Clustering of stock time series data and price prediction

YUHAO KANG

5/8/2018

Predictability of stock markets

- Efficient market hypothesis (Fama, 1970)
- $y(t)$=$g(t) + c(t)$ (Hodrick & Prescott, 1997)

Trend component Cyclic component

- If the short-range correlation of $c(t)$ can be utilized? (Nair, 2017)

Time series data

	Date	Open	High	Low	Close	Volume	OpenInt
1	2012-12-13	15	15	15	15	100	0
2	2012-12-14	19	30	19	30	144600	0
3	2012-12-17	31.5	65	31.5	65	68600	0
4	2012-12-18	65	89	65	80	43600	0
5	2012-12-19	80	84	78	84	24000	0
6	2012-12-20	84	84	80	80.25	33300	0
7	2012-12-21	80.5	81.5	75	80	20700	0
8	2012-12-24	81	93.5	80	80	3700	0
9	2012-12-26	80	90	77	77	56100	0
10	2012-12-27	75	75.02	55	59	54300	0
11	2012-12-28	58.5	75	56.5	72.35	94400	0
12	2012-12-31	72	90	72	82	43500	0
13	2013-01-02	83.5	85	75.75	76	26000	0
14	2013-01-03	75.5	83.9	73.5	81.25	80800	0
15	2013-01-04	81.5	86	81.5	85	26000	0
16	2013-01-07	86	87.75	79.8	79.8	31800	0
17	2013-01-08	79	83.99	78.25	80	21800	0
18	2013-01-09	80.75	86.25	80.75	84	53100	0
19	2013-01-10	86.8	99.49	86	98	38900	0
20	2013-01-11	100	107	97.15	106.5	37000	0
21	2013-01-14	111	120	110	120	42000	0
22	2013-01-15	117.05	129.49	108	110	97600	0

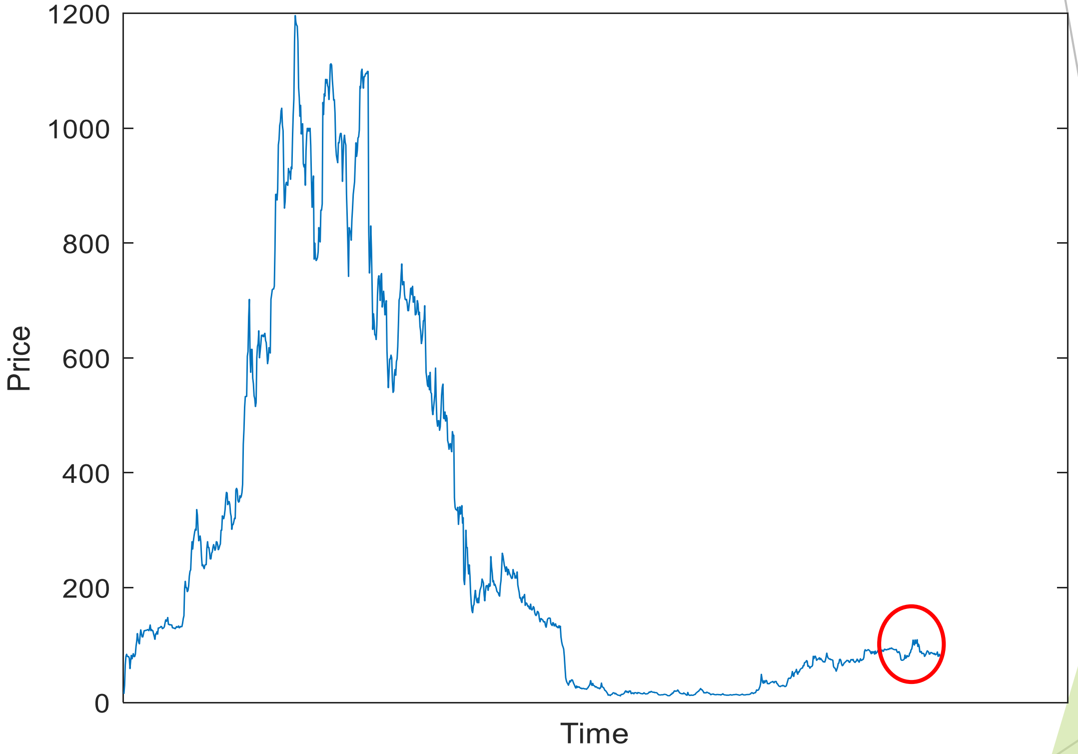

Fig. 1. Time series data of AAMC. (Left panel) original data, here we only focus on the close column. (Right panel) run chart based on the data, time range is 2012/12/13—2017/11/10.

Aim

- 30-days-ahead series $p(t) = [y(t), y(t-1) \dots y(t-29)]$

- 5-days-after data $r(t) = [y(t+1), y(t+2) \dots y(t+5)]$

Training dataset

- $$[P\ R] = \begin{bmatrix} p(t) & r(t) \\ p(t-1) & r(t-1) \\ \vdots & \vdots \\ p(t-29) & r(t-29) \end{bmatrix}$$

Input features $V_1 \sim V_{30}$

Responses $V_{31} \sim V_{35}$

Problem: In real application, amount of input features may be huge!

Regression trees

```
Call:
rpart(formula = V31 ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 +
    V9 + V10 + V11 + V12 + V13 + V14 + V15 + V16 + V17 + V18 +
    V19 + V20 + V21 + V22 + V23 + V24 + V25 + V26 + V27 + V28 +
    V29 + V30, data = xt, method = "anova")
  n= 30

          CP nsplit rel error    xerror      xstd
1 0.62179905      0 1.0000000 1.0972339 0.3162066
2 0.03983376      1 0.3782010 0.5438167 0.1511534
3 0.01000000      2 0.3383672 0.5446017 0.1501308

Variable importance
 V7  V8  V1  V6  V2  V5 V17 V18 V19  V9
 30  15  14  14  11  11   2   1   1   1

Node number 1: 30 observations,    complexity param=0.621799
  mean=2.542333, MSE=0.01068456
  left son=2 (22 obs) right son=3 (8 obs)
  Primary splits:
      V7  < 2.475  to the right, improve=0.6217990, (0 missing)
      V24 < 2.595  to the left,  improve=0.4437202, (0 missing)
      V6  < 2.475  to the right, improve=0.4130582, (0 missing)
      V8  < 2.505  to the right, improve=0.3743746, (0 missing)
      V5  < 2.475  to the right, improve=0.3675202, (0 missing)
  Surrogate splits:
      V1 < 2.705  to the left,  agree=0.867, adj=0.500, (0 split)
      V6 < 2.465  to the right, agree=0.867, adj=0.500, (0 split)
      V8 < 2.465  to the right, agree=0.867, adj=0.500, (0 split)
      V2 < 2.73   to the left,  agree=0.833, adj=0.375, (0 split)
      V5 < 2.455  to the right, agree=0.833, adj=0.375, (0 split)

Node number 2: 22 observations,    complexity param=0.03983376
  mean=2.493182, MSE=0.001548967
  left son=4 (11 obs) right son=5 (11 obs)
  Primary splits:
      V17 < 2.525  to the right, improve=0.3746832, (0 missing)
      V18 < 2.55   to the right, improve=0.2946512, (0 missing)
      V9  < 2.61   to the left,  improve=0.2653112, (0 missing)
      V11 < 2.61   to the left,  improve=0.2166822, (0 missing)
      V16 < 2.5825 to the right, improve=0.2042735, (0 missing)
  Surrogate splits:
      V9  < 2.61   to the left,  agree=0.818, adj=0.636, (0 split)
      V18 < 2.55   to the right, agree=0.818, adj=0.636, (0 split)
      V19 < 2.55   to the right, agree=0.818, adj=0.636, (0 split)
      V7  < 2.595  to the left,  agree=0.773, adj=0.545, (0 split)
      V8  < 2.595  to the left,  agree=0.773, adj=0.545, (0 split)
```

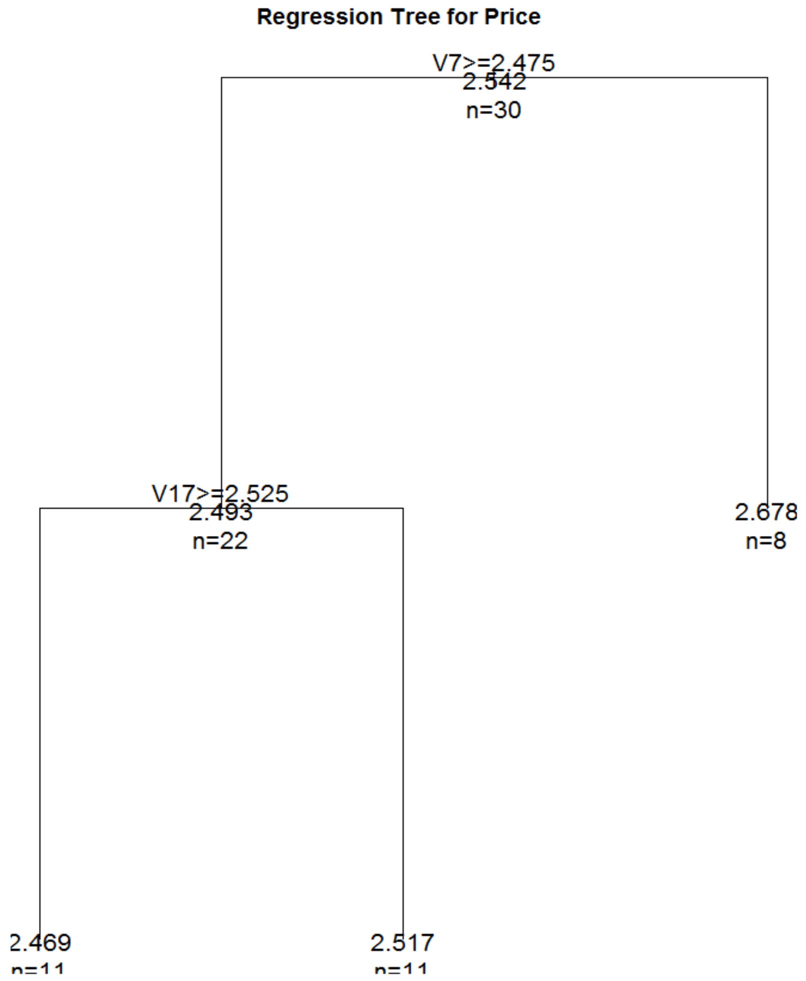

Fig. 2. Regression trees

Principle inputs tt $\subseteq \{V_1 \sim V_{30}\}$

Reduce Dimensionality

- $[P\ R] = \begin{bmatrix} p(t) & r(t) \\ p(t-1) & r(t-1) \\ \vdots & \vdots \\ p(t-29) & r(t-29) \end{bmatrix}$
- $P_{reduced} = \{V_i\ for\ i\ in\ tt\} \subseteq P$
- $final = [P_{reduced}\ \ R]$

SOM and k-means clustering

- $final = [P_{reduced} \ \ R]$, try 6 clusters

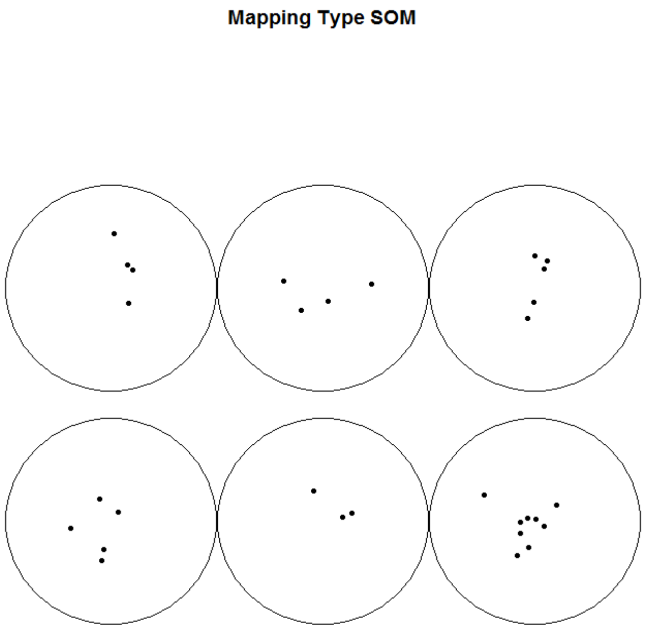

Fig.3. SOM mapping of dataset *final*

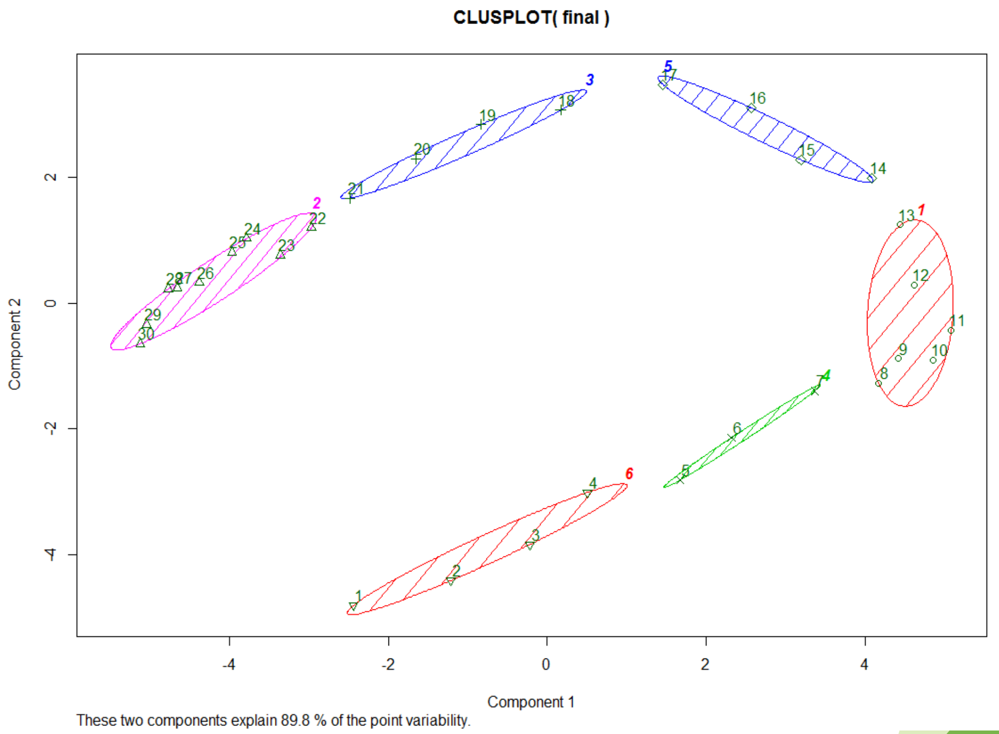

Fig.4. k-means clustering

Centroids and prediction

- $C_{i=1\sim6} = [P_{ci} \ \ R_{ci}]$
- If $\max(R_c) > mean(P_c)$,   price up
- Else, price down
- $\begin{matrix} Input \\ W(t = 30\sim1) \end{matrix} \Longrightarrow W_{reduced}(t \subseteq tt) \xRightarrow{\text{Distance matrix}}$ Closest $P_{ci} \Longrightarrow$ Prediction

Conclusion

- Results vary from various samples, performance is not stable
- SOM and k-means make no big difference
- Influence factors: cluster number, data size, time interval

	prediction	reality	correctness
1	1	1	TRUE
2	1	1	TRUE
3	1	1	TRUE
4	0	1	FALSE
5	0	1	FALSE
6	0	1	FALSE
7	0	1	FALSE
8	0	1	FALSE
9	0	1	FALSE
10	0	1	FALSE
11	0	1	FALSE
12	0	1	FALSE
13	0	1	FALSE
14	0	1	FALSE
15	0	1	FALSE
16	0	0	TRUE
17	0	0	TRUE
18	0	0	TRUE
19	0	0	TRUE
20	0	0	TRUE

	prediction	reality	correctness
1	1	1	TRUE
2	0	1	FALSE
3	0	1	FALSE
4	0	1	FALSE
5	0	1	FALSE
6	0	1	FALSE
7	0	1	FALSE
8	0	1	FALSE
9	0	1	FALSE
10	0	1	FALSE
11	0	1	FALSE
12	0	1	FALSE
13	0	1	FALSE
14	0	1	FALSE
15	0	1	FALSE
16	0	0	TRUE
17	0	0	TRUE
18	0	0	TRUE
19	0	0	TRUE
20	0	0	TRUE

Fig.5. Prediction of AAPL. Left panel is k-means method; right panel is SOM

Thanks!