Beyond summary performance metrics for forecast selection and combination

Nikolaos Kourentzes[a]

Ivan Svetunkov[a]

Stephan Kolassa[a,b]

[a]Lancaster Centre for Marketing Analytics and Forecasting; [b]SAP

International Symposium on Forecasting 2018 - Boulder

LUMS

Marketing Analytics
and Forecasting

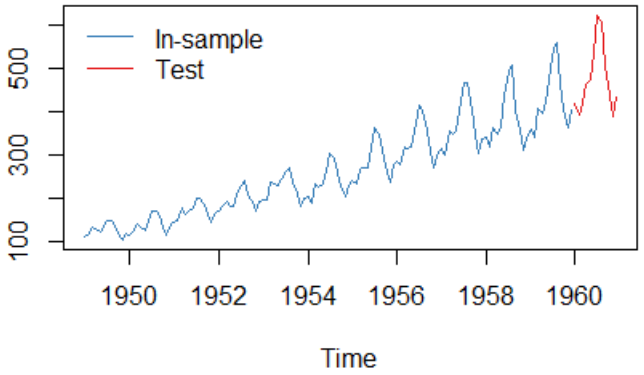
The problem

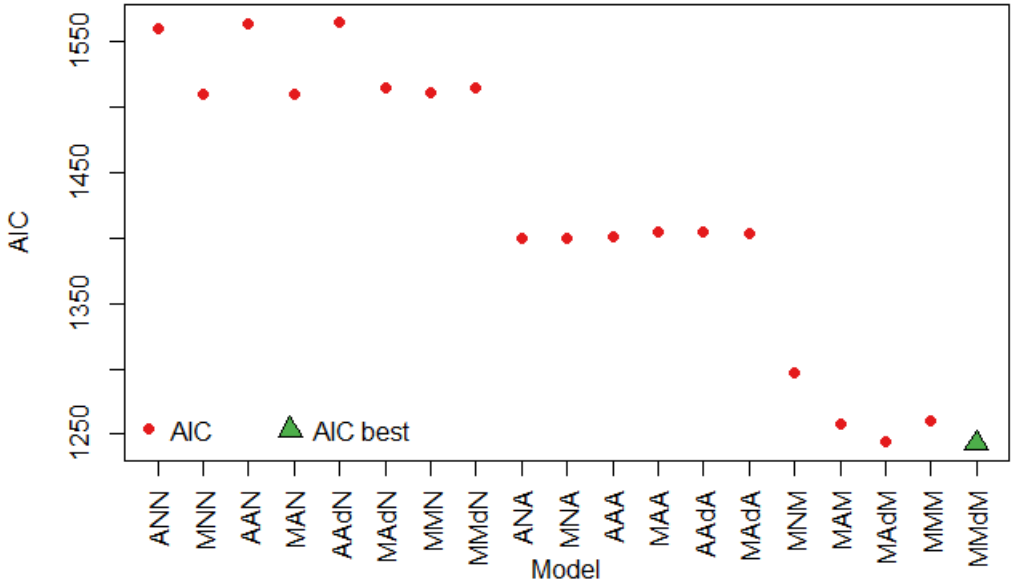
- Forecast combination is widely regarded to have many merits over selection (Elliott and Timmermann, 2016), yet appropriate selection can lead to substantial gains and convenience (Fildes and Petropoulos, 2015) → but there are many uncertainties.
- Selection of best forecast. Do you have a model?
  - Yes! Information criteria (Hyndman et al., 2002; Burnham and Anderson, 2003);
  - No, then cross-validate (Fildes and Petropoulos, 2015; Kourentzes et al., 2018).
- Combination of forecasts:
  - Optimal weights → issues with estimation of weights (Smith and Wallis, 2009; Claeskens et al., 2016), but at times still worthwhile (Elliott, 2011).
  - Fixed weights → empirically work very well, weights can be inferred in many ways (AIC; Kolassa, 2011, cross-validation; Kourentzes et al., 2018; etc.).
- Pooling, in between selection and combination (Aiolfi and Timmermann, 2006; Elliott, 2011; Kourentzes et al., 2018). In between optimal and fixed weights.
- Most rely on (appropriate) summary performance metrics → but there is evidence that this may be inadequate, from meta-learning (Lemke and Gabrys, 2010) or judgemental forecast selection (Petropoulos et al., 2018).

CMAF

An example

- The infamous Airline Passenger time series – retain last year as test.
- Fit exponential smoothing models and get their AIC.
- Select according to best AIC (Hyndman et al., 2002)
- … or combine with AIC weights (Kolassa, 2011)



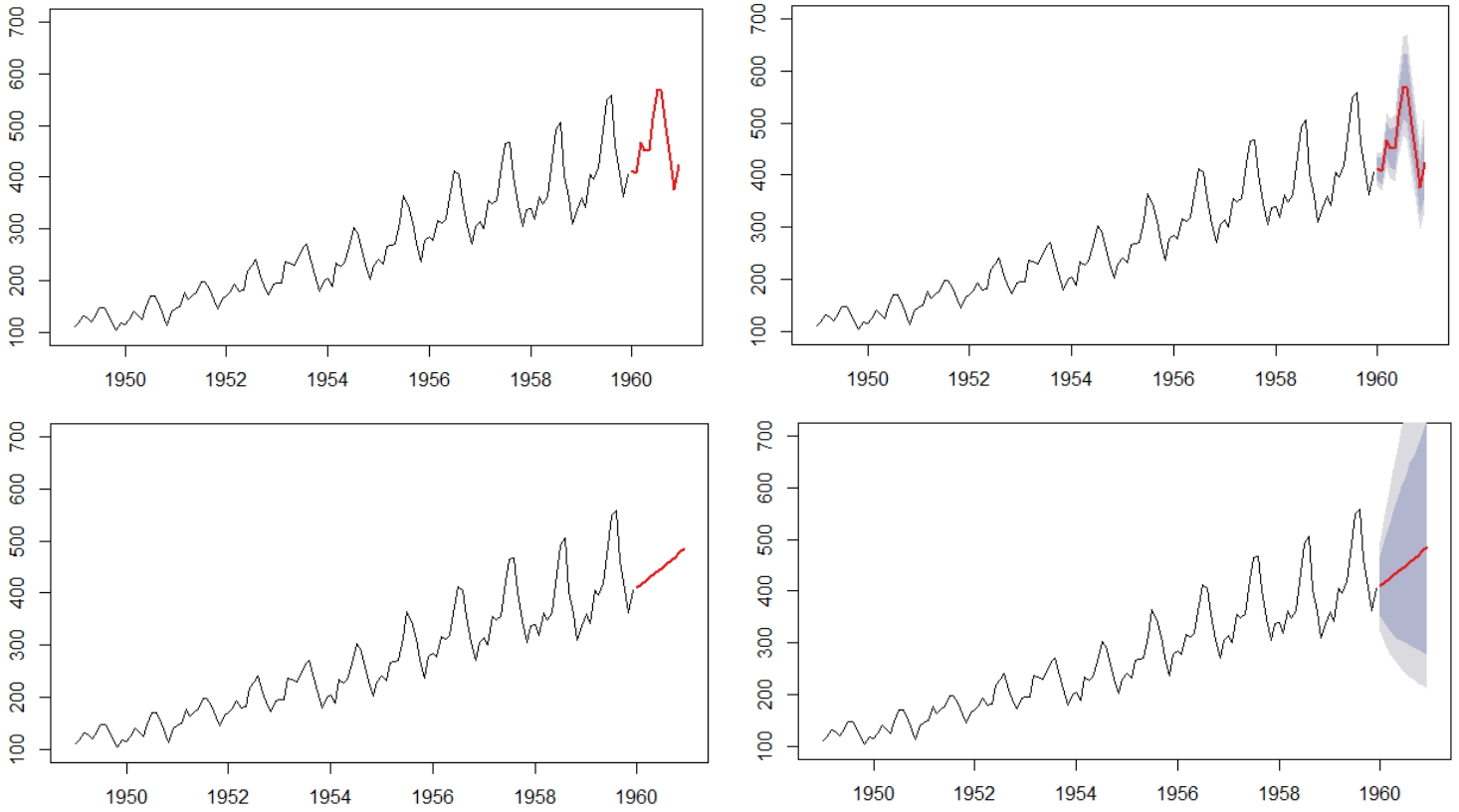


- How different is the best from the second best?
- What about sampling/estimation uncertainty?

CMAF

An example

- We are quite content that we need prediction intervals to take decisions on forecasts



- Yet, for model selection we are using summary performance metrics, akin to point forecasts, ignoring any higher moments!

CMAF

Beyond summary metrics

- The AIC is calculated as

$$\text{AIC} = 2k + \mathcal{L}^*$$

- For the state space exponential smoothing models the double negative log likelihood is

$$\mathcal{L}^* = n \ln \left(\sum_{t=1}^{n} \varepsilon_t^2 \right) + 2 \sum_{t=1}^{n} \ln |r(\boldsymbol{x}_{t-1})|$$

This summary can conceal important information for modelling

- So we propose the following modification

$$\mathcal{L}_t^* = n \ln \left(n\varepsilon_t^2 \right) + 2 \sum_{t=1}^{n} \ln |r(\boldsymbol{x}_{t-1})|$$

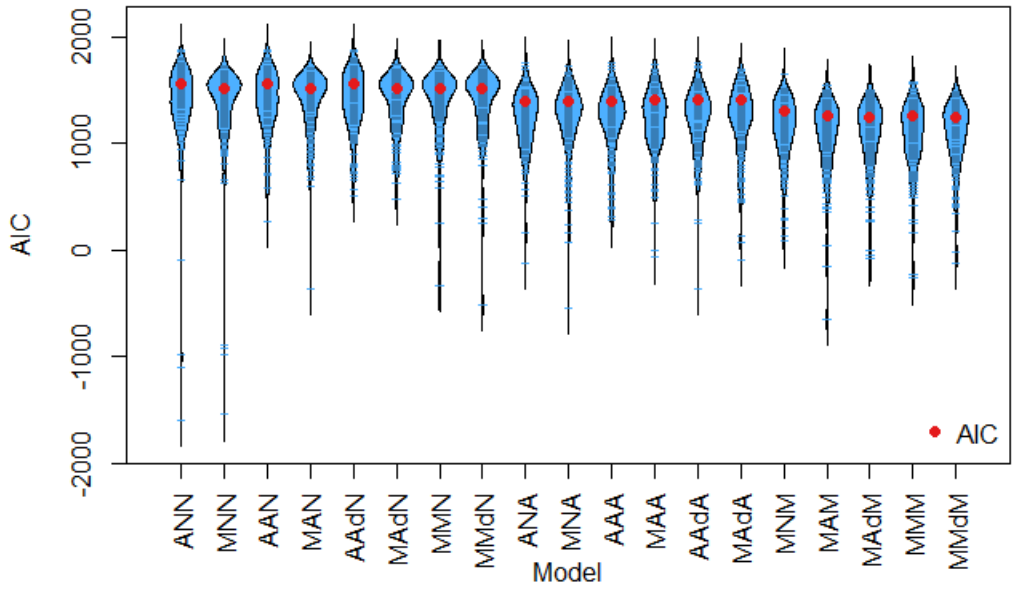
- And a point AIC:

$$\text{pAIC}_t = 2k + \mathcal{L}_t^*$$

CMAF

pAIC distributions

- Now we have distributions to compare



- We like distributions: gauge the uncertainty in choosing a specific model over alternatives and when combination is appealing!
- pAIC distributions are not normal and can look peculiar, so we rely on non-parametric statistics.

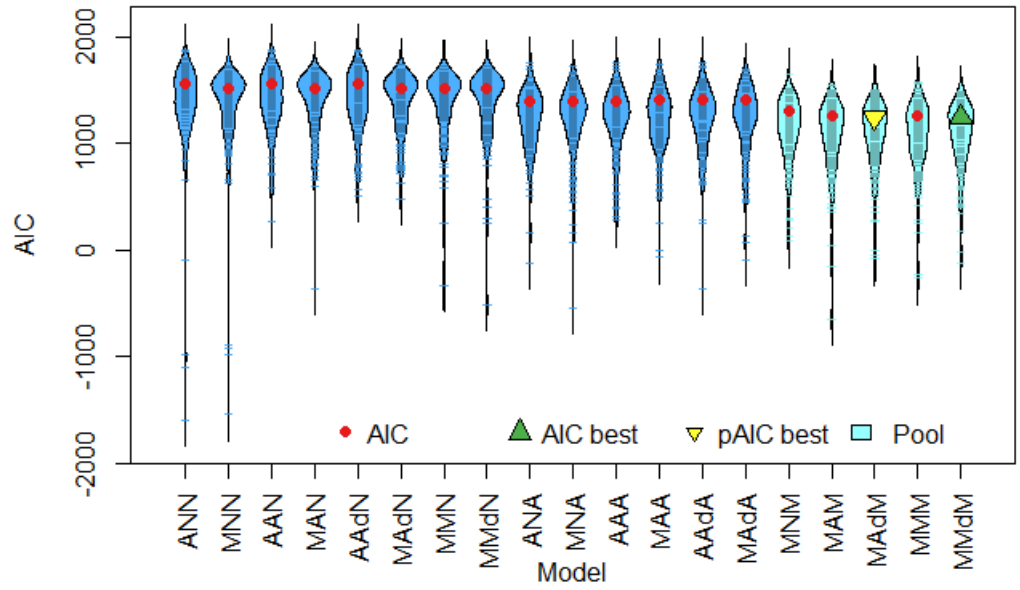
CMAF

Model choice and combination

- We can calculate mean pAIC ranks – the ranks retain the information about the comparative location and dispersion of the distributions.
- We can:
  - Select the best ranking model
  - Combine them with rank based weights: $w_i = \frac{r_i^{-1}}{\sum_{i=1}^{m} r_i^{-1}}$
- pAIC allows for even more informed combinations. When there no evidence that distributions are statistically different then combine, otherwise drop poorly performing models → use non-parametric tests

CMAF

Model choice and combination



Method	MAE
AIC selection	21.637
pAIC selection	22.805
AIC combination	22.032
pAIC combination	12.318

- AIC selects ETS(M,Md,M), pAIC selects ETS(M,Ad,M) → quite similar
- pAIC pooling indicates the time series has **some** trend and multiplicative seasonality. That is exactly right! We can still debate about the trend in this time series, but surely it is not generated by additive damped or multiplicative damped in an ETS framework.
- pAIC allows to transition in a data driven way between model selection and combination → pooling.

CMAF

Predictive model choice and combination

- We select/combine forecasts using past statistics → is that wise?
- Considering the instability of model selection in practical situations, i.e. how the selected model may change each period, what we are doing has some parallels with random walk forecasting → we do not consider any dynamics in the model selection.
- pAIC is ordered in time and therefore is a time series itself, we can forecast it
  - What model ranks best in the future period of interest?
  - Predictively select the best model or form model pools predictively!

Method	MAE
AIC selection	21.637
pAIC selection	22.805
AIC combination	22.032
pAIC combination	12.318
Pred. pAIC selection	19.141
Pred. pAIC comb.	19.141

It picked up only 1 model, ETS(M,M,M)

CMAF

Empirical evaluation

- Four model selection approaches
  - Sel.AIC – select using AIC
  - Sel.CV – select using MSE cross-validated error (20% validation sample)
  - Sel.Rank – select using pAIC
  - Sel.RankP – select using predicted pAIC
- Ten combination approaches
  - Comb.Mean – combine using unweighted average
  - Comb.Median – combine using unweighted median
  - Comb.AIC – combine using AIC weights
  - Comb.CV – combine using cross-validation weights
  - Comb.Rank.pAIC – combine using pAIC
  - Comb.Rank.AIC – combine using pAIC pools and AIC weights
  - Comb.Rank.CV – combine using pAIC pools and CV weights
  - + 3 options for predictive pAIC.

CMAF

Empirical evaluation

- Evaluate on M3 and two additional dataset – total 3555 series
- Rolling origin evaluation
- Compare using Average Relative Mean Absolute Error (Davydenko and Fildes, 2013)

$$\text{AvgRelMAE}_i = \sqrt[q]{\prod_{r=1}^{q} \left(\frac{\text{MAE}_{i,r}}{\text{MAE}_{b,r}} \right)},$$

$$\text{MAE} = \frac{1}{oh} \sum_{j=1}^{o} \sum_{t=1}^{h} |y_{t+j-1} - \hat{y}_{t+j-1}|$$

Dataset	Test set	Horizon	No. of Series
M3 - Yearly	6	4	645
M3 - Quarterly	8	4	756
M3 - Monthly	18	12	1428
M3 - Other	18	12	174
FMCG	52	13	229
FRED	36	12	323

CMAF

Results

Scheme	M3				FMCG	FRED	Overall
	Yearly	Quarterly	Monthly	Other			
Sel.AIC	1.000	1.000	1.000	1.000	1.000	**1.000**	1.000
Sel.CV	1.014	1.008	1.002	1.052	1.007	1.027	1.018
Sel.Rank	**0.976**	**0.983**	**0.986**	**0.997**	**0.992**	1.009	**0.991**
Sel.RankP	0.979	1.017	1.022	1.103	1.006	1.049	1.029

Best benchmark

Best pAIC method

- Sel.Rank improves over Sel.AIC – small differences, but consistent
- Sel.RankP does not work, apart from the yearly M3 case that selects between fewer models.

CMAF

Results

Scheme	M3				FMCG	FRED	Overall
	Yearly	Quarterly	Monthly	Other			
Sel.AIC	1.000	1.000	1.000	1.000	1.000	**1.000**	1.000
Sel.CV	1.014	1.008	1.002	1.052	1.007	1.027	1.018
Sel.Rank	**0.976**	**0.983**	**0.986**	**0.997**	**0.992**	1.009	**0.991**
Sel.RankP	0.979	1.017	1.022	1.103	1.006	1.049	1.029
Comb.Mean	0.980	1.046	1.013	1.023	0.991	1.065	1.019
Comb.Median	0.962	1.010	0.972	1.016	0.993	1.011	0.994
Comb.AIC	0.967	0.983	0.976	0.979	0.992	0.998	0.982
Comb.CV	0.962	0.976	0.963	1.009	0.987	1.006	0.984
Comb.Rank.pAIC	**0.952**	0.979	0.961	0.967	0.984	1.001	0.974
Comb.Rank.AIC	0.966	0.982	0.976	0.974	0.991	0.998	0.981
Comb.Rank.CV	0.954	0.969	0.953	0.994	0.983	0.987	0.973
Comb.RankP.pAIC	0.948	1.103	0.981	1.073	**0.985**	1.045	1.021
Comb.RankP.AIC	0.974	0.982	0.975	**0.982**	0.992	1.001	0.984
Comb.RankP.CV	0.955	**0.975**	**0.954**	1.008	0.989	**0.989**	**0.978**

Best benchmark

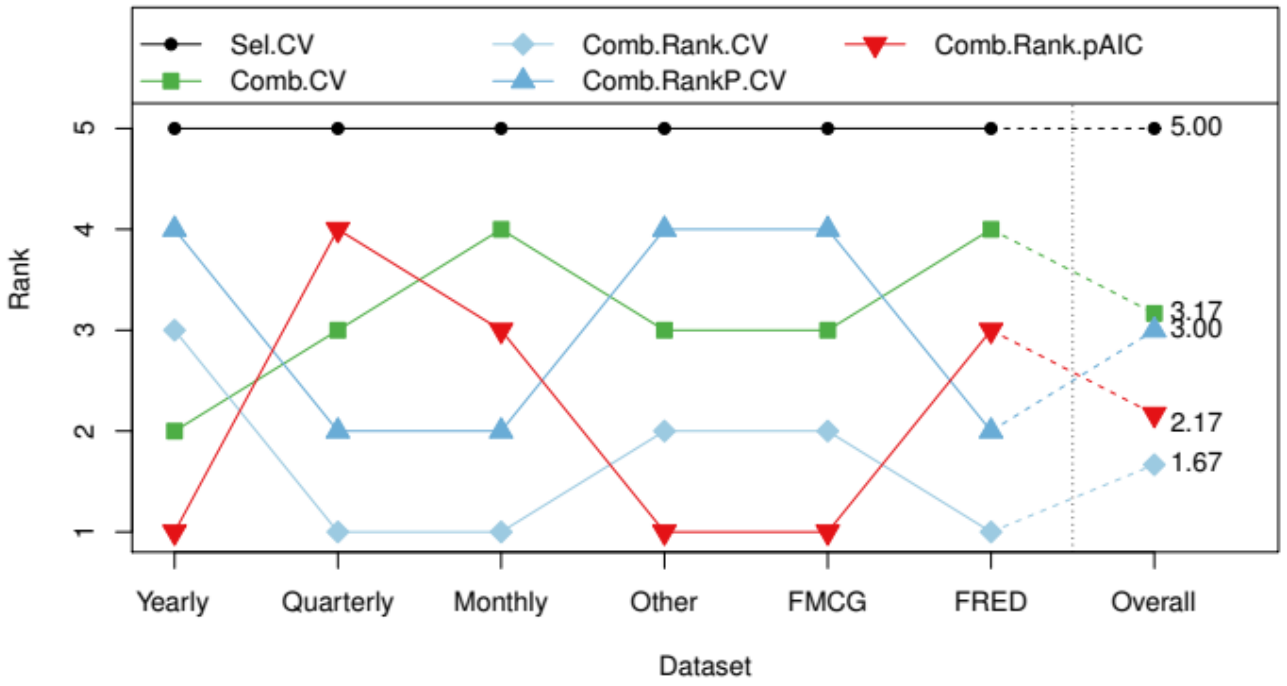
Best pAIC method

- pAIC pools consistently best.
- Rank.pAIC simple and very accurate.
- Rank.CV expensive, marginally best.
- RankP.CV not best, but beats all benchmarks

CMAF

Results

- Compare in terms of ranks Sel.CV, Comb.CV, Comb.Rank.CV, Comb.RankP.CV and the recommended Comb.Rank.pAIC.



- Comb.CV better than Sel.CV always – that is known
- Comb.RankP.CV one step worse than Comb.Rank.CV, comparable/better to Comb.CV
- Comb.Rank.pAIC works very well and is cheaper than Comb.Rank.CV

CMAF

Findings

- There is merit into going beyond summary performance statistics for forecast selection and combination. Pooling using pAIC is consistently better than benchmarks and very easy/fast to calculate.

- Pooling is a natural outcomes of using distributions. There is no dichotomy between selection and combination.

- Trivial to extend to other information criteria and cross-validated errors – the latter will have sample size issues and will be quite slow, but may be necessary.

- Predictive selection/combination: not best, but better than all benchmarks → we argue that there is value in this line of thinking:

  Build a model of model selection/combination!

  … and then a model of model of model selection/combination!

CMAF

Thank you for your attention!
Questions?

Working paper available!

Nikolaos Kourentzes (@nkourentz)

email: nikolaos@kourentzes.com

blog: http://nikolaos.kourentzes.com

Marketing Analytics and Forecasting

Lancaster University Management School