Analyzing and Correcting Biased Machine Learning-Based Tuning of Weight Shrinkage in Forecast Combination

Veronika Wachslander (✉)[0009−0009−0440−6749]

Catholic University of Eichstaett-Ingolstadt, Ingolstadt, Germany
veronika.wachslander@ku.de

Abstract. A forecast combination typically corresponds to a weighted average of individual forecasts and aims at increasing predictive accuracy. Application fields include business, economics, information systems such as recommender systems and financial portfolios. One popular weighting approach used in various studies is to learn weights optimal on past data (optimal weights) and shrink them towards equal weights to mitigate overfitting. The required shrinkage hyperparameter is usually tuned by machine learning-based techniques like K-fold cross-validation (CV). This paper shows that CV-tuned shrinkage levels are generally biased: Depending on the characteristics (parameters) of training forecast data (e.g. number of forecasters, error correlations, spread in predictive ability, training set size, number of CV-folds), such approaches lead to systematic over- or undershrinkage. The impact of different parameters on these biases is studied on large sets of synthetically generated data and a model is trained to predict the bias (direction and degree) by using data characteristics as features. This model is evaluated for its ability to correct biases on various sets of synthetic data, where the corrected weights lead to improved predictive accuracy across a range of data characteristics. Codes are available at https://github.com/VeronikaWachslander/shrinkage-tuning-bias.

Keywords: Weight Shrinkage · Hyperparameter Tuning · Debiasing · Forecast Combination · Machine Learning.

1 Introduction

The combination of forecasts provided by different models or humans is a technique used in business, economics and other fields to generate more accurate and reliable predictions. Typical applications of forecast combinations are predictions of economic growth, inflation rates or electricity demand [12, 14, 23, 25].

Besides business contexts, the approaches can be applied to hybrid recommender systems (information systems that combine e.g. different purchase or movie recommendations) or used for financial portfolio optimization [17, 18, 20].

While various disciplines conduct research regarding combination approaches and numerous methods already exist (see [11, 28]), there is still no generalizable cross-domain suggestion how to determine the combination weights.

Two weighting schemes often considered as benchmarks or bases for more sophisticated approaches are optimal weights (OW) – introduced in Bates and Granger [4] for two forecasters and later extended (see, e.g., [27]) – and equal weights (EW). The OW are estimated on past forecast errors (training set), leading to weights that minimize the mean squared error (MSE) on this dataset.

However, numerous studies [2, 14, 15] show that, on unseen data, this and other more sophisticated weighting schemes are mostly outperformed by the simple average that assigns EW to all forecasters. This superiority of EW, named *forecast combination puzzle* by Stock and Watson [25], is typically explained by the consideration of random variations in training data when estimating OW (see, e.g., [5, 9, 10]). These structures do not necessarily exist on unseen data, so the OW overfit the training set – especially in the case of small training sizes. In contrast, EW completely ignore potential differences in forecast ability.

Therefore, OW and EW can be seen as two opposite approaches and a compromise between these weighting schemes could be beneficial. In various studies [1, 12, 25], this is achieved by shrinking OW towards EW controlled by a shrinkage parameter. However, there is no rule how to derive the optimal shrinkage level, i.e. the one resulting in the minimum combination error on unseen data.

As the shrinkage level can be considered a hyperparameter, cross-validation (CV) can be used for the tuning as applied by Schulz and Setzer [22], Schulz et al. [21], as well as Diebold and Shin [13]. However, Schulz et al. detected slight deviations between the CV-optimal and the truly optimal shrinkage, while Diebold and Shin recommend to combine only a few forecasters and assign EW.

According to Schulz and Setzer [22], these findings might be (at least partly) due to the determination of inappropriate, typically too high shrinkage levels with CV – in particular with small training sets and low to medium error correlations among the forecasters (with moderate differences in predictive ability).

This paper addresses the phenomenon of shrinkage biases when the level of shrinkage from OW towards EW is determined by CV, with all forecasters being included in the combination. Since determining weights is particularly challenging if available (past forecast) data is very limited, which is usually the case in business and economics, the paper mainly focuses on small datasets.

First, the impact of various data- and CV-related properties on shrinkage biases is analyzed in detail. Second, a regression tree is developed that predicts shrinkage biases and serves as model to correct CV-determined shrinkage levels. Third, this correction model is evaluated and discussed.

In addition, this paper aims to raise critical awareness of using CV for hyperparameter tuning in any machine learning task and to encourage reviews of the tuning results.

The paper is structured as follows: Section 2 introduces weight shrinkage and Section 3 CV-based tuning along with notation. In Section 4, the experimental design for data generation and tuning as well as evaluation procedures are explained, while Section 5 provides analytical insights. Section 6 presents a model to predict and correct tuning biases, which is evaluated in Section 7. Finally, Section 8 draws conclusions and presents an outlook on future work.

2 Shrinking Optimal to Equal Combination Weights

Assume $J \in \mathbb{N}, J > 1$ forecasters are available and f_{ij} denotes the i-th forecast of the j-th forecaster, with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, J\}$. A combined forecast for the i-th observation x_i can be calculated as $\sum_{j=1}^{J} w_j \cdot f_{ij}$, i.e. by assigning a weight $w_j \in \mathbb{R}$ to forecaster j for all $j \in \{1, \ldots, J\}$, multiplying the weight by this forecaster's prediction for x_i and adding up these weighted predictions.

One question is which weighting scheme to use, with $\sum_{j=1}^{J} w_j = 1$ regardless of the specific choice. The simplest weighting technique assigns EW to all forecasters by setting $w_j = J^{-1}$ for every $j \in \{1, \ldots, J\}$. The corresponding weight vector $\boldsymbol{w}^{EW} \in \mathbb{R}^J$ can be defined as $\boldsymbol{w}^{EW} = J^{-1} \cdot \mathbf{1}$, with the column vector $\mathbf{1} \in \mathbb{R}^J$ containing one in each entry and $\mathbf{1}' \cdot \boldsymbol{w}^{EW} = 1$ as required.

Another common approach is the calculation of OW based on a matrix $\boldsymbol{E} \in \mathbb{R}^{n \times J}$ with n past errors per forecaster j. This means, each entry $e_{ij} \in \boldsymbol{E}$ represents the difference between the actual value of the observation x_i and its prediction provided by the forecaster j, calculated as $e_{ij} = x_i - f_{ij}$.

For each forecaster, efficient forecasts are assumed, i.e. multivariate normally distributed forecast errors with a mean of zero and an error covariance matrix $\boldsymbol{\Sigma}_e$, which is typically unknown and estimated as $\hat{\boldsymbol{\Sigma}}_e = n^{-1} \cdot \boldsymbol{E}'\boldsymbol{E}$.

The vector $\hat{\boldsymbol{w}}^{OW} \in \mathbb{R}^J$ contains the OW and is estimated as shown in (1), with $\hat{\boldsymbol{\Sigma}}_e^{-1}$ denoting the inverse of $\hat{\boldsymbol{\Sigma}}_e$, and ensuring $\mathbf{1}' \cdot \hat{\boldsymbol{w}}^{OW} = 1$ (see [27]).

$$\hat{\boldsymbol{w}}^{OW} = \frac{\hat{\boldsymbol{\Sigma}}_e^{-1}\mathbf{1}}{\mathbf{1}'\hat{\boldsymbol{\Sigma}}_e^{-1}\mathbf{1}} \tag{1}$$

As discussed, OW are typically prone to overfitting data structures and thus, do not fit well unseen data, while EW consider all forecasts to be equally reliable. To obtain less overfitted weights that still consider differences in forecast ability, OW are shrunk towards EW by a shrinkage parameter $0 \leq \lambda \leq 1$, with $\lambda = 0$ resulting in OW and $\lambda = 1$ in EW (100% shrinkage), as shown in (2) (see [22]).

$$\hat{\boldsymbol{w}}^{\lambda} = (1 - \lambda) \cdot \hat{\boldsymbol{w}}^{OW} + \lambda \cdot \boldsymbol{w}^{EW} \tag{2}$$

The optimal shrinkage level λ^* leads to weights $\hat{\boldsymbol{w}}^{\lambda^*} = (\hat{w}_1^{\lambda^*}, \ldots, \hat{w}_J^{\lambda^*})$ that minimize the MSE (shown in (3)) on unseen data $\{x_1, \ldots, x_n\}$.

$$MSE(\hat{\boldsymbol{w}}^{\lambda}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \Big(x_i - \sum_{j=1}^{J} \hat{w}_j^{\lambda} \cdot f_{ij} \Big)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \Big(\sum_{j=1}^{J} \hat{w}_j^{\lambda} \cdot e_{ij} \Big)^2 \tag{3}$$

Typically, the CV-optimal shrinkage λ_{CV}^* will deviate from the truly optimal shrinkage λ_{true}^*. In the following, this deviation (in percentage points $\%P$) is called *shrinkage bias* and calculated as provided in (4). A bias $B(\lambda_{CV}^*) > 0$ corresponds to overshrinkage, while $B(\lambda_{CV}^*) < 0$ means undershrinkage.[1]

$$B(\lambda_{CV}^*) = \lambda_{CV}^* - \lambda_{true}^* \tag{4}$$

The next section explains cross-validation and its use for shrinkage tuning.

[1] E.g., if $\lambda_{CV}^* = 0.25 = 25\%$ and $\lambda_{true}^* = 0.13 = 13\%$, then $B(\lambda_{CV}^*) = 0.12 = 12\%P$.

3 Cross-Validation and Shrinkage Tuning

A frequently used resampling technique to both estimate the performance of a model on unseen data and tune hyperparameters is K-fold cross-validation, which randomly divides the training data into K (almost) equally sized, pairwise disjoint subsets (called *folds*). The model is trained on the calibration set consisting of $K-1$ folds, and tested on the remaining fold that serves as validation set. This is repeated until each fold has been part of the calibration set $K-1$ times and represented the validation set once (for more details see, e.g., [16, 19]).

Typically, the errors (with respect to the chosen error measure) on the validation sets are averaged over the K iterations to estimate the overall performance on unseen data. This can be done for different hyperparameter values and the one resulting in the lowest overall error is selected for the final model.

Different variants of K-fold CV are distinguished, depending on the value of K: The number of folds is either set to a value like the frequently recommended ones $K = 5$ or $K = 10$ (see [3, 7]), or the number of observations n belonging to each fold is fixed as with Leave-One-Out (LOO) and Leave-Two-Out (LTO) CV – i.e., the training set is split into n folds for LOO and $n/2$ folds for LTO.[2]

However, there is no general rule to choose the number of folds K, as there is a bias–variance trade-off regarding this decision [16, 19]; rather, Zhang and Yang [29] explain that the specific task the CV is used for should be considered.

In the context of weight shrinkage tuning, OW are estimated on the calibration sets and shrunk towards EW on the respective validation sets. Finally, the shrinkage level with the lowest average MSE on these validation sets is chosen.

For a low number of folds, the calibration sets typically differ strongly from each other and also from the full training set. The OW estimated on a calibration set will overfit its structure due to significantly lower data amounts compared to the full training set and thus, will be quite different from the ones learned on the other calibration sets and on the full training set. Since overfitted OW will not reflect the structure of the corresponding validation set well, these will be shrunk strongly, resulting in weights near EW that can be highly biased.

In contrast, for LOO, the calibration sets and thus, the estimations for OW will be similar and also close to the full training set and its estimated OW. However, each validation set contains only one observation, which determines the respective MSE values, so the results could have high variance.

In summary, the number of folds can severely affect the accuracy of combined forecasts, as the estimation of OW and the shrinkage determination are sensitive to the underlying dataset. For this reason, datasets with varying characteristics are generated and the shrinkage biases are studied for different numbers of folds K, which leads to various scenarios (i.e., parameter constellations).

The next section describes the data generation and the procedures for analyzing shrinkage biases and evaluating the later introduced shrinkage correction.

[2] Unlike the typically exhaustive Leave-p-out CV that creates folds for all $\binom{n}{p}$ combinations of the training observations as for example described in [8], the term LTO is used in this paper to describe one random division of the training set into $n/2$ folds.

4 Experimental Design

4.1 Data Generation

For analyzing, predicting and correcting shrinkage biases, synthetic error datasets are generated, as this allows to identify general relations and characteristics without uncontrollable random effects (and is, e.g., also done in [10, 21, 24, 26]).

The synthetic errors are drawn from different multivariate normal distributions with mean zero and covariance matrices that are calculated using predefined forecasters' variances and fixed values for the pairwise error correlation among the forecasters.

Various data samples are generated with varying numbers of forecasters, different pairwise error correlations, and different error variance structures to allow for comprehensive analyses. Each data sample contains 20,000 error observations for each of the $J \in \{5, 8, 10, 12, 15\}$ forecasters.

The pairwise error correlations are either set to $\rho \in \{0.1, 0.2, \ldots, 0.9\}$, identical for all pairs, or identical within two nearly equally sized groups, but slightly different between the groups.[3] Therefore, the range of correlations $\Delta\rho$ (difference between maximum and minimum correlation) takes the values $\Delta\rho \in \{0, 0.2\}$.

Further, the error variances of the forecasters increase from $\sigma_1^2 = 1$ to $\sigma_J^2 \in \{1.2, 1.5, 2, 4, 9\}$, either linearly with $\sigma_j^2 = \sigma_{j-1}^2 + \frac{\sigma_J^2 - \sigma_1^2}{J-1}$ or in a quadratic fashion with $\sigma_j^2 = \left(\sigma_{j-1} + \frac{\sigma_J - \sigma_1}{J-1}\right)^2$ for $j \in \{2, \ldots, J\}$. In addition, $\Delta\sigma^2 = \sigma_J^2 - \sigma_1^2$ represents the range of variances and thus, $\Delta\sigma^2 \in \{0.2, 0.5, 1, 3, 8\}$.

For analyzing the shrinkage bias on limited data, small training sets $\boldsymbol{E}_{train}$, with $n \in \{10, 20, \ldots, 100, 125, 150, 175, 200\}$ error observations per forecaster, are randomly drawn from the generated samples. The respective observations, that are not part of $\boldsymbol{E}_{train}$, form a large test set $\boldsymbol{E}_{test}$, which ensures stable parameter values that approach those of the data generation.[4]

$\boldsymbol{E}_{train}$ is used to estimate OW and apply K-fold CV to tune the shrinkage hyperparameter, resulting in the CV-optimal shrinkage level λ_{CV}^*, while $\boldsymbol{E}_{test}$ serves as unseen data to identify the truly optimal shrinkage level λ_{true}^* and to calculate the resulting shrinkage bias $B(\lambda_{CV}^*)$.

The shrinkage tuning biases are studied for $K \in \{2, 5, 10, n/2, n\}$ CV-folds, with $K = n/2$ folds corresponding to LTO and $K = n$ folds to LOO.

Table 1 (shown in Subsection 4.3) summarizes the parameters and values used to generate datasets for the analyses and predictions of shrinkage biases.

A detailed explanation of the procedure to tune the shrinkage level by CV and calculate the resulting shrinkage bias is provided in Algorithm 1.

The procedure is repeated 250 times for each scenario with new, randomly drawn error data in each repetition to obtain reliable measures.

[3] As an example for the second case, $\rho = 0.1$ in one group and $\rho = 0.3$ in the other, while the correlations between the groups receive a value of $\rho = 0.2$.

[4] For estimating OW, a sufficient amount of training observations is required. E.g., for 15 forecasters and 2-fold CV, at least 30 observations per forecaster are needed.

Algorithm 1 Shrinkage Determination by CV and Shrinkage Bias Calculation.

1: ***Initialization:*** Set shrinkage values $\lambda_s = 0.01 \cdot s$ with $s \in \{0, \dots, 100\}$.
Set the number of folds K.
Generate error sample matrices $\boldsymbol{E}_{train}$ and $\boldsymbol{E}_{test}$.
2: Split $\boldsymbol{E}_{train}$ (its rows) into K (almost) equally sized, pairwise disjoint folds.
3: **For** $k = 1, ..., K$ **do:**
– Set fold k as validation $\boldsymbol{E}_{val}^{(k)}$ and $\boldsymbol{E}_{cal}^{(k)} = \boldsymbol{E}_{train} \setminus \boldsymbol{E}_{val}^{(k)}$ as calibration set.
– Estimate $\hat{\boldsymbol{w}}^{OW_{cal}(k)}$ as $\hat{\boldsymbol{w}}^{OW}$ on $\boldsymbol{E}_{cal}^{(k)}$ by (1), with $\boldsymbol{E} = \boldsymbol{E}_{cal}^{(k)}$ for $\hat{\boldsymbol{\Sigma}}_e$.
– **For** $s = 0, \dots, 100$ **do:**
• Calculate $\hat{\boldsymbol{w}}^{\lambda_s(k)}$ using (2) with $\lambda = \lambda_s$ and $\hat{\boldsymbol{w}}^{OW} = \hat{\boldsymbol{w}}^{OW_{cal}(k)}$.
• Apply $\hat{\boldsymbol{w}}^{\lambda_s(k)}$ to $\boldsymbol{E}_{val}^{(k)}$ and calculate the MSE value $MSE_s^{(k)}$ by (3).
End For.
End For.
4: **For** $s = 0, \dots, 100$ **do:**
– Calculate the mean MSE for λ_s, $MSE_s^{[val]} = \frac{1}{K} \cdot \sum_{k=1}^{K} MSE_s^{(k)}$.
End For.
5: Identify the λ_s producing $min(MSE_s^{[val]})$ as CV-optimal shrinkage λ_{CV}^*.
6: Estimate $\hat{\boldsymbol{w}}^{OW_{train}}$ as $\hat{\boldsymbol{w}}^{OW}$ on $\boldsymbol{E}_{train}$ by (1), with $\boldsymbol{E} = \boldsymbol{E}_{train}$ for $\hat{\boldsymbol{\Sigma}}_e$.
7: **For** $s = 0, \dots, 100$ **do:**
– Calculate $\hat{\boldsymbol{w}}^{\lambda_s}$ using (2) with $\lambda = \lambda_s$ and $\hat{\boldsymbol{w}}^{OW} = \hat{\boldsymbol{w}}^{OW_{train}}$.
– Apply $\hat{\boldsymbol{w}}^{\lambda_s}$ to $\boldsymbol{E}_{test}$ and calculate the MSE value $MSE_s^{[test]}$ by (3).
End For.
8: Identify the λ_s producing $min(MSE_s^{[test]})$ as truly optimal shrinkage λ_{true}^*.
9: Calculate the shrinkage bias $B(\lambda_{CV}^*)$ using (4).

The results are compiled into a dataset comprising more than 13 million cases, which is used for the analyses and to develop the correction model.[5]

However, the representation of some parameters is slightly modified and additional parameters are created to show the analytical results and to develop a model for predicting shrinkage biases, as will be explained next.

4.2 Parameter Estimation and Representation

Besides the already discussed variables, there are two more used to predict the shrinkage bias. These are n_{cal} and $n_{cal}^{\%}$, corresponding to the number and share of training observations being part of each calibration set, which depend on the number of folds (e.g., for LOO, the values are $n_{cal} = n-1$ and $n_{cal}^{\%} = (n-1)/n$).

Further, it can be distinguished between observable parameters, which are J, n, K, n_{cal} and $n_{cal}^{\%}$, as their values can be directly observed, and estimable ones. The estimable parameters are $\rho, \Delta\rho$ and $\Delta\sigma^2$, as their values depend on the respective randomly drawn dataset, so these need to be estimated.

While the values of the estimable variables will deviate on the datasets to a negligible extent from those set for the data generation due to the large size,

[5] See the codes at https://github.com/VeronikaWachslander/shrinkage-tuning-bias.

their estimation on limited amounts of training data introduces uncertainty and bears the risk of strong deviations, which will impact the weight determination.

For this reason and to enable an application of the later introduced bias prediction model to datasets with other parameter values than studied here, the estimable variables are modified or binned, i.e. the values are assigned to groups.

The spread in predictive ability is reflected by the range of estimated forecasters' variances $\Delta\sigma^2$ and grouped as shown in (5) based on the studied values.

$$\Delta\sigma^2 = \begin{cases} \textit{tiny} & \Delta\sigma^2 < 0.45 \\ \textit{low} & 0.45 \leq \Delta\sigma^2 < 0.95 \\ \textit{medium} & 0.95 \leq \Delta\sigma^2 < 2.50 \\ \textit{high} & 2.50 \leq \Delta\sigma^2 < 6.00 \\ \textit{extreme} & 6.00 \leq \Delta\sigma^2 \end{cases} \tag{5}$$

In addition, the pairwise error correlations among the forecasters are estimated, with $\Delta\rho$ corresponding to their interquartile range and ρ to their mean, which is assigned to one of the categories shown in (6).

$$\rho = \begin{cases} \textit{weak} & \rho < 0.25 \\ \textit{moderate} & 0.25 \leq \rho < 0.55 \\ \textit{strong} & 0.55 \leq \rho < 0.75 \\ \textit{extreme} & 0.75 \leq \rho \end{cases} \tag{6}$$

Since the bias prediction model will also be assessed regarding its ability to correct shrinkage biases, the next subsection describes the evaluation.

4.3 Evaluation Setting

The bias predictions of the later introduced model can be treated as shrinkage correction factors CV_C, and the corrected shrinkage levels $\lambda^*_{CV_C}$ can be expected to improve the weight determination for known or precisely estimated data characteristics and parameter values.

However, as the values of the estimable variables are typically not known and to be estimated on limited training data, the model is evaluated for its ability to correct shrinkage biases in case of uncertainties in parameter estimation.

For the evaluation, new synthetic datasets are generated, with the parameter values provided in Table 1 and including scenarios, which are not part of the database used to learn the model to check for a more general validity.

The application and evaluation of the shrinkage correction are formally described in Algorithm 2, which is a continuation of Algorithm 1.

The shrinkage correction is repeated 250 times for each scenario with new, randomly drawn error data in each repetition, so the evaluation database contains $720,000$ cases.

The next section discusses shrinkage biases regarding different parameters.

Table 1. Parameters and Values for Bias Analyses and Evaluation of Correction.

Parameter	Description	Analyzed Values	Evaluated Values
n	Size of Training Set	$10, 20, \ldots, 100, 125, \ldots, 200$	$25, 50, 100, 200$
J	Number of Forecasters	$5, 8, 10, 12, 15$	$4, 9$
K	Number of CV-Folds	$2, 5, 10, LTO, LOO$	$2, 5, LOO$
ρ	Pairwise Correlation	$0.1, 0.2, \ldots, 0.8, 0.9$	$0.15, 0.3, \ldots, 0.75, 0.9$
$\Delta\sigma^2$	Range of Variances	$0.2, 0.5, 1, 3, 8$	$0.1, 0.25, 0.7, 2, 5, 7$
$\Delta\rho$	Range of Correlations	$0, 0.2$	$0, 0.3$

Algorithm 2 Application and Evaluation of Shrinkage Correction.

10: Calculate $\hat{\boldsymbol{w}}^{\lambda^*_{CV}}$ using (2) with $\lambda = \lambda^*_{CV}$ and $\hat{\boldsymbol{w}}^{OW} = \hat{\boldsymbol{w}}^{OW_{train}}$.
11: Estimate $\rho, \Delta\rho$ and $\Delta\sigma^2$ on the provided training set $\boldsymbol{E}_{train}$.
12: Apply the bias prediction model to determine the correction factor CV_C.
13: Calculate the corrected shrinkage level $\lambda^*_{CV_C} = \lambda^*_{CV} - CV_C$.
14: Calculate $\hat{\boldsymbol{w}}^{\lambda^*_{CV_C}}$ using (2) with $\lambda = \lambda^*_{CV_C}$ and $\hat{\boldsymbol{w}}^{OW} = \hat{\boldsymbol{w}}^{OW_{train}}$.
15: Calculate $MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV}})$ and $MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV_C}})$ on $\boldsymbol{E}_{test}$ using (3).
16: Calculate the percentage MSE deviation by $\frac{MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV_C}}) - MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV}})}{MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV}})}$.
(*Example:* If $MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV_C}}) = 0.40$ and $MSE(\hat{\boldsymbol{w}}^{\lambda^*_{CV}}) = 0.50$, the percentage deviation is $-0.20 = -20\%$, which corresponds to a MSE reduction of 20%.)

5 Analytical Insights into Weight Shrinkage Biases

This section provides analytical insights into shrinkage biases, as these will vary depending on the parameter values of the datasets and the number of CV-folds.

5.1 Impact of Variables on Shrinkage Biases

Table 2 provides shrinkage biases regarding the number of CV-folds K and forecasters J, their variance range $\Delta\sigma^2$ and pairwise correlations ρ with range $\Delta\rho$.

The values represent the shrinkage bias (written in black if positive and in gray if negative), averaged over the different training sizes n and 250 repetitions per scenario, with a darker cell background reflecting a stronger bias.

As an example, the mean shrinkage bias for datasets with $\Delta\sigma^2 = medium$, $\rho = weak$ and $\Delta\rho < 0.1$ equals $B(\lambda^*_{CV}) = 18.25(\%P)$ when applying 2-fold CV, so the resulting shrinkage λ^*_{CV} is on average 18.25 percentage points higher than the truly optimal shrinkage λ^*_{true}.

At first glance, most scenarios suffer from overshrinkage (i.e. $B(\lambda^*_{CV}) > 0$), while undershrinkage appears only for boundary values of $\Delta\sigma^2$ and ρ: The majority of scenarios with $B(\lambda^*_{CV}) < 0$ shows little spread in forecast ability along with identical, rather weaker pairwise correlations and will be studied in Subsection 5.2, while negligible undershrinkage can be observed for scenarios belonging to the category *extreme* for both $\Delta\sigma^2$ and ρ (or $\rho = strong$) regardless of $\Delta\rho$.

Table 2. Mean Shrinkage Bias (in $\%P$) for Selected Values of Forecasters J, Folds K, Variance Range $\Delta\sigma^2$ and Correlation ρ, Differentiated by Range of Correlations $\Delta\rho$.

		$\Delta\rho < 0.1$						$\Delta\rho \geq 0.1$					
$\Delta\sigma^2$	J	5			15			5			15		
	K / ρ	2	5	LOO	2	5	LOO	2	5	LOO	2	5	LOO
tiny	*extreme, strong, moderate, weak*	-6.62	-5.58	-4.88	-4.71	-6.94	-7.70	14.20	11.62	11.01	21.54	8.65	3.88
		-3.53	-2.75	-2.11	-2.34	-5.47	-6.38	15.01	12.04	11.13	21.25	8.87	4.15
		2.82	2.43	3.15	2.03	-1.77	-3.39	17.99	13.73	12.31	22.03	9.16	3.98
		12.50	10.56	9.79	12.21	4.52	1.72	18.83	13.13	10.59	22.28	8.71	3.58
low		8.25	7.32	7.05	1.85	-1.95	-3.34	20.20	15.27	13.79	22.29	8.53	3.64
		13.01	11.00	10.19	8.78	2.32	-0.02	19.64	13.84	11.75	23.30	9.56	4.37
		18.67	13.96	12.70	16.81	7.16	3.15	19.70	12.93	10.49	22.73	8.88	3.62
		18.14	11.73	9.16	22.61	9.14	4.11	16.35	9.18	6.48	20.63	7.87	3.00
medium		18.25	13.80	11.91	11.59	4.50	1.77	19.64	12.93	10.51	23.35	8.86	3.66
		19.72	14.02	11.58	19.09	7.99	3.66	18.79	11.54	8.91	23.13	8.95	3.74
		18.90	11.48	9.02	22.83	9.18	3.93	15.93	8.75	6.24	20.63	7.40	2.51
		12.83	6.06	3.96	20.23	7.72	3.02	11.78	5.57	3.59	17.03	6.40	2.12
high		16.46	9.07	6.61	22.77	9.45	4.39	13.92	7.43	5.28	21.32	7.64	2.90
		13.45	6.76	4.61	22.65	8.73	3.54	11.55	5.07	3.12	19.13	7.18	2.75
		9.26	3.72	2.08	17.84	6.56	2.11	8.74	3.10	1.67	14.40	5.26	1.41
		5.60	1.18	0.34	11.44	4.31	1.05	5.75	1.19	0.20	10.27	4.05	0.94
extreme		9.66	3.73	2.14	21.45	8.05	3.04	7.86	2.72	1.38	16.65	5.90	1.77
		7.06	2.12	0.94	16.93	6.30	2.10	6.42	1.73	0.66	13.86	5.42	1.67
		4.36	0.74	-0.06	11.34	4.15	0.83	4.34	0.63	-0.12	9.50	3.55	0.61
		2.68	-0.15	-0.59	6.21	2.41	0.22	2.94	-0.09	-0.59	6.27	2.37	0.21

However, in case of differing pairwise correlations, strong overshrinkage can be observed for $\Delta\sigma^2 = tiny$: The differences in ρ might be used to out-balance errors, so the calibration sets might be overfitted and strong shrinkage is required.

In contrast, the following relations can be identified for scenarios with at least *medium* differences in $\Delta\sigma^2$ independent of $\Delta\rho$.

First, the bias typically decreases with increasing variance range. Since differing weights are increasingly beneficial for larger $\Delta\sigma^2$, less shrinkage is required, so λ^*_{CV} and λ^*_{true} will be lower and closer to each other.

Second, overshrinkage typically decreases for increasing pairwise correlations, as weights learned on calibration sets might indeed be more extreme than on the full training set, but nevertheless, do not need to be shrunk strongly to EW, as complementary weights are then increasingly beneficial for out-balancing errors.

Third, overshrinkage usually increases with J for scenarios with larger $\Delta\sigma^2$. This seems reasonable, as estimating more weights increases the uncertainty and thus, stronger shrinkage of the more extreme weights is required.

However, opposite relations can be observed for $\Delta\sigma^2 \in \{tiny, low\}$, as the bias increases with ρ and with decreasing J in many scenarios:

Regarding the number of forecasters, λ^*_{true} is usually much higher for many forecasters than for a few, whereas λ^*_{CV} is generally quite high for comparable

variances and increases to a lower extent than λ^*_{true} (and thus, the bias decreases) for increasing J.

Regarding increasing correlations, more differentiated weights will be learned on the calibration sets that require stronger shrinkage, as similar weights would be more appropriate for such small differences in predictive ability.

In line with expectations, increasing K generally reduces overshrinkage, as the calibration sets contain more observations, which leads to increasingly similar and stable estimations of OW. These weights will also be more similar to those estimated on $\boldsymbol{E}_{train}$, which enables a more accurate shrinkage determination.

The next subsection focuses on scenarios with identical pairwise correlations ($\Delta\rho < 0.1$), as these show over- and undershrinkage. In addition, the general development of shrinkage levels is discussed with regards to the training size.

5.2 Shrinkage Levels and Bias Development

This subsection discusses the development of the truly optimal (λ^*_{true}) and the 5-fold CV-optimal (λ^*_{CV}) shrinkage and the respective biases shown in Fig. 1.

The solid lines represent λ^*_{CV} and the dashed ones the corresponding values of λ^*_{true} identified on $\boldsymbol{E}_{test}$ in the same color for the different categories of ranges in variance $\Delta\sigma^2$. The development is shown over the training size n, differentiated by the strength of constant correlations ρ and averaged over all forecasters J.

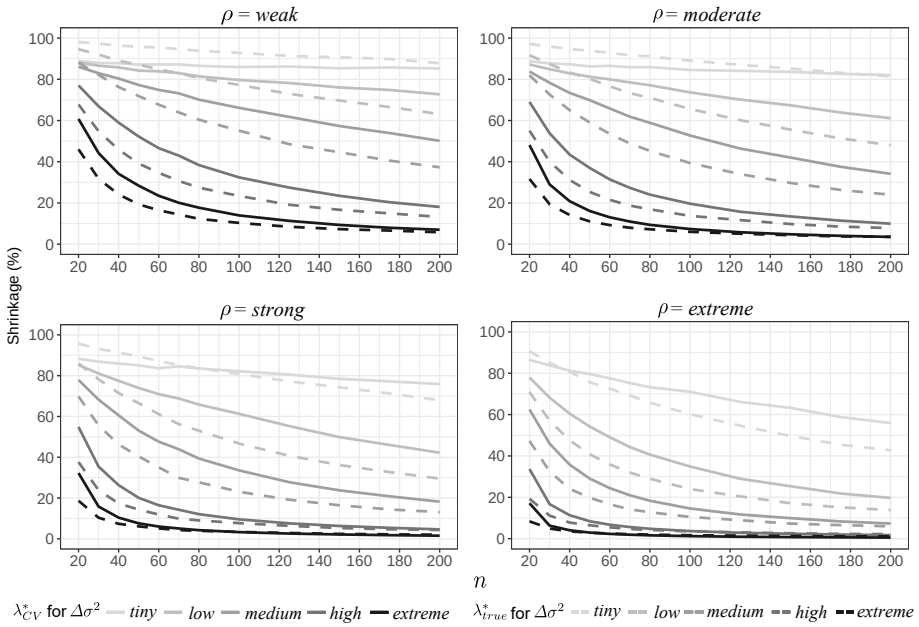

Fig. 1. Development of 5-Fold CV- (λ^*_{CV}) and Truly Optimal (λ^*_{true}) Shrinkage.

Obviously, both λ^*_{CV} and λ^*_{true} decrease with increasing n. This matches expectations, as larger amounts of training data typically represent the structures of the full dataset more precisely, which enables a more reliable estimation of OW on the full training set and the calibration sets, so less shrinkage is required.

Furthermore, λ^*_{CV} and λ^*_{true} decrease with increasing $\Delta\sigma^2$ and ρ, as differentiated weights seem reasonable for stronger differences in predictive ability and almost complementary weights can be assigned to highly correlated forecasters to neutralize error patterns.

Besides the shrinkage levels λ^*_{CV} and λ^*_{true}, the respective shrinkage bias can be observed as vertical distance between their corresponding lines.

According to the previous findings, overshrinkage is dominating, while undershrinkage is particularly pronounced for similar variances together with lower correlations. However, with increasing ρ, the bias development of scenarios with $\Delta\sigma^2 \in \{tiny, low\}$ approaches those with $\Delta\sigma^2 \in \{medium, high, extreme\}$.

In addition, the development of shrinkage biases regarding the training size n can be observed. For $\Delta\sigma^2 \in \{high, extreme\}$, the bias is typically higher for smaller n, as taking away parts of already small training sets fosters extreme (overfitted) OW. These will not fit well the validation sets and will be shrunk strongly to EW, whereas the OW learned on $\boldsymbol{E}_{train}$ will be less extreme, with no need to shrink as much as CV suggests. For increasing n, the OW estimated on the calibration sets will be less overfitted, and λ^*_{CV} will be closer to λ^*_{true}.

In contrast, for comparable variances, the appearing undershrinkage decreases with increasing n, but can turn into overshrinkage. Focusing on $\Delta\sigma^2 = tiny$, λ^*_{true} is near 100% for low n, as learning weights on small training sets does then not provide any benefits compared to assigning EW (i.e., full shrinkage). However, λ^*_{CV} is too low for $n = 20$ and does not decrease to the same extent as λ^*_{true} for increasing n, which reduces undershrinkage, but can lead to overshrinkage.

Finally, the findings in [22] can be confirmed, as overshrinkage appears to be higher for weaker ρ and lower n in case of distinguishable variances.

Based on these analyses, a bias prediction model is developed in Section 6.

6 Prediction Model for Weight Shrinkage Biases

The discussions above indicate that CV-based shrinkage tuning mostly leads to biased shrinkage levels, whereby the bias degree and direction are influenced by various parameters and their values.

Based on all studied scenarios, a CART regression tree [6] is learned to predict the shrinkage bias (target variable), depending on the CV-variant used and data characteristics, as trees can handle numerical as well as categorical variables with linear and non-linear relationships and also incorporate interaction effects.

The final tree, tuned by 5-fold CV regarding the *complexity parameter*, has a depth of 21 and considers all variables discussed in Section 4. For reasons of conciseness, Fig. 2 shows a simplified version (the first few splits).[6]

[6] The complete regression tree model can be generated using the code available online.

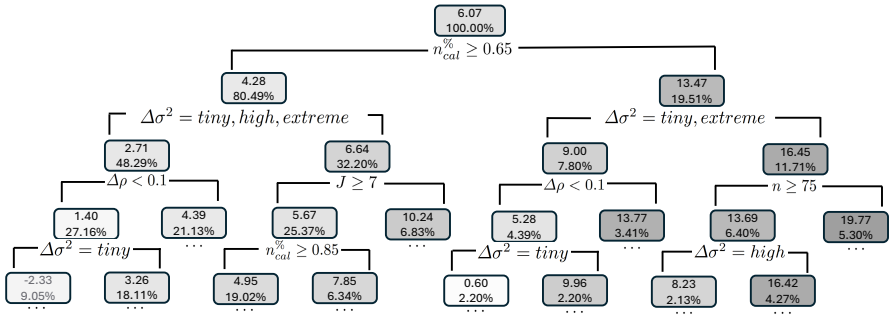

Fig. 2. Simplified Version of Regression Tree to Predict and Correct Shrinkage Biases.

The first value displayed in a node corresponds to the mean shrinkage bias of all cases assigned to this node during the learning process, while the second value indicates the share of the dataset used to learn the tree that belongs to the respective node. In addition, a darker node color reflects a stronger bias, with positive bias values written in black and negative ones in gray.

Starting with the root node, the mean shrinkage bias of all studied scenarios equals $6.07\%P$. All scenarios with calibration sets containing at least 65% of the available training observations ($n^{\%}_{cal} \geq 0.65$) are assigned to the left child node and their mean bias is $4.28\%P$. The remaining scenarios (i.e., those with $n^{\%}_{cal} < 0.65$) take the right branch and have a mean bias of $13.47\%P$.

Next, the range of variances $\Delta\sigma^2$ serves as splitting variable for both root child nodes. Considering the left child node, all scenarios with $n^{\%}_{cal} \geq 0.65$ and $\Delta\sigma^2 \in \{tiny, high, extreme\}$ take the left branch, while the ones with $n^{\%}_{cal} \geq 0.65$ and $\Delta\sigma^2 \in \{low, medium\}$ are assigned to the right. The splitting process continues until a branch terminates in a leaf node (no more outgoing branches).

The most left branches of both root child nodes use similar splitting criteria (at least for the first three splits). Also, the bias continuously decreases for these cases and is negative ($-2.33\%P$) for $n^{\%}_{cal} \geq 0.65, \Delta\rho < 0.1, \Delta\sigma^2 = tiny$, but slightly positive ($0.60\%P$) otherwise. In addition, the shrinkage bias is lower for $\Delta\sigma^2 = tiny$ than for $\Delta\sigma^2 \in \{(high), extreme\}$ in case of $\Delta\rho < 0.1$.

Also in line with the previous findings, the availability of more training observations per fold ($n^{\%}_{cal}$) reduces overshrinkage and the model predicts a higher bias on average for less forecasters, if $\Delta\sigma^2 \in \{low, medium\}$ and $n^{\%}_{cal} \geq 0.65$.

Considering $\Delta\rho$, similar pairwise error correlations ($\Delta\rho < 0.1$) lead to a lower shrinkage bias than differing ones ($\Delta\rho \geq 0.1$) when averaging over $\Delta\sigma^2 \in \{tiny, (high), extreme\}$ due to the much higher overshrinkage for $\Delta\sigma^2 = tiny$ in the case of $\Delta\rho \geq 0.1$ compared to $\Delta\rho < 0.1$.

Since the regression tree provides predictions for the shrinkage bias, it can serve as bias correction model. For this purpose, the CV-optimal shrinkage level λ^*_{CV} is determined by Algorithm 1 for the provided dataset denoted as $\boldsymbol{E}_{train}$ and the number of folds K randomly chosen.

Subsequently, the tree is used to predict the bias, depending on the observed and estimated parameter values of $\boldsymbol{E}_{train}$. This bias prediction is treated as correction factor CV_C and subtracted from λ^*_{CV}, resulting in the corrected CV-optimal shrinkage $\lambda^*_{CV_C}$.

The correction can be expected to improve shrinkage tuning if parameter values are known or precisely estimated. However, as their estimation on limited training samples can deviate from those of the complete dataset and/or test data, the correction model is evaluated by applications to provided training data.

7 Evaluation of Shrinkage Correction

This section evaluates the bias prediction model for its ability to correct shrinkage biases, if parameter values are estimated on limited training data and are therefore subject to uncertainty.

For this purpose, new synthetic samples (with $20,000$ multivariate error observations per set) are generated, with $J \in \{4, 9\}$ forecasters, variance ranges $\Delta\sigma^2 \in \{0.1, 0.25, 0.7, 2, 5, 7\}$, and pairwise correlations $\rho \in \{0.15, 0.3, 0.45, 0.6, 0.75, 0.9\}$ with ranges $\Delta\rho \in \{0, 0.3\}$.

Per sample, small training sets $\boldsymbol{E}_{train}$ containing $n \in \{25, 50, 100, 200\}$ observations per forecaster are randomly drawn, while the respective remaining observations form a large test set $\boldsymbol{E}_{test}$.

Table 3 shows the mean percentage deviation of the *MSE* on $\boldsymbol{E}_{test}$ (rounded to two digits) for various scenarios when shrinking OW towards EW by the corrected shrinkage $\lambda^*_{CV_C}$ instead of λ^*_{CV}, tuned on $K \in \{2, 5, LOO\}$ CV-folds.[7]

The evaluation results are averaged over $\Delta\rho, J$ and 250 repetitions per scenario and faceted by $n, K, \Delta\sigma^2$ and ρ. For example, considering datasets with $\Delta\sigma^2 = medium$ and $\rho = extreme$, for which $n = 25$ training observations are available and 2-fold CV is applied. In this scenario, applying the correction model can reduce the *MSE* by 7.99% on average.

Overall, the correction is beneficial (i.e., negative *MSE* deviation, highlighted in **bold**) for the majority of scenarios.

However, in scenarios with comparable variances, the correction typically leads to worse results. This could be due to the fact that deviations between estimated variances or correlations and the actual (true) values are particularly critical with smaller $\Delta\sigma^2$: According to Table 2, different categories for ρ and $\Delta\rho$ are associated with different degrees and even opposite directions of biases, which can also be expected for bias predictions, resulting in unreliable corrections.

Also, OW are typically shrunk strongly to EW for lower $\Delta\sigma^2$ (see Fig. 1), so it is suggested to simply assign EW if comparable variances are assumed.

For stronger spread in variances, the following relations are observed: First, the correction is successful for scenarios with few folds (even for smaller spread in variances), in which overshrinkage is typically higher. This seems reasonable, as the *MSE* on test data might be large in case of strong overshrinkage, so the correction has great potential for improvements.

[7] For a detailed description of the procedure, see Algorithm 2 in Subsection 4.3.

Table 3. Mean Percentage Deviation of *MSE* with $\lambda^*_{CV_C}$ from *MSE* with λ^*_{CV} for Selected Values of Training Size n, Folds K, Variance Range $\Delta\sigma^2$ and Correlation ρ.

$\Delta\sigma^2$	K	2				5				*LOO*			
	n \ ρ	*weak*	*moderate*	*strong*	*extreme*	*weak*	*moderate*	*strong*	*extreme*	*weak*	*moderate*	*strong*	*extreme*
tiny	25	1.92	1.73	1.58	1.31	1.33	1.53	1.63	0.93	1.05	1.28	1.35	0.78
	50	1.61	0.79	0.44	0.37	1.09	0.64	0.61	0.29	0.96	0.66	0.59	0.29
	100	0.62	0.15	**-0.01**	0.04	0.43	0.26	0.25	0.14	0.32	0.27	0.24	0.17
	200	0.12	**-0.03**	**-0.07**	**-0.07**	0.11	0.07	0.08	0.04	0.08	0.08	0.09	0.07
low	25	1.73	0.80	**-0.81**	**-1.47**	1.35	1.37	1.00	0.08	1.11	1.18	1.02	0.33
	50	1.13	0.07	**-0.54**	**-1.69**	0.87	0.45	0.35	**-0.31**	0.80	0.49	0.41	**-0.08**
	100	0.23	**-0.32**	**-0.56**	**-1.05**	0.18	0.05	0.01	**-0.22**	0.16	0.13	0.10	**-0.08**
	200	**-0.06**	**-0.24**	**-0.26**	**-0.44**	0.02	**-0.02**	**-0.01**	**-0.10**	0.05	0.02	0.00	**-0.06**
medium	25	0.35	**-1.28**	**-5.12**	**-7.99**	0.36	0.07	**-0.49**	**-2.11**	0.57	0.48	0.35	**-1.15**
	50	**-0.48**	**-1.61**	**-1.84**	**-3.64**	0.21	**-0.16**	**-0.01**	**-1.14**	0.33	0.02	0.14	**-0.70**
	100	**-0.58**	**-0.78**	**-0.68**	**-1.13**	**-0.04**	**-0.12**	**-0.10**	**-0.37**	0.03	0.01	**-0.04**	**-0.19**
	200	**-0.34**	**-0.29**	**-0.17**	**-0.37**	**-0.04**	**-0.01**	**-0.01**	**-0.13**	0.01	0.02	**-0.01**	**-0.08**
high	25	**-3.62**	**-6.61**	**-9.15**	**-11.78**	**-0.58**	**-0.78**	**-0.97**	**-2.18**	0.04	**-0.08**	**-0.09**	**-1.34**
	50	**-2.03**	**-2.24**	**-1.64**	**-2.46**	**-0.33**	**-0.32**	**-0.20**	**-0.59**	**-0.08**	**-0.16**	**-0.14**	**-0.26**
	100	**-0.63**	**-0.68**	**-0.44**	**-0.55**	**-0.05**	**-0.11**	**-0.13**	0.06	0.01	**-0.03**	**-0.08**	0.08
	200	**-0.23**	**-0.15**	**-0.08**	**-0.13**	0.01	**-0.01**	**-0.01**	0.07	0.01	**-0.00**	**-0.01**	0.08
extreme	25	**-4.90**	**-7.69**	**-9.42**	**-10.60**	**-0.73**	**-0.91**	**-1.16**	**-1.92**	**-0.05**	**-0.20**	**-0.42**	**-1.23**
	50	**-1.75**	**-1.89**	**-1.32**	**-1.87**	**-0.31**	**-0.29**	**-0.31**	**-0.29**	**-0.15**	**-0.21**	**-0.19**	**-0.04**
	100	**-0.43**	**-0.49**	**-0.33**	**-0.30**	**-0.08**	**-0.09**	**-0.09**	0.12	**-0.03**	**-0.03**	**-0.07**	0.17
	200	**-0.17**	**-0.09**	**-0.05**	**-0.05**	**-0.01**	**-0.01**	**-0.00**	0.11	**-0.00**	**-0.00**	**-0.01**	0.12

Second, the extent of correction typically decreases with increasing training size: With more training data, overshrinkage decreases, and so does the achievable *MSE* reduction, because the *CV*-determined shrinkage will already be closer to the truly optimal one. However, the correction is not necessarily beneficial with (too) small training set sizes together with weaker correlations.

Third, although overshrinkage typically decreases with increasing correlations, the magnitude of correction increases with ρ, so the determination of correction factors appears to be increasingly precise with stronger correlations.

Nevertheless, when relating the *MSE* to the number of folds used for the shrinkage tuning, *LOO* clearly dominates – in 69 of the 80 evaluated scenarios, the lowest mean *MSE* is achieved with corrected or uncorrected *LOO*. For the remaining scenarios (mainly belonging to $\Delta\sigma^2 \in \{tiny, low\}$ together with $n = 25$ or $\rho = weak$, for which assigning *EW* is suggested), uncorrected 2-fold *CV* performs slightly better (i.e., on average around 0.56% lower *MSE*) than *LOO*.

However, *LOO* can be computationally burdensome for large training sets, and in case of $\Delta\sigma^2 \in \{medium, high, extreme\}$ and at least $n = 100$ training observations, the *MSE* obtained with shrinkage levels that are determined on less folds (and additionally corrected if suggested by Table 3) is just slightly higher (on average 0.20% for $K = 2$ and 0.06% with $K = 5$) than with *LOO*.

The final section draws conclusions based on the analytical insights as well as the evaluation of shrinkage corrections and provides an outlook on future work.

8 Conclusions and Future Work

After discussing shrinkage biases for various data and CV-related characteristics as well as evaluating the introduced bias prediction model regarding its ability to correct shrinkage biases, the following conclusions are drawn.

First, for comparable predictive abilities of forecasters, it seems reasonable to assign EW, as learning weights and tuning (and correcting) shrinkage levels introduces uncertainties that outweigh the benefits of estimating weights.

Second, contrary to the general suggestion of 5- or 10-fold CV, practitioners should rather choose (corrected) LOO for weight shrinkage tuning: In scenarios with distinguishable forecaster variances, shrinkage levels tuned by LOO typically dominate the ones tuned on less folds, as taking away larger data parts can seriously affect the estimation of OW. Therefore, LOO is suggested for shrinkage tuning, with an additional correction in case of large spread in predictive ability.

However, if less CV-folds are chosen, e.g. due to the computational effort of LOO, the shrinkage should be corrected for distinguishable forecast abilities.

Third, researchers and applicants of any domain should be aware of estimation and tuning biases resulting from machine learning-based approaches and thus, use methods to detect and correct these tuning errors.

Future work will analyze additional data properties and develop advanced shrinkage corrections, as variance and correlation structures of empirical datasets do not necessarily follow easily identifiable patterns. The corrections will be compared to established combination techniques on synthetic or real-world datasets.

In addition, other machine learning techniques such as bootstrap aggregating (bagging) will be examined for biases in shrinkage tuning and their correction.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Aiolfi, M., Timmermann, A.: Persistence in forecasting performance and conditional combination strategies. J. Econom. **135**(1–2), 31–53 (2006). https://doi.org/10.1016/j.jeconom.2005.07.015
2. Aksu, C., Gunter, S.I.: An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts. Int. J. Forecast. **8**(1), 27–43 (1992). https://doi.org/10.1016/0169-2070(92)90005-T
3. Arlot, S., Lerasle, M.: Choice of V for V-fold cross-validation in least-squares density estimation. J. Mach. Learn. Res. **17**(1), 7256–7305 (2016). http://jmlr.org/papers/volume17/14-296/14-296.pdf
4. Bates, J.M., Granger, C.W.J.: The combination of forecasts. J. Oper. Res. Soc. **20**(4), 451–468 (1969). https://doi.org/10.2307/3008764

5. Blanc, S.M., Setzer, T.: When to choose the simple average in forecast combination. J. Bus. Res. **69**(10), 3951–3962 (2016). https://doi.org/10.1016/j.jbusres.2016.05.013
6. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: Classification and regression trees. 1st edn. Chapman and Hall, New York (1984). https://doi.org/10.1201/9781315139470
7. Breiman, L., Spector, P.: Submodel selection and evaluation in regression. The x-random case. Int. Stat. Rev. **60**(3), 291–319 (1992). https://doi.org/10.2307/1403680
8. Celisse, A.: Optimal cross-validation in density estimation with the L2-loss. Ann. Statist. **42**(5), 1879–1910 (2014). https://doi.org/10.1214/14-AOS1240
9. Chan, F., Pauwels, L.L.: Some theoretical results on forecast combinations. Int. J. Forecast. **34**(1), 64–74 (2018). https://doi.org/10.1016/j.ijforecast.2017.08.005
10. Claeskens, G., Magnus, J.R., Vasnev, A.L., Wang, W.: The forecast combination puzzle: A simple theoretical explanation. Int. J. Forecast. **32**(3), 754–762 (2016). https://doi.org/10.1016/j.ijforecast.2015.12.005
11. Clemen, R.T.: Combining forecasts: A review and annotated bibliography. Int. J. Forecast. **5**(4), 559–583 (1989). https://doi.org/10.1016/0169-2070(89)90012-5
12. Diebold, F.X., Pauly, P.: The use of prior information in forecast combination. Int. J. Forecast. **6**(4), 503–508 (1990). https://doi.org/10.1016/0169-2070(90)90028-A
13. Diebold, F.X., Shin, M.: Machine learning for regularized survey forecast combination: partially-egalitarian LASSO and its derivatives. Int. J. Forecast. **35**(4), 1679–1691 (2019). https://doi.org/10.1016/j.ijforecast.2018.09.006
14. Genre, V., Kenny, G., Meyler, A., Timmermann, A.: Combining expert forecasts: Can anything beat the simple average? Int. J. Forecast. **29**(1), 108–121 (2013). https://doi.org/10.1016/j.ijforecast.2012.06.004
15. Graefe, A., Küchenhoff, H., Stierle, V., Riedl, B.: Limitations of ensemble bayesian model averaging for forecasting social science problems. Int. J. Forecast. **31**(3), 943–951 (2015). https://doi.org/10.1016/j.ijforecast.2014.12.001
16. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction. 1st edn. Springer, Berlin (2009). https://doi.org/10.1007/b94608
17. Haubner, N., Setzer, T.: Hybrid recommender systems for next purchase prediction based on optimal combination weights. In: Wirtschaftsinformatik 2021 Proceedings. (2021). https://aisel.aisnet.org/wi2021/RDataScience/Track09/1/
18. Jahrer, M. Töscher, A., Legenstein, R.: Combining predictions for accurate recommender systems. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 693–702 (2010). https://doi.org/10.1145/1835804.1835893
19. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning: with applications in R. 2nd edn. Springer, New York (2021). https://doi.org/10.1007/978-1-0716-1418-1
20. Schanbacher, P.: Combining Portfolio Models. ANN ECON FINANC **15**(2), 433–455 (2014). http://aeconf.com/Articles/Nov2014/aef150208.pdf
21. Schulz, F., Setzer, T., Balla, N.: Linear hybrid shrinkage of weights for forecast selection and combination. In: Proceedings of the 55th Hawaii International Conference on System Sciences. pp. 2125–2134 (2022). https://doi.org/10.24251/HICSS.2022.267
22. Schulz, F., Setzer, T.: Shrinkage of weights towards subset selection in forecast combination. Available at SSRN: https://ssrn.com/abstract=4485995 (2023)

23. Smith, D.G.C.: Combination of forecasts in electricity demand prediction. J. Forecast. **8**, 349–356 (1989). https://doi.org/10.1002/for.3980080316
24. Smith, J., Wallis, K.F.: A simple explanation of the forecast combination puzzle. Oxf. Bull. Econ. Stat. **71**(3), 331–355 (2009). https://doi.org/10.1111/j.1468-0084.2008.00541.x
25. Stock, J.H., Watson, M.W.: Combination forecasts of output growth in a seven-country data set. J. Forecast. **23**(6), 405–430 (2004). https://doi.org/10.1002/for.928
26. Thompson, R., Qian, Y., Vasnev, A.L.: Flexible global forecast combinations. Omega **126**, (2024). https://doi.org/10.1016/j.omega.2024.103073
27. Timmermann, A.: Forecast combinations. In: Handbook of Economic Forecasting. 1st edn. pp. 135–196. Elsevier (2006). https://doi.org/10.1016/S1574-0706(05)01004-9
28. Wang, X., Hyndman, R.J., Li, F., Kang, Y.: Forecast combinations: An over 50-year review. Int. J. Forecast. **39**(4), 1518–1547 (2023). https://doi.org/10.1016/j.ijforecast.2022.11.005
29. Zhang, Y., Yang, Y.: Cross-validation for selecting a model selection procedure. J. Econom. **187**(1), 95–112 (2015). https://doi.org/10.1016/j.jeconom.2015.02.006