MONASH University

MONASH BUSINESS SCHOOL

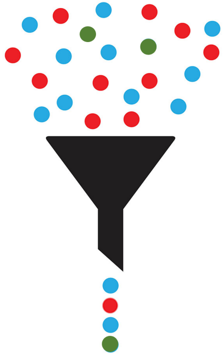
Another look at forecast trimming for combinations: robustness, accuracy and diversity

Xiaoqian Wang

In collaboration with: Yanfei Kang & Feng Li

Department of Econometrics and Business Statistics

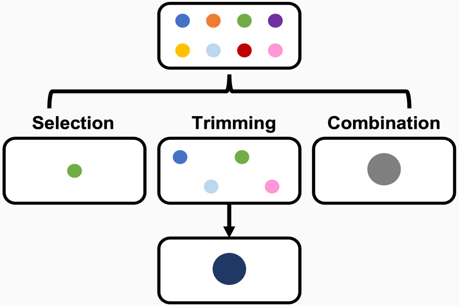
MONASH University



Outline

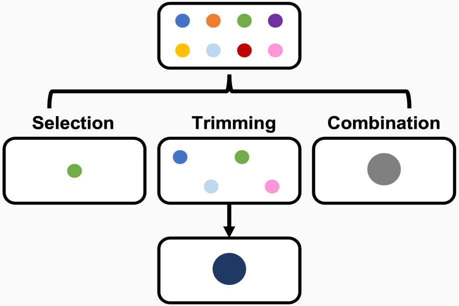
Outline

Forecast trimming



Forecast trimming: combine a subset of individual forecasts.

Forecast trimming



Forecast trimming: combine a subset of individual forecasts.

Two Questions

- Why use forecast trimming?
- How to use forecast trimming?

Why use forecast trimming?

Forecast selection:

- data uncertainty, model uncertainty, and parameter uncertainty

Forecast combination:

- quality of the forecast pool
- estimation of combination weights

Why use forecast trimming?

Forecast selection:

- data uncertainty, model uncertainty, and parameter uncertainty

Forecast combination:

- quality of the forecast pool
- estimation of combination weights

Forecast trimming:

- weight estimation error vs. return when including additional forecasts
- risk of an outlier forecast creeping into the pool

Why use forecast trimming?

Forecast selection:

- data uncertainty, model uncertainty, and parameter uncertainty

Forecast combination:

- quality of the forecast pool
- estimation of combination weights

Forecast trimming:

- weight estimation error vs. return when including additional forecasts
- risk of an outlier forecast creeping into the pool

Principle

- Many could be better than all

How to use forecast trimming?

Three key characteristics of a good forecast pool:

- **Robustness**

  How robust an individual forecast is to pattern evolution

- **Accuracy**

  Forecast error of an individual forecast

- **Diversity**

  Independent information contained in the component forecasts

Research gap

Robustness

- Lichtendahl & Winkler (2020): highlight the importance of robustness

Accuracy

- Kourentzes et al. (2019): ‘forecast islands’
- literature on the ‘wisdom of crowds’: ‘select-crowd’ strategy

Diversity

- Cang & Yu (2014): use mutual information and try all possible combinations
- Lichtendahl & Winkler (2020): screen out individual forecasts with low accuracy and highly correlated errors, respectively

Research gap

Robustness

- Lichtendahl & Winkler (2020): highlight the importance of robustness

Accuracy

- Kourentzes et al. (2019): ‘forecast islands’
- literature on the ‘wisdom of crowds’: ‘select-crowd’ strategy

Diversity

- Cang & Yu (2014): use mutual information and try all possible combinations
- Lichtendahl & Winkler (2020): screen out individual forecasts with low accuracy and highly correlated errors, respectively

Main Objective

- Forecast trimming algorithm addressing robustness, accuracy, and diversity simultaneously

Outline

Robustness, accuracy, and diversity

Robustness

- $\sigma_i^2 = \mathrm{Var}(|f_{i,h} - y_h|)$, where $1 \leqslant h \leqslant H$

Accuracy

- $\mathrm{MSE}_i = \frac{1}{H}\sum_{h=1}^{H}\left(f_{i,h} - y_h\right)^2$

Diversity

- $\mathrm{MSEC}_{i,j} = \frac{1}{H}\sum_{h=1}^{H}\left(f_{i,h} - f_{j,h}\right)^2$ (Thomson et al., 2019; Kang et al., 2022)
  - a larger value indicates a higher degree of diversity
  - be averaged to characterize the overall diversity
  - diversity between a pair & interaction with the rest

Accuracy-diversity trade-off

Toy Example

Select three individuals from the forecast pool $\{-5, 1, 2, 4\}$. The true value is 0.

- $A = \{1, 2, 4\}$
- $D = \{-5, 2, 4\}$
- Best $= \{-5, 1, 4\}$ (simple averaging)

Accuracy-diversity trade-off

Toy Example

Select three individuals from the forecast pool $\{-5, 1, 2, 4\}$. The true value is 0.

- $A = \{1, 2, 4\}$
- $D = \{-5, 2, 4\}$
- Best $= \{-5, 1, 4\}$ (simple averaging)

Accuracy-Diversity Trade-off (ADT)

$$\mathrm{ADT} = \mathrm{AvgMSE} - \kappa\,\mathrm{AvgMSEC}$$

$$= \underbrace{\frac{1}{M}\sum_{i=1}^{M}\mathrm{MSE}_i}_{\text{mean level of accuracy}} \quad - \underbrace{\kappa\frac{1}{M^2}\sum_{i=1}^{M-1}\sum_{j=2, j>i}^{M}\mathrm{MSEC}_{i,j}}_{\text{overall diversity}}$$

- κ is a scale factor and $\kappa \in [0, 1]$

The RAD algorithm

We first divide the in-sample data into D_{train} and D_{valid}.

1. Set the initial selected individual forecaster set $\mathbb{S} = \{1, 2, \ldots, i, \ldots, M\}$.
2. Apply Tukey's fences approach to exclude from $\mathbb{S}$ the individuals that lack robustness.
3. Calculate the ADT criterion of $\mathbb{S}$ based on forecasts and actual values on D_{valid}.
4. For each i in $\mathbb{S}$, calculate the ADT value of the remaining set after removing i from $\mathbb{S}$, and find $\mathrm{Min}_i \mathrm{ADT}(\mathbb{S} \backslash \{i\})$ among all i.
5. Exclude from the forecaster set $\mathbb{S}$ the individual forecasters corresponding to the minimum ADT value $\mathrm{Min}_i \mathrm{ADT}(\mathbb{S} \backslash \{i\})$.
6. Calculate the ADT value for the updated $\mathbb{S}$.
7. Repeat Steps 4-6 until there is non-significant reduction of the ADT value for $\mathbb{S}$ compared to the previous one or until $\mathbb{S}$ contains only two forecasters.

Benchmark algorithms

Algorithm	Description	Robustness	Accuracy	Diversity
None	Do not trim any individuals from the original forecast pool.			
R	Exclude only the individuals that lack robustness.	✓		
A	Exclude only the individuals with relatively low forecast accuracy from the original forecast pool.		✓	
D	Exclude only the individuals whose departure would result in a significant increase in AvgMSEC from the original forecast pool.			✓
RAD	Address robustness, accuracy and diversity simultaneously when implementing forecast trimming.	✓	✓	✓
AutoRAD	The only difference from the RAD algorithm is that the scale factor κ is automatically identified as the one that yields an optimal subset with the minimum MSE value of the simple average among all pre-set values of κ.	✓	✓	✓

Outline

Data and design

Data: the M, M3, and M4 competition data (103,826 series)

- yearly, quarterly, monthly, weekly, daily, and hourly time series
- forecast horizons are 1, 4, 12, 52, 7, and 168
- remove short and constant time series

Data and design

Data: the M, M3, and M4 competition data (103,826 series)

- yearly, quarterly, monthly, weekly, daily, and hourly time series
- forecast horizons are 1, 4, 12, 52, 7, and 168
- remove short and constant time series

Forecast pool: a set of ETS models

Data and design

Data: the M, M3, and M4 competition data (103,826 series)

- yearly, quarterly, monthly, weekly, daily, and hourly time series
- forecast horizons are 1, 4, 12, 52, 7, and 168
- remove short and constant time series

Forecast pool: a set of ETS models

Pre-processing: exclude models with unreasonable prediction intervals

Data and design

Data: the M, M3, and M4 competition data (103,826 series)

- yearly, quarterly, monthly, weekly, daily, and hourly time series
- forecast horizons are 1, 4, 12, 52, 7, and 168
- remove short and constant time series

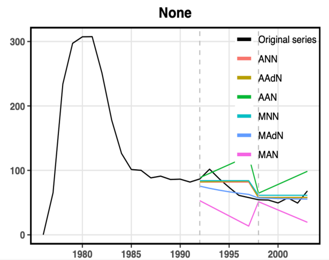
Forecast pool: a set of ETS models

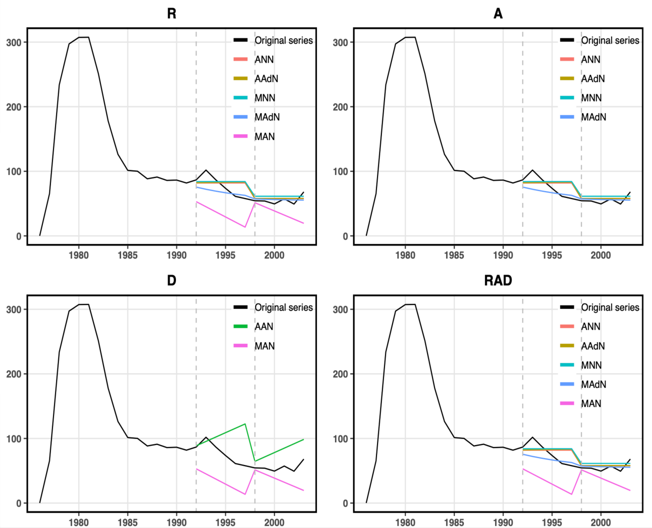
Pre-processing: exclude models with unreasonable prediction intervals

Combination method: simple averaging

- the choice of weight estimation schemes is subjective
- surprising robustness and superior forecasting performance

Trimming example

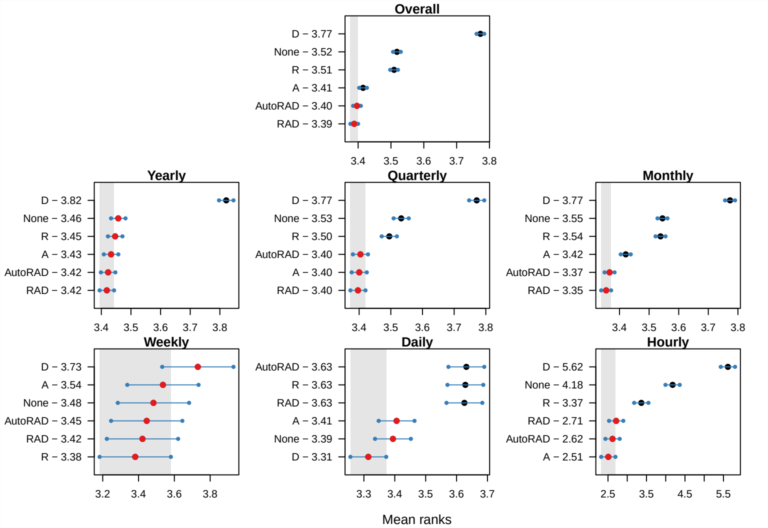




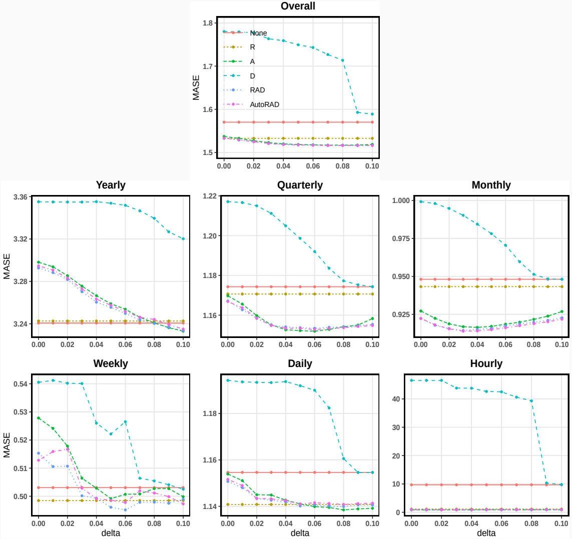
Forecast combination results

Data set	Measure	Simple Average					
		None	R	A	D	RAD	AutoRAD
M	MASE	1.693	1.685	**1.598**	1.751	1.600	1.601
	sMAPE	16.157	16.062	**15.242**	16.663	15.484	15.246
	MSIS	**18.702**	18.739	19.398	19.249	19.044	19.228
	Coverage	0.877	0.874	0.852	**0.879**	0.858	0.854
	Upper coverage	0.916	0.915	0.908	**0.917**	0.911	0.909
	Spread	0.980	0.974	**0.875**	1.004	0.889	**0.875**
	Bias	0.071	0.071	**0.058**	0.071	**0.058**	**0.058**
M3	MASE	1.387	**1.383**	1.401	1.443	1.399	1.399
	sMAPE	13.399	**13.355**	13.401	13.997	13.383	13.371
	MSIS	**11.424**	11.444	13.373	11.682	13.103	13.181
	Coverage	0.928	0.927	0.905	**0.931**	0.911	0.909
	Upper coverage	0.948	0.948	0.939	**0.950**	0.942	0.942
	Spread	0.844	0.838	**0.785**	0.890	0.798	0.792
	Bias	0.014	0.013	**0.003**	0.013	**0.003**	**0.003**
M4	MASE	1.574	1.535	1.521	1.758	**1.520**	**1.520**
	sMAPE	12.284	12.239	12.154	12.708	**12.148**	12.149
	MSIS	24.729	18.005	14.300	48.813	**14.219**	14.245
	Coverage	**0.933**	0.932	0.918	0.929	0.921	0.920
	Upper coverage	**0.954**	**0.954**	0.951	0.950	0.952	0.952
	Spread	1.408	1.105	**0.892**	2.461	0.904	0.898
	Bias	0.027	0.033	0.021	**0.010**	0.022	0.022
Overall	MASE	1.570	1.533	1.519	1.749	**1.518**	**1.518**
	sMAPE	12.352	12.306	12.218	12.782	12.214	**12.212**
	MSIS	24.308	17.834	14.324	47.516	**14.235**	14.264
	Coverage	**0.933**	0.931	0.917	0.929	0.921	0.919
	Upper coverage	**0.953**	**0.953**	0.950	0.950	0.952	0.951
	Spread	1.389	1.097	**0.889**	2.404	0.901	0.895
	Bias	0.027	0.033	0.021	**0.011**	0.022	0.022

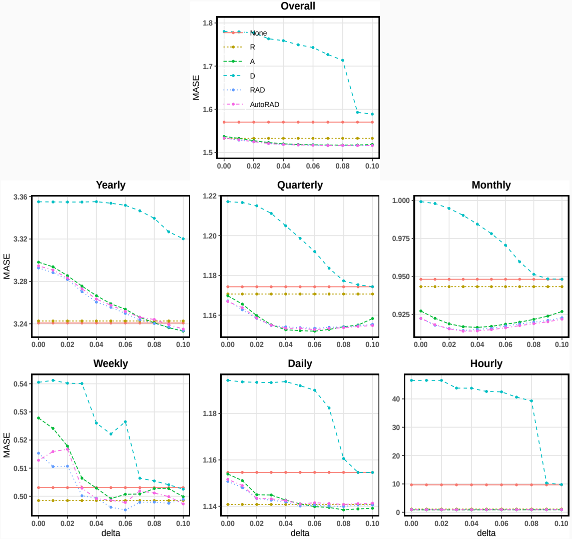
MCB tests for each data frequency



The effect of the level parameter



The effect of the level parameter



- Overall, RAD and AutoRAD are superior to other four trimming algorithms across all values of δ.
- A value of δ in the region between 0.04 and 0.06 seems to work well for seasonal series.
- The average performance gap between RAD (or AutoRAD) and A is relatively small.

Relative diversity

Aim

- For a given pool, explore the importance of the degree of diversity relative to accuracy on the selection of trimming algorithm.

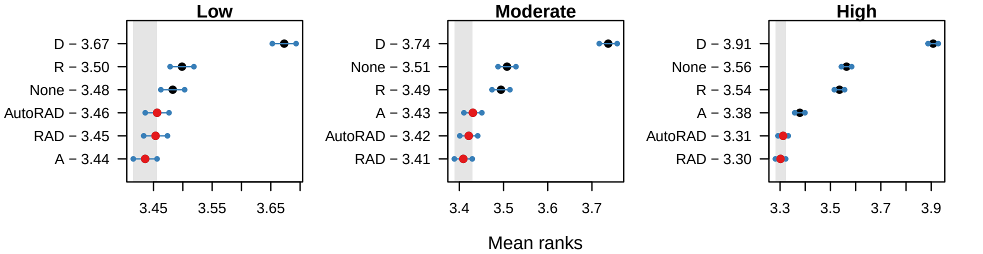
RelDiv (Relative Diversity)

$$\text{RelDiv} = \frac{\mathrm{AvgMSEC}}{\mathrm{AvgMSE}} = \frac{\sum_{i=1}^{M-1} \sum_{j=2, j>i}^{M} \left[\frac{1}{H} \sum_{h=1}^{H} \left(f_{i,h} - f_{j,h}\right)^2\right]}{M \sum_{i=1}^{M} \left[\frac{1}{H} \sum_{h=1}^{H} \left(f_{i,h} - y_h\right)^2\right]}$$

- comparable between series with different units
- allows to average the RelDiv values across time series

RAD/AutoRAD vs. A

- Remove the instances in which both algorithms identify the same optimal subset from the given forecast pool.
- Split the time series with regard to different levels of RelDiv (low, moderate, and high levels) using Q1 (0.2) and Q3 (0.5) of RelDiv.



Outline

Conclusions

- RAD addresses robustness, accuracy, and diversity simultaneously.
- ADT is used to achieve a trade-off between accuracy and diversity.
- Good performance and robustness.
- Simple guidelines for selecting forecast trimming algorithm.

Guidelines

1. Not always have to address the diversity issue
2. RelDiv < 0.2, A is preferred
3. RelDiv > 0.5, RAD and AutoRAD are preferred

Conclusions

- RAD addresses robustness, accuracy, and diversity simultaneously.
- ADT is used to achieve a trade-off between accuracy and diversity.
- Good performance and robustness.
- Simple guidelines for selecting forecast trimming algorithm.

Guidelines

1. Not always have to address the diversity issue
2. RelDiv < 0.2, A is preferred
3. RelDiv > 0.5, RAD and AutoRAD are preferred

References

- Cang, S., & Yu, H. (2014). A combination selection algorithm on forecasting. European Journal of Operational Research, 234(1), 127-139.
- Kang, Y., Cao, W., Petropoulos, F., & Li, F. (2022). Forecast with forecasts: Diversity matters. European Journal of Operational Research, 301(1), 180-190.
- Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. International Journal of Production Economics, 209, 226-235.
- Lichtendahl Jr, K. C., & Winkler, R. L. (2020). Why do some combinations perform better than others?. International Journal of Forecasting, 36(1), 142-149.
- Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: Performance and coherence. International Journal of Forecasting, 35(2), 474-484.

THANK YOU

Find me at ...

xqnwang.rbind.io
@Xia0qianWang
@xqnwang
xiaoqian.wang@monash.edu