# Arbitrated Ensemble

## for Time Series Forecasting

Vítor Cerqueira, **Luís Torgo**, Fábio Pinto and Carlos Soares

INESCTEC
University of Porto

Jozef Stefan Institute
January, 2018

# Outline

# Outline

# Time Series Forecasting

## Motivation for forecasting

- Time series is an important topic in several research communities;
- Uncertainty due to dynamic characteristics of many phenomena;

## Examples

- financial analysts forecast the behaviour of stock prices for economic profit;
- Intelligent transportation systems forecast the short-term traffic flow to enhance the operational efficiency in road networks

# State-of-the-art

## A plethora of contributions

- Over the last few decades the research community developed several methods for forecasting;

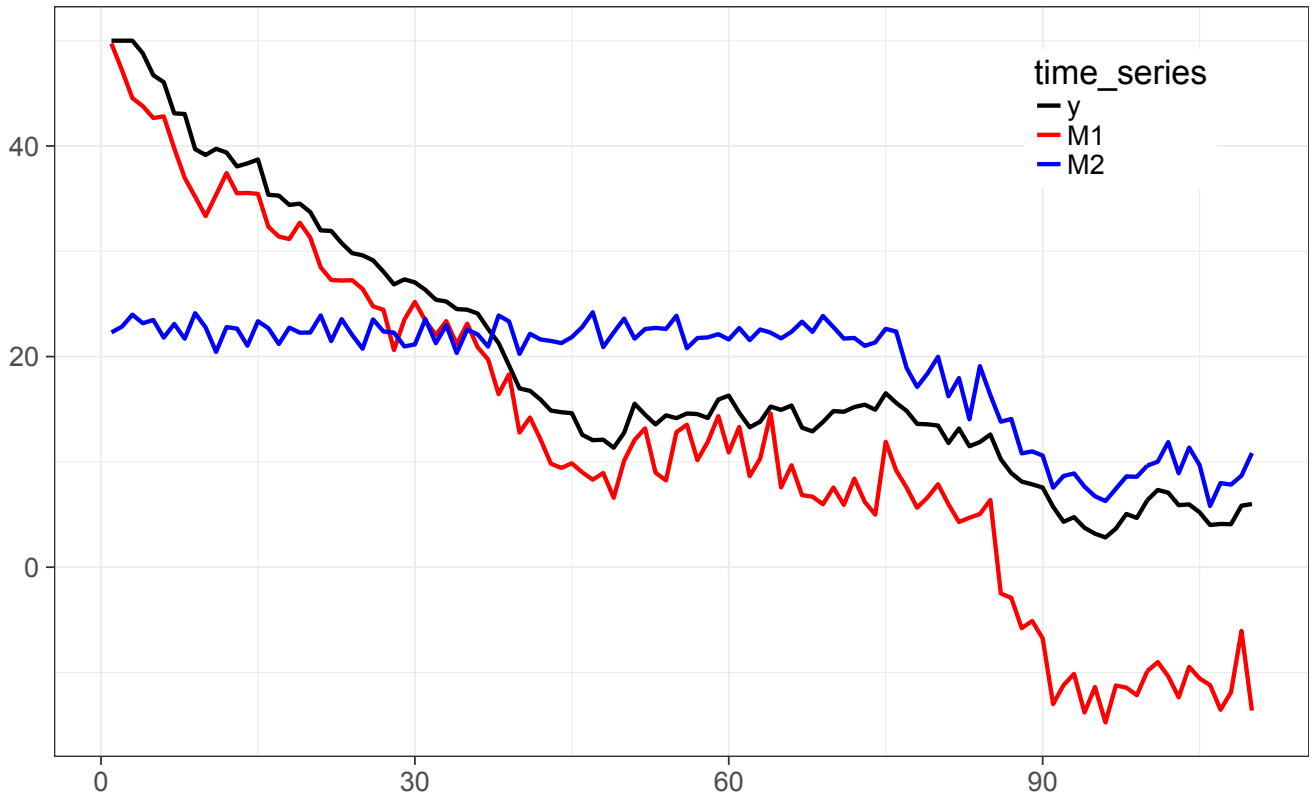
## Forecasting approaches

- Typically, methods are designed to cope with temporal dependencies among observations;
- Time series often comprise non-stationarities and time-evolving complex structures:
  - For example, recurring structures are common due to factors such as seasonality (Gama & Kosina, 2009)

## Combining experts

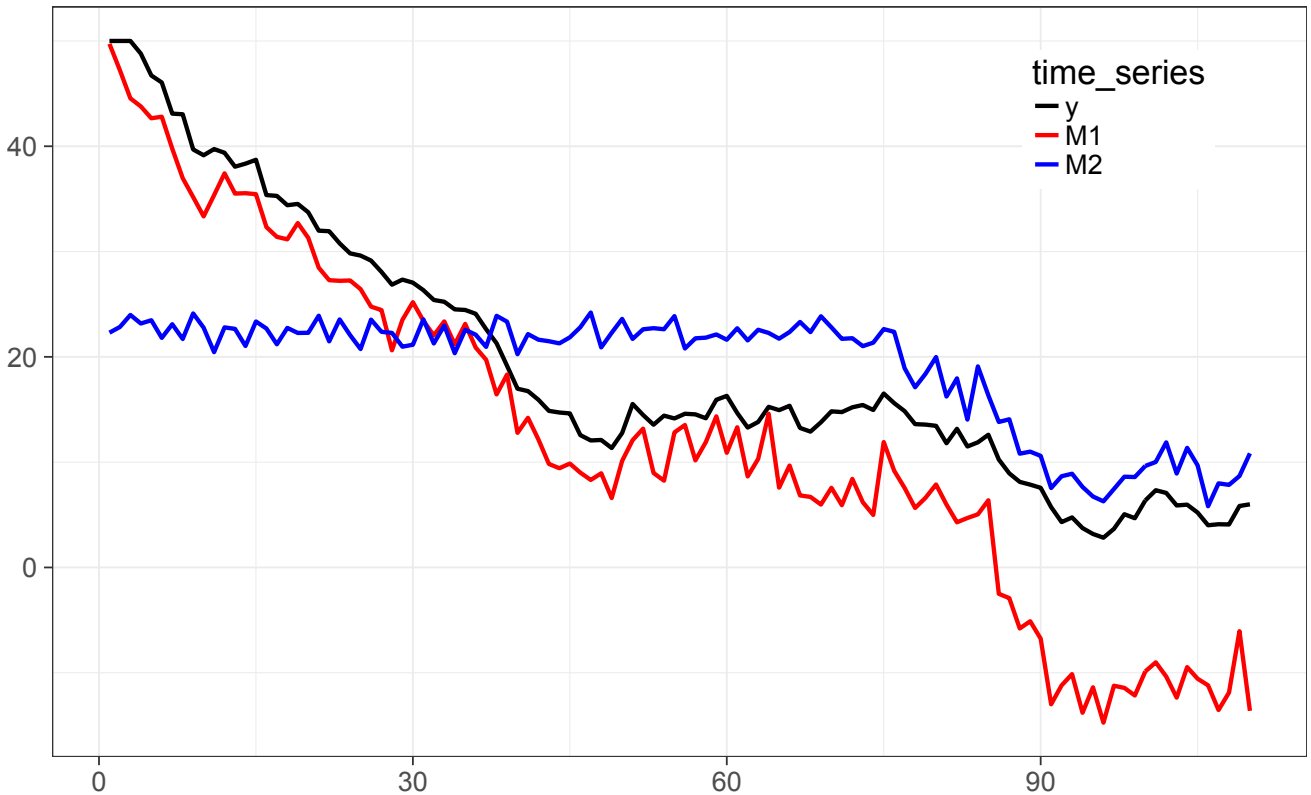
- Selecting the weights of each individual forecaster in the combination rule is known to be a difficult task;
- In time-dependent domains, the loss of each experts is tracked over time and used to adaptively combine them
  - in a window of recent observations (Newbold and Granger, 1974), or other mechanism that promotes recency;
- The idea is that recent observations are more similar to the one we intend to predict, and thus they are considered more relevant;
- Metalearning is also commonly used, for example approaches similar to stacking

# Forecasters' varying relative performance




- Different forecasting models have different areas of expertise and a varying relative performance (Timmermann, 2006);

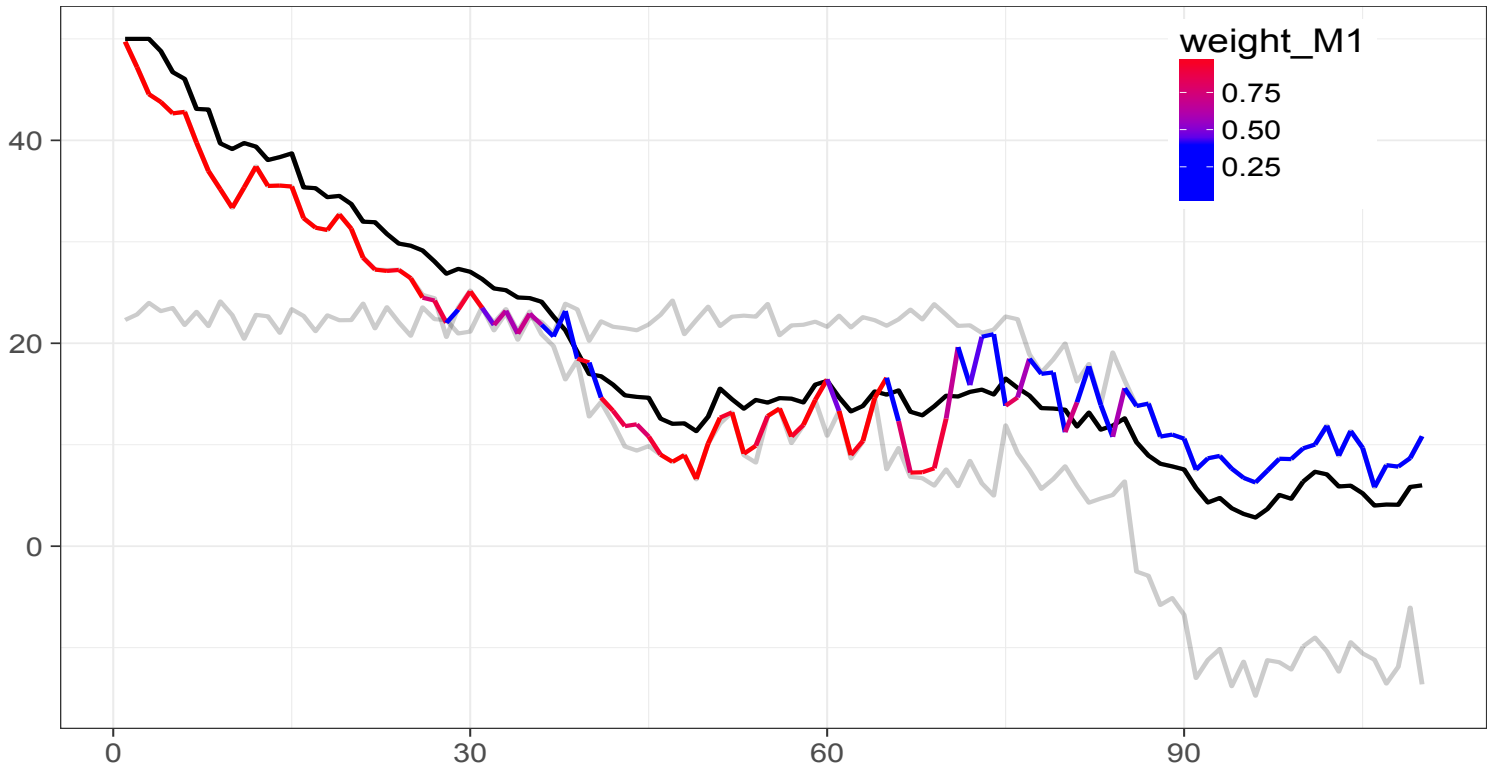
# Main goal




## Main goal

Dynamically combine forecasting models across the time series

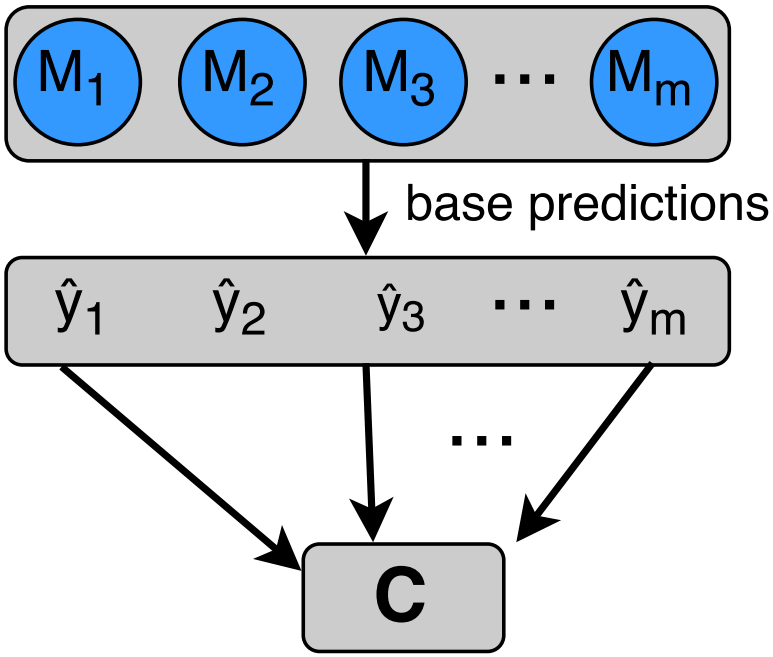
# Our proposal




## A metalearning method ...

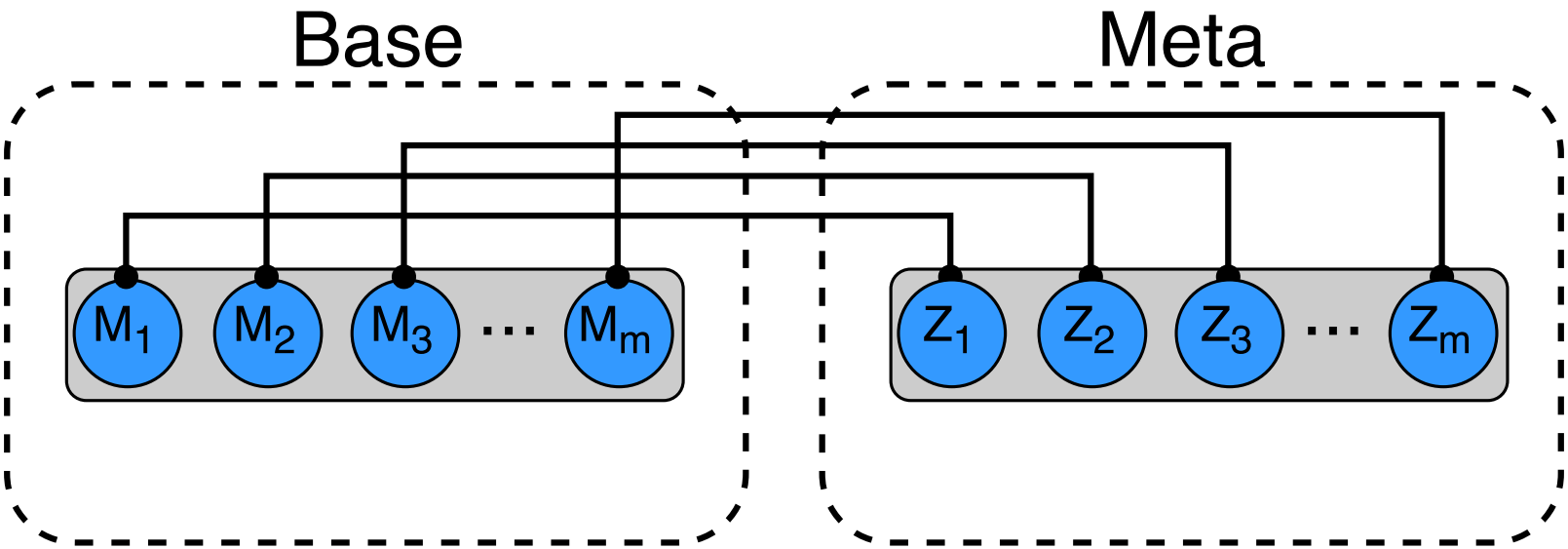
- for dynamic environments
- that performs dynamic weighing at each time step
- and handles the inter-dependency among experts;

# Outline



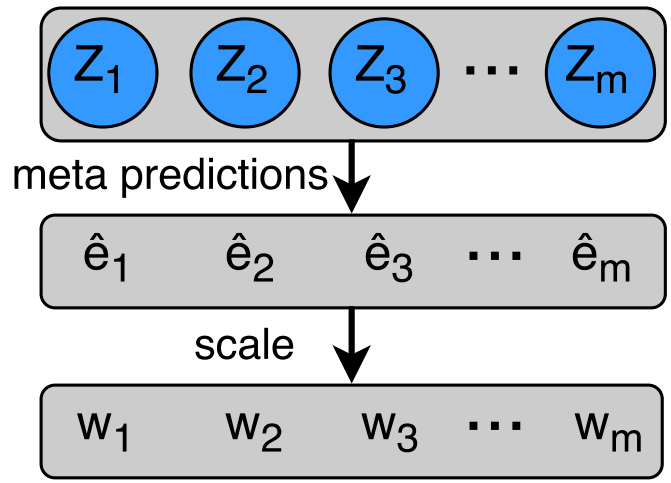

1. Training base models: $Y_{t+1} = f(Y_t, ..., Y_{t-K})$
2. Combine them using some criterion **C**

# Arbitration



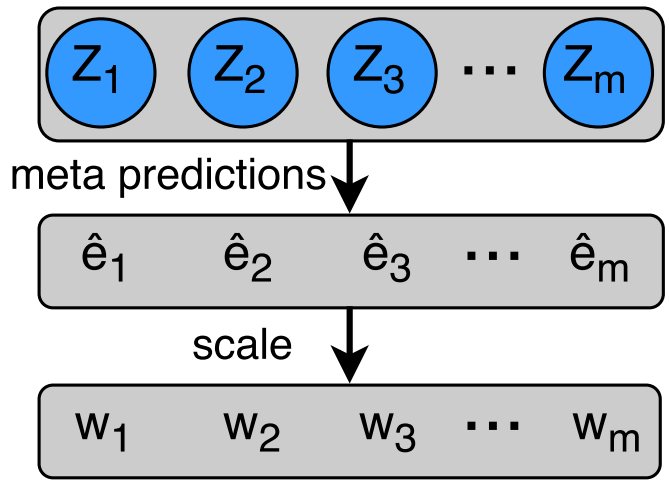

- Metalearning with a meta-learner **for each** base-learner
  - The meta-learners model the error of the base-learners

# Arbitration




2. Training meta-models: $e = f(Y_t, ..., Y_{t-K})$
   A regression analysis to model how the error **e** of a base learner relates to the different dynamics of a time series

# Arbitration




- Arbitrating vs Windowing
  - better long-range modeling
    - recurring concepts

# Related Work

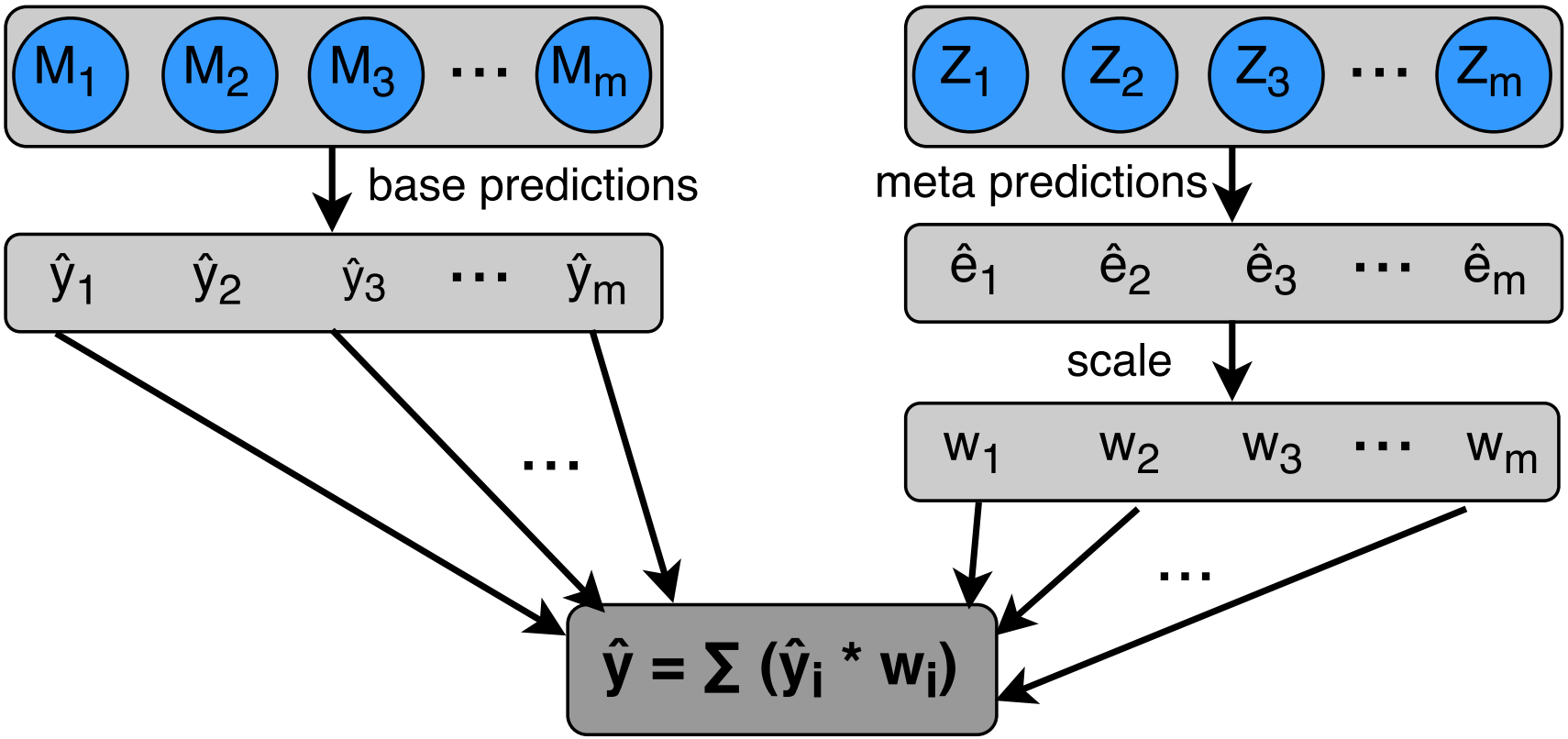
## Arbitration of classifiers

- Arbitrating was originally proposed for dynamic classifier selection
  - a classifier is selected based on confidence estimates
- We re-work this idea and apply it to forecasting

## Mixture of experts

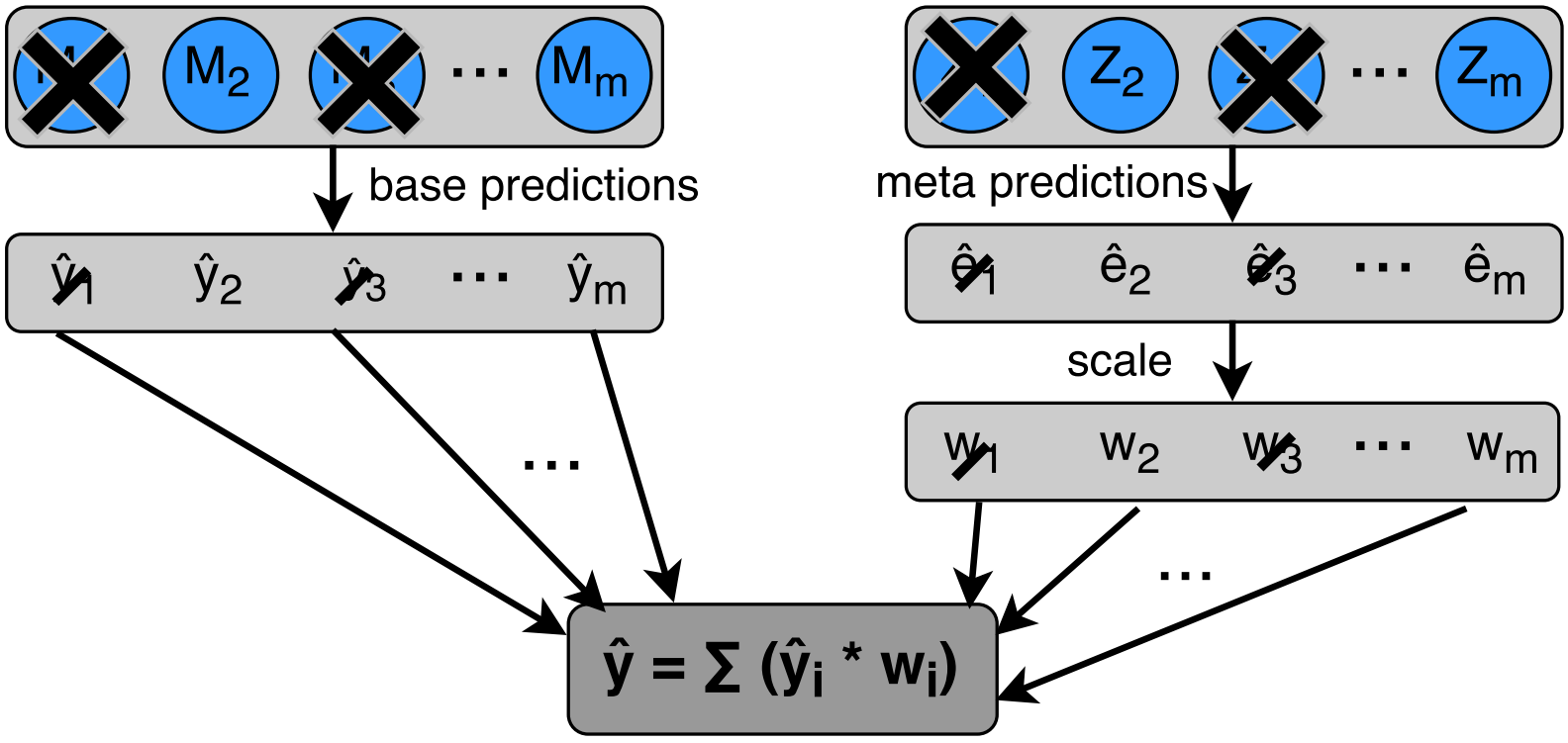
- Arbitrating is similar to mixture of experts
- The main difference is the way expert's weights are computed.
  - A gating network as opposed to error forecasts
- And how diversity is handled
  - incremental specialization versus ensemble heterogeneity

# At each time-step



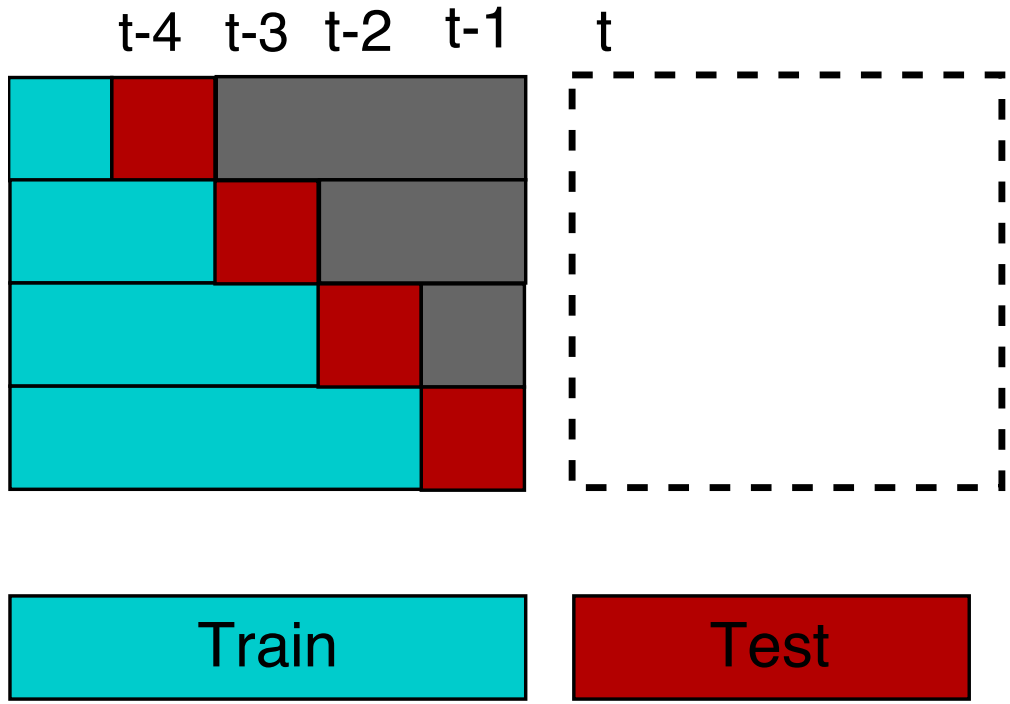

- Each $Z_i$ predicts the loss to be incurred by $M_i$ which is weighed accordingly

# At each time-step




- Dynamic ensemble composition: **suspending** models with poor recent performance from the combination.
  - At each point, the top $\lambda\%$ of models in the last $\Omega$ observations are arbitrated.

- Typical metalearning approaches for dynamic model selection or combination, only start the metalearning layer at run-time;




- Prequential in blocks procedure before runtime
  - To generate unbiased predictions for metalearning;
  - Improving the data efficiency, since the available data is used to fit both the base models and the meta models

# Expert's inter-dependence

## Arbitrating model

- Arbitrating analyses each expert separately over the data space through a regression analysis of their error
  - This can be advantageous in terms of deployment but it can also be a limitation
  - It does not directly model the inter-dependencies among experts

## Diversity encouragement

- Diversity among experts is known to be a critical component in ensemble learning
- Jacobs (1985) points out that ensemble methods require:
  - training procedures that result in relatively independent experts
  - aggregation methods that explicitly or implicitly model the expert's interdependence

# Expert's inter-dependence

## Diversity encouragement

- Diversity among experts is known to be a critical component in ensemble learning
- Jacobs (1985) points out that ensemble methods require:
  - training procedures that result in relatively independent experts
  - aggregation methods that explicitly or implicitly model the expert's interdependence

## Our approach

- We tackle the first issue by employing heterogeneous ensembles;
- The second issue is addressed explicitly by re-weighting the experts at each prediction according to their recent correlation

# Controlling expert's redundancy

## An example from the information retrieval literature

- a list of documents is ranked to answer a query by maximising a function that couples the relevance and redundancy of documents (e.g. Maximal Marginal Relevance);
- the value of the second most relevant document (with respect to a given query) also depends on its redundancy to the most relevant document.
- Essentially, the point is to emphasise the novelty of information in the document set and enhance their complementarity.

# Controlling expert's redundancy

## Controlling the redundancy of experts

- We employ an analogous approach for the combination of experts.
- We use the Pearson correlation among the output of the experts to quantify their redundancy.
  - computed in a window of recent observations to cope with eventual non-stationarities of time series

# Outline

# Time series

## Time series data

Experiments were carried out using 62 real-world time series from several domains

- size ranges from 750 to 3000
  - which is considerable... only 3.5% of the 100.000 in the M4 forecasting competition have over 750 observations.

# Time series modeling

## Time delay embedding

- We transform the time series into an Euclidean space using time delay embedding
- The embedding dimension was estimated according to the false nearest neighbors method

## Summary statistics

Besides the embedding vectors we also use summary statistics to further summarise the dynamics of the time series

- Mean; standard deviation; kurtosis; skewness; serial correlation; long-range dependence; chaos; trend

# Experimental setup

## Evaluation

- Approaches are evaluated using out-of-sample in multiple testing periods
  - has been shown to provide robust performance estimation for time series forecasting tasks (Cerqueira et al., 2017)

Cerqueira, V.; Torgo, L.; Smailovic, J.; Mozetic, I. (2017): A comparative study of performanceestimation methods for time series forecasting. In: Proc. 4th Intl. Conf. on Data Science and Advanced Analytics, pp. 529-538. IEEE (2017).

# Experimental setup

## Ensemble heterogeneity

- We focus on heterogeneous ensembles
  - Our assumption is that this approach is useful to cope with the different dynamic regimes of time series.

## Experts and arbiters

- We used 50 experts (and, accordingly, 50 arbiters) in the main experiments.
  - We will analyse this choice afterwards
- Experts: SVR; Multi-layer Perceptron; Gaussian Processes; Linear regression; Random Forests; Generalised Boosted Regression; PPR; MARS; Rule-based regression
- Arbiters are all random forests

# Experimental Setup

## Dynamic ensemble composition

$\lambda = \Omega = 50$

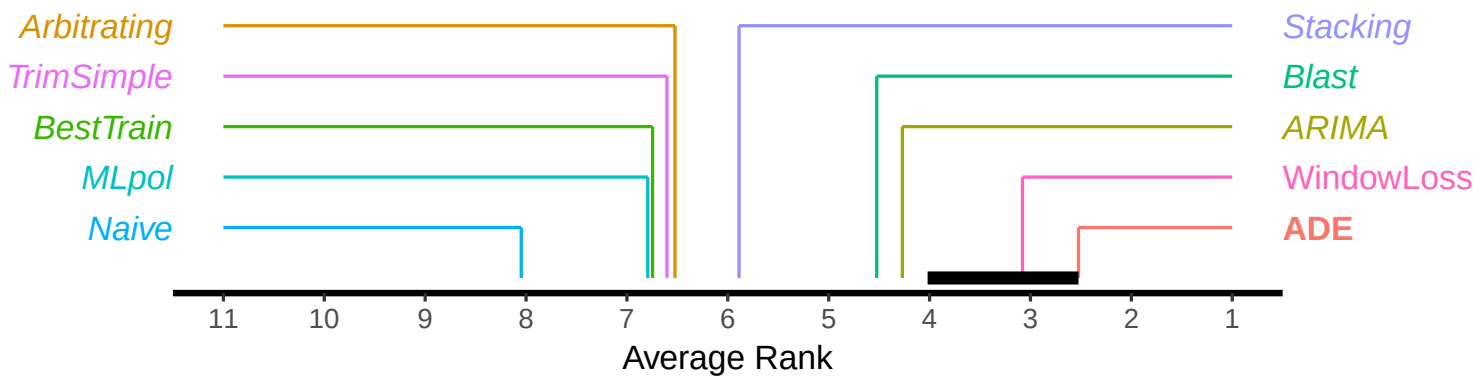
50% of the models with top performance in the last 50 observations are arbitrated in the upcoming prediction

## *Dynamic combiner* approach

- Experts are fit on the training data, and dynamically combined at run-time;
- Arbiters are fit using out-of-bag observations from training data

# Results: comparing ADE to state of the art approaches

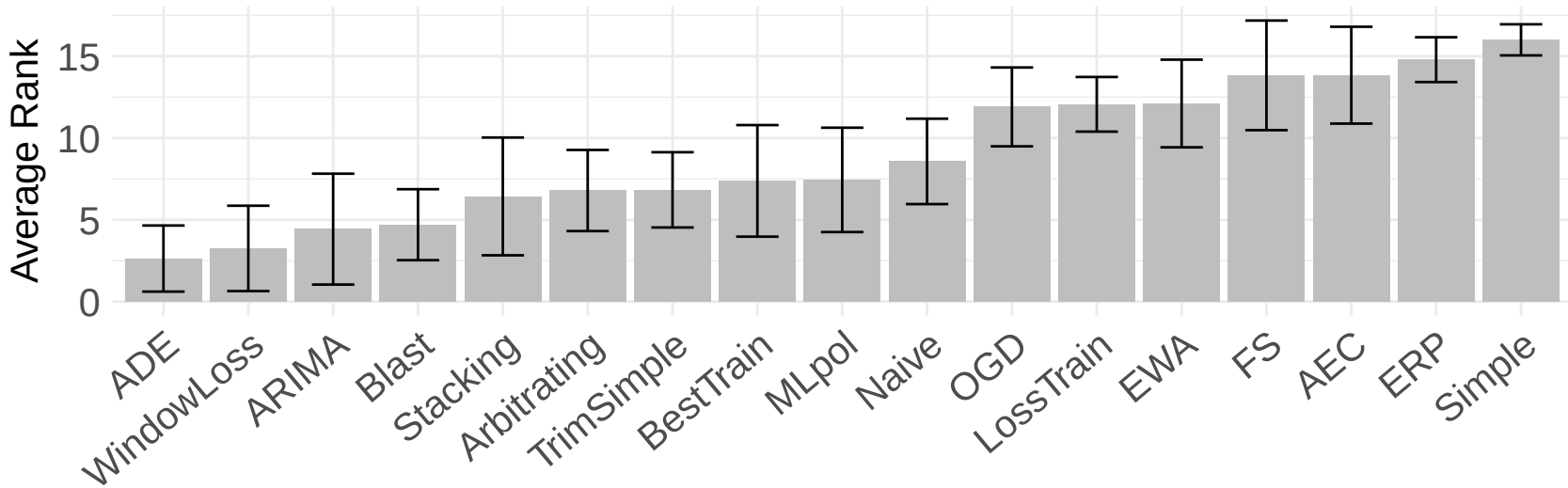
**Q1:** How does the performance of ADE relates to that of the state of the art for combining forecasting models?




- ADE shows a significant improvement over several approaches

# Results: comparing ADE to state of the art approaches

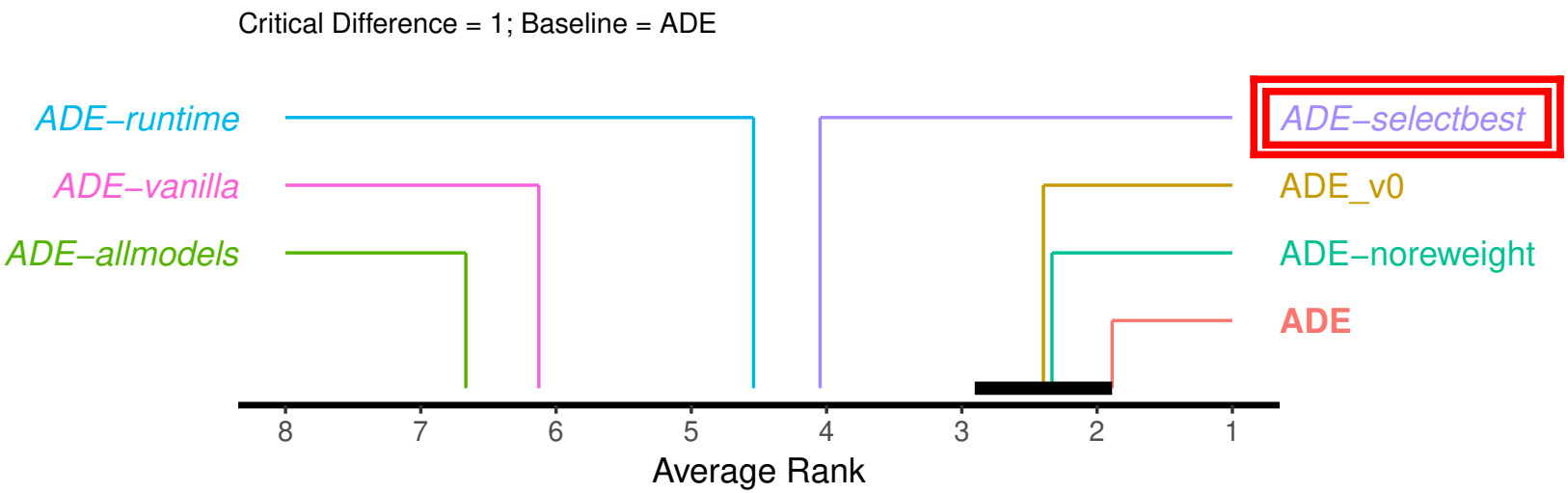
**Q1:** How does the performance of ADE relates to that of the state of the art for combining forecasting models?




- ADE shows a significant improvement over several approaches

# Results: combination versus selection

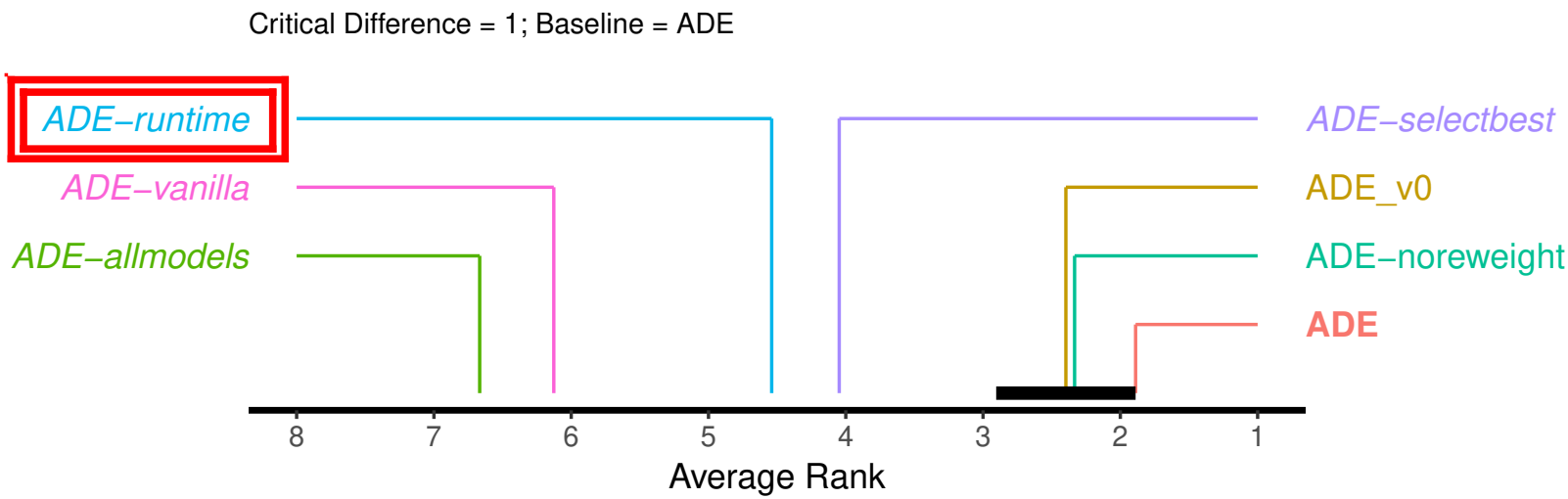
**Q2:** Is it beneficial to use a weighing scheme in the arbitrating approach instead of selecting the predicted to be the best forecaster?




- Combining forecasters shows a superior average rank (as several prior works suggested)

# Results: out-of-bag observations

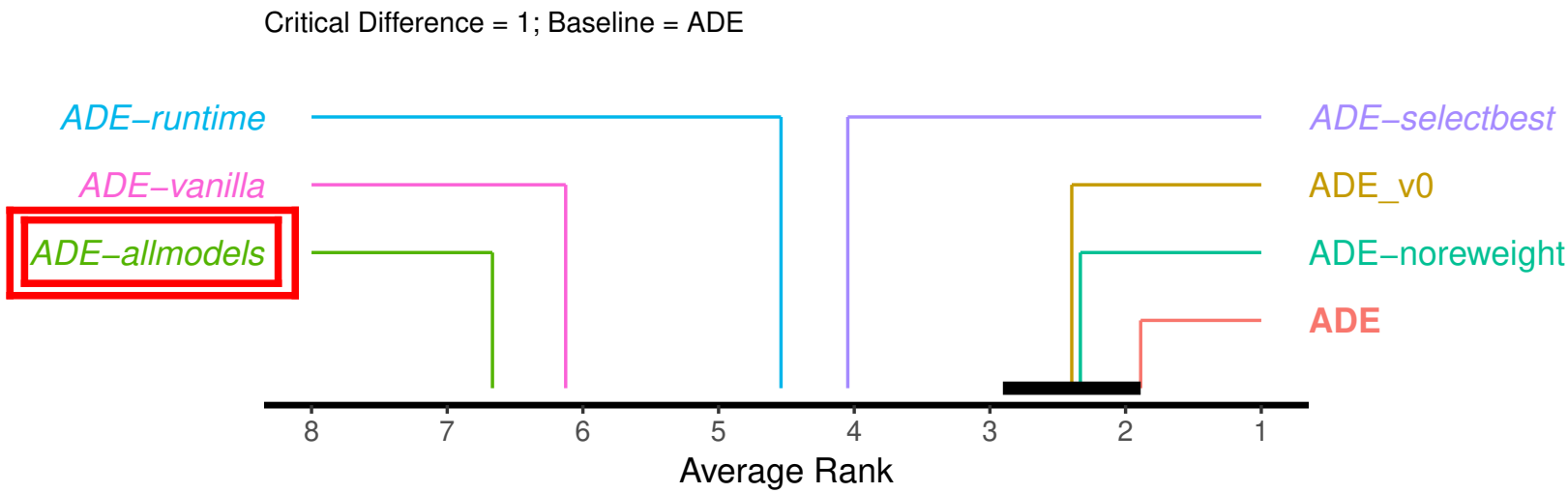
**Q3:** Is it beneficial to use out of bag predictions to train the arbiters?




- Use out of bag predictions for fitting the meta-models

# Results: dynamic ensemble composition

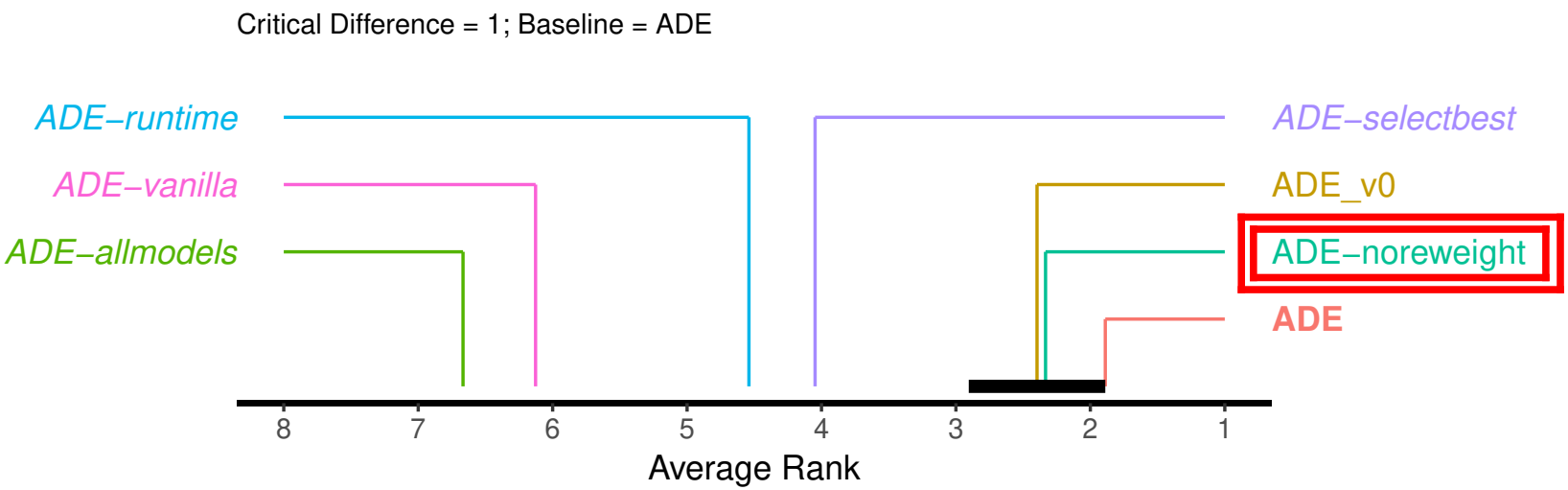
**Q4:** How does the performance of ADE varies by the introduction of a committee, where poor recent base learners are discarded from the upcoming prediction, as opposed to weighing all the models?




- Use a dynamic composition of experts according to recent performance

# Results: deployment strategies

**Q5:** What is the impact of the re-weighting procedure that re-weights the experts according to their recent correlation?




- Control the redundancy of predictions during aggregation using re-weighting of experts

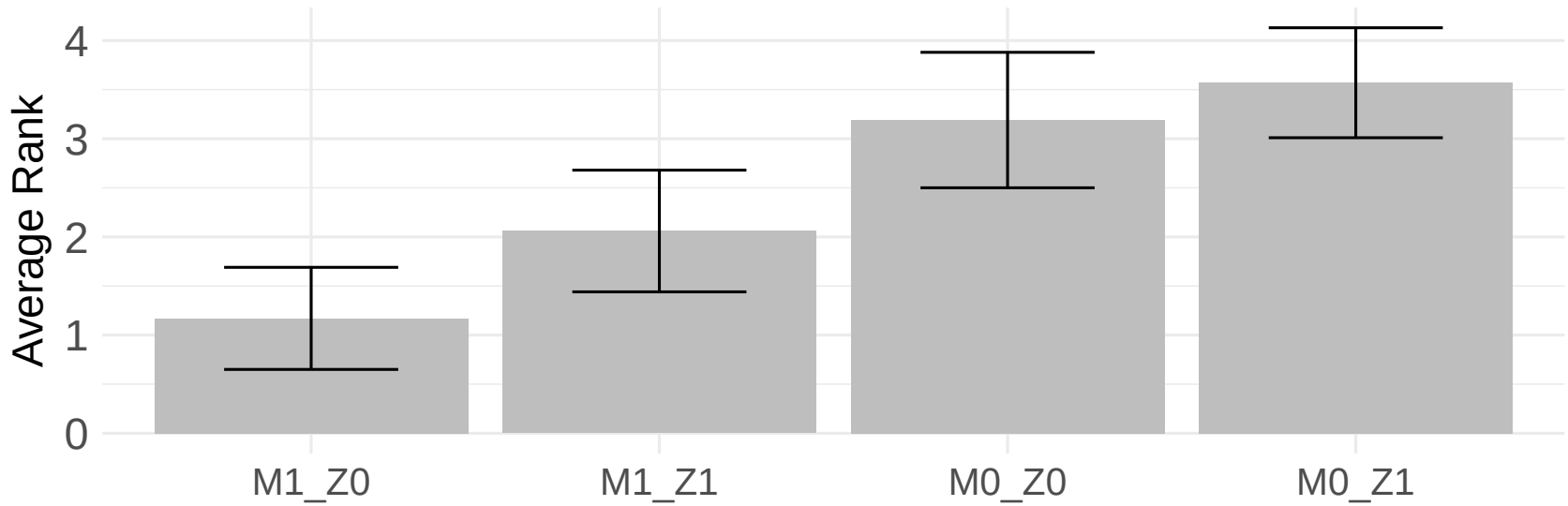
# Results: deployment strategies

**Q6:** How does the performance of ADE varies by using different updating strategies for the base and meta models?

- **M0_Z0**: both experts (M) and arbiters (Z) are trained in the training set and not updated during test time (as reported in the main experiments);
- **M0_Z1**: M is trained only in the training data but Z is re-trained every $\Delta$ observations;
- **M1_Z0**: M is re-trained every $\Delta$ observations but Z is trained only in the training data;
- **M1_Z1**: Both M and Z are re-trained every $\Delta$ observations, which is particularly interesting if the models in M are typical online methods (e.g. ARIMA);

# Results: sensitivity analysis

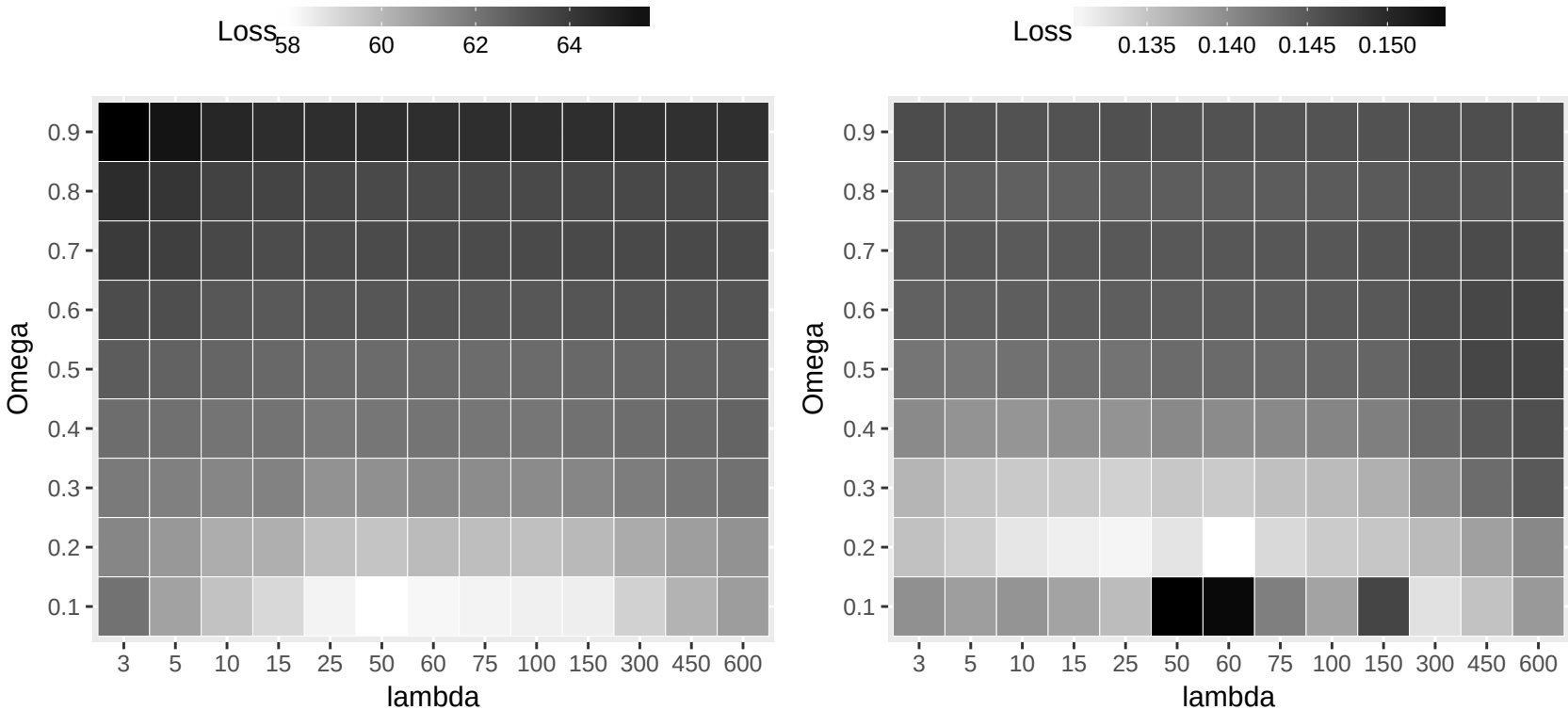
**Q6:** How does the performance of ADE vary by using different updating strategies for the base and meta models?




- Update the experts during run-time

# Results: sensitivity analysis on $\Omega$ and $\lambda$

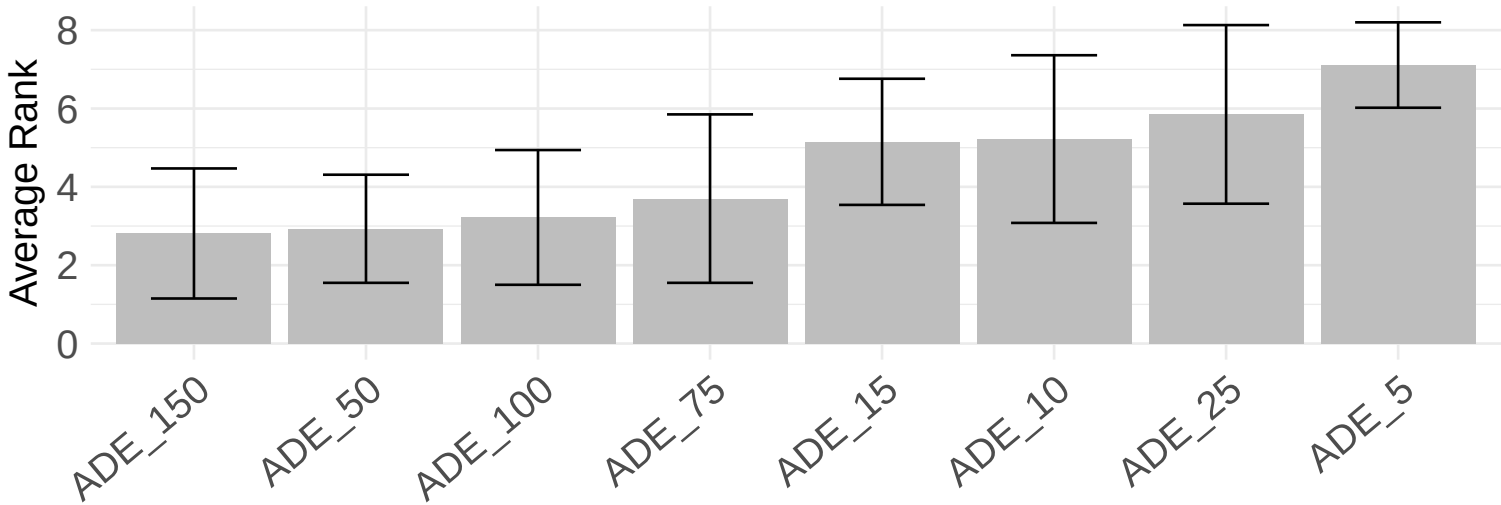
**Q7a:** How sensitive is ADE to the parameters $\Omega$ (size of committe) and $\lambda$ (window size of performance analysis for arbitrating)?




- Although highly dependent on data and ensemble composition, there are regions where the model clearly performs better

# Results: sensitivity analysis on ensemble size

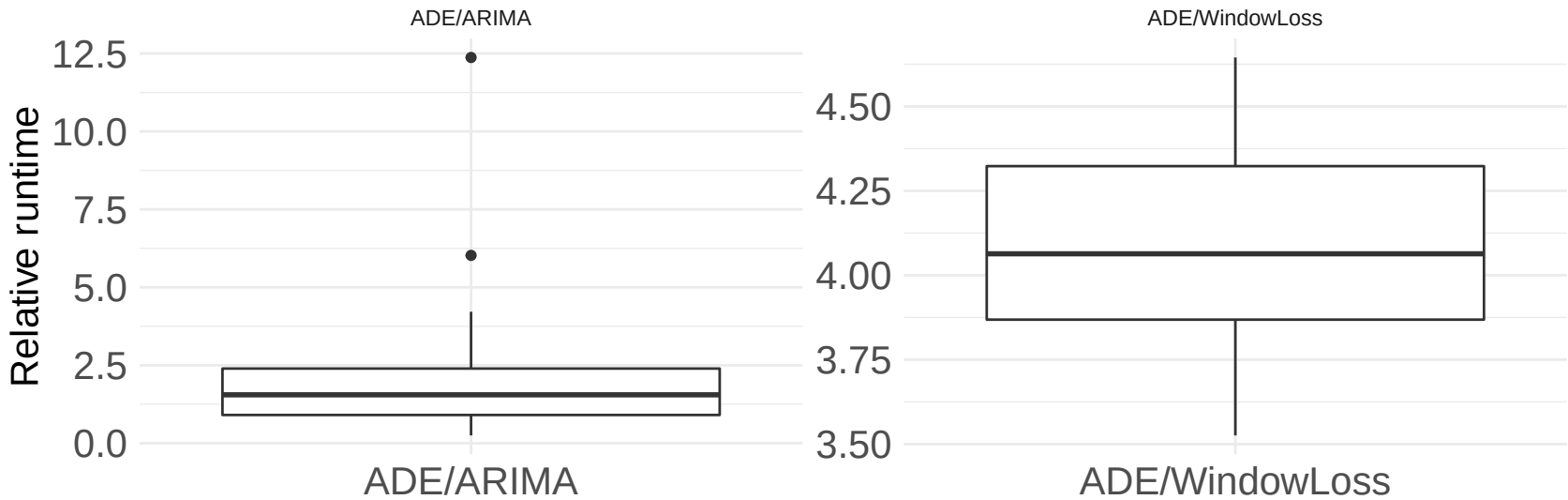
**Q7b:** How sensitive is ADE to the size of the ensemble in terms of the number of experts?




- Increasing experts leads to better performance, but the difference is indistinguishable for more than 50 experts

# Results: sensitivity analysis

**Q8:** How does it scale in comparison to other state of the art approaches for combination of forecasters?




- ADE scales worse than state of the art approaches

# Outline

# Conclusions

## Scientific Contributions

- A dynamic arbitrated and heterogeneous ensemble for time series forecasting;
- An approach for generating data for metalearning
  - using prequential in blocks
- An approach for controlling the redundancy among expert's and handling their inter-dependence

## Future Work

- Improve scalability
  - e.g. multi-target regression using a single arbiter
- Study how to cover data spaces where the model underperforms

Available as an R package

# tsensembler

vitor.cerqueira@fe.up.pt