The Anatomy of Out-of-Sample Forecasting Accuracy

Daniel Borup[1] Phillipe Goulet Coulombe[2] David Rapach[3]
Erik Christian Montes Schütte[1] Sander Schwenk-Nebbe[1]

[1]Aarhus University [2]UQAM [3]Federal Reserve Bank of Atlanta

12th ECB Conference on Forecasting Techniques

Frankfurt

June 13, 2023

- **Disclaimer** $\Rightarrow$ The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any errors are the authors' responsibility.

- **Large datasets** (ie, "big data") and **machine learning**
  - These are growing in importance in macro/finance for out-of-sample time-series forecasting
- Large datasets
  - Provide greater capacity to incorporate relevant signals
- Machine learning
  - Tools to guard against **overfitting**
    - Improve out-of-sample performance with many predictors
  - Accommodates general nonlinearities
    - Random forests, boosted trees, neural networks

- Macro forecasting (eg, inflation, output/employment growth, unemployment rate, initial claims, recessions)
  - Li & Chen (2014), Exterkate et al (2016), Medeiros & Mendes (2016), Döpke, Fritsche & Pierdzioch (2017), Kim & Swanson (2018), Smeekes & Wijler (2018), Medeiros et al (2021), Vrontos, Gelakis & Vrontos (2021), Yousuf & Ng (2021), Borup & Schütte (2022), Goulet Coulombe (2022), Goulet Coulombe et al (2022), Hauzenberger, Huber & Klieber (2023), Borup, Rapach & Schütte (forthcoming),
- Financial forecasting (eg, stock returns)
  - Chinco, Clark-Joseph & Ye (2019), Rapach et al (2019), Freyberger, Neuhierl & Weber (2020), Gu, Kelly & Xiu (2020), Kozak, Nagel & Santosh (2020), Bryzgalova, Pelger & Zhu (2021), Avramov, Cheng & Metzker (forthcoming), Cong et al (2022), Dong et al (2022), Avramov, Cheng & Metzker (forthcoming), Chen, Pelger & Zhu (forthcoming)

- In addition to out-of-sample forecasting accuracy, the **interpretation** of fitted prediction models is important
  - Which predictors are the most relevant for determining the forecasts generated by fitted models?
  - How do the predictors contribute to out-of-sample forecasting accuracy?
- Interpretation helps users to wrap their minds around forecasting models, so they're not simply "black boxes"
  - Gain insight into empirically important economic mechanisms to guide the assessment/development of theories
  - Provide more comprehensible advice to policymakers

- Existing model-agnostic tools for model interpretation
  - How the model's predictions vary with the individual predictors
    - Partial dependence plots (PDP, Friedman 2001)
    - Shapley values (Shapley 1953, Štrumbelj & Kononenko 2014, Lundberg & Lee 2017)
    - Individual conditional expectation (ICE) plots (Goldstein et al 2015)
    - Local interpretable model-agnostic explanations (LIME, Ribeiro, Singh & Guestrin 2016)
    - Accumulated local effects (ALE, Apley & Zhu 2020)
  - Variable/feature importance
    - PDP based (Greenwell, Boehmke & McCarthy 2018)
    - Permutations (Fisher, Rudin & Domenici 2019)
    - Shapley based (Casalicchio, Molnar & Bischl 2019)

- Existing tools are more appropriate for cross-sectional data
- We propose Shapley-based metrics for **time-series** data
  - In-/out-of-sample variable importance
    - iShapley-VI/oShapley-VI
  - Out-of-sample **performance-based Shapley value** (PBSV)
    - Measures the contributions of individual predictors to out-of-sample forecasting accuracy based on a loss function
    - Identifies the predictors most responsible for a model's out-of-sample forecasting performance
- PBSV applies to
  - Any fitted prediction model (eg, linear/nonlinear, parametric/nonparametric)
  - Any loss function (eg, MSE/RMSE/MAE)

- Empirical application ⇒ forecasting US inflation
  - Large dataset and machine-learning methods
  - Machine-learning forecasts consistently outperform a conventional AR benchmark for horizons of 1 to 12 months
  - Close correspondence between iShapley-VI and oShapley-VI
  - In many cases, there is a relatively close correspondence between the in-sample iShapley-VI and out-of-sample PBSV
  - However, there are also a number of discrepancies between individual predictor relevance according to iShapley-VI and PBSV
    - Warning ⇒ in-sample importance of a predictor in determining the predicted target values doesn't necessarily align with its role in determining out-of-sample forecasting accuracy (even when a forecasting model performs well)

- Time-series context
- Index individual predictors by p
  - Index set of predictors $\Rightarrow S = \{1, \ldots, P\}$
- Period-t vector of predictors $\Rightarrow \boldsymbol{x}_t = \begin{bmatrix} x_{1,t} & \cdots & x_{P,t} \end{bmatrix}'$
- Prediction model $\Rightarrow \underbrace{y_{t+1:t+h}}_{\text{target}} = f(\boldsymbol{x}_t) + \varepsilon_{t+1:t+h}$
  - $y_{t+1:t+h} = \frac{1}{h} \sum_{k=1}^{h} y_{t+k}$ ($h \Rightarrow$ forecast horizon)
  - $f(\boldsymbol{x}_t) \Rightarrow$ conditional mean (ie, prediction) function
  - $\varepsilon_{t+1:t+h} \Rightarrow$ zero-mean disturbance term

- Training sample/window $\Rightarrow W_i = \{t_{i,\text{start}}, \ldots, t_{i,\text{end}}\}$
  - Set of observations used to train the prediction model based on window W_i
- Fitted prediction model $\Rightarrow \hat{f}(\boldsymbol{x}_t ; W_i, h)$
  - Evaluated at instance $\boldsymbol{x}_t$
  - Trained using data from window W_i
  - For forecast horizon h

- We interpret fitted prediction models with Shapley values
  - For model interpretation, Shapley values use coalitional game theory to utilize the analogy between the predictors/players in a cooperative game earning payoffs
    - Payoff $\Rightarrow$ predictor's contribution to a model's prediction
    - Shapley values fairly allocate the predictors' contributions to the prediction
- Štrumbelj & Kononenko (2014) use Shapley values to interpret prediction models
  - We modify their approach for time-series prediction
- Aim of the Shapley value in a time-series context
  - Quantify the marginal contribution of the predictor $x_{t,p}$ to $\hat{f}(\boldsymbol{x}_t\,; W_i, h)$, conditional on the presence of all of the other predictors $(S \setminus \{p\})$

- Shapley value for predictor $p \Rightarrow$

$$\phi_p(\boldsymbol{x}_t\,; W_i, h) = \sum_{Q \subseteq S \setminus \{p\}} \frac{|Q|!(P-|Q|-1)!}{P!} \left[\xi_{Q\cup\{p\}}(\boldsymbol{x}_t\,; W_i, h) - \xi_Q(\boldsymbol{x}_t\,; W_i, h)\right]$$

  - $Q \Rightarrow$ subset of predictors (ie, coalition)
  - $Q \subseteq S \setminus \{p\} \Rightarrow$ set of all possible coalitions of $P-1$ predictors excluding p
  - $|Q| \Rightarrow$ cardinality of Q
  - $|Q|!(P-|Q|-1)!/P! \Rightarrow$ combinatorial weight
  - Value function
    - $\xi_Q(\boldsymbol{x}_t\,; W_i, h) = \mathbb{E}\left[\hat{f} \,\middle|\, X_{t,j} = x_{t,j}\,\forall\, j \in Q\,; W_i, h\right]$
      - Model prediction conditional on the predictors in coalition Q

- $\xi_{Q \cup \{p\}}(\boldsymbol{x}_t ; W_i, h) - \xi_Q(\boldsymbol{x}_t ; W_i, h) \Rightarrow \Delta$ in value function
  - Δ in the model prediction, conditional on the predictors in coalition Q, when predictor p is included in the conditioning information set
  - Compute for all possible coalitions of $P-1$ predictors that exclude predictor p and then take a weighted average
- Objective $\Rightarrow$ isolate the marginal contribution of predictor p
  - Shapley values use coalitions to control for the other predictors when measuring the contribution of predictor p to the prediction corresponding to instance $\boldsymbol{x}_t$
  - Integrate out the predictors not in a coalition when taking the expectation in the value function
  - Finally, take a weighted average of the Δs in the value function for all possible coalitions of $P-1$ predictors excluding p

- Efficiency (AKA local accuracy)
  - $\sum_{p \in S} \phi_p(\boldsymbol{x}_t \,;\, W_i, h) = \hat{f}(\boldsymbol{x}_t \,;\, W_i, h) - \mathbb{E}\left[\hat{f} \,;\, W_i, h\right]$
    - $\mathbb{E}\left[\hat{f} \,;\, W_i, h\right] \Rightarrow$ baseline prediction corresponding to the unconditional expectation of $\hat{f}$
  - Can exactly decompose the model prediction corresponding to instance $\boldsymbol{x}_t$ (in terms of the deviation from the baseline prediction) into the sum of the Shapley values for the individual predictors for that instance
- Missingness
  - $\forall\, R \subseteq S \setminus \{p\} : \xi_{R \cup \{p\}}(\boldsymbol{x}_t \,;\, W_i, h) = \xi_R(\boldsymbol{x}_t \,;\, W_i, h) \Rightarrow \phi_p(\boldsymbol{x}_t \,;\, W_i, h) = 0$

- Symmetry
  - $\forall\, R \subseteq S \setminus \{p, q\} : \xi_{R \cup \{p\}}(\boldsymbol{x}_t\,;\, W_i, h) = \xi_{R \cup \{q\}}(\boldsymbol{x}_t\,;\, W_i, h) \Rightarrow \phi_p(\boldsymbol{x}_t\,;\, W_i, h) = \phi_q(\boldsymbol{x}_t\,;\, W_i, h)$
- Missingness and symmetry are intuitively appealing
- Linearity
  - For any real numbers c_1 and c_2 and models $\hat{f}(\boldsymbol{x}_t\,;\, W_i, h)$ and $\hat{f}'(\boldsymbol{x}_t\,;\, W_i, h)$, $\phi_p\Big(c_1\Big[\hat{f}(\boldsymbol{x}_t\,;\, W_i, h) + c_2\hat{f}'(\boldsymbol{x}_t\,;\, W_i, h)\Big]\Big) = c_1\phi_p\Big(\hat{f}(\boldsymbol{x}_t\,;\, W_i, h)\Big) + c_1c_2\phi_p\Big(\hat{f}'(\boldsymbol{x}_t\,;\, W_i, h)\Big)$
  - Useful for computing Shapley values for ensembles of prediction models

- Practically infeasible to compute exact Shapley values for even a moderate number of predictors
- Štrumbelj & Kononenko (2014) develop an algorithm for estimating Shapley values
  - Builds on the Castro, Gómez & Tejada (2009) sampling-based approach
  - We use a refined version of the algorithm

- $\phi_p(\boldsymbol{x}_t\,; W_i, h) \Rightarrow$ can express equivalently as

$$\frac{1}{P!} \sum_{\mathcal{O} \in \pi(P)} \left[\xi_{\text{Pre}_p(\mathcal{O}) \cup \{p\}}(\boldsymbol{x}_t\,; W_i, h) - \xi_{\text{Pre}_p(\mathcal{O})}(\boldsymbol{x}_t\,; W_i, h)\right]$$

  - $\mathcal{O} \Rightarrow$ ordered permutation for the predictor indices in S
  - $\pi(P) \Rightarrow$ set of all ordered permutations for S
  - $\text{Pre}_p(\mathcal{O}) \Rightarrow$ set of indices that precede p in $\mathcal{O}$
- Estimate $\phi_p(\boldsymbol{x}_t\,; W_i, h)$ via an algorithm
  - Make a random draw m with replacement for ordered permutation from $\pi(P) \Rightarrow \mathcal{O}_m$

- For a random draw m, compute

$$\theta_{p,m}(\boldsymbol{x}_t\,; W_i, h) = \frac{1}{|W_i|} \sum_{s \in W_i} \Big[\hat{f}(\boldsymbol{x}_{j,t} : j \in \text{Pre}_p(\mathcal{O}_m) \cup \{p\},\ \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m)\,; W_i, h) - \hat{f}(\boldsymbol{x}_{j,t} : j \in \text{Pre}_p(\mathcal{O}_m),\ \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m) \cup \{p\}\,; W_i, h) \Big]$$

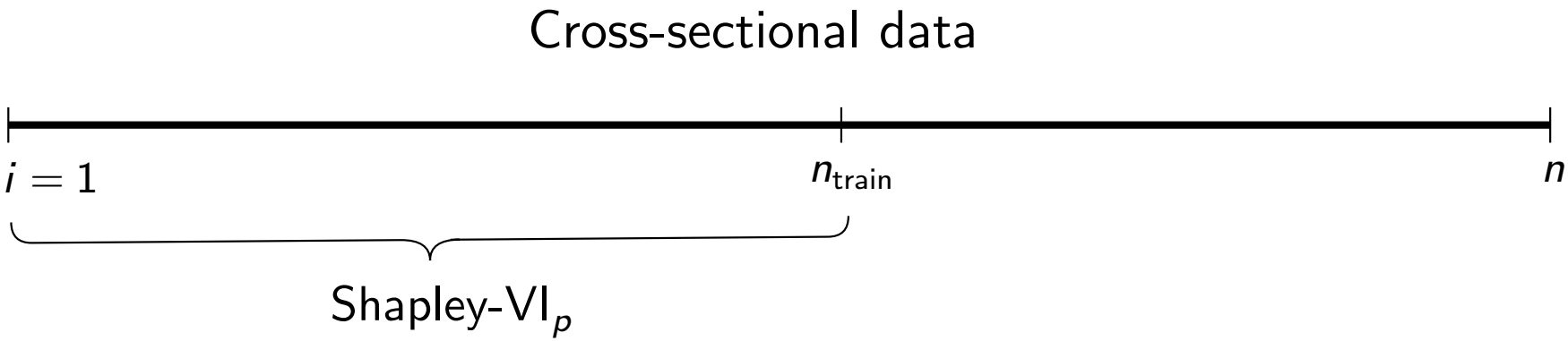
  - $\text{Post}_p(\mathcal{O}) \Rightarrow$ set of indices that follow p in $\mathcal{O}$

- Approximate the effect of removing predictors not in the coalition by replacing them with background data from the training sample (Štrumbelj & Kononenko 2014, Lundberg & Lee 2017)

- $\hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h) = \frac{1}{2M} \sum_{m=1}^{2M} \theta_{p,m}(\boldsymbol{x}_t\,; W_i, h)$
  - $M \Rightarrow$ number of random draws

- Increase the algorithm's computational/estimation efficiency
  - Compute $\theta_{p,m}(\boldsymbol{x}_t\,; W_i, h)$ for each predictor $p \in S$ for a given random draw m (Castro, Gómez & Tejada 2009)
  - Antithetical sampling $\Rightarrow$ reverse the order of $\mathcal{O}_m$ to reduce the variance (Mitchell et al 2022)
- Efficiency property holds for $\hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h)$
  - $\sum_{p \in S} \hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h) = \hat{f}(\boldsymbol{x}_t\,; W_i, h) - \underbrace{\bar{\hat{f}}(W_i, h)}_{\hat{\phi}_\emptyset(W_i, h)}$
    - $\hat{\phi}_\emptyset(W_i, h) \Rightarrow$ average in-sample prediction for the model trained using sample W_i, (ie, baseline/unconditional forecast based on the empty coalition set)

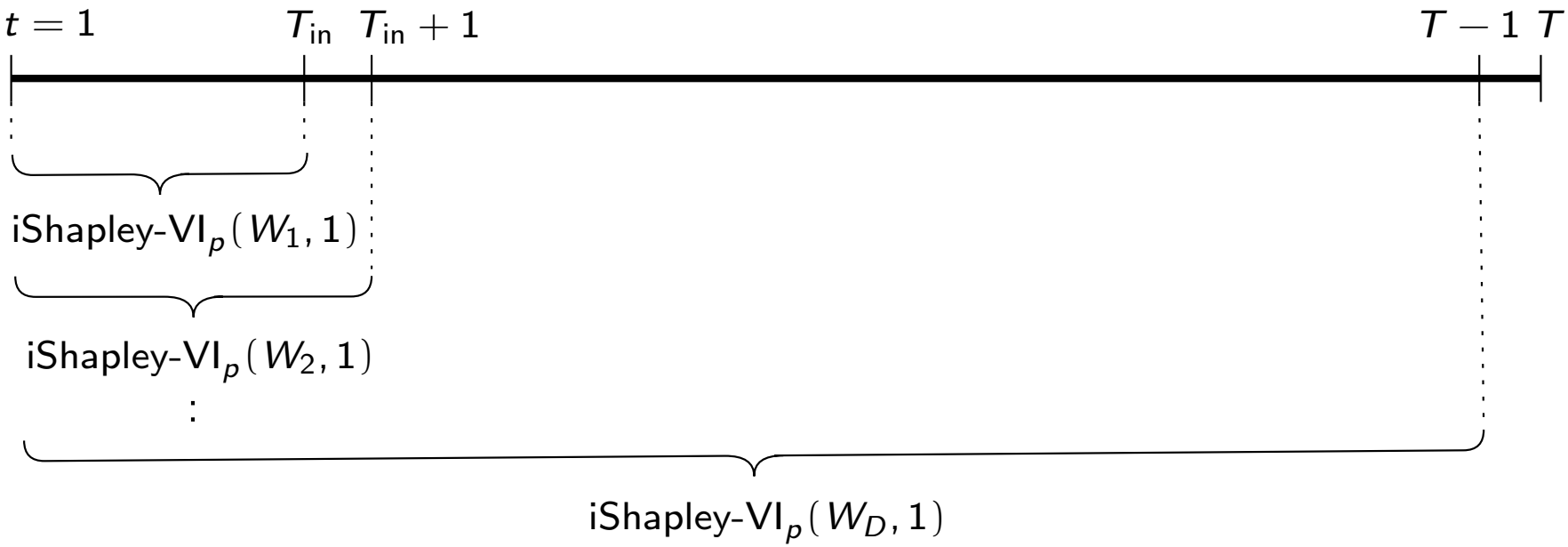
- $\hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h)$ for $p \in S$ and $t \in W_i$
  - Local measure $\Rightarrow$ contribution of predictor p to the prediction corresponding to instance $\boldsymbol{x}_t$ in the training sample
- Also want a global measure of predictor p's importance
  - Take the average of the absolute values of the Shapley values for p across all of the training sample observations
    - Shapley-VI$_p(W_i, h) = \frac{1}{|W_i|} \sum_{t \in W_i} \left| \hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h) \right|$
  - Popular metric for assessing predictor importance
    - Straightforward if there is only a single training sample, but we typically have multiple training samples for time-series data
    - More appropriate for cross-sectional data


Cross-sectional data
$i = 1$
n_{train}
n
Shapley-VI$_p$

- Typically retrain a prediction model on a regular basis using data available at the time of forecast formation with an expanding/rolling window
- Total available observations $\Rightarrow t = 1, \ldots, T$
  - Initial in-sample period $\Rightarrow$ ends in $t = T_{\text{in}}$
  - Out-of-sample period $\Rightarrow$ remaining $T - T_{\text{in}} = D$ observations
- Sequence of out-of-sample forecasts
  - $\underbrace{\hat{y}_{T_{\text{in}}+1:T_{\text{in}}+h},\ \hat{y}_{T_{\text{in}}+2:T_{\text{in}}+h+1},\ \ldots,\ \hat{y}_{T-(h-1):T}}_{D-(h-1)\text{ time-series forecasts}}$
- Set of training samples $\Rightarrow W = \{W_1, \ldots, W_{D-(h-1)}\}$
  - Based on an expanding/rolling window

- Interested in the importance of a predictor for the entire sequence of fitted models used to generate the sequence of time-series forecasts
- Take the average Shapley-VI$_p(W_i, h)$ across all of the training samples used to generate the sequence of time-series forecasts
  - iShapley-VI$_p(W, h) = \frac{1}{|W|} \sum_{i \in W}$ Shapley-VI$_p(W_i, h)$

Expanding window, $h = 1$



$$\text{iShapley-VI}_p(W, 1) = \frac{1}{D} \sum_{i=1}^{D} \text{iShapley-VI}_p(W_i, 1)$$

- ith forecast $\Rightarrow \hat{y}_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)} = \hat{f}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)$
  - $i = 1, \ldots, D-(h-1)$
  - $\boldsymbol{x}_{T_{\text{in}}+(i-1)} \Rightarrow$ vector of predictors plugged into the fitted prediction model trained with W_i used to compute the ith out-of-sample forecast
- Shapley value for predictor p and ith out-of-sample forecast
  - $\phi_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right) =$

$$\frac{1}{P!} \sum_{\mathcal{O} \in \pi(P)} \left[\xi_{\text{Pre}_p(\mathcal{O})\cup\{p\}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right) - \xi_{\text{Pre}_p(\mathcal{O})}\left(\boldsymbol{x}_t; W_i, h\right)\right]$$

- Modify the algorithm to estimate $\phi_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)$
- For a random draw $m \Rightarrow$

$$\theta_{p,m}^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right) =$$

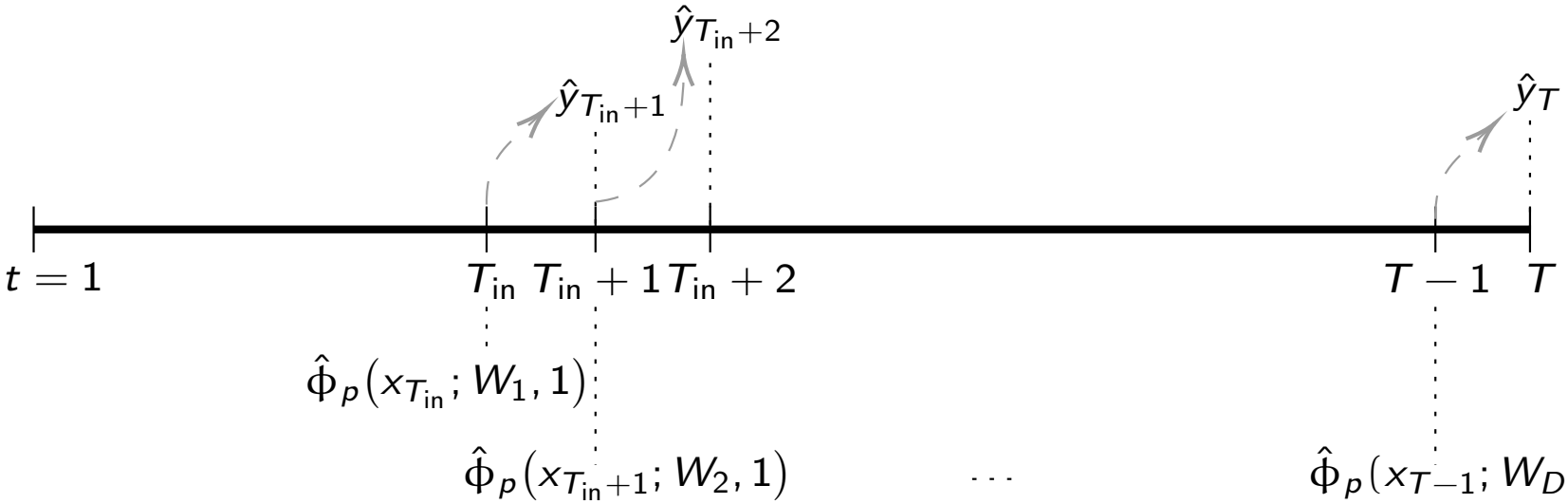
$$\frac{1}{|W_i|}\sum_{s \in W_i}\Big[\hat{f}\left(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m) \cup \{p\},\ \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m); W_i, h\right) -$$

$$\hat{f}\left(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m),\ \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m) \cup \{p\}; W_i, h\right)\Big]$$

  - Approximate the effect of removing predictors not in the coalition by replacing them with background data from the training sample (W_i)
    - Thus, we remain "true to the model"
- $\hat{\phi}_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right) = \frac{1}{2M}\sum_{m=1}^{2M}\theta_{p,m}^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)$
  - Contribution of predictor p to the ith out-of-sample forecast

- Interested in the importance of a predictor for the entire sequence of time-series forecasts
- Average $\left|\hat{\phi}_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)\right|$ across the sequence of time-series forecasts
  - oShapley-VI$_p(W, h) = \frac{1}{|W|}\sum_{i \in W}\left|\hat{\phi}_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)\right|$
- Compare iShapley-VI$_p(W, h)$ to oShapley-VI$_p(W, h)$
  - Compare the relative importance of predictor p over the in-sample/out-of-sample periods according to the Shapley-based variable importance measures

Expanding window, $h = 1$



$$\text{oShapley-VI}_p(W, 1) = \frac{1}{D}\sum_{i=1}^{D}\left|\hat{\phi}_p\left(x_{T_{\text{in}}+(i-1)}; W_i, 1\right)\right|$$

- Main methodological contribution $\Rightarrow$ global PBSV$_p$
  - Average loss over the out-of-sample period attributable to individual predictor $p \in S$
    - Takes realized—not just fitted—value into account
    - Key insight $\Rightarrow$ wrap the loss function around the forecast in $\theta_{p,m}^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}, W_i, h\right)$
    - Anatomizes out-of-sample forecasting accuracy
- Generic loss function
  - $L\Big(\underbrace{y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}}_{\text{realized value}}, \underbrace{\hat{f}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)}_{\text{forecast}}\Big)$

- Modify the algorithm to compute the local PBSV_p
- For a random draw $m \Rightarrow$

$$\theta^{\text{out}}_{p,m}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h, L\right) =$$

$$L\left(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \frac{1}{|W_i|}\sum_{s\in W_i}\hat{f}\left(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m)\cup\{p\}, \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m); W_i, h\right)\right) -$$

$$L\left(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \frac{1}{|W_i|}\sum_{s\in W_i}\hat{f}\left(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m), \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m)\cup\{p\}; W_i, h\right)\right)$$

  - Continue to remain true to the model by approximating the effect of removing predictors not in the coalition by replacing them with background data from the training sample (W_i)
- $\hat{\phi}^{\text{out}}_p\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h, L\right) = \frac{1}{2M}\sum_{m=1}^{2M}\theta^{\text{out}}_{p,m}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h, L\right)$
  - Contribution of predictor p to the loss corresponding to the ith out-of-sample forecast

- Efficiency property holds for $\hat{\Phi}_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h, L\right)$
  - $\sum_{p \in S} \hat{\Phi}_p^{\text{out}}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h, L\right) =$

    $L\left(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \hat{f}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)\right) - \hat{\Phi}_{\emptyset}^{\text{out}}(W_i, h, L)$
    - $\hat{\Phi}_{\emptyset}^{\text{out}}(W_i, h, L) \Rightarrow$ loss for the baseline or unconditional prediction based on the empty coalition set
- Shapley value logic $\Rightarrow$ local PBSV_p fairly allocates the loss among the predictors for the ith out-of-sample forecast

- Global $\text{PBSV}_p \Rightarrow$ for the sequence of out-of-sample forecasts
- Modify the algorithm to estimate the global PBSV_p
- For a random draw $m \Rightarrow$

$$\theta^{\text{out}}_{p,m}(W, h, L) =$$

$$\frac{1}{|W|}\sum_{i \in W} L\left(\text{realized}_i, \frac{1}{|W_i|}\sum_{s \in W_i} \hat{f}\left(\boldsymbol{x}_{j, T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m) \cup \{p\}, \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m); W_i, h\right)\right) -$$

$$\frac{1}{|W|}\sum_{i \in W} L\left(\text{realized}_i, \frac{1}{|W_i|}\sum_{s \in W_i} \hat{f}\left(\boldsymbol{x}_{j, T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m), \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m) \cup \{p\}; W_i, h\right)\right)$$

- $\hat{\phi}^{\text{out}}_p(W, h, L) = \frac{1}{2M}\sum_{m=1}^{2M} \theta^{\text{out}}_{p,m}(W, h, L)$
- Efficiency property holds for $\hat{\phi}^{\text{out}}_p(W, h, L)$
  - $\sum_{p \in S} \hat{\phi}^{\text{out}}_p(W, h, L) =$
    $\frac{1}{|W|}\sum_{i \in W} L\left(\text{realized}_i, \hat{f}\left(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\right)\right) - \hat{\phi}^{\text{out}}_{\emptyset}(W, h, L)$

- Global PBSV_p for $p \in S$
  - Decomposes the loss for the sequence of out-of-sample forecasts into the contributions of each of the P predictors
  - Shapley value logic $\Rightarrow$ fairly assesses how the individual predictors contribute to out-of-sample forecasting accuracy
  - Applies to any fitted prediction model
    - Linear/nonlinear, parametric/nonparametric
    - Just need the inputs for the `predict` function
  - Applies to any loss function
    - MSE/RMSE/MAE
    - R^2_{OS} (Fama & French 1989, Campbell & Thompson 2008)
- `Python` package to compute $\text{PBSV}_p \Rightarrow$ `anatomy`

- Nonlinear machine-learning methods improve inflation forecasts, especially at longer horizons (eg, Medeiros et al 2021, Goulet Coulombe 2022, Goulet Coulombe et al 2022, Hauzenberger, Huber & Klieber 2023)
- Target $\Rightarrow \pi_{t+1:t+h} = \frac{1}{h}\sum_{k=1}^{h} \pi_{t+k}$
  - $\pi_t = \log(\text{CPI}_t) - \log(\text{CPI}_{t-1})$
- Benchmark forecast $\Rightarrow$ AR model
  - Select the AR lag length via the BIC (max lag of 12)
  - Standard benchmark in macro, including for inflation (eg, Kotchoni, Leroux & Stevanovic 2019, Medeiros et al 2021)

- Prediction model $\Rightarrow \pi_{t+1:t+h} = f\left(\underbrace{\boldsymbol{\pi}_t^{\text{AR}}, \boldsymbol{w}_t, \boldsymbol{w}_t^{\text{MA}(q)}}_{\boldsymbol{x}_t}\right) + \varepsilon_{t+1:t+h}$
  - AR component $\Rightarrow \boldsymbol{\pi}_t^{\text{AR}} = \begin{bmatrix} \pi_t & \cdots & \pi_{t-L} \end{bmatrix}'$
    - Set $L = 11$ (corresponding to 12 lags of inflation)
  - Vector of predictors $\Rightarrow \boldsymbol{w}_t$
  - MAs of predictors (Goulet Coulombe et al 2021)
    - $\boldsymbol{w}_t^{\text{MA}(q)} = \frac{1}{q}\sum_{k=1}^{q} \boldsymbol{w}_{t-(k-1)} \Rightarrow$ set $q = 3$
- Linear model $\Rightarrow \pi_{t+1:t+h} = \alpha + \boldsymbol{x}_t' \boldsymbol{\beta} + \varepsilon_{t+1:t+h}$
- OLS forecast $\Rightarrow \hat{\pi}_{t+1:t+h}^{\text{OLS}} = \hat{\alpha}^{\text{OLS}} + \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}^{\text{OLS}}$
  - Prone to overfitting $\Rightarrow$ poor out-of-sample performance

- Principal component regression (Stock & Watson 2002)
  - Dimension-reduction technique
  - PCR forecast $\Rightarrow \hat{\pi}^{\text{PCR}}_{t+1:t+h} = \hat{\alpha}^{\text{OLS}}_{z} + \hat{\boldsymbol{z}}_t' \hat{\boldsymbol{\beta}}^{\text{OLS}}_{z}$
    - $\hat{z}_t \Rightarrow$ vector of first $C \ll P$ PCs from $\boldsymbol{x}_t$
  - Select L (max $= 11$) and C (max $= 10$) via $\bar{R}^2$
- Elastic net (Zhou & Hastie 2005) estimation of linear model
  - Refinement of the seminal LASSO (Tibshirani 1996)
  - Penalized regression $\Rightarrow$ shrink the estimated slope coefficients toward zero to guard against overfitting
    - Includes both ℓ_1 (LASSO) and ℓ_2 (ridge) components in the penalty term
  - ENet forecast $\Rightarrow \hat{\pi}^{\text{ENet}}_{t+1:t+h} = \hat{\alpha}^{\text{ENet}} + \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}^{\text{ENet}}$

- Random forest (Breiman 2001)
  - Regression tree forecast $\Rightarrow \hat{\pi}^{RT}_{t+1:t+h} = \sum_{u=1}^{U} \bar{\pi}_u 1_u(\boldsymbol{x}_t\,;\boldsymbol{\eta}_u)$
    - $U \Rightarrow$ number of "terminal nodes" (ie, "leaves")
    - $1_u(\boldsymbol{x}_t\,;\boldsymbol{\eta}_u) = 1$ if $\boldsymbol{x}_t \in R_u(\boldsymbol{\eta}_u)$ for the uth region denoted by R_u
    - $\bar{\pi}_u \Rightarrow$ average value of the target observations in R_u for the training sample based on data through t
  - RF forecast $\Rightarrow \hat{\pi}^{RF}_{t+1:t+h} = \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{u=1}^{U} \bar{\pi}_u^{(b)} 1_u^{(b)}(\boldsymbol{x}_t\,;\boldsymbol{\eta}_u)\right]$
    - To reduce variance, average many "deep" regression tree forecasts based on bootstrapped samples
  - Strong record in macro forecasting (eg, Medeiros et al 2021, Borup & Schütte 2022, Goulet Coulombe et al 2022)

- XGBoost (Chen & Guestrin 2016) $\Rightarrow$ boosted tree
  - Regression tree approach based on gradient boosting (Breiman 1997, Friedman 2001)
  - Additive prediction function $\Rightarrow \hat{f}(\boldsymbol{x}_t ; \hat{\boldsymbol{\eta}}) = \sum_{j=1}^{J} \hat{f}_j(\boldsymbol{x}_t ; \hat{\boldsymbol{\eta}}_j)$
    - $\hat{f}_j(\boldsymbol{x}_t ; \hat{\boldsymbol{\eta}}_j) \Rightarrow$ "weak" learner with a low variance but a potentially sizable bias
    - Add another tree that is trained using the residuals from the previous function in the sequence to improve the fit (ie, reduce the bias)
  - Stochastic gradient boosting (Friedman 2002)
    - Makes boosting more robust by training each element in the sequence on a randomly drawn (without replacement) subsample of the data

- Neural network $\Rightarrow$ can approximate any smooth function
  - Input layer $\Rightarrow$ predictors
  - $L \geqslant 1$ hidden layers, each with P_l neurons $\Rightarrow$ each neuron takes signals from the neurons in the previous layer and generates a new signal via a nonlinear activation function
    - $h_m^{(l)} = g\left(\omega_{m,0}^{(l)} + \sum_{j=1}^{P_{l-1}} \omega_{m,j}^{(l)} h_j^{(l-1)}\right)$
    - ReLU activation function $\Rightarrow g(x) = \max\{x, 0\}$
  - Output layer $\Rightarrow$ takes signals from the last hidden layer and converts them to a prediction
    - $\hat{\pi}_{t+1:t+h}^{\text{NN}} = \omega_0^{(L+1)} + \sum_{j=1}^{P_L} \omega_j^{(L+1)} h_j^{(L)}$
  - Ensemble $\Rightarrow$ average of shallow/deep NN forecasts

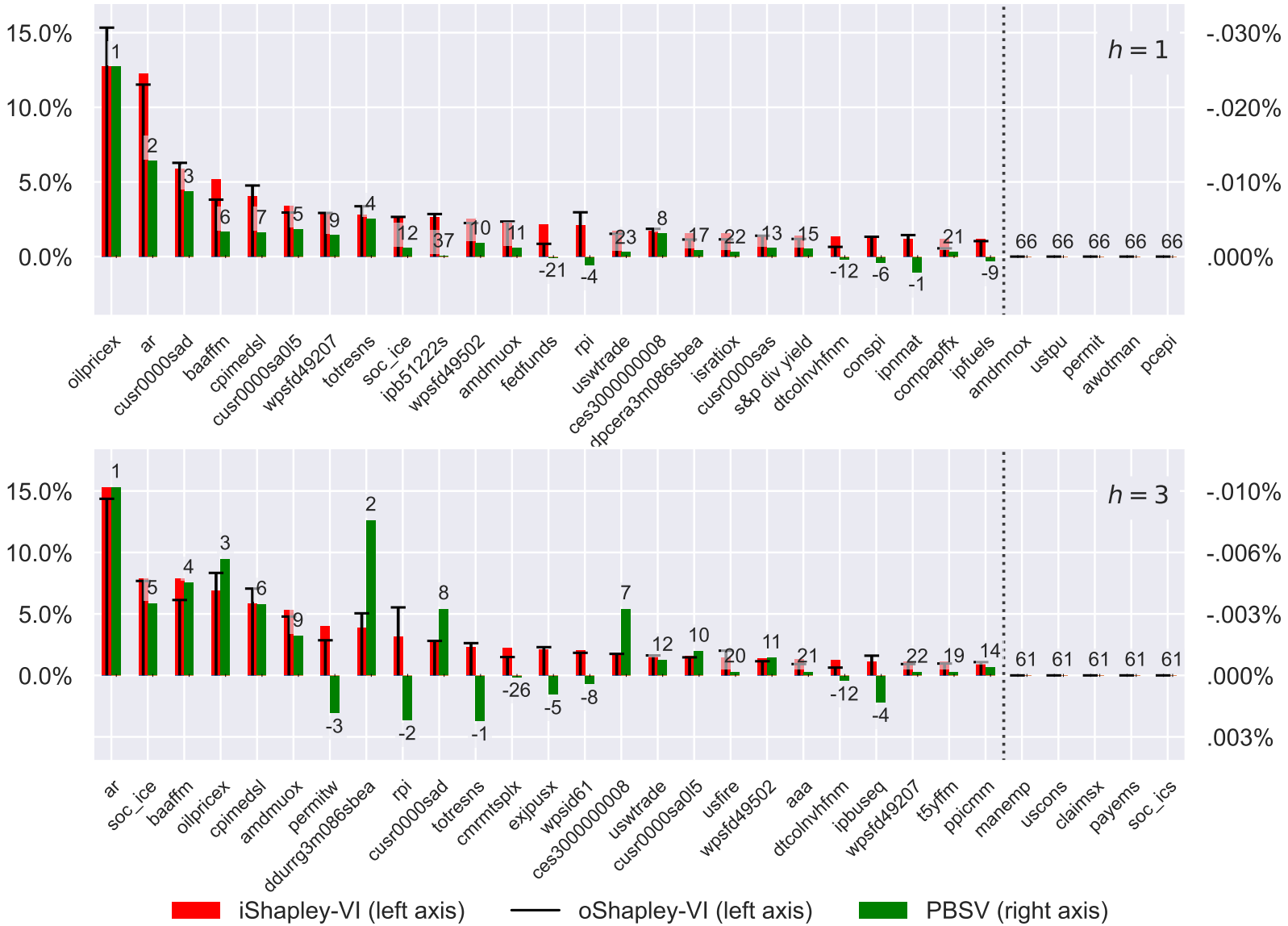
- Ensemble forecasts
  - Ensemble-linear $\Rightarrow$ average of PCR/ENet
  - Ensemble-nonlinear $\Rightarrow$ average of RF/XGBoost/NN
  - Ensemble-all $\Rightarrow$ average of PCR/ENet/RF/XGBoost/NN

- 118 variables from FRED-MD (McCracken & Ng 2016)
  - Output/income, labor market, housing, consumption/orders/inventories, money/credit, interest/exchange rates, prices, stock market
- 3 variables from Univ of Michigan Survey of Consumers
  - Index of consumer sentiment, index of consumer expectations, index of current economic conditions
- Initial in-sample period $\Rightarrow$ 1960:01–1989:12
- Out-of-sample period $\Rightarrow$ 1990:01–2022:12
- All forecasts are based on a rolling estimation window

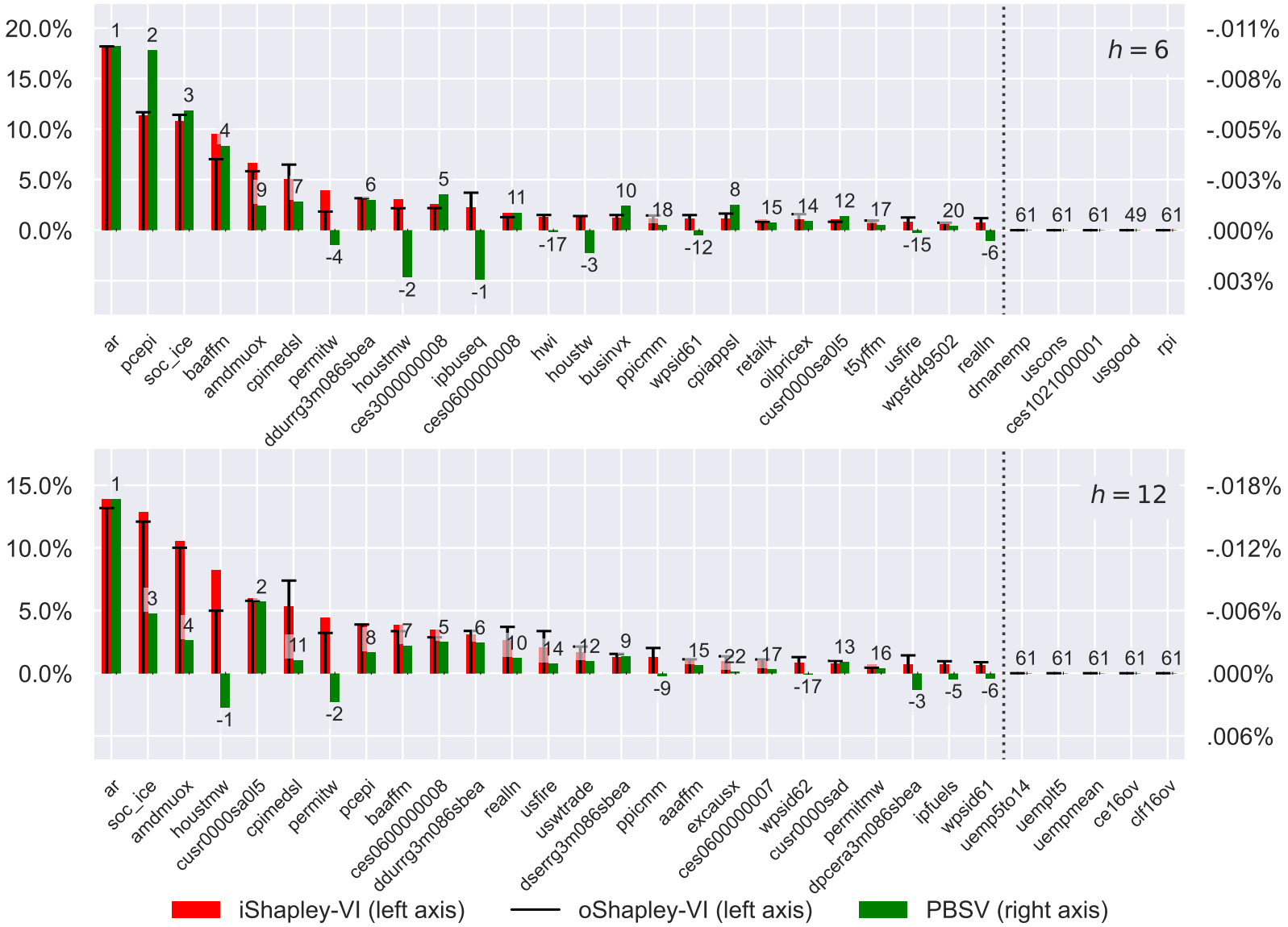
RMSE ratios vis-à-vis AR benchmark

Forecast	$h = 1$	$h = 3$	$h = 6$	$h = 12$
AR RMSE	0.26%	0.23%	0.20%	0.16%
PCR	1.08	1.01	0.96	0.92**
ENet	0.93**	0.95*	0.96	0.94
Random forest	0.96	0.97	0.92*	0.82***
XGBoost	1.00	0.98	0.91**	0.85***
Neural network	0.94**	0.93**	0.94	0.83***
Ensemble-linear	0.96	0.96	0.93*	0.90**
Ensemble-nonlinear	0.93**	0.93**	0.93**	0.81***
Ensemble-all	0.93**	0.93**	0.90**	0.84***

ENet $\Rightarrow \hat{\Phi}_p^{\text{out}}(W, h, \text{RMSE})$



ENet $\Rightarrow \hat{\phi}_p^{\text{out}}(W, h, \text{RMSE})$ (cont'd)

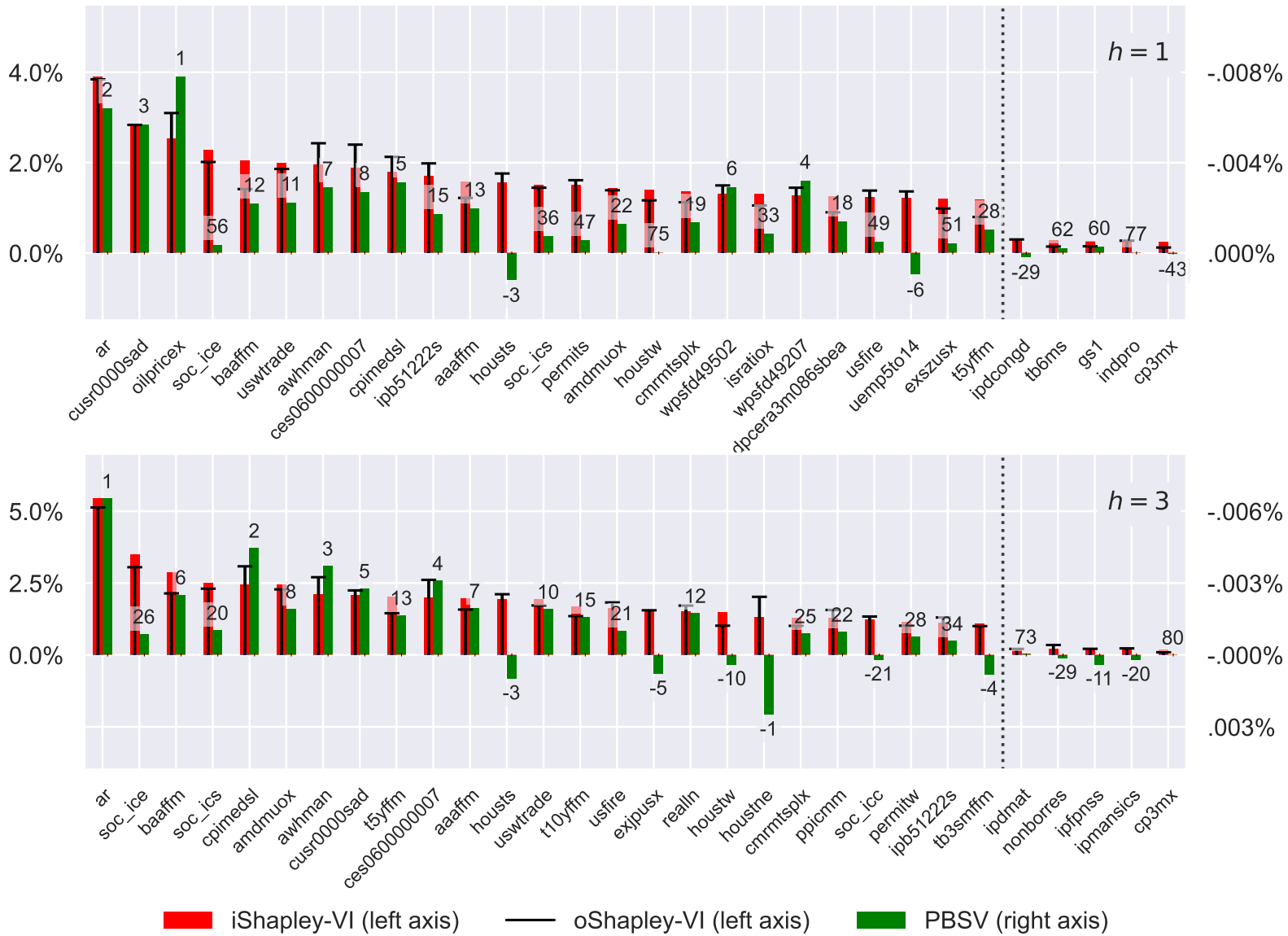


- iShapley-VI/oShapley-VI $\Rightarrow$ reasonably close correspondence between the in-sample/out-of-sample variable importance
  - Perhaps not surprising $\Rightarrow$ in-sample/out-of-sample predictions are based on the same fitted models
- Considerable accord across the in-sample iShapley-VI and the out-of-sample PBSV for a number of predictors
  - AR component (`ar`)
  - Price of oil (`oilpricex`)
  - CPI: durables (`cusr0000sad`)
  - Index of consumer expectations (`soc_ice`)
  - PCE price index: durable goods (`ddurrg3m086sbea`)
  - CPI: medical services (`cpimedsl`)
  - Interest-rate spread (`baaffm`)

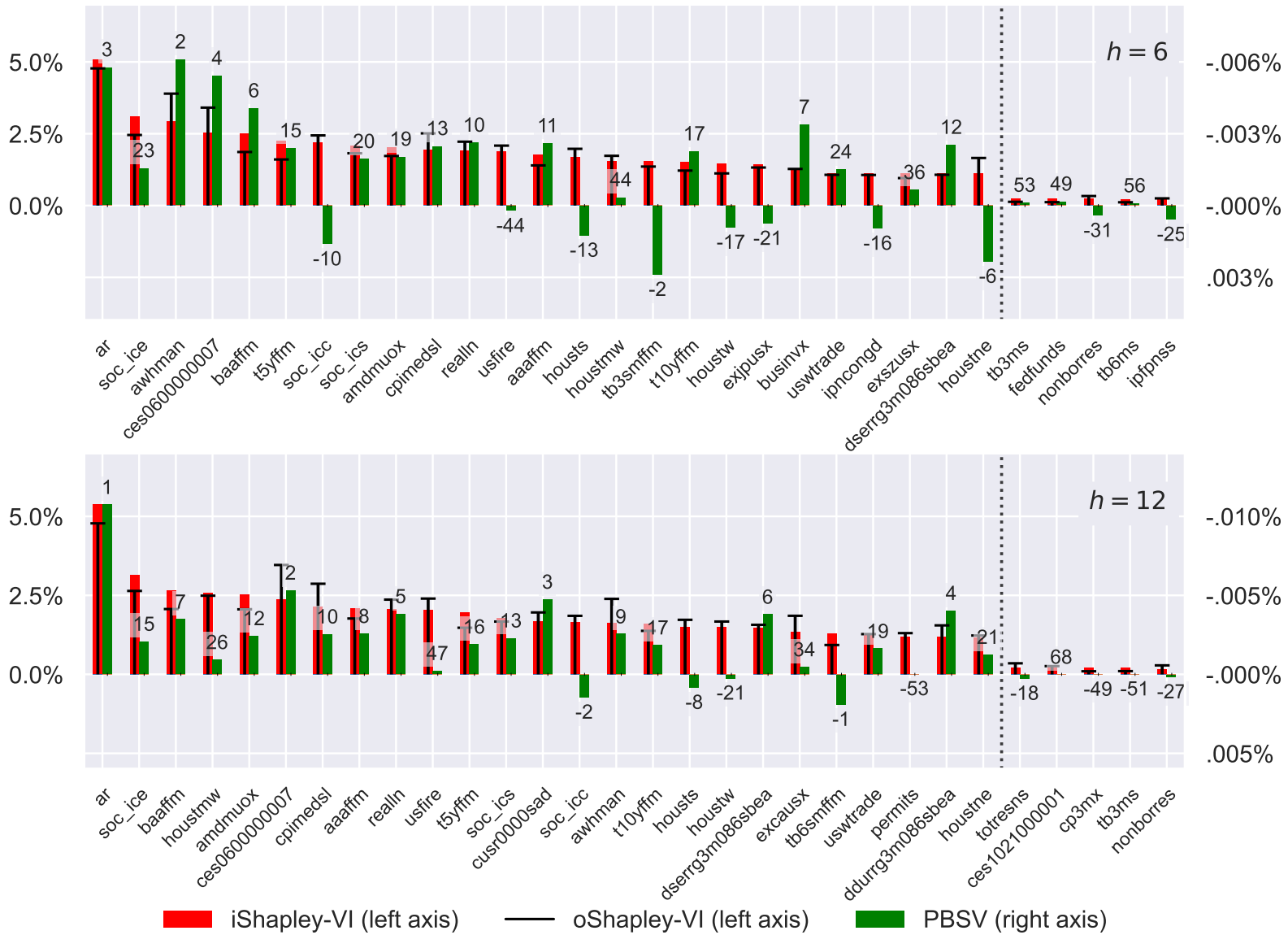
- But there are also major points of discord between iShapley-VI/PBSV
  - Real personal income (`rpi`)
  - Industrial production: materials (`ipmat`)
  - Total reserves of depository institutions (`totresns`)
  - New housing permits: West (`permitw`)
  - Industrial production: business equipment (`ipbuseq`)
  - Housing starts: Midwest and West (`houstmw` and `houstw`)
  - Real estate loans (`realln`)
  - Real manufacturing and trade sales (`dpcera3m086sbea`)
  - Industrial production: fuels (`ipfuels`)
  - PPI: intermediate materials (`wpsid61`)

Neural Network PBSVs

Neural network $\Rightarrow \hat{\Phi}_p^{\text{out}}(W, h, \text{RMSE})$



Neural network $\Rightarrow \hat{\Phi}_p^{\text{out}}(W, h, \text{RMSE})$ (cont'd)

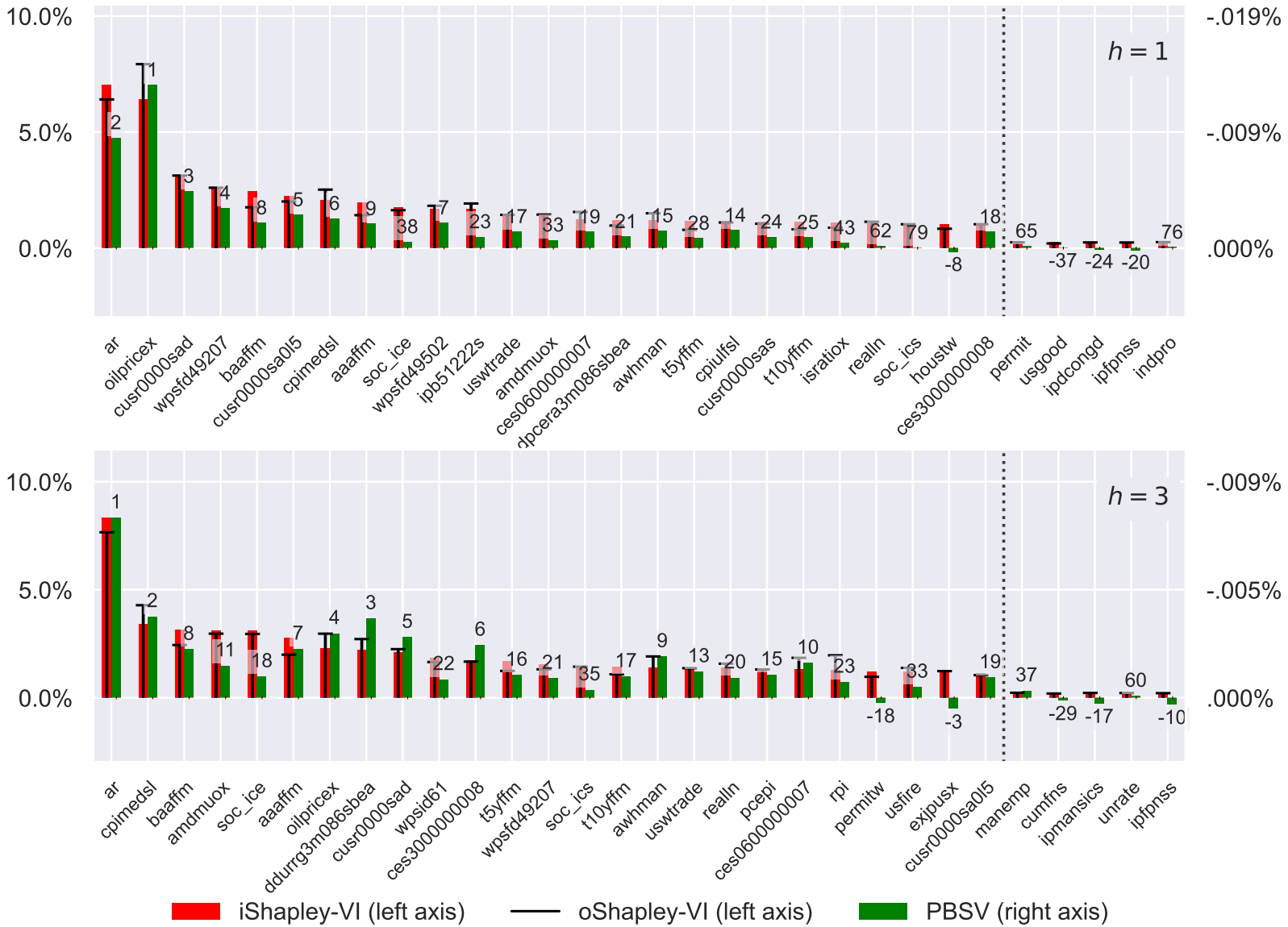


- iShapley-VI$_p$/oShapley-VI$_p$ align relatively closely
- Considerable correspondence between the in-sample iShapley-VI and the out-of-sample PBSV for many predictors
  - AR component
  - Price of oil
  - CPI: durables
  - Average weekly hours in manufacturing (`awhman`)
  - CPI: medical services

- Nevertheless, there are important divergences between iShapley-VI/PBSV for a number of predictors
  - Housing starts: South and Northeast (`housts` and `houstne`)
  - Two interest rate spreads (`tb3smffm` and `tb6smffm`)
  - Yen-USD exchange rate (`exjpusx`)
  - Number of unemployed for 5–14 weeks (`uemp5to14`)
  - Index of current economic conditions (`soc_icc`)

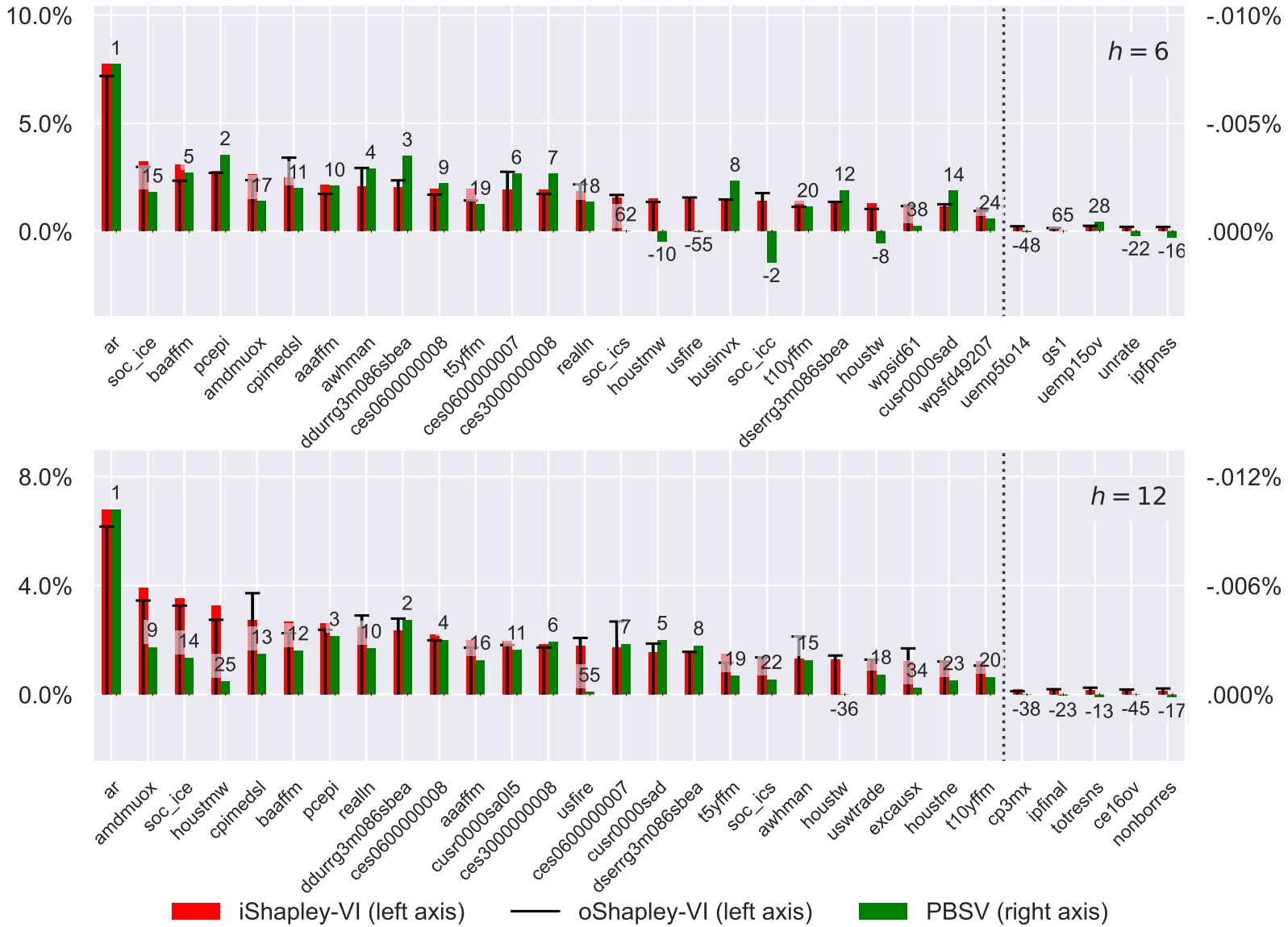
Ensemble-All PBSVs

Ensemble-all $\Rightarrow \hat{\Phi}_p^{\text{out}}(W, h, \text{RMSE})$



Ensemble-All PBSVs

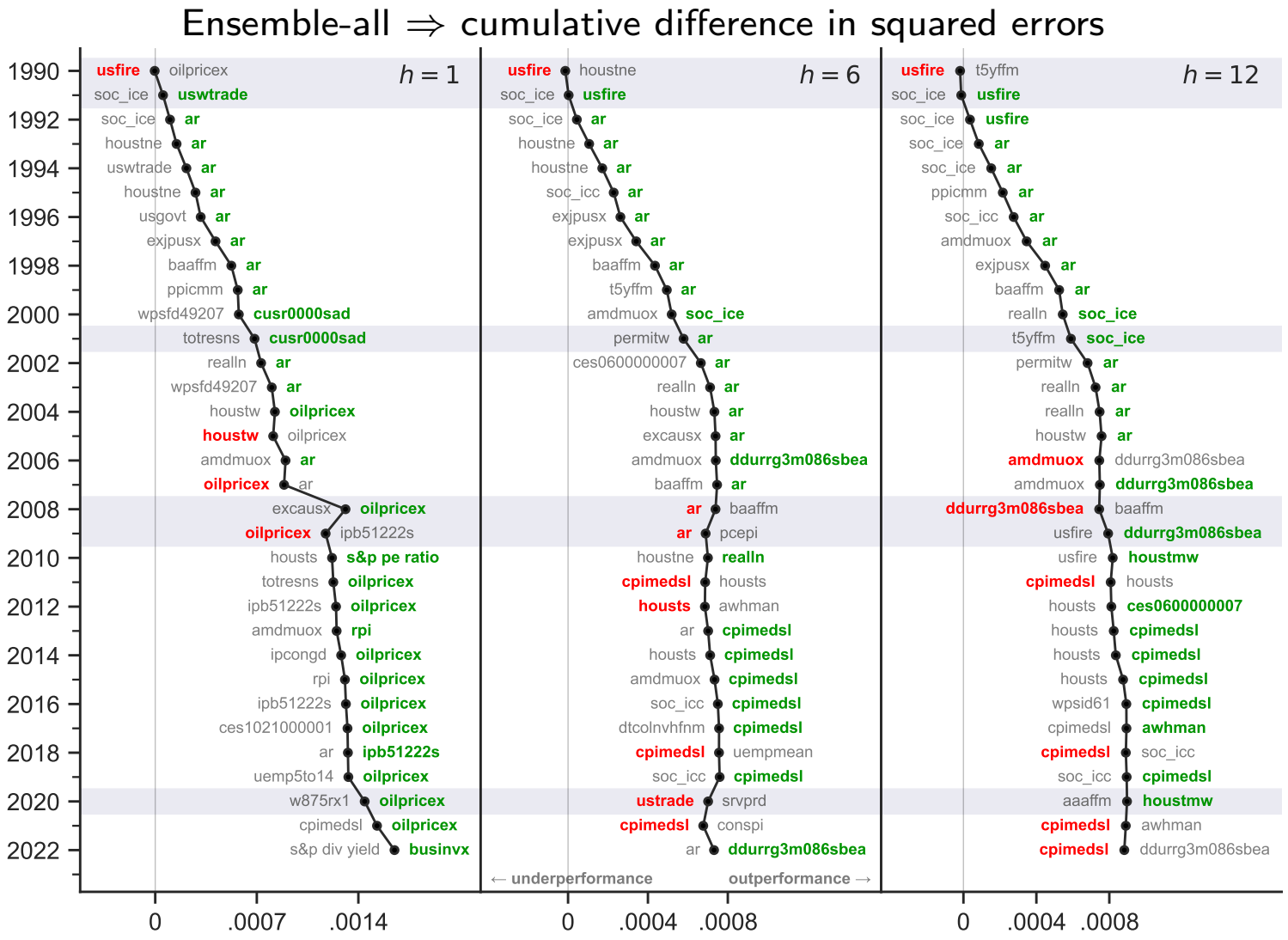
Ensemble-all $\Rightarrow \hat{\Phi}_p^{\text{out}}(W, h, \text{RMSE})$



- iShapley-VI/oShapley-VI match up quite closely
- Generally greater alignment between oShapley-VI/PBSV than for the ENet and neural network
- Still a few noteworthy disparities between iShapley-VI/PBSV
  - Housing starts: West ($h = 1, 6$)
  - Yen-USD exchange rate ($h = 3$)
  - Index of current economic conditions ($h = 6$)
  - Housing starts: Midwest ($h = 6$)

- Cumulative difference in squared errors (Goyal & Welch 2008)
  - Ascertain whether a competing forecast is more accurate than a naïve forecast for any subsample of out-of-sample period
  - If the curve lies more to the right (left) at the end of the interval corresponding to the subsample relative to the beginning, then the competing (naïve) forecast is more accurate in terms of MSE for the subsample
  - Abbreviation to the right (left) of the curve indicates the predictor that contributes the most to positive (negative) performance during a 12-month subsample

Ensemble-All Cumulative Difference in Squared Errors



- Ensemble-all consistently outperforms the naïve forecast
- For numerous 12-month periods before the Great Recession in 2008, the AR component is the predictor most responsible for the outperformance of ensemble-all
- For $h = 1$, there are 11 12-month periods for which the price of oil is the predictor most responsible for the outperformance of ensemble-all
  - Including during the Great Recession and the recent COVID-19 recession as well as the inflation surge starting in mid 2021
  - Consistent with the important influence of energy prices on short-run CPI fluctuations

- CPI: medial services is the leading predictor in terms of the outperformance of ensemble-all for 6 (5) of the 12-month subsamples for $h = 6$ ($h = 12$)
  - Bils & Klenow (2004) and Bryan & Meyer (2010) rank medical services among the stickiest components of the CPI (in terms of its low frequency of price adjustment)
  - Important component in the Atlanta Fed's Sticky-Price CPI
- Multiple cases for $h = 1, 6$ for which housing starts: West is the predictor that contributes the most to negative performance

- Existing model interpretation tools are typically well-suited for forecasts based on cross-sectional data
- We develop Shapley-based metrics for interpreting time-series forecasting models in macro/finance
  - iShapley-VI/oShapley-VI/PBSV
- Main methodological contribution $\Rightarrow$ PBSV
  - Measures the contributions of individual predictors to the out-of-sample loss
  - Anatomizes out-of-sample forecasting accuracy
    - Allows researchers to quantify the roles of predictors in time-series forecasting models along perhaps the most relevant dimension—namely, their contributions to out-of-sample forecasting accuracy

- Empirical application ⇒ forecasting US inflation with a large dataset and machine learning models
  - Close correspondence between iShapley-Vi/oShapley-VI
  - Rankings of many predictors also accord reasonably closely based on in-sample iShapley-VI and out-of-sample PBSV
  - However, there are also substantial discrepancies between iShapley-VI/PBSV
  - Warning ⇒ predictors that are important for determining a model's predicted values aren't necessarily those primarily responsible for the model's out-of-sample forecasting accuracy (even when a forecasting model performs well)
- We created the `Python` package `anatomy` to compute oShapley-VI/PBSV ⇒ check it out