Adaptive Combinations of Large Model Pools

Nicos Pavlidis

Department of Management Science

Marketing Analytics and Forecasting

Lancaster University Management School

CMAF Friday Forecasting Talks

Motivation

Question: How to *optimally* forecast a time-series in the presence of different candidate models?

- Problem can be considered a superset of **model selection**
- Optimality criterion: MSE, Kullback-Leibler divergence, forecast error (?)
- Combine point or density forecasts (?)
- Are individual models considered exogenous (?)
- Static or dynamic environment (?)

Background

Model Combination has long history in **many** areas:

- Econometrics/ Statistics: Bates & Granger (1969)
  - Bayesian/ Dynamic Moving Average
  - Density combination (Hall & Mitchell 2007)
  - Shrinkage
- Machine Learning:
  - Stacking (Breiman 1996)
  - Prediction with expert advice (Littlestone, 1992)
- Computer science/ AI: Neural Networks
- Change point detection

Different works impose different sets of assumptions and thus consider different instances of this problem

Our problem

Main characteristics:

- Interested in both model estimation and combination
- Number of candidate models is very large: $K \gg T$
- Little or no prior knowledge about appropriate model(s)
- Model parameters **and** optimal model can change
- Interpretability

Dynamic linear models with forgetting

Each model (k) is a dynamic linear model (Kalman filter)

$$\begin{aligned} \theta_t^{(k)} &= \theta_{t-1}^{(k)} + \omega_t^{(k)}, & \omega_t^{(k)} &\sim \mathcal{N}\left(0, W_t^{(k)}\right) \\ y_t &= x_t^{(k)\top}\theta_t^{(k)} + \varepsilon_t^{(k)}, & \varepsilon_t^{(k)} &\sim \mathcal{N}\left(0, \sigma_t^{2(k)}\right) \end{aligned}$$

Why linear models?

Surprisingly effective; Well understood; Uncertainty quantification; Computationally cheap

What does this mean?

Coefficients follow a *random walk*: Allows adaptation in response to different types of change

Dynamic linear models with forgetting

Each model (k) is a dynamic linear model (Kalman filter)

$$\begin{aligned} \theta_t^{(k)} &= \theta_{t-1}^{(k)} + \omega_t^{(k)}, \qquad & \omega_t^{(k)} \sim \mathcal{N}\left(0, W_t^{(k)}\right) \\ y_t &= x_t^{(k)\top}\theta_t^{(k)} + \varepsilon_t^{(k)}, & \varepsilon_t^{(k)} \sim \mathcal{N}\left(0, \sigma_t^{2(k)}\right) \end{aligned}$$

Unknown quantities?

$\sigma_t^{2(k)}$: Observation noise variance easy to estimate

$W_t^{(k)}$ difficult/ expensive to estimate

Solution: DLM with forgetting (discounting) – Very widely used

Forgetting (discount) factor: $\lambda \in (0, 1]$

Meaning: Periods of high uncertainty about $\theta_t^{(k)}$ coincide with periods of high coefficient variability

Adaptive Forgetting DLM

Central Idea:

Coefficients depend on choice of λ

Identify λ that minimises **expected** forecast error

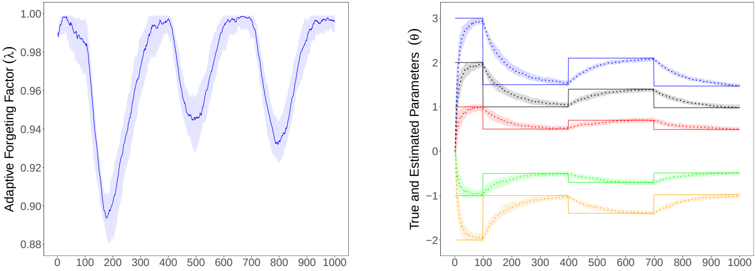
$$\lambda^{\star} = \underset{\lambda \in (0,1]}{\arg\min}\, \mathbb{E}_{Y,X} \left\{ \frac{1}{2} \left[y_{t+1} - x_{t+1}^{(k)\top} \theta_t^{(k)} \right]^2 \right\}$$

Expectation above is not accessible so how can this be achieved?

Stochastic optimisation (approximation): Observed error at each time-step is unbiased estimator of this expectation

A lot of research in this area related largely to large NNs

Adaptive Forgetting DLM



(a) Median and IQR of SGD estimates of λ_t

(b) Evolution of θ_t and median and IQR of $\hat{\theta}_t$

Figure: Coefficient vector subject to abrupt changes of different magnitude

No fixed value of λ is appropriate for entire time-series

Prediction with Expert Advice

Online learning problem

Predict: $\hat{y}_{t+1} = \sum_{k=1}^{K} w_{t,k}\, \hat{y}_{t+1}^{(k)}$

Observe y_{t+1}

Update weight of each expert according to **loss**: $\ell(y_{t+1}, \hat{y}_{t+1}^{(k)})$

Objective: Minimise **regret** w.r.t. best sequence of M experts (each optimal in segment of time series)

$$R_T = \sum_{t=1}^{T} \ell(y_t, \hat{y}_t) - \sum_{m=1}^{M} \min_{k \in K} \left\{ \sum_{t=t(m)}^{t(m+1)-1} \ell(y_t, \hat{y}_t^{(k)}) \right\}$$

Unknown: M and $1 = t(1) < \cdots < t(M) < t(M+1) = T$

Prediction with Expert Advice

Objective: Minimise **regret** w.r.t. best sequence of M experts (each optimal in segment of time series)

$$R_T = \sum_{t=1}^{T} \ell(y_t, \hat{y}_t) - \sum_{m=1}^{M} \min_{k \in K} \left\{ \sum_{t=t(m)}^{t(m+1)-1} \ell(y_t, \hat{y}_t^{(k)}) \right\}$$

Unknown: M and $1 = t(1) < \cdots < t(M) < t(M+1) = T$

No assumptions on time-series y_t or forecasts $\hat{y}_t^{(k)}$

Assumptions on **loss function** ℓ (typically ℓ is bounded)

`ConfHedge`: First algorithm with known bound on R_T for finite T and **unbounded** ℓ assuming changes

No user-defined parameters

Simple example

Simulate from 3 AR(1) models $T=300$

$$
\begin{aligned}
M_1 &: y_{1,t} = 0.75 + 0.25\, y_{1,t-1} + \varepsilon_1, && \varepsilon_1 \sim \mathcal{N}(0, 3.75)\\
M_2 &: y_{2,t} = 0.375\, y_{2,t-1} + \varepsilon_2, && \varepsilon_2 \sim \mathcal{N}(0, 0.55)\\
M_3 &: y_{3,t} = -1.1 + 0.56\, y_{3,t-1} + \varepsilon_3, && \varepsilon_3 \sim \mathcal{N}(0, 0.0618)
\end{aligned}
$$

Observed time series: $y_t = \begin{cases} y_{1,t}, & \text{if } t \in [1, 100], \\ y_{2,t}, & \text{if } t \in (100, 200], \\ y_{3,t}, & \text{if } t \in (200, 300]. \end{cases}$

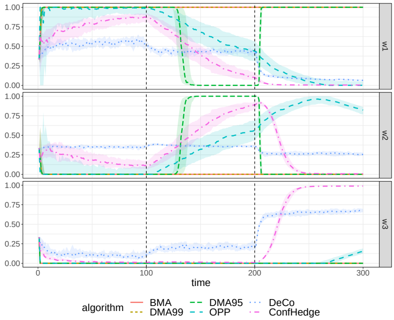
Combination methods

1. Bayesian/ Dynamic Model Averaging:

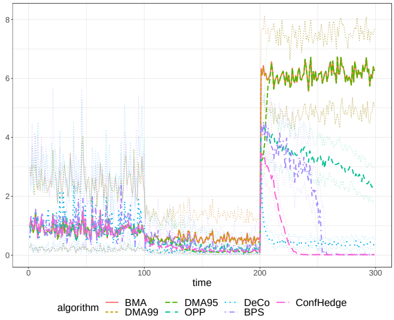
$$p(M_k|t) \propto \prod_{i=1}^{t-1} p(y_{i+1}|M_k, i)^{\alpha^{t-1}}$$

2. Optimal Prediction Pools (Hall & Mitchell, 2007):
   Minimise KL-divergence of true density from combination density
3. Density Combination (Billio et al. 2013):
   Employs SMC to track evolution of weights
4. Bayesian Predictive Synthesis (McAlinn & West, 2019):
   DLMs to track evolution of weights

Simple example



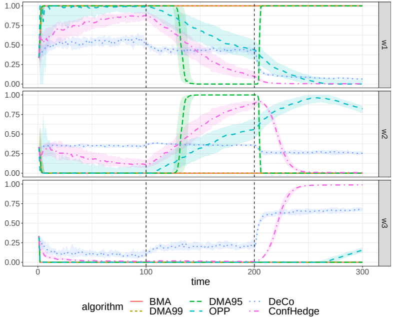
(a) Evolution of weights



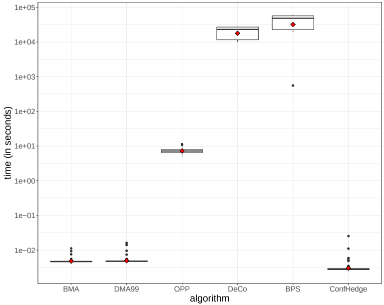
(b) Forecast error

Figure: Evolution of weights and one-step-ahead squared forecast error

Simple example



(a) Evolution of weights



(b) Time

Figure: Evolution of weights and computational time

UK regional housing markets

Data: Quarterly seasonally adjusted regional real house price indices: 1982:Q1 – 2017:Q4 (provided by NationWide)

Predictors:

Regional

- Price-to-income ratio (affordability)
- Income growth
- Unemployment rate
- Labour force growth

National

- Real mortgage rate
- Spread between yields on long/ short-term government securities
- Growth in industrial production
- Number of housing starts
- Growth in real consumption
- Credit Conditions Index (CCI)
- House price uncertainty (HPU)

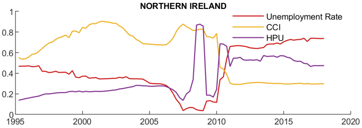
K is of the order 10^3 or 10^4; $T = 143$

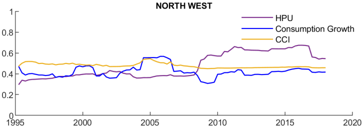
UK regional housing markets

Region	AR(1)	AMC	DeCo	BPS	HM	eDMA	$DMA_{0.99}$	BMA
EA	103.99	**81.20**● ⋆ ○ †	89.60	102.53	94.46	104.75	104.25	110.19
EM	70.94	56.89● ⋆ †	66.12	68.94	68.24	**56.01**∗	59.45	65.19
GL	115.35	**86.97**● ○ †	93.92	109.94	87.94	92.83	93.10	87.11
NI	287.22	**260.94**● ⋆ †	276.51	276.31	276.35	284.25	268.09	311.38
NT	165.35	**120.18**● ⋆ ○ †	148.99	163.97	135.04	126.18	131.19	128.43
NW	64.51	59.85●	60.10	64.49	60.77	59.22	65.86	**58.64**∗
OM	54.23	52.03● ○	**49.77** ∗	56.33	49.78 ∗	56.43	61.89	52.29
OSE	63.14	56.61● ⋆ ○	52.44	60.76	**51.91**	62.60	65.54	57.83
SC	75.73	73.32● ⋆	72.90	79.76	73.77	**70.44**∗	77.06	75.97
SW	62.15	59.88 ● ⋆	58.45	61.19	**57.08**∗	58.63 ∗	58.88 ∗	63.39
WM	53.14	**45.13** ● ⋆ ○ †	49.08	52.57	50.27	51.74	50.69	52.63
WW	145.73	120.00● †	132.61	131.09	125.27	**117.91**	134.09	118.95
YH	91.30	72.11● †	81.92	84.78	77.39	72.79	80.69	**71.44**

DeCo, BPS and HM combine 5 models: M1- AR(1); M2 - AR(1), Contiguous Regions; M3 - AR(1), INCOME, Real Mortgage Rate; M4 - AR(1), Housing Starts; M5 - AR(1), HPU. Forecasting model with lowest MSFE in bold. For the AMC results, ●, ⋆, ○ and † indicate rejection of the null hypothesis of Clark & West (2007) test at the 5% significance level for the AR(1), BMA, eDMA and DeCo benchmarks respectively. For the rest of the models, ∗ indicates rejection of the null hypothesis at 5% significance level for the AMC benchmark. Evaluation period: 1995:Q1 - 2017:Q4.

Uncertainty Index





Key Findings:

Variable importance differs across regions and over time

House Price Uncertainty: Constructed by text analysis of articles of 5 major newspapers obtained from LexisNexis

One of very few variables found important across all regions

HPU Movements coincide with recessions, and events such as the Brexit referendum and withdrawal aggreement negotiations

Conclusions

- Model combination remains interesting
- Consider both model estimation and combination
- Emphasise interpretability both in models and combinations
- Online adaptation of very large models pools is challenging
- Work in ML not yet considered in forecasting