation into ranges of higher vs. medium correlation and using the number of forecaster as second splitting criterion.

We observe that scenarios with $Correlation >= 0.71$ are dominated by *eLASSO* and *OW*, where for lower numbers of forecasters of eight or ten *OW* dominates, and *eLASSO* dominates when $Forecaster >= 13$, i.e., for high numbers of forecasters. This observation seems reasonable, as *OW* performs comparably well in scenarios with high correlations. In contrast, with a higher number of forecasters relative to the training sample size, the estimation uncertainty and hence overfitting increases and a correction (shrinkage) of learned weights might be necessary [3, 11, 16]).

In scenarios with moderate correlation ($Correlation < 0.71$), where generally stronger shrinkage is recommended in the literature, and high numbers of forecaster, i.e. $Forecaster >= 11$, *LHS_P* mostly wins. Considering ten forecasters and moderate correlation, *Linear Shrinkage* wins 28 of a total of 60 scenarios with these treatment combination. In summary, the treatment combinations described include scenarios where shrinkage is expected to have the observed positive effect and the results are generally consistent with previous findings in the literature.

The more interesting observations, revealing a more granular picture for the remaining 20% of all scenarios, are found further down the tree, where CART additionally uses the variables *Variance* and *Train*. In the case of high correlation and a number of twelve forecasters, the wins are split between *LHS_P*, *eLASSO* and *peLASSO (eRidge)*. Similarly, in the case of moderate correlation and a number of eight forecasters, wins are split between *LHS_P* and *peLASSO (eRidge)*, but also *Linear Shrinkage*. Thereby, at a smaller training sample size, i.e. $Train < 45$, and high variance, i.e. $Variance >= 2.5$, *LHS_P* dominates, with *Linear Shrinkage* dominating in case of smaller variances.

These results are reasonable, since smaller amounts of training data are likely to require high shrinkage values to avoid overfitting. As ability dispersion increases, additional selection might be beneficial, while more equalization of weights between all forecasters might be necessary when dispersion is equal.

Further, in the case of moderate correlation, for training sample sizes of 45 and above, a correlation smaller than 0.59 as well as a variance smaller than 4.5, *LHS_P* dominates, whereby in the two other combinations *peLASSO (eRidge)* dominates.

In the following subsection, we will provide more in-depth analysis to interpret these more complex results by studying the shrinkage levels determined by the different combination methods. Specifically, we study whether keeping the non-selected forecasts in the set (but shrinking them towards zero) provides benefits.

5.3. Analyses of Shrinkage Levels

As shown above, *LHS_P* performs particularly well in scenarios with moderate correlations among forecasts and a number of forecasts greater than eleven, whereas in scenarios with higher correlation and high numbers of forecasters *eLASSO* dominates. Interesting questions are why *eLASSO* tends to outperform *LHS_P* in high correlation scenarios, *LHS_P eLASSO* in moderate correlation scenarios, and also why *peLASSO (eRidge)* dominates *LHS_P* in moderate correlation scenarios with lower numbers of forecasters ($Forecaster = 8$), very high spread in ability ($Variance >= 4.5$) and a high number of training data ($Train > 45$).

To shed light on this phenomena, we first filter the results according to a specific treatment combination, namely $Correlation = 0.9$ and $Forecaster = 16$, and compare the average weights, shrinkage levels and numbers of selected forecasters of the node winner (i.e. *eLASSO*) versus *LHS_P*. To gain further insights into the optimal shrinkage level, we include the actual OW (henceforth: *OW (act.)*) calculated by the true covariance matrix of the forecasters' errors from the simulation process as shown in (1), and analyze the difference between the average weights of the methods and the average *OW (act.)*.

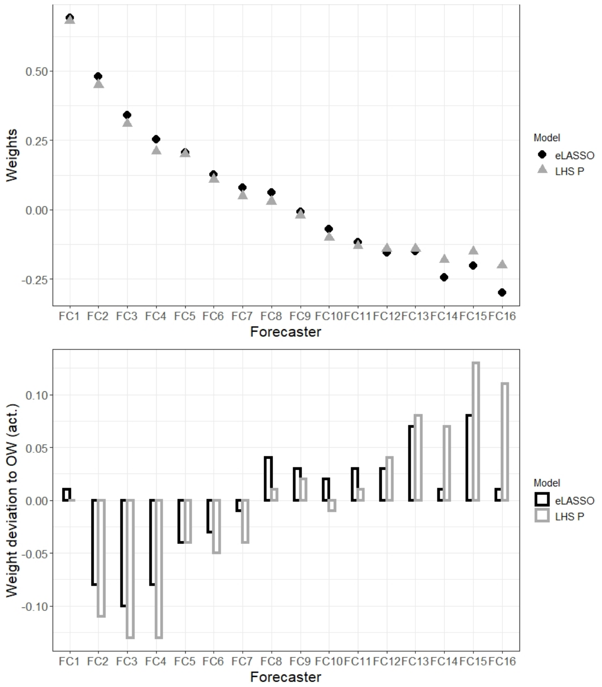

Figure 4. Learned weights by $eLASSO$ and LHS_P and weight deviations to $OW(act.)$ in scenarios with 16 forecasters and high correlation of 0.9

The average weights of the methods *eLASSO* and *LHS_P* for all scenarios with $Correlation = 0.9$ and $Forecaster = 16$ are plotted in Figure 4 with the colors black and dark gray, respectively. The weight deviations of *eLASSO* and *LHS_P* from *OW (act.)* are further displayed below in bars with the respective color. As known from recent studies, in scenarios of high correlations *OW* are sensitive to even minor changes in the forecasters' variances leading to the assignment of high positive weights to forecasters with lower variances and negative weights to bad-performing forecasters [25]. Due to the low shrinkage levels and the geometrically descending variance simulation between forecasters, the resulting high spread in weights can be seen in Figure 4.

The average MSE on test data for all scenarios is 0.43 for *eLASSO* and 0.475 for *LHS_P*, whereby *LHS_P* selects 1.885 forecaster on average. We observe a small relative shrinkage level of around 3% for *eLASSO*, and moderate relative shrinkage levels of around 32% for *LHS_P*. Comparing the estimated weights of *eLASSO* and *LHS_P*, the estimated weights of *eLASSO* are on average closer to the *OW (act.)* for well-performing forecasters ($FC1$ - $FC6$), while *LHS_P* estimated the weights slightly better for moderate forecasters ($FC8$ - $FC11$). The main difference in weights is observed for the individually worst-performing forecasters ($FC14$ - $FC16$). Here, *eLASSO* assigns negative weights to these forecasters, while *LHS_P* shrinks them towards zero. By reducing the magnitude of the negative weights, *LHS_P* loses the ability to neutralize errors by combining predictions that simultaneously overestimate or underestimate the true value, which might overall result in increased test MSE.

A reason for the slight over-shrinkage of bad-performing forecasts could be the high spread of the estimated *OW*, which increases the slope between the *OW* and the shrinkage target in terms of the mean or zero for forecasters with high absolute weights. The steep slope, in turn, leads to larger weight losses or gains per shrinkage step for all forecasters due to the linear shrinkage in *LHS_P*, while *eLASSO* may correct individual forecaster's weights to varying degrees and benefit from the non-linear shrinkage behavior.

Similar to the analysis above, we now filter the result data for moderate $Correlation = 0.4$ and $Forecaster = 16$. The average weights of the methods *eLASSO* and *LHS_P* as well as their deviations to the *OW (act.)* within the filtered scenarios are plotted in Figure 5 with the colors black and dark gray, respectively.

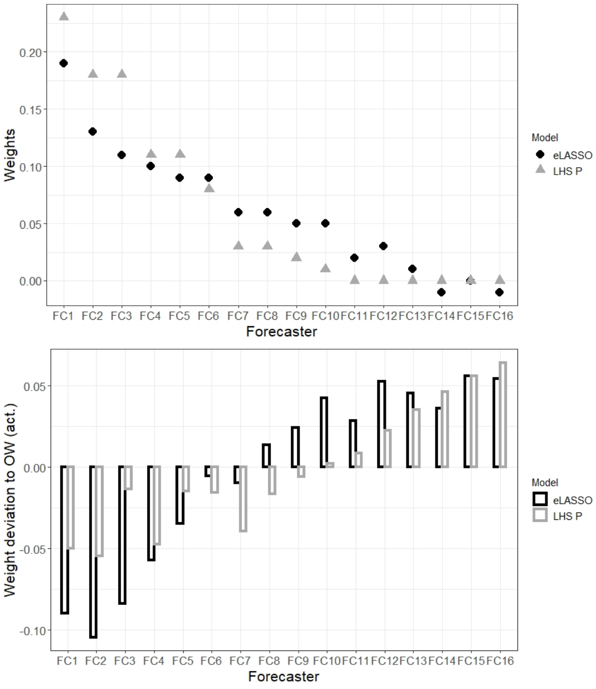

Figure 5. Learned weights by $eLASSO$ and LHS_P and weight deviations to $OW(act.)$ in scenarios with 16 forecasters and moderate correlation of 0.4

The average MSE on test data for all scenarios is 0.664 for *LHS_P* and 0.706 for *eLASSO*, with shrinkage ratios of relatively 95% for *LHS_P* and 23.5% for *eLASSO* indicating that more shrinkage is required in areas of moderate correlations.

The required higher shrinkage seems to be unfavorable for *eLASSO*. As *eLASSO* increasingly pushes forecasters' weights toward their mean value, too little weights seem to be assigned to well-performing forecasters ($FC1$ - $FC6$) and too much weight to moderate- and poor-performing forecasters ($FC7$ - $FC16$). This behavior leads to higher average deviations to *OW (act.)*. *LHS_P*, on the other hand, focuses on a subset of well-performing forecasters, while moderately to poorly performing forecasters are shrunk toward low positive weights or zero. As displayed in the bar chart, assigning more weighting to well-performing forecasters on average leads to lower deviations from *OW (act.)* for well-performing forecasters compared to *eLASSO*. This finding underlines that shrinking toward a higher mean may be beneficial if the right forecasters are selected. Although *LHS_P* reveals slight deviations to the *OW (act.)* for poor-performing forecasters, they are shown to be acceptable as the impact of cancellation of errors by combination is lower with moderate correlation.

Compared to the second best method, *peLASSO (eRidge)*, with an average test MSE of 0.689, the reason for the dominance of *peLASSO (eRidge)* in some scenarios could be explained. Although 96% of the selected forecasters are congruent in both methods, *LHS_P* selects on average the 5.2 best forecasters, while *peLASSO (eRidge)* selects more diverse 10.4 forecasters. With a higher number of forecasters ($Forecaster = 16$), the probability of selecting unfavorable forecasters increases, especially for smaller datasets. While the forecasters in *peLASSO (eRidge)* are completely removed, *LHS_P* can still benefit from forecasters by assigning at least some weight to them even if no optimal selection was made. This could be a reason why *peLASSO (eRidge)* shows slight MSE losses on test data compared to *LHS_P* for $Forecaster = 16$, but dominated *LHS_P* for the scenarios in $Forecaster = 8$, $Variance >= 4.5$ and $Train > 45$. In these cases, it is easier for *peLASSO (eRidge)* to select the correct forecasters, and the complete removal of forecasters may result in improvements in terms of MSE.

6. Conclusion and Future Work

We introduced Linear Hybrid Shrinkage (*LHS*) for forecast combination and selection. *LHS* first selects forecasts based on a prior information criterion, and then only shrinks the selection to equality, but the remaining forecasters towards zero. For forecast selection, we proposed to rank forecasters either based on their individual forecast ability (MSE on training data) or variable importance measures (based on their contribution to the accuracy of a combined forecast).

The results show improvements over existing weighting approaches in specific scenarios over *LASSO*-based shrinkage approaches, which can likely be attributed to the latter property of *LHS*, namely the ability of shrinking worse performing forecasts toward zero, but not completely shrinking them out of the group. In addition, using information criteria to rank forecasters might provide additional benefits over cross-validation-based out-shrinkage in case of lower amounts of forecasts. A more detailed investigation of the advantages of information criteria-based selection will be the subject of our future research.

Future work will also be related to further studying the conditions under which *LHS* can be expected to lead to lower MSE than alternative approaches. In addition, we plan to explore different forecast selection and ranking criteria, such as Shapley values or the penalization of model complexity using criteria such as the Bayesian or Akaike information criterion. A further, promising direction of future research is also the usage of different starting values or initial values for shrinkage than in-sample *OW*, as well as the application of non-linear shrinkage instead of the proportional shrinkage currently implemented in *LHS*. Finally, we are working on approaches to prune the vector of potential selection candidates to reduce the computational costs of the conducted cross-validation.

References

[1] D. Reid, “Combining Three Estimates of Gross Domestic Product,” *Economica*, pp. 431 – 444, 1968.

[2] J. Bates and C. Granger, “The Combination of Forecasts,” *Journal of the Operational Research Society*, vol. 20, no. 4, pp. 451 – 468, 1969.

[3] D. Schmittlein, J. Kim, and D. Morrison, “Combining Forecasts: Operational Adjustments to Theoretically Optimal Rules,” *Management Science*, vol. 36, no. 9, pp. 1044 – 1056, 1990.

[4] S. Makridakis and M. Hibon, “The M3-Competition: Results, Conclusions and Implications,” *International Journal of Forecasting*, vol. 16, no. 4, pp. 451 – 476, 2000.

[5] J. Stock and M. Watson, “Combination Forecasts of Output Growth in a Seven-Country Data Set,” *Journal of Forecasting*, vol. 23, no. 6, pp. 405 – 430, 2004.

[6] A. Aiolfi and A. Timmermann, “Persistence in Forecasting Performance and Conditional Combination Strategies,” *Journal of Econometrics*, vol. 135, pp. 31 – 53, 2006.

[7] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “The M4 Competition: 100,000 time series and 61 forecasting methods,” *International Journal of Forecasting*, vol. 36, no. 1, pp. 54 – 74, 2020.

[8] A. Laouafi, M. Mordjaoui, S. Haddad, T. Boukelia, and A. Ganouche, “Online electricity demand forecasting based on an effective forecast combination methodology,” *Electric Power Systems Research*, vol. 148, pp. 35 – 47, 2017.

[9] C. Baumeister and L. Kilian, “Forecasting the real price of oil in a changing world: a forecast combination approach,” *Journal of Business and Economic Statistics*, vol. 33, no. 3, pp. 338 – 351, 2015.

[10] F. Diebold and P. Pauly, “The use of prior information in forecast combination,” *International Journal of Forecasting*, vol. 6, no. 4, pp. 503 – 508, 1990.

[11] S. Blanc and T. Setzer, “When to choose the simple average in forecast combination,” *Journal of Business Research*, vol. 69, no. 10, pp. 3951 – 3962, 2016.

[12] S. Blanc and T. Setzer, “Bias–Variance Trade-Off and Shrinkage of Weights in Forecast Combination,” *Management Science*, vol. 66, no. 12, pp. 5485 – 6064, 2020.

[13] A. Mannes, J. Soll, and R. Larrick, “The Wisdom of Select Crowds.,” *Journal of Personality and Social Psychology,*, vol. 107, no. 2, pp. 276 – 299, 2014.

[14] A. Inoue and L. Kilian, “On the selection of forecasting models,” *Journal of Econometrics*, vol. 130, no. 2, pp. 273 – 306, 2006.

[15] F. Diebold and M. Shin, “Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives,” *International Journal of Forecasting*, vol. 35, no. 4, pp. 1679 – 1691, 2019.

[16] A. Timmermann, “Forecast Combinations,” *Handbook of Economic Forecasting*, vol. 1, pp. 135 – 196, 2006.

[17] R. L. Winkler and S. Makridakis, “The combination of forecasts,” *Journal of the Royal Statistical Society: Series A (General)*, vol. 146, no. 2, pp. 150 – 157, 1983.

[18] C. Granger and Y. Jeon, “Thick modeling,” *Economic Modelling*, vol. 21, no. 2, pp. 323 – 343, 2004.

[19] M. Aiolfi and C. A. Favero, “Model uncertainty, thick modeling and the predictability of stock returns,” *Journal of Forecasting*, vol. 24, no. 4, pp. 233 – 254, 2005.

[20] B. Gregorutti, B. Michel, and P. Saint-Pierre, “Correlation and variable importance in random forests,” *Statistics and Computing*, vol. 27, no. 3, pp. 659 – 678, 2017.

[21] A. Fisher, C. Rudin, and F. Dominici, “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously,” *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1 – 81, 2019.

[22] B. A. Taylor and R. J. Tibshirani, *genlasso: Path Algorithm for Generalized Lasso Problems*, 2020. R package version 1.5.

[23] T. Therneau and B. Atkinson, *rpart: Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

[24] S. Milborrow, *rpart.plot: Plot ’rpart’ Models: An Enhanced Version of ’plot.rpart’*, 2020. R package version 3.0.9.

[25] P. Radchenko, A. L. Vasnev, and W. Wang, “Too similar to combine? On negative weights in forecast combination.,” *working paper*, 2020.