# The Hedged Random Forest

Elliot Beck[1], Damian Kozbur[1] and Michael Wolf[1,2]

[1]University of Zurich

[2]ADIA Lab

# Outline

# Outline

# Motivation

The random forest is one of the most popular and widely employed tools for supervised machine learning.

It can be used for both classification and regression tasks;
in this paper, the focus will be on regression only.

In its standard form, the crux of the random forest is to use an equal-weighted ensemble of (decorrelated) tree-based predictors.

Instead, we suggest a more general weighting scheme that borrows certain ideas from the related problem of financial portfolio selection.

# Outline

# Setting and Criterion

Setting:

- The goal is to forecast (or predict) a random variable $y \in \mathbb{R}$
- Available is a set of attributes (or regressors) $x \in \mathbb{R}^d$
- A generic forecast is denoted by $\hat{f}$
- The criterion is the mean-squared error (MSE):

$$\mathrm{MSE}(\hat{f}) := \mathbb{E}(y - \hat{f}(x))^2$$

Letting

- $\mathrm{Bias}(\hat{f}) := \mathbb{E}(y - \hat{f}(x))$ and
- $\mathrm{Var}(\hat{f}) := \mathbb{V}\mathrm{ar}(y - \hat{f}(x)) = \mathbb{E}((y - \hat{f}(x))^2 - \mathrm{Bias}^2(\hat{f})$

we have the well-known decomposition

$$\mathrm{MSE}(\hat{f}) = \mathrm{Bias}^2(\hat{f}) + \mathrm{Var}(\hat{f})$$

Note: The oracle minimizing the MSE is $\hat{f}_{\mathrm{or}}(x) := \mathbb{E}(y|x)$.

# Generic Framework ⊇ Random Forest

Forecast-combination problem:

- Have an ensemble of $p$ forecast methods, denoted by $\{\mathcal{M}_j\}_{j=1}^{p}$ (which may involve fitted parameters)
- "Hard to beat" is the equal-weighted forecast

$$\hat{f}_{\text{EW}}(x) := \frac{1}{p}\sum_{j=1}^{p}\mathcal{M}_j(x)$$

- Interested in weighted forecasts of the kind

$$\hat{f}_w(x) := \sum_{j=1}^{p} w_j \mathcal{M}_j(x) \quad \text{with} \quad \sum_{j=1}^{p} w_j = 1$$

where negative weights are allowed

# The Oracle Solution

Vector of forecast errors $e := (e_1, \ldots, e_p)'$, with $e_j := y - \mathcal{M}_j(x)$, has

$$\mu := \mathbb{E}(e) \quad \text{and} \quad \Sigma := \mathbb{V}\text{ar}(e)$$

The MSE of a weighted forecast $\hat{f}_w$ is then given by

$$\text{MSE}(\hat{f}_w) = (w'\mu)^2 + w'\Sigma w$$

The oracle forecast combination is thus the solution to

$$\begin{aligned} &\min_w \ (w'\mu)^2 + w'\Sigma w \\ \text{s.t.} \quad &w'\mathbf{1} = 1 \end{aligned}$$

# The Feasible Solution

Based on training data, compute estimators $\hat{\mu}$ and $\hat{\Sigma}$.

A feasible forecast combination is thus the solution to

$$\begin{aligned} \min_{w} \; & (w'\hat{\mu})^2 + w'\hat{\Sigma}w \\ \text{s.t.} \quad & w'\mathbf{1} = 1 \quad \text{and} \\ & \|w\|_1 \leq \kappa \end{aligned}$$

The final constraint with $\kappa \in [1, \infty]$ is a gross-exposure constraint:

- Motivated by the related problem of financial portfolio selection
- $\kappa = 1$ enforces non-negative weights ("no short-selling")
- $\kappa = \infty$ corresponds to dropping the constraint
- Values $\kappa \in [1.5, 2.5]$ tend to work well in practice

Results in a hedged forecast combination.

# Impressive Asymptotic Optimality Theory

# Astounding Finite-Sample Guarantees

# Outline

# Details

Estimation of $\mu$ and $\Sigma$:

- Training data set $\{v_i\}_{i=1}^n$ with $v_i := (y_i, x_i')'$
- We only consider i.i.d. data in this paper
- Train $p = 500$ trees using the "ranger library" with default choices
- Extract the forecasts of each tree $\mathcal{M}_j$ on the entire training set
  $\Longrightarrow$ residual matrix $R$ of size $n \times p$
- $\hat{\mu}$ is obtained as the (column-wise) sample mean of $R$
- $\hat{\Sigma}$ is obtained by applying nonlinear shrinkage to $R$;
  but can also consider the sample covariance matrix

The default choice for the gross-exposure constraint is $\kappa = 2$.

# Possible Concerns

"Using residuals, that is, in-sample errors leads to an underestimation (in magnitude) of $\mu$ and $\Sigma$."

True, but the feasible solution $\hat{w}$ does not change if the inputs $(\hat{\mu}, \hat{\Sigma})$ are replaced by $(c\hat{\mu}, c^2\hat{\Sigma})$ for any $c > 0 \Longrightarrow$ scale invariance.

"To estimate $\mu$ and $\Sigma$ one should only use out-of-bag (OOB) residuals."

This could be done to estimate $\mu$ but not to estimate $\Sigma$,
since nonlinear shrinkage needs a 'full' rectangular matrix R.

Also, scale invariance alleviates this concern.

# Outline

# Data

Use the 17 numerical benchmark data sets of Grinsztajn et al. (2022).

Accessible on the official openML website `www.openml.org`, which also offers comprehensive metadata and descriptions.

For large data sets, we follow Grinsztajn et al. (2022) by selecting 10,000 observations at random.

# Data Sets Used

| Name | # Observations | # Attributes |
|---|---|---|
| ailerons | 13,750 | 734 |
| Bike_Sharing_Demand | 17,379 | 12 |
| brazilian_houses | 10,692 | 10 |
| cpu_act | 8,192 | 22 |
| diamonds | 53,940 | 10 |
| elevators | 16,599 | 12 |
| house_16H | 22,784 | 17 |
| house_sales | 21,613 | 18 |
| houses | 20,640 | 9 |
| isolet | 7,797 | 614 |
| medical_charges | 163,065 | 4 |
| MiamiHousing2016 | 13,932 | 14 |
| nyc-taxi-green-dec-2016 | 581,835 | 19 |
| pol | 15,000 | 49 |
| sulfur | 10,081 | 7 |
| superconduct | 21,263 | 80 |
| wine_quality | 6,497 | 12 |

# Performance Evaluation

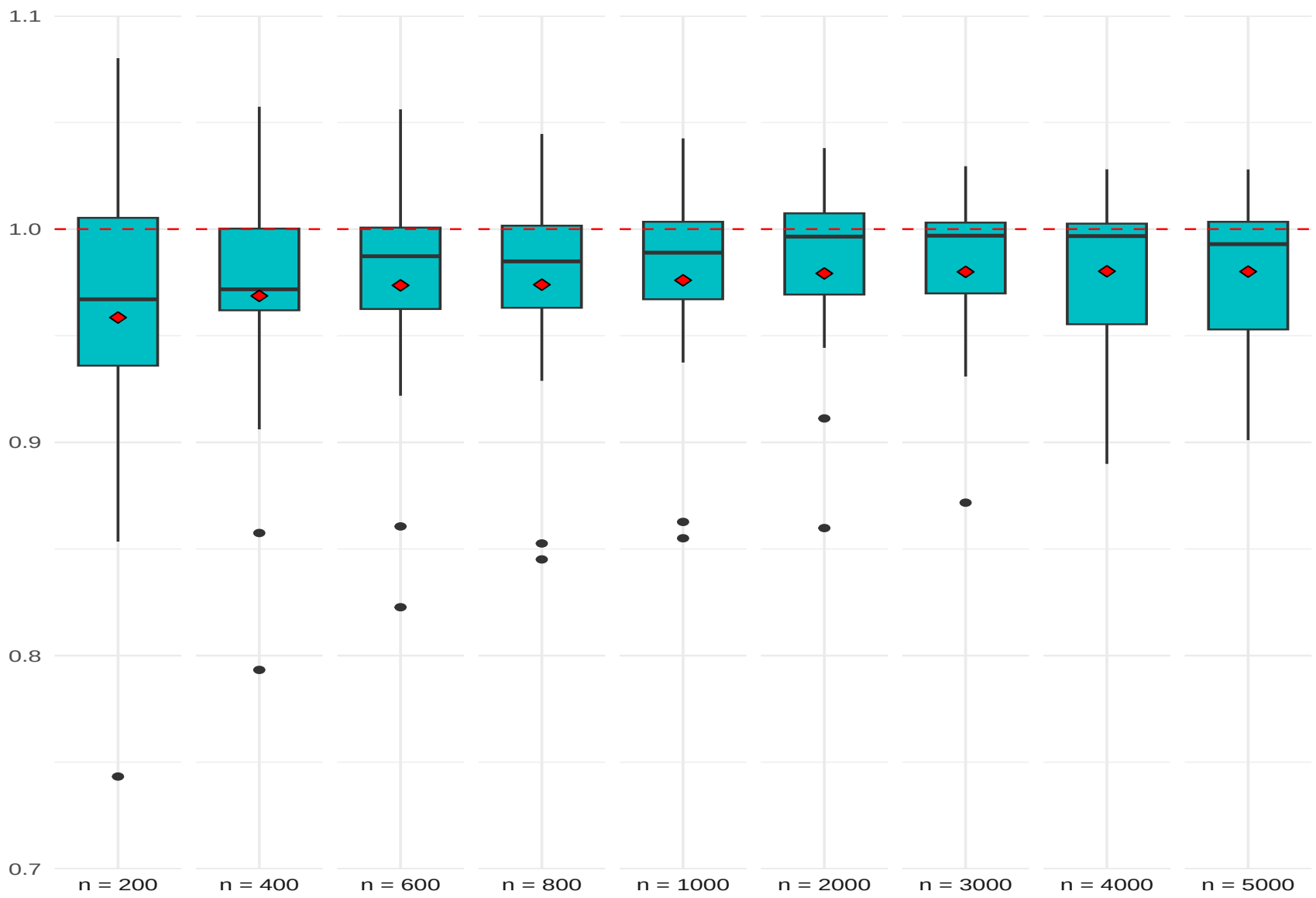
Prescription:

- Partition a data set into a training set and a test set by selecting $n$ observations at random as the training set
- Any forecasting method then yields a MSE on the test set
- Repeat this process $B = 100$ times
- Obtain the following RMSE ratio:

$$\mathrm{RMSE}_{\mathrm{HRF/RF}} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B} \mathrm{MSE}_{\mathrm{HRF,b}}}}{\sqrt{\frac{1}{B}\sum_{b=1}^{B} \mathrm{MSE}_{\mathrm{RF,b}}}}$$

- Get 17 such ratios for each $n \Longrightarrow$ boxplot
- Use $n \in \{200, 400, 600, 800, 1000, 2000, 3000, 4000, 5000\}$
- Line up the corresponding boxplots horizontally

# HRF vs. RF


1.1
1.0
0.9
0.8
0.7
n = 200
n = 400
n = 600
n = 800
n = 1000
n = 2000
n = 3000
n = 4000
n = 5000

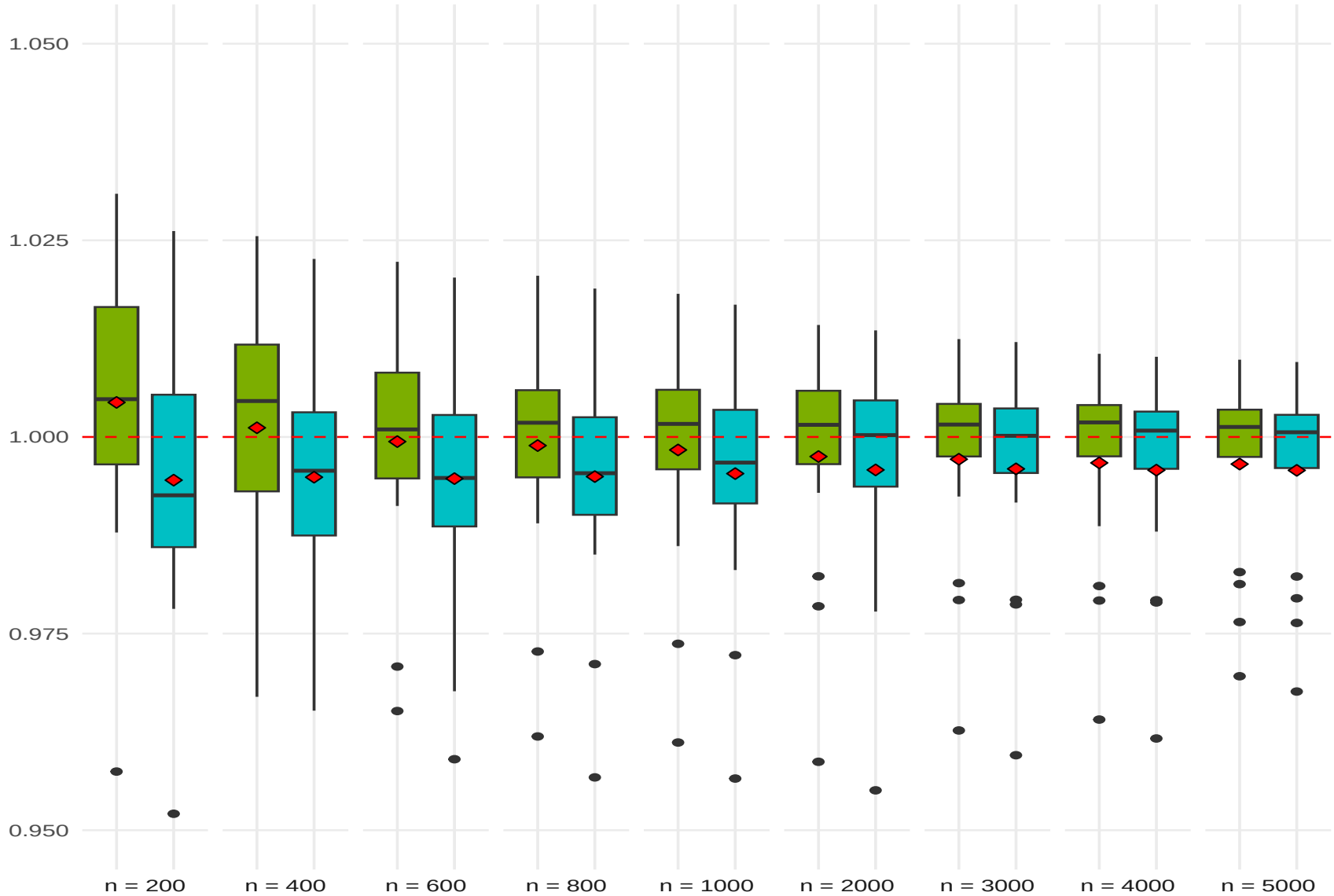
# Negative Weights and Shrinkage

Look at

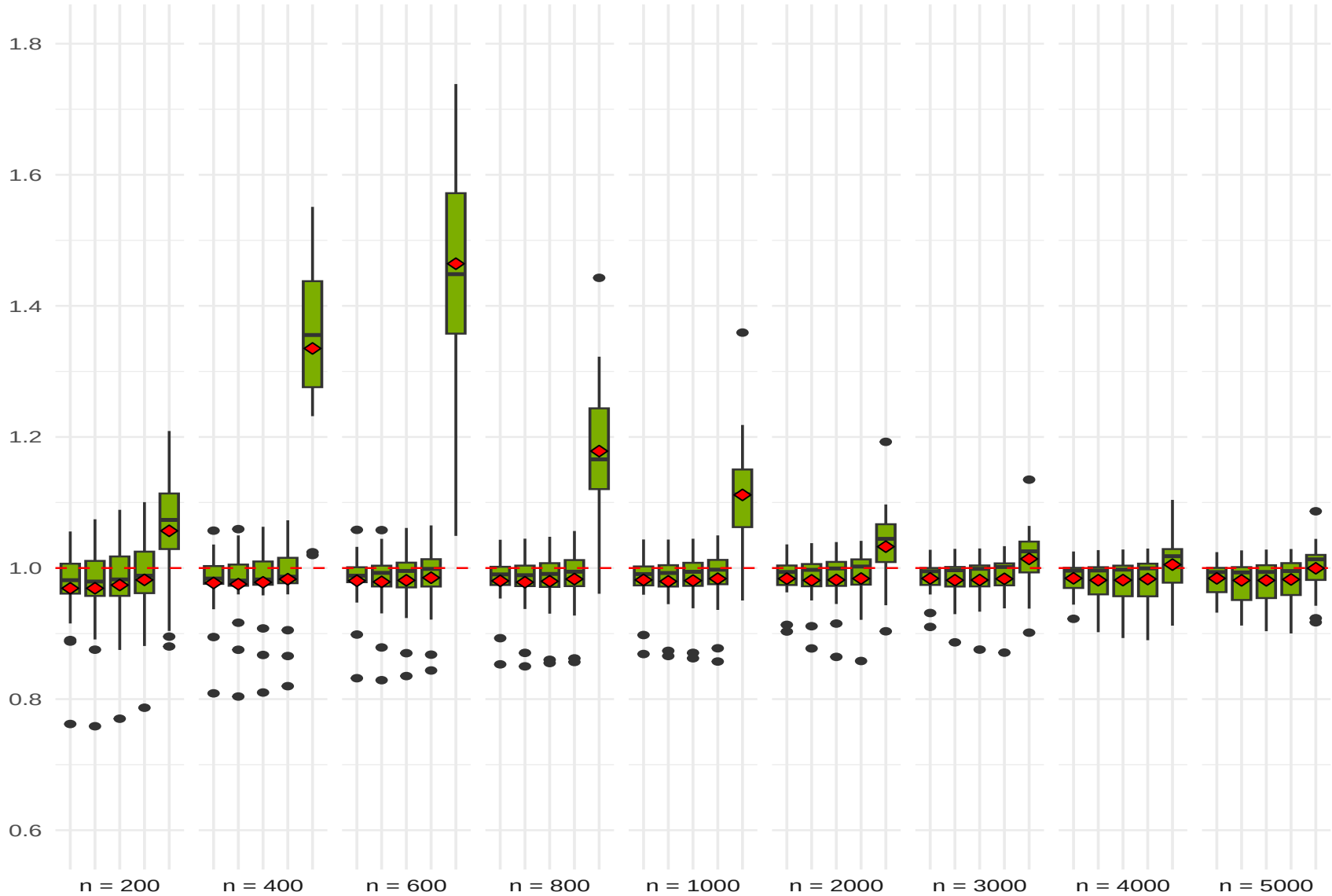
$$\mathrm{RMSE}_{\kappa=2/\kappa=1} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B} \mathrm{MSE}_{\mathrm{HRF,b},\kappa=2}}}{\sqrt{\frac{1}{B}\sum_{b=1}^{B} \mathrm{MSE}_{\mathrm{HRF,b},\kappa=1}}}$$

for $\hat{\Sigma} \in \{$Sample Covariance matrix, Nonlinear Shrinkage$\}$.

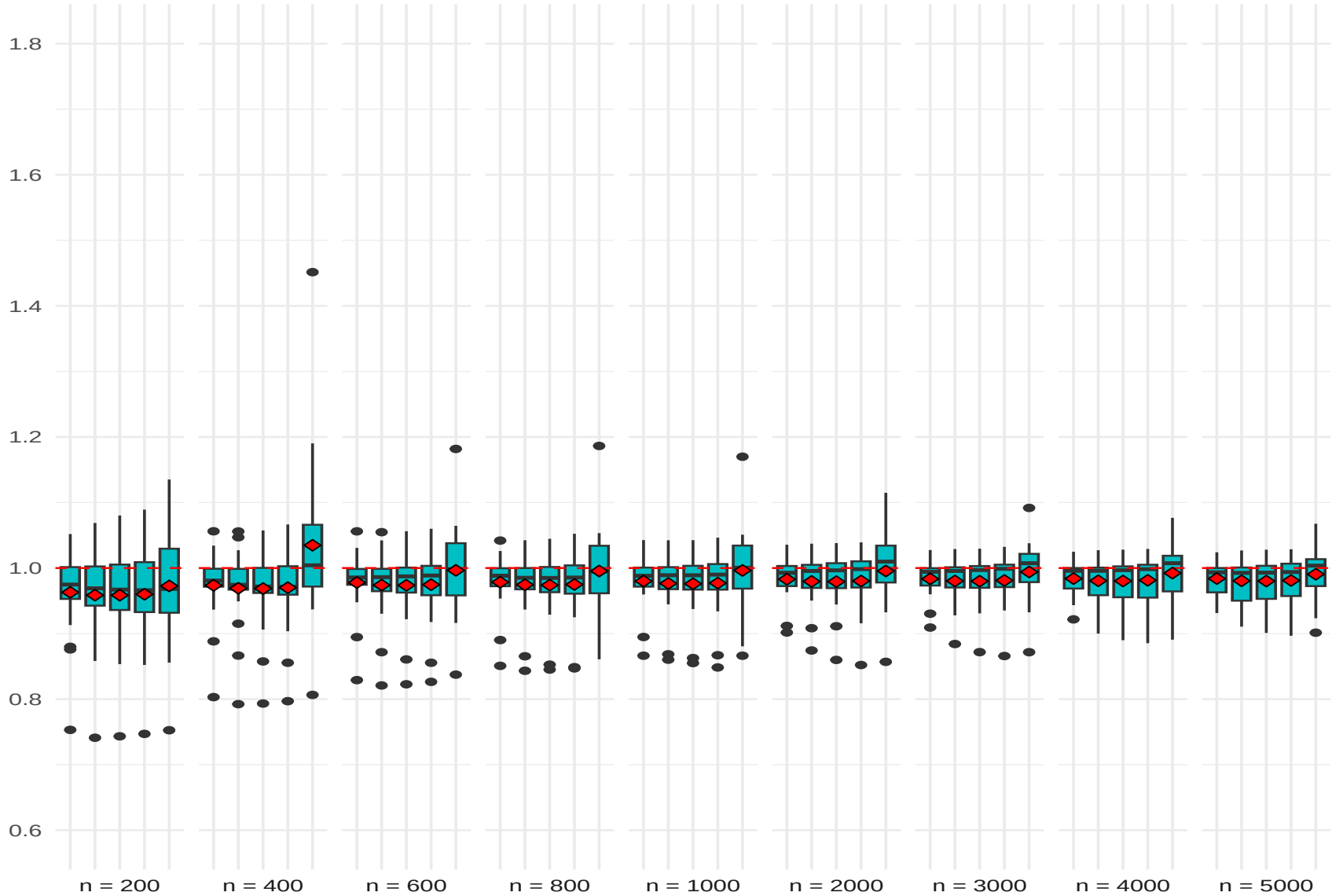
# $\kappa = 2$ vs. $\kappa = 1$ for Sample and NL Shrinkage



# $\kappa \in \{1, 1.5, 2, 2.5, \infty\}$ for Sample



# $\kappa \in \{1, 1.5, 2, 2.5, \infty\}$ for NL Shrinkage



# Comparison with Alternative Proposals

Found two alternative proposals for a weighted random forest:

- Winham et al. (2013)
- Pham and Olfsson (2020)

Proposed for classification, but can be adapted to regression.

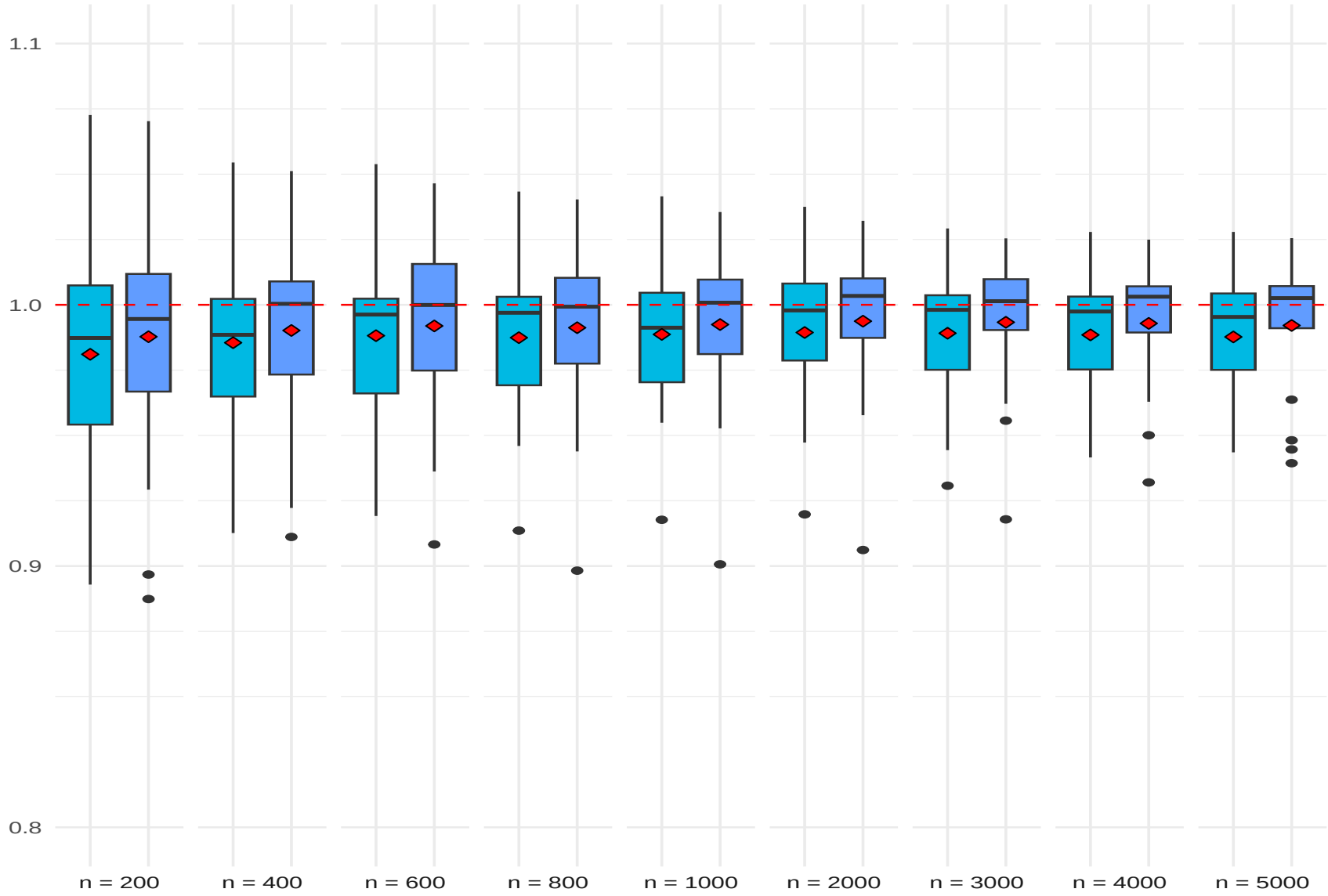
Common characteristics:

- Only use out-of-bag (OOB) residuals
- Rank performances of individual trees
- Neglect "covariances"
- All weights are strictly positive

Look at

$$\mathrm{RMSE}_{\mathrm{HRF/AP}} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B} \mathrm{MSE}_{\mathrm{HRF,b}}}}{\sqrt{\frac{1}{B}\sum_{b=1}^{B} \mathrm{MSE}_{\mathrm{AP,b}}}}$$

# Comparison with Alternative Proposals



# Outline

# Conclusions

We have proposed a generic methodology to 'estimate' optimal forecast combinations under the MSE criterion.

The methodology allows for negative weights but provides a hedge by imposing an upper bound on the sum of the absolute weights.

We apply this methodology to the random forest where the individual forecast methods correspond to the tree-based forecasts.

The hedged random forest (HRF) outperforms the standard RF and two weighted competitors on 17 real-life benchmark data sets.

Future research:

- HRF with time-series data
- HRF for classification

# References


Chen, X., Yu, D., and Zhang, X. (2023). Optimal weighted random forests. arXiv Preprint arXiv:2305.10042.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520.

Pham, H. and Olafsson, S. (2020). On Césaro averages for weighted trees in the random forest. *Journal of Classification*, 37(1):223–236.

Winham, S. J., Freimuth, R. R., and Biernacka, J. M. (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):496–505.