

Using The Yield Curve To Forecast Economic Growth

Parley Ruogu Yang



DEPARTMENT OF
STATISTICS

Modelling with Big Data and Machine Learning Conference
4-5 Nov 2019, Bank of England

“

I do not define time, space, place and motion, as being well known to all

”

— Sir Issac Newton (1687)

Overview

- The yield curve has a well-performing ability to forecast the real GDP growth in the US, c.f. professional forecasters and pure ARIMA models.
- Results depend largely on the estimation and forecasting techniques employed.
- Statistical learning methods play a role in validating and choosing which particular model to use.
- Remark: this talk is leaning towards the statistical methods instead of the current concerns on recession. For this reason the motivation part contains a hands-on example which helps to motivate and introduce the methodology. (Click here to see an example)

Motivation

① Macro & Finance Literature:

$$g_{t,t+k} = \alpha + \beta S_t + \varepsilon_t$$

- Which k fits / forecasts better?
- How to define S_t ? — Variable selection

② Time Series Literature:

- Window-based estimation & forecasting

③ Statistical Learning Literature:

- Bias-Variance trade-off in estimation
- Model selection
- Loss / learning function

An example: asymptotic analysis

Consider a Data Generating Process (DGP) over time
 $\{1, \dots, T_2\} = \{1, \dots, T_1\} \cup \{T_1 + 1, \dots, T_2\}$, with ergodic time series
 $x_t, y_t \in \mathbb{R} \forall t$ with the following evolution:

$$\forall t, \forall j \in \{1, 2\} \quad \varepsilon_{j,t} \sim iidN(0, 1) \quad (1a)$$

$$\text{When } 1 \leq t \leq T_1 \quad y_t = \alpha_1 + \beta_1 x_t + \sigma_1 \varepsilon_{1,t} \quad (1b)$$

$$\text{When } T_1 + 1 \leq t \leq T_2 \quad y_t = \alpha_2 + \beta_2 x_t + \sigma_2 \varepsilon_{2,t} \quad (1c)$$

We concern about the forecast

$$y_{t+1|t} := \mathbb{E}[y_{t+1} | x_{t+1}, y_t, x_t, y_{t-1}, x_{t-1}, \dots].$$

We write the pooled OLS estimation at time t as β_t . When $t < T_1$, we have $\mathbb{E}[\hat{\beta}_t] = \beta_1$ and conveniently $\mathbb{E}[(y_{t+1} - y_{t+1|t})] = 0$ and $\mathbb{E}[(y_{t+1} - y_{t+1|t})^2] = \sigma_1^2$. However, when $t > T_1$, we have

$$\mathbb{E}[\hat{\beta}_t] = \frac{\beta_1 \sum_{\tau=1}^{T_1} (x_\tau - \bar{x})^2 + \beta_2 \sum_{\tau=T_1+1}^t (x_\tau - \bar{x})^2}{\sum_{\tau=1}^t (x_\tau - \bar{x})^2} \quad (2)$$

Thus

$$\mathbb{E}[(y_{t+1} - y_{t+1|t})] = \mathbb{E}[\alpha_2 - \hat{\alpha}_t] + \mathbb{E}[\beta_2 - \hat{\beta}_t]x_{t+1} \quad (3)$$

Now suppose $T_2 \rightarrow \infty$ with the process description 1, then $\mathbb{E}[\hat{\beta}_t] \rightarrow \beta_2$, assuming ergodicity. Thus $\mathbb{E}[(y_{t+1} - y_{t+1|t})] \rightarrow 0$ and $\mathbb{E}[(y_{t+1} - y_{t+1|t})^2] \rightarrow \sigma_2^2$.

- However, if we expand T_2 by admitting the change in the frequency of switching between regimes, then we probably would still have significant bias and larger-than-desired MSE.
- For example, let $T \rightarrow \infty$ with the set $\{1, \dots, T\} = A \cup B$ where A contains some of the points and B contains the remaining of the points. While in A we have the DGP evolving equation 1b, and in B we have equation 1c as the evolution of the datapoint, then $\hat{\beta}_t \xrightarrow{P} a\beta_1 + (1 - a)\beta_2$ where $\frac{|A|}{T} \xrightarrow{P} a$ is assumed to exist.

Small window estimation:

- Transition Period $TP(w) := \{T_1 + 1, \dots, T_1 + w\}$
- Stable Period $SP(w) := \{T_1 + w + 1, \dots, T_2\}$
- When $t \in TP(w)$, we get

$$\mathbb{E}[\hat{\beta}_t] = \frac{\beta_1 \sum_{\tau=t-w}^{T_1} (x_\tau - \bar{x})^2 + \beta_2 \sum_{\tau=T_1+1}^t (x_\tau - \bar{x})^2}{\sum_{\tau=t-w}^t (x_\tau - \bar{x})^2} \quad (4)$$

- When $t \in SP(w)$ we get $\mathbb{E}[\hat{\beta}_t] = \beta_2$.

Large window estimation:

- $TP(w)$ covers almost all time.
- $SP(w)$ covers little time.

If we now expand T by admitting the change in the frequency of switching as previously described, then (with regularity assumption)

$$\left(\frac{nw}{T_n}\right)^{-1} \max(|\hat{\beta}_t - \beta_1|, |\hat{\beta}_t - \beta_2|) \xrightarrow{P} |\beta_1 - \beta_2| \quad (5)$$

While the pooled OLS can only achieve

$$(\max(a, 1 - a))^{-1} \max(|\hat{\beta}_t - \beta_1|, |\hat{\beta}_t - \beta_2|) \xrightarrow{P} |\beta_1 - \beta_2| \quad (6)$$

An example: simulation

Consider $\{x_t, y_t\}_{t=1}^T$ to be drawn from the following distribution and relationship:

$$x_t \sim N(5, 1) \quad \forall t \quad (7a)$$

$$\varepsilon_{j,t} \sim N(0, 1) \quad \forall j, t \quad (7b)$$

$$y_t = \alpha_1 + \beta_1 x_t + \sigma_1 \varepsilon_{1,t} \quad \text{when } t \in A \quad (7c)$$

$$y_t = \alpha_2 + \beta_2 x_t + \theta x_t^2 + \sigma_2 \varepsilon_{2,t} \quad \text{when } t \in B \quad (7d)$$

$$y_t = \alpha_3 + \delta x_t^3 + \sigma_3 \varepsilon_{3,t} \quad \text{when } t \in C \quad (7e)$$

Here A, B, C are partition sets for $\{1, \dots, T\}$.

Consider the following specific parameters: $T = 600$,

$\alpha = (2, -2, -1)$, $\beta = (1, -1)$, $\theta = 3$, $\delta = 1$, $\sigma = (10, 20, 20)$.

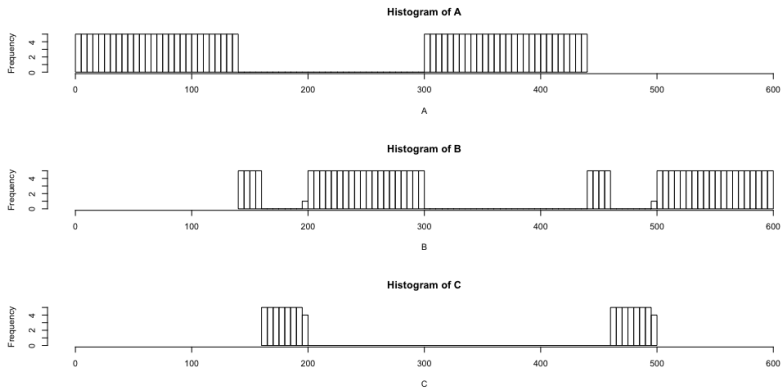


Figure: Partition of the dataset

We consider five models:

$$y_t = \alpha + \beta x_t + \sigma \varepsilon_{1,t} \quad (\text{pooled OLS}) \quad (8a)$$

$$y_t = \alpha + \beta x_t + \sigma \varepsilon_{2,t} \quad (w = 20) \quad (8b)$$

$$y_t = \alpha + \beta x_t + \sigma \varepsilon_{3,t} \quad (w = 50) \quad (8c)$$

$$y_t = \alpha + \beta x_t + \theta x_t^2 + \sigma \varepsilon_{4,t} \quad (w = 20) \quad (8d)$$

$$y_t = \alpha + \beta x_t + \theta x_t^2 + \sigma \varepsilon_{5,t} \quad (w = 50) \quad (8e)$$

At every time $t \in \{100, \dots, T - 1\}$, we estimate the five models and record their forecasts $y_{t+1|t}$. MAE and MSE are then recorded after the iterative process.

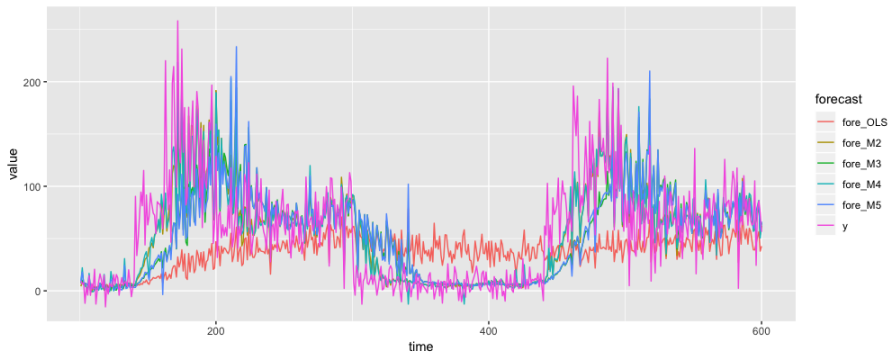


Figure: Forecasting result

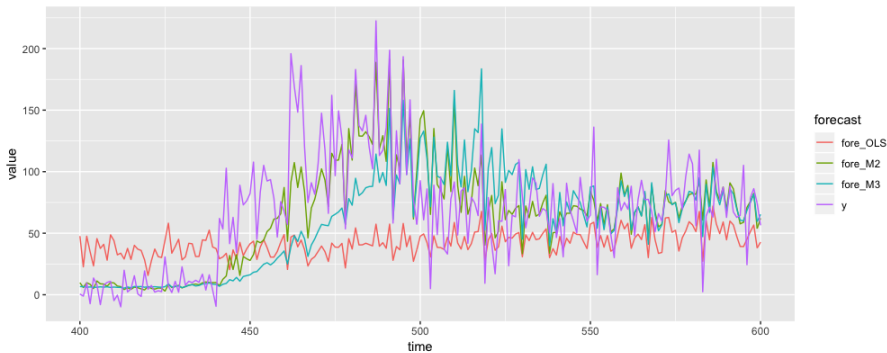


Figure: Forecasting result (models b and c)

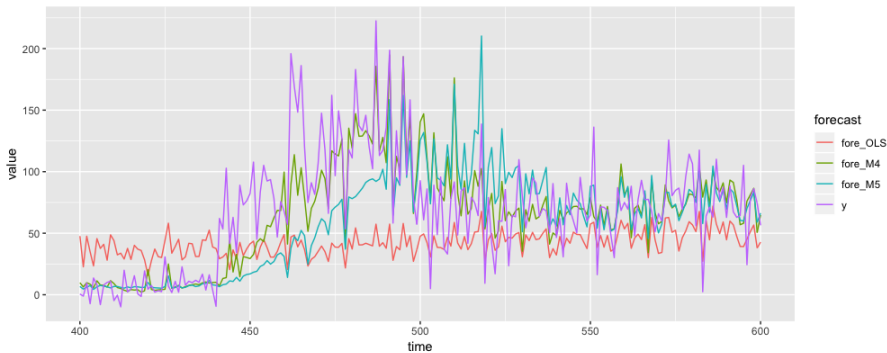


Figure: Forecasting result (models d and e)

Model selection over time

- For each $t \geq 100$, repeat:
 - Run the five models, collect the output of loss function for each model.
 - Pick the model which minimises the loss function at time t and call such a forecast the forecast from the learning model at time t .

Specification of the learning function $L(\{|y_{\tau+1} - y_{\tau}|\}_{100 \leq \tau \leq t-1})$:

$$L(\{x_{\tau}\}_{100 \leq \tau \leq t-1}) = \sum_{100 \leq \tau \leq t-1} \lambda^{t-1-\tau} l(x_{\tau}) \quad (9)$$

where

$$l(x) = l_{\delta, \epsilon}(x) = \begin{cases} (\epsilon - \delta)x + \frac{\delta^2 - \epsilon^2}{2} & \text{if } x > \epsilon \\ \frac{(x - \delta)^2}{2} & \text{if } \delta < x \leq \epsilon \\ 0 & \text{if } x \leq \delta \end{cases} \quad (10)$$

Figure: Model selection over time

Results:

Model	8a	8b	8c	8d	8e
MAE	40.3	19.6	28.6	20.2	29.0
MSE	2828.4	817.2	1662.0	848.7	1706.5

Model	$\lambda = 0.9, \delta = 20, \epsilon = 50$	$\lambda = 0.7, \delta = 5, \epsilon = 25$
MAE	19.0	16.8
MSE	728.5	574.3

Data

- Time indexing: $\{1976Q3, \dots, 2019Q1\} \cong \{1, \dots, T\}$ with $T = 171$.
- Interest rate vector $\mathbf{x}_t \in \mathbb{R}^9$ contains:
 - Effective Federal Funds Rate and 3-month US Interbank Rate.
 - US Treasury yields of the following durations: 3, 12, 24, 36, 60, 84, and 120 months.
- Growth rate of GDP defined as: $k \in \{1, \dots, 12\}$,

$$g_{t,t+k} = \frac{GDP_{t+k} - GDP_t}{GDP_t} \times \frac{400}{k}$$

- For a given k , an information set up to time t is $\Phi_t = \{\mathbf{x}_\tau | 1 \leq \tau \leq t\} \cup \{g_{\tau,\tau+k} | 1 \leq \tau \leq t-k\}$.
- Dickey-Fuller Test (individually) checked for stationarity.
- Comparing the results against SPF forecasts. (k up to 5)

General setting

- Ultimate aim: $\hat{g}_{t,t+k} = f(\Phi_t; \theta_t, \eta_t)$
- $M = \{f(\cdot; \theta, \eta) | \theta \in \Theta, \eta \in H\}$ is then a collection of functions that f can choose from.
 - Model groups 1 to 6: H is a singleton and $\Theta = \mathbb{R}^n$
 - Model groups 7 to 9: H finite and Θ depends on the specification of $\eta \in H$
- Estimation (M_1 to M_6): OLS to estimate the fit

$$g_{\tau, \tau+k} = f(\theta) + \varepsilon_{\tau}, \quad \varepsilon_{\tau} \sim iidN(0, \sigma^2), \quad p \leq \tau \leq t - k$$

Then take the estimated θ as θ_t .

N.B. p depends on the window method specification.

- Assess the forecasts by MAE and MSE.

Methodology (M_1 & $M_{2,w}$)

- Equation of interest for model groups 1 and 2:

$$f(\Phi_t; \alpha_t, \beta_t) = \alpha_t + \beta_t S_t \quad (11)$$

- M_1 : expanding window size estimation & forecast for $t \geq 61$.
- $M_{2,w}$: fixed window size estimation for $w \in \{20, 28, \dots, 124, 132\}$, and forecast during $t \in \{w + k + 1, \dots, 171 - k\}$.

Result (M_1 & $M_{2,w}$)

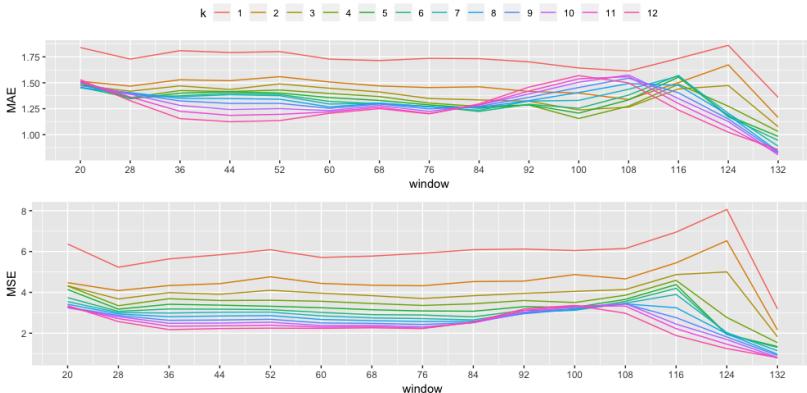


Figure: MAE (top) and MSE (bottom) for different window sizes (w) and lags (k) in the model group 2.

k	$MAE(M_1)$	$MAE(M_2)$		$MSE(M_1)$	$MSE(M_2)$	
		minimum	mean		minimum	mean
1	1.77	1.36	1.72	6.10	3.19	5.94
2	1.67	1.17	1.47	5.29	2.16	4.53
3	1.64	1.08	1.38	5.23	1.82	3.94
4	1.67	1.03	1.33	5.28	1.53	3.48
5	1.72	0.98	1.32	5.19	1.34	3.19
6	1.69	0.94	1.31	4.98	1.30	3.01
7	1.64	0.89	1.32	4.65	1.14	2.88
8	1.55	0.84	1.32	4.15	0.97	2.74
9	1.45	0.83	1.31	3.61	0.91	2.61
10	1.37	0.81	1.30	3.13	0.81	2.52
11	1.27	0.82	1.28	2.68	0.79	2.45
12	1.19	0.86	1.26	2.30	0.80	2.35

Table: MAE (columns 2 to 4) and MSE (columns 5 to 7) for different k from model group 1 and 2. Columns 3 and 6 take the minimum over 15 window sizes and columns 4 and 7 take the mean over 15 window sizes in model group 2.

Methodology ($M_{3,i,j,w}$ to $M_{6,i,j,w}$)

- Define vector **short**_{*t*} as a vector of short term interest rates, in particular, the Federal Funds Rate, 3-month Interbank Rate, 3-month, 12-month, and 24-month Treasury yields.
- Define vector **long**_{*t*} as a vector of long term interest rates consisted of 120-, 84-, 60-, and 36-month Treasury yields.

Models 3 to 6:

$$f(\Phi_t; \alpha_t, \beta_{1,t}, \beta_{2,t}) = \alpha_t + \beta_{1,t} \mathbf{long}_{t,j} + \beta_{2,t} \mathbf{short}_{t,i}$$

$$\begin{aligned} f(\Phi_t; \alpha_t, \beta_{1,t}, \beta_{2,t}, \phi_t) &= \frac{\alpha_t(1 - \phi_t^k)}{1 - \phi_t} \\ &\quad + \sum_{l=0}^{k-1} \left(\beta_{1,t} \phi_t^l \mathbf{long}_{t-l,j} + \beta_{2,t} \phi_t^l \mathbf{short}_{t-l,i} \right) \\ &\quad + \phi_t^k g_{t-k,t} \end{aligned}$$

$$f(\Phi_t; \alpha_t, \beta_{1,t}, \beta_{2,t}) = \alpha_t + \beta_{1,t} \mathbf{long}_{t-1,j} + \beta_{2,t} \mathbf{short}_{t-1,i}$$

$$\begin{aligned} f(\Phi_t; \alpha_t, \beta_{1,t}, \beta_{2,t}, \phi_t) &= \frac{\alpha_t(1 - \phi_t^k)}{1 - \phi_t} \\ &\quad + \sum_{l=0}^{k-1} \left(\beta_{1,t} \phi_t^l \mathbf{long}_{t-l-1,j} + \beta_{2,t} \phi_t^l \mathbf{short}_{t-l-1,i} \right) \\ &\quad + \phi_t^k g_{t-k,t} \end{aligned}$$

Results ($M_{3,i,j,w}$ to $M_{6,i,j,w}$)

Present $\min_{l,j} MAE(M_{l,i,j,w})$ for each k, w ; likewise for MSE.

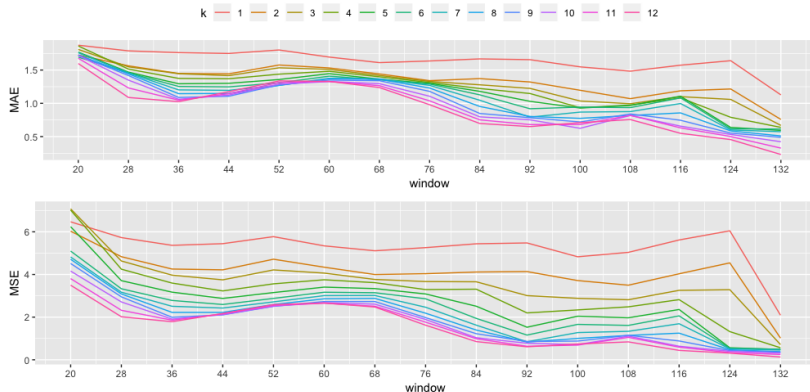


Figure: MAE (top) and MSE (bottom) for different window sizes and different k for the best models in model groups 3 to 6.

Comparison against model group 2.

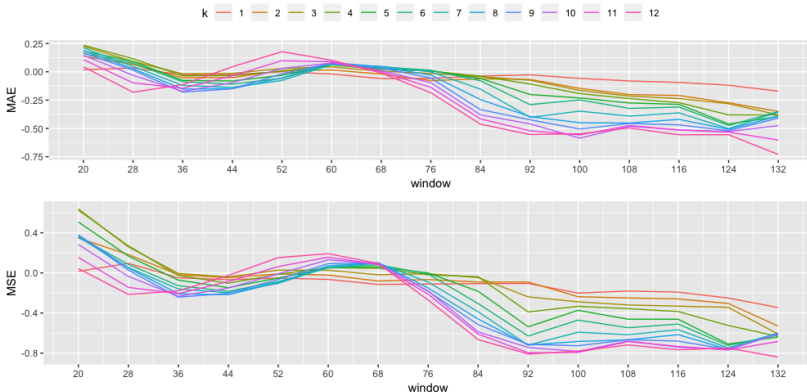


Figure: Proportional comparison of the MAE (top) and MSE (bottom) obtained by models in groups 2 and 3-6, across different k and w . A negative number means the model from groups 3-6 yields lower MAE or MSE compared to group 2, and vice versa.

Comparison against SPF.

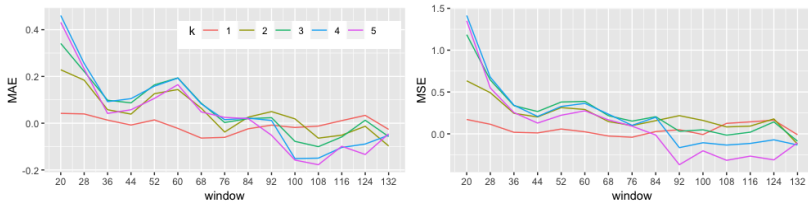


Figure: Proportional comparison of the MAE (top) and MSE (bottom) obtained by the best model in groups 3-6 and the SPF, across different k and w .

Methodology (M_7 to M_9)

- Two main questions:
 - How to ensure we pick the "right" or "almost right" model so that we achieve the minimum?
 - Can we do better? Dynamically picking up models that do well historically?
- For any given (k, w) , at any time $t \geq w + 2k + 1$, there are 4 model groups available, each containing 20 models given by (i, j) . Now, for these total of 80 models which generate forecasts, an assessment is made at time t . Call such assessment $L(\{g_{\tau, \tau+k} - \hat{g}_{\tau, \tau+k}\}_{w+k+1 \leq \tau \leq t-k})$ a loss function. Optimisation is then done through minimising the loss function, and thereafter forecast.

Algorithm for the models. (Labelled as Algo 3.3 in the paper)

Algorithm 3.3: Model groups 7 to 9

1. For $k \in \{1, \dots, 5\}$, $w \in \{20, 28, \dots, 132\}$, repeat:
 - (a) For each $t \in \{w + 2k + 1, \dots, 171 - k\}$, repeat:
 - i. For each $\iota \in \{3, 4, 5, 6\}$, $i \in \{1, 2, 3, 4, 5\}$, $j \in \{1, 2, 3, 4\}$, repeat:
 - A. Over the period $\{w + k + 1, \dots, t - k\}$, run the required OLS estimation such that the forecast based on the model $M_{\iota, i, j, w}$, can be generated, then forecast.
 - B. Compute the output of loss function for each $M_{\iota, i, j, w}$.
 - ii. Pick the model, say M^* which minimises the loss function.
 - iii. Use M^* as the model to make forecast at time t ,^a and call such a forecast the forecast from the learning model at time t .
 - (b) Collect the MAE and MSE for the overall forecasts from the learning model.

^aRunning the relevant OLS regression prior to the forecast of M^* is also required.

- M_7 : A relatively naïve way:

$$L(\{g_{\tau, \tau+k} - \hat{g}_{\tau, \tau+k}\}_{w+k+1 \leq \tau \leq t-k}) = |g_{t-k, t} - \hat{g}_{t-k, t}|$$

- M_8 : Full-history learning:

$$L(\dots) = \sum_{w+k+1 \leq \tau \leq t-k} l(|g_{\tau, \tau+k} - \hat{g}_{\tau, \tau+k}|)$$

- M_9 : Discounted-history learning:

$$L(\dots) = \sum_{w+k+1 \leq \tau \leq t-k} \lambda^{t-k-\tau} l(|g_{\tau, \tau+k} - \hat{g}_{\tau, \tau+k}|)$$

where $\lambda \in (0, 1]$

The design of $l(\cdot)$:

- Vapnik (2000): $j \in \{1, 2\}$

$$l(x) = l_{\epsilon}(x) = \mathbb{1}[x > \epsilon](x - \epsilon)^j$$

- Huber (1964):

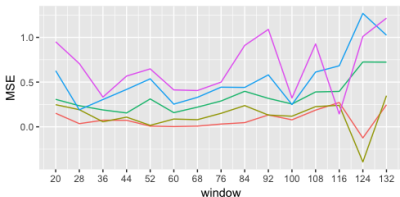
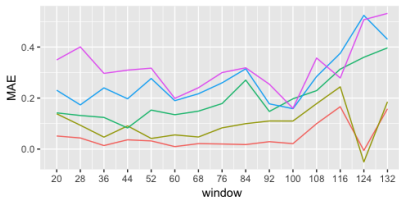
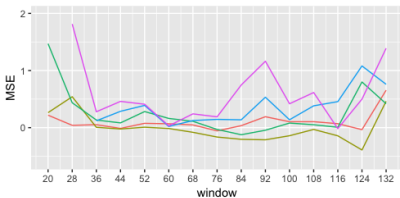
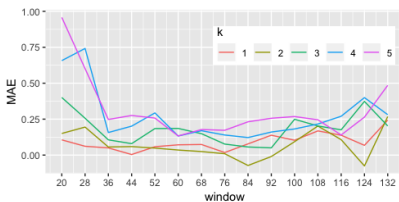
$$l(x) = l_{H,\epsilon}(x) = \begin{cases} \epsilon x - \frac{\epsilon^2}{2} & \text{if } x > \epsilon \\ \frac{x^2}{2} & \text{if } x \leq \epsilon \end{cases}$$

- Another way (introduced in the motivation section):

$$l(x) = l_{\delta,\epsilon}(x) = \begin{cases} (\epsilon - \delta)x + \frac{\delta^2 - \epsilon^2}{2} & \text{if } x > \epsilon \\ \frac{(x - \delta)^2}{2} & \text{if } \delta < x \leq \epsilon \\ 0 & \text{if } x \leq \delta \end{cases}$$

Results (M_9)

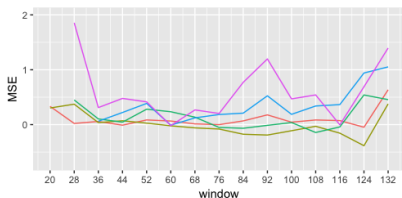
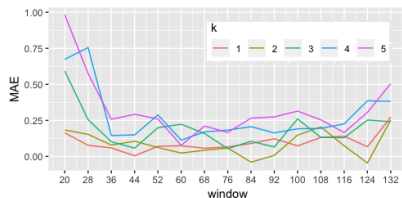
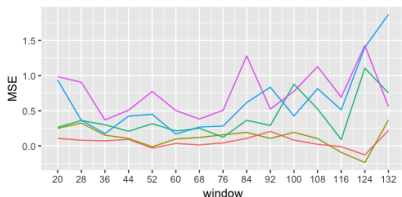
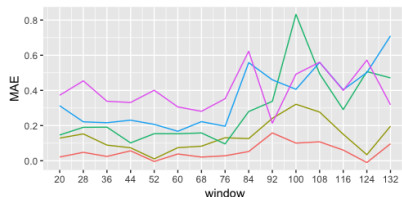
- M_7 and M_8 work badly, minor achievement can be seen occasionally.
- Let $M_{9,1,w}(\lambda, \epsilon)$ to be the model which employs $I_{H,\epsilon}$ as the specification of I .
- Let $M_{9,2,w}(\lambda, \epsilon, \delta)$ to be the model which employs $I_{\delta,\epsilon}$ as the specification of I .



Left to right: Proportional comparison of the MAE (left) and MSE (right) obtained by each model and the best model from groups 3-6, across different k and w .

Top to bottom: $M_{9,1}(0.5, 0.5)$ and $M_{9,1}(0.9, 0.5)$.

Note: three outliers in the top right plot are dropped.



Left to right: Proportional comparison of the MAE (left) and MSE (right) obtained by each model and the best model from groups 3-6, across different k and w .

Top to bottom: $M_{9,2}(0.75, 2.5, 0.5)$ and $M_{9,2}(0.7, 2, 0.7)$.

Improvement counts

	k	1	2	3	4	5
$M_{9,1}(0.5, 0.5)$	MAE	0	3	0	0	0
	MSE	3	10	3	0	1
$M_{9,1}(0.9, 0.5)$	MAE	1	1	0	0	0
	MSE	1	1	0	0	0

	k	1	2	3	4	5
$M_{9,2}(0.75, 2.5, 0.5)$	MAE	2	0	0	0	0
	MSE	3	3	0	0	0
$M_{9,2}(0.7, 2, 0.7)$	MAE	0	2	0	0	0
	MSE	3	9	5	1	2

Example output

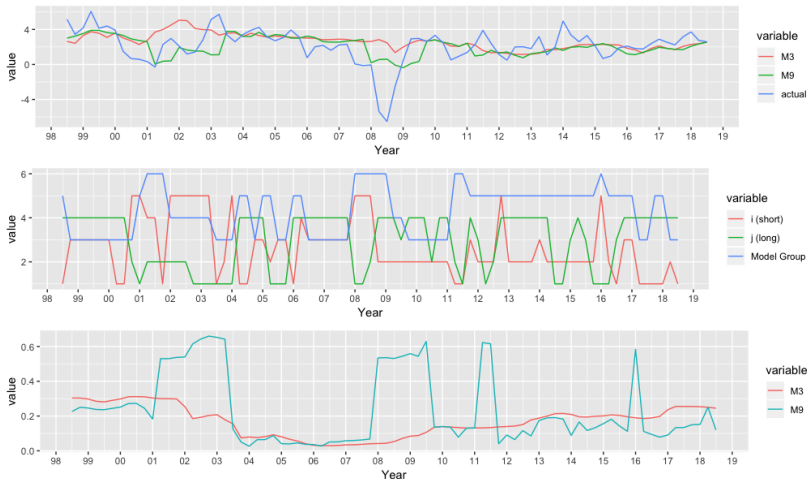


Figure: From top to bottom: $\hat{g}_{t,t+2}$ from different models and the actual $g_{t,t+2}$; the corresponding model that model 9 chooses over time; R^2 for the estimations of different models at each time.

Conclusion

- ① Which k ?
 - The larger the k , the less forecasting error it makes.
 - Small k can well outperform SPF forecasts.
 - Learning functions help to reduce "structural break" in the betterment.
- ② Estimation and forecasting methods:
 - Variety in variable selection & window size methods bear fruit to the improvement.
 - Workhorse to the learning algorithms.
- ③ Future:
 - Engagement with macro & finance literature for variable selection and functional forms.
 - Asymptotics for learning function choices. (Harder ones than the initial example).

“

The only function of economic forecasting is to make astrology look respectable.

”

— Professor Ezra Solomon (1985)