Non-Linear Hybrid Shrinkage of Weights for Forecast Selection and Combination

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

Non-Linear Hybrid Shrinkage of Weights for Forecast Selection and Combination

Felix Schulz[1]

[1]Catholic University of Eichstätt-Ingolstadt, Ingolstadt School of Management, Ingolstadt, Germany
{Felix.Schulz@ku.de}

Abstract. *We introduce Non-Linear Hybrid Shrinkage (NLHS) as a holistic model for forecast combination, shrinkage and selection. NLHS first determines the selection of forecasters based on information criteria such as forward feature selection and stores the selection status of forecasters in a selection vector. Depending on the selection status, the estimated optimal weights of the forecasters are either shrunk to zero or equal weights by the least absolute shrinkage and selection operator (LASSO). Among benchmark algorithms such as simple average, optimal weights, or linear and LASSO-based shrinkage methods, NLHS is superior for a larger number of forecasters, as shown in simulation-based experiments.*

Keywords: forecast combination, shrinkage, forward feature selection

1 Introduction

Forecasts are essential for economic life (e.g., [1–3]). Accordingly, myriad forecasting methods have been developed, with research showing that combinations of forecasts steadily improve accuracy over pure statistical or machine learning methods, as confirmed in representative forecasting competitions [4–6].

In a combination of individual forecasts, the goal is to learn weights to minimize the out-of-sample Mean Squared Error (MSE). The most common weighting schemes here are the simple average (*SA*) of available forecasts, and the least squares solution that determines so-called optimal weights (*OW*), i.e., weights that would have minimized the MSE on observed errors (e.g., [7–10]). Weighting-based forecast combination is a type of ensemble learning, namely stacking, where the outcomes of models are used as inputs to other models [11]. In combination, typically errors observed out-of-sample with individual prediction models are used as input for the combination model to learn weights based on their error covariance relationships.

Statistically, *OW* and *SA* are of special interest, as the former aims to reduce the inherent error due to model assumptions (bias), while the latter reduces the sample-based error (variance) [12]. To mitigate the trade-off between the high bias of *SA* and the high variance of *OW*, some papers have been published that shrink *OW* proportionally to *SA* with promising results (e.g., [13, 14]). In contrast, other studies show that forecaster performance can be highly persistent, supporting the idea of selecting and combining only a few forecasters from a group of forecasters [15]. Many studies confirm this approach

17th International Conference on Wirtschaftsinformatik
February 2022, Nürnberg, Germany

by showing scenarios in which a group of forecasters is superior to the combination of the whole group in terms of forecast accuracy (e.g., [16–18]).

The main gap in the forecast combination literature is the combination of optimal weighting, required shrinkage, and potential selection in a holistic model. Diebold and Shin [19] have taken a first step to address this gap with their so-called partially-egalitarian LASSO (*peLASSO*), a two-stage model that first selects forecasts based on *LASSO*, learns *OW* for the survivors, and shrinks the survivors' *OW*s toward *SA*. Although inspired by this approach, we believe that selection via *LASSO* may have drawbacks, since in a situation with high correlation among forecasters, which is common in practice, the selection process may be random [20]. Further, the complete exclusion of forecasters may lead to accuracy losses, since the exclusion could, for instance, neglect interaction and suppressor effects between forecasts. Research has already shown, that even the inclusion of weaker forecasts can lead to an improvement in accuracy [21]. Therefore, we propose Non-Linear Hybrid Shrinkage (*NLHS*). *NLHS* is also a forecast combination and selection method, with the difference that we use information criteria from statistical learning theory to initially select forecasts, while selected forecasts are not entirely removed from the selection, but their in-sample *OW* are used as starting points for shrinkage. Depending on the selection status of the forecasters, the forecasters are either shrunk to equality or zero.

2 Forecast combination, shrinkage and selection

Let f_{it} be a vector of forecasts of the i-th forecast model from $i = 1, ..., k$ competing unbiased forecasts for the period $t = 1, ..., n$ and e_{it} be the associated forecast error, calculated as the difference between the actual outcome of the forecast event y_t and f_{it}. A linear combination of forecasts can be derived by fw whereby f is the $n \times k$ forecast matrix and w the $k \times 1$ weight vector for the forecasts with weights following the properties of $w \in R^k$ and $\sum_{i=1}^{k} w_i = 1$. *OW*, $\hat{w}^o = (\hat{w}_1^o, \ldots, \hat{w}_k^o)$, can be estimated by simply linearly regressing the error of the k-th forecast e_{kt}, in the training sample on the differences to the other errors, as shown in (1) [22, 23].

$$e_{kt} = \hat{w}_1^o(e_{kt} - e_{1t}) + ... + \hat{w}_{k-1}^o(e_{kt} - e_{k-1,t}) \tag{1}$$

As weights sum up to one, $\hat{w}_k^o$ is then computed as $1 - \sum_{i=1}^{k-1} \hat{w}_i^o$. The derived weights from (1) are thereby coined optimal as they aim to minimize the in-sample squared error (bias). In contrast to *OW*, predefined weight vectors like the *SA*, $w^s = \frac{1}{k}\iota$, minimize the sampling-based variance in combined predictions with ι as a column vector of k ones. The combination of both approaches can be expressed in the weight vector $\hat{w}^\lambda$ as illustrated in (2) and proposed in [9]. Thereby, λ is used as a shrinkage parameter to trade-off the bias of *OW* and the variance of *SA* whereby $\lambda = 0$ corresponds to *OW* and $\lambda = 1$ to *SA*.

$$\hat{w}^\lambda = \lambda w^s + (1 - \lambda)\,\hat{w}^o \tag{2}$$

The union of forecast combination, selection and shrinkage is proposed in the method *peLASSO* [19]. In its two-step implementation, the best set of forecasts is first selected using standard *LASSO* as shown in (3).

$$\hat{w}^{LASSO} = \arg\min_{w}(\sum_{t=1}^{n}(\sum_{i=1}^{k} w_i e_{it})^2 + \lambda \sum_{i=1}^{k} |w_i|) \tag{3}$$

In step 2, estimated combining weights of k' selected forecasts are either set directly to $\frac{1}{k'}$, called *peLASSO (Avg.)*, or shrunk towards $\frac{1}{k'}$ using *OW* of k' survivors as baseline. Shrinkage using a L1 regularization, called *peLASSO (eLASSO)*, is shown in (4), while alternatively a L2 regularization term can be used, called *peLASSO (eRidge)*.

$$\hat{w}^{eLASSO} = \arg\min_{w}(\sum_{t=1}^{n}(\sum_{i=1}^{k'} w_i e_{it})^2 + \lambda \sum_{i=1}^{k'} |w_i - \frac{1}{k'}|) \qquad s.t. \sum_{i=1}^{k'} w_i = 1 \tag{4}$$

Please note, that (4) can also be used as an alternative to (2) to non-linearly shrink all k forecasters towards equality.

3 Non-Linear Hybrid Shrinkage

We present Non-Linear Hybrid Shrinkage (*NLHS*) for forecast combination and selection as a solution to an optimization problem that uses L1 regularization with the purpose of shrinking the estimated weights w_i to a predefined vector w_i^{sel}, as shown in (5).

$$\hat{w}^{NLHS} = \arg\min_{w}(\sum_{t=1}^{n}(\sum_{i=1}^{k} w_i e_{it})^2 + \lambda \sum_{i=1}^{k} |w_i - w_i^{sel}|) \qquad s.t. \sum_{i=1}^{k} w_i = 1 \tag{5}$$

Thereby, w_i^{sel} serves as a selection vector and equals $\frac{v}{\iota' v}$ with v as a vector of size k and ι as a column vector of k ones. Each element in v is representing a forecaster i and is assigning a value of one if the particular i-th forecaster is selected by an information criterion, and zero otherwise. Since $\iota' v$ computes the vector sum and thus the number of k' selected forecasters, w_i^{sel} contains either the value $\frac{1}{k'}$ or zero for each forecaster i depending on its selection status. *NLHS* has accordingly two hyperparameters that can be set by cross-validation. First, the vector w_i^{sel}, or v, which is responsible for the selection of forecasters, and λ, which determines the required shrinkage. For the latter, $\lambda = 0$ corresponds to *OW*, while with increased λ deviations from the selection vector w_i^{sel} are penalized, pushing the weight of forecaster i non-linearly either to $\frac{1}{k'}$ or zero.

For forecast selection, we propose two strategies based on information criteria to rank the forecasters and select the forecasters that achieve the highest gains for the combination in terms of accuracy improvement.

First, we propose selection based on forecaster performance, e.g. based on the forecasters' in-sample MSE-values or other loss functions [24]. For this purpose, the forecasters are ranked according to their MSE. Next, forecasters are selected incrementally, first selecting the best performing forecaster on the training data, then the two best, and so on. This results in k different forms of v to be tested using cross-validation in (5).

Second, we present selection based on forward feature selection for generating the vector v. Forward feature selection (aka stepwise regression) is a well-known method from classification, in which variables are selected from a set of variables [25]. In the first iteration, the MSEs of all forecasters are evaluated and the forecaster with the lowest MSE on the training data is selected. The best forecaster is assigned the value one in v, the others are zeroed. Based on the first selection, in the second iteration *OW* combinations are formed with the remaining forecasters and evaluated on in-sample data. The remaining forecaster with the largest possible improvement in in-sample forecast accuracy is added to generate the next shape of v. The last steps are performed until all k forecasters are included in the selection. As our first selection strategy ignores any correlations between forecasts, but focuses solely on the performance aspect of forecasters, we expect the second method to have an advantage as interaction effects between forecasters are considered, leading to a more effective consideration of, e.g., balancing effects between forecasters' errors.

4 Experimental Evaluation

For evaluating *NLHS*, we simulate different scenarios by setting a variety of parameters as follows: number of forecasters with $k \in \{8,\ 12,\ 16\}$, size of training data with $n \in \{20,\ 30, ...,\ 70\}$, pairwise correlations between forecasters with $\rho \in \{0.3,\ 0.6,\ 0.9\}$ and forecaster error variances of $\sigma_p^2 \in \{2,\ 4\}$. Thereby, the variance of the p-th forecast is decreasing geometrically by $(\sigma_p^2)^{\frac{i-1}{k-1}}$ for the i-th forecast, with $i \in \{1, ...,\ k-1\}$. To measure the asymptotic performance of the learned weights to the actual optimal weights of the population, a test set size of 5,000 data points was chosen. As commonly assumed, our simulation errors follow a multivariate normal distribution with mean zero. Hence, the error variance of an individual forecast is the forecast's MSE. In total, 108 treatment combinations are tested, with each scenario repeated ten times to increase robustness. Besides *NLHS* using forecaster performance (*NLHS Performance*) or forward feature selection (*NLHS FFS*) for selection, the benchmark methods *OW*, *SA*, *Linear Shrinkage* as proposed in (2), non-linear shrinkage via *eLASSO* without prior selection of forecasters, and *peLASSO (Avg.)* and *peLASSO (eRidge)* are implemented.

Before presenting preliminary results, an example shrinkage path of *NLHS* in a scenario with $k = 8$, $n = 40$, $\rho = 0.9$, and $\sigma_p^2 = 4$ is shown in Figure 1.

The weights at a shrinkage level of 0% correspond to *OW*, while with increased shrinkage the weights are non-linearly shifted to the selection vector w_i^{sel}. Since only the first forecaster *FC1* (in black) was selected in the cross-validation based procedure, at a shrinkage level of 100%, the weight of *FC1* equals one, while the others (in gray) are shrunk out of the selection, i.e., correspond to a weight of zero.

5 Preliminary Results

Preliminary results are provided in Table 1. The table includes, for all scenarios by method and by number of forecasters, the average test MSE, the optimal lambda $\lambda_{opt}^{\%}$

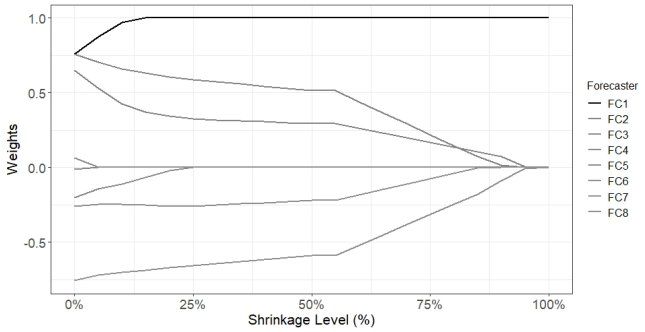

Figure 1. Example shrinkage path of weights in *NLHS*

relative to the λ required to approach full shrinkage, and the number of forecasters selected k'.

Table 1. Preliminary results per method split by number of forecasters

Method	k = 8			k = 12			k = 16		
	MSE	$\lambda^{\%}_{opt}$	k'	MSE	$\lambda^{\%}_{opt}$	k'	MSE	$\lambda^{\%}_{opt}$	k'
OW	0.774		8	0.797		12	1.100		16
SA	1.150		8	1.120		12	1.110		16
Linear Shrinkage	0.733	40.9%	8	0.701	47.9%	12	0.722	54.4%	16
eLASSO	0.738	25.7%	8	0.700	29.0%	12	0.676	22.6%	16
NLHS Performance	0.738	84.4%	3.5	0.663	80.9%	4.2	0.629	79.8%	5.7
NLHS FFS	0.752	82.7%	3.5	0.701	78.5%	3.8	0.667	79.1%	4.6
peLASSO (Avg.)	1.130	8.9%	5.6	1.060	8.3%	7.6	1.040	9.8%	8.8
peLASSO (eRidge)	0.731	18.3%	5.6	0.687	41.1%	7.6	0.683	16.4%	8.8

Preliminary results show interesting findings. Among them, selection and shrinkage approaches show better results than *OW* and *SA*, especially with an increased number of forecasts, as can be seen from the growing distances between the MSE of selection and shrinkage approaches and the MSE of *OW*. That is because of the limited training data leading to overfitted *OW*. In contrast, shrinkage approaches like *Linear Shrinkage* and *eLASSO* can adapt *OW* by pushing them toward the *SA* and improve MSE. The combination of selection and shrinkage shows further benefits as in *NLHS*-based methods, where the MSE even decreases with increasing k, leading to the two best methods at $k = 16$ with MSE values of 0.629 and 0.667, respectively. Comparing selection and shrinkage approaches, the *NLHS*-based methods select on average half as many forecasters as the *peLASSO* methods, yielding slight improvements in forecast accuracy.

6 Conclusion and future work

This paper presented a hybrid model for combining and selecting forecasters. First results are promising and motivate a deeper analysis of the results, the inclusion of additional scenarios, and investigation of further selection criteria like backward elimination.

References

1. Newbold, P., Bos, T.: Introductory business economic forecasting. South-Western Publ., 2 edn. (1994)
2. Hanke, J.E., Wichern, D.W.: Business Forecasting. Pearson, 9 edn. (2009)
3. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts, 2 edn. (2018)
4. Makridakis, S., Hibon, M.: The M3-competition: results, conclusions and implications. International Journal of Forecasting 16(4), 451–476 (2000)
5. Andrawis, R., Atiya, A., El-Shishiny, H.: Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. International Journal of Forecasting 27(3), 672–688 (2011)
6. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36(1), 54–74 (2020)
7. Bates, J., Granger, C.: The Combination of Forecasts. Journal of the Operational Research Society 20(4), 451 – 468 (1969)
8. Clemen, R.: Combining Forecasts: A Review and Annotated Bibliography. International Journal of Forecasting 5(4), 559–583 (1989)
9. Stock, J., Watson, M.: Combination Forecasts of Output Growth in a Seven-Country Data Set. Journal of Forecasting 23(6), 405–430 (2004)
10. Genre, V., Kenny, G., Meyler, A., Timmermann, A.: Combining Expert Forecasts: Can Anything Beat the Simple Average? International Journal of Forecasting 29(1), 108–121 (2013)
11. Wolpert, D.: Stacked generalization. Neural networks 5(2), 241–259 (1992)
12. Blanc, S., Setzer, T.: Bias–Variance Trade-Off and Shrinkage of Weights in Forecast Combination. Management Science 66(12), 5720–5737 (2020)
13. Diebold, F., Pauly, P.: The use of prior information in forecast combination. International Journal of Forecasting 6(4), 503–508 (1990)
14. Blanc, S., Setzer, T.: When to choose the simple average in forecast combination. Journal of Business Research 69(10), 3951–3962 (2016)
15. Aiolfi, A., Timmermann, A.: Persistence in Forecasting Performance and Conditional Combination Strategies. Journal of Econometrics 135, 31–53 (2006)
16. Mannes, A., Soll, J., Larrick, R.: The Wisdom of Select Crowds. Journal of Personality and Social Psychology 209, 276–299 (2014)
17. Budescu, D., Chen, E.: Identifying Expertise to Extract the Wisdom of Crowds. Management Science 61(2), 267 – 280 (2015)
18. Kourentzes, N., Barrow, D., Petropoulos, F.: Another look at forecast selection and combination: Evidence from forecast pooling. International Journal of Production Economics 107(2), 226–235 (2019)
19. Diebold, F., Shin, M.: Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. International Journal of Forecasting 35(4), 1679–1691 (2019)
20. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67(2), 301–320 (2005)
21. Geweke, J., Amisano, G.: Optimal prediction pools. Journal of Econometrics 164(1), 130–141 (2011)
22. Granger, C., Ramanathan, R.: Improved Methods of Combining Forecasts. Journal of Forecasting 3(2), 197–204 (1984)
23. Timmermann, A.: Forecast Combinations. Handbook of Economic Forecasting 1, 135–196 (2006)

24. Schulz, F., Setzer, T., Balla, N.: Linear Hybrid Shrinkage of Weights for Forecast Selection and Combination. Proceedings of the 55th Hawaii International Conference on System Sciences (2021), forthcoming
25. Kumar, V., Minz, S.: Feature selection: a literature review. SmartCR 4(3), 211–229 (2014)