Combination of hidden markov model and MLE or James-Stein analysis with two-sigma method for stock selection

Xu siao
ID:20377629
Department of mathematics
HKUST

1 Introduction

The hidden Markov model (HMM) is typically used to predict the hidden regimes of observation data. Therefore, this model finds applications in many different areas, such as speech recognition systems, computational molecular biology and financial market predictions. Maximal likelihood estimate is one of the most popular statistical tool with application in almost every area in modern discipline. Later on, scholars extend MLE further to James-Stein estimate, which performs better than MLE in terms of mean square error at some cost of precision in high dimension data.

2 Problem

In my project, I explore the possibility of combining HMM and MLE or J-S for stock selection. My method is based on following six assumptions:

(1) Macroeconomic regimes have some effects on stock performances
(2) The stocks share the similar performance in similar macroeconomic environment
(3) MLE and J-S are both good tool for estimating mean of stock price
(4) J-S perhaps is better than MLE in terms of estimating mean of stock price since J-S has smaller mean square error than MLE, meaning J-S estimator is closer to real stock price mean in comparison with MLE
(5) Stock price follows normal distribution, at least locally normal distribution (follows normal distribution in each specific time interval)
(6) 2-sigma is a simple and useful method for ensuring open position, closing position and clearing position

If above six assumptions make sense, which means, based on HMM, we can predict the macroeconomic performance of next state, then we look back the historical data and figure out the dates with the same macroeconomic environment. Extracting stock price from dates with same macroeconomic regimes, we compute mean price and, along with two-sigma method, establish benchmark. Finally, by comparing benchmark and current stock price, we decide our strategy for next state.

3 methodology

3.1 Hidden markov model

The hidden markov model is a model that can capture the hidden states of observation data. An observation at time t of an HMM has a certain probability distribution corresponding to a possible state.
The basic elements of a hidden markov model are:

- Length of observation data, T
- Number of states, N
- Number of symbols per state, M
- Observation sequence, O = $\{\mathbf{O_t}, \boldsymbol{t = 1, 2, 3, \dots, T}\}$
- Hidden state sequence, Q = $\{\mathbf{q_t}, \boldsymbol{t = 1, 2, 3, \dots, T}\}$
- Symbol values of each state, $\{\mathbf{S_i}, \boldsymbol{i = 1, 2, 3, \dots, N}\}$
- Transitional matrix, A = $(\mathbf{a_{ij}}), \boldsymbol{where\ a_{ij} = P(q_t = S_j \,|q_{t-1} = S_i)}$
- Vector of initial probability of being in state 1, $\mathbf{p_i} = \boldsymbol{P(q_1 = S_i), i = 1, 2, 3, \dots, N}$
- Observation probability matrix, B = $(\mathbf{b_{ik}}), \boldsymbol{where\ b_{ik} = P(O_t = v_k \,|q_t = S_i), i = 1, 2, .., N\ and\ k = 1, 2, 3, \dots, M}$

There are three sets of algorithm which solves out three problems: $\boldsymbol{(1)\ the\ probability\ of\ obsevations, P(O|\lambda)}$ $\boldsymbol{(2)\ choose\ the\ best\ corresponding\ state\ sequence\ Q = \{q_1, q_2, q_3, \dots, q_T\}}$ $\boldsymbol{(3)\ calibrate\ the\ best\ HMM\ parameter\ \lambda(A, B, p)\ to\ maximize\ P(O|\lambda)}$

Forward algorithm:

1: initialization: for i = 1,2,⋯,N

$$\boldsymbol{\alpha_{t=1}(i) = p_i b_i(O_1)}$$

2: recursion: for t =2,3,⋯,T and for j = 1,2,⋯,N

$$\boldsymbol{\alpha_t(j) = [\sum_{i=1}^{N} \alpha_{t-1}(i)\alpha_{ij}]b_j(O_t)}$$

3: output:

$$\mathbf{P(O|\lambda)} = \boldsymbol{\sum_{i=1}^{N} \alpha_T(i)}$$

The Viterbi algorithm

1: initialization:

$$\sigma_1(j) = p_j b_j(O_1),\ j = 1, 2, \dots, N$$
$$\gamma_1(j) = 0$$

2: recursion: for 2 ≤ t ≤ T, 1 ≤ j ≤ N

$$\sigma_t(j) = \max_i [\sigma_{t-1}(i)\alpha_{ij}] b_j(O_{t+1})$$

$$\gamma_t(j) = argmax_i[\sigma_{t1}(i)\alpha_{ij}]$$

3: output:

$$q_t^* = \gamma_{t+1}(q_{t+1}^*), t = T-1, \dots, 1$$

$$q_T^* = argmax_i[\gamma_T(i)]$$

The Baum-Welch algorithm:

1: initialization: input parameters λ, the tolerance tol, and a real number del

2: repeat until del < tol

Calculate P(O|λ) using forward algorithm

Calculate new parameters λ^*: $for\ 1 < i < N$

$$p_i^* = \gamma_1(i)$$

$$a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},\ 1 \le j \le N$$

$$b_{ik}^* = \frac{\sum_{t=1, O_t = v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)},\ 1 \le k \le M$$

Calculate del = |P(O,λ^*) – P(O,λ)|
Update $\lambda = \lambda^*$

3: output: parameters λ

3.2 MLE and James-Stein Estimate

Suppose X is a p*n data matrix following multivariate normal distribution and each column in X is independent and identically distributed with each other, then:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu^*$$

$$\hat{\sum_{n}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n^*)(X_i - \mu_n^*)\hat{}T$$

James-Stein estimator:

$$\mu_{JS} = \left(1 - \frac{\sigma^2(p-2)}{||\mu_{MLE}||}\right)\mu_{MLE}$$

3.3 two-sigma method

(1) mean adding two standard deviations as upper bound

upper bound = μ + 2δ

(2) mean subtracting two standard deviations as lower bound

lower bound = μ + 2δ

4 data source

(1) all of stocks data in hushen300 from 20070131 to 20161230 extracted from financial terminal wande
(2) hushen300 index in years from 200701 to 201612 extracted from wande
(3) consumer price index in China from 200701 to 201612 extracted from wande
(4) industrial production index in China from 200701 to 201612 extracted from wande
(5) short-period bond index in China from 200701 to 201612 extracted from wande

5 two steps and results

The first step is to find regimes of macro variables, and the second step is to make stock selection. Among these four macroeconomic indicators, CPI and INDPRO are monthly data, while bond index and the hushen300 index are daily data. Thus, we use monthly frequency to take advantage of fresh data. Furthermore, using monthly data will give a moderate rotation rate, the average of time (months) a stock stays in the portfolio, for

our composite portfolio.

5.1 first step

First, we calibrate HMM's parameters, using the Baum–Welch algorithm, and one of the four macroeconomic variables above. We then use the obtained parameters to predict the corresponding hidden regimes of each economic indicator using the Viterbi algorithm. We use monthly historical data of the variables from January 2007 to December 2013 to predict next month, namely, January 2014 and repeat the process over and over and get all regimes for first macroeconomic indicator. So on and so forth, we repeat the process until four macroeconomic regimes for January 2014 to December 2016 are figured out.

date	cpi	hushen_300	ipi	delta(interest)
2014-01	bad	good	bad	good
2014-02	bad	good	bad	good
2014-03	bad	good	bad	bad
2014-04	bad	good	bad	good
2014-05	bad	good	bad	good
2014-06	bad	good	bad	good
2014-07	bad	good	bad	bad
2014-08	bad	good	bad	good
2014-09	bad	good	bad	good
2014-10	bad	good	bad	good
2014-11	bad	good	bad	good
2014-12	bad	bad	bad	good
2015-01	bad	bad	bad	good
2015-02	bad	bad	bad	good
2015-03	bad	bad	bad	good
2015-04	bad	bad	bad	good
2015-05	bad	bad	bad	good
2015-06	bad	bad	bad	bad
2015-07	bad	bad	bad	bad
2015-08	bad	bad	bad	bad
2015-09	bad	bad	bad	good
2015-10	bad	bad	bad	good
2015-11	bad	bad	bad	bad
2015-12	bad	bad	bad	good
2016-01	bad	good	bad	bad
2016-02	bad	good	bad	bad
2016-03	bad	bad	bad	good
2016-04	bad	bad	bad	good
2016-05	bad	bad	bad	good
2016-06	bad	bad	bad	good
2016-07	bad	bad	bad	good
2016-08	bad	bad	bad	bad

2016-09	bad	bad	bad	good
2016-10	bad	bad	good	good
2016-11	bad	bad	good	good
2016-12	bad	bad	good	good

5.2 second step

Each month, after predicting economic regimes of the four macroeconomic variables (CPI, INDPRO, hushen300, bond index) for a period from the target month back to the past in Step 1, we look back in history for periods with similar regimes as those of the next month. For example, if the predicted regimes for the next month of inflation, industrial production index, market index and bond index are good, bad, bad, good, respectively, we will look from recent time back to the past to find the months that these four variables had the same regimes good, bad, bad, good.
And by using MLE and J-S to determine the mean of stock price in these periods (since macroeconomic performance in these dates are similar to each other, the data in these periods are the most representative). With the use of two-sigma method, we find the upper bound and lower bond of our stock price mean. Finally, we check if our current stock price is higher, lower or within the established stock price bound interval for making our decision. For example, if today's closed price for one specific stock is lower than the lower bound of our predicted price interval, we buy in that stock. if today's closed price for one specific stock is higher than the upper bound of our predicted price interval, we sell out that stock (if that stock is in our portfolio, otherwise do nothing). If today's closed price for one specific stock is within price boundary, we do nothing whatever the specific stock is within or without our portfolio. And the same logic applies to each stock. In my project, I choose ten stocks and stick to them for implementing my strategy.

The following is a small tablet for J-S estimator (of course, the complete data table is much bigger):

	000001.SZ	000002.SZ	000008.SZ	000009.SZ	000060.SZ
	平安银行	万科 A	神州高铁	中国宝安	中金岭南
2014-01	8.756734	4.936819	5.003517	5.779992	5.825485
2014-02	10.46968	6.763383	7.054081	6.511112	5.651222
2014-03	10.46968	6.763383	7.054081	6.511112	5.651222
2014-04	10.40379	6.700675	7.100178	6.652666	5.613187
2014-05	7.99884	4.541495	4.639414	5.341614	5.251743
2014-06	10.26555	6.684519	7.154309	6.819909	5.514117
2014-07	10.26555	6.684519	7.154309	6.819909	5.514117
2014-08	10.16632	6.68719	7.140458	6.888272	5.515506
2014-09	10.11635	6.751793	7.132485	6.893932	5.576751
2014-10	na	na	na	na	na

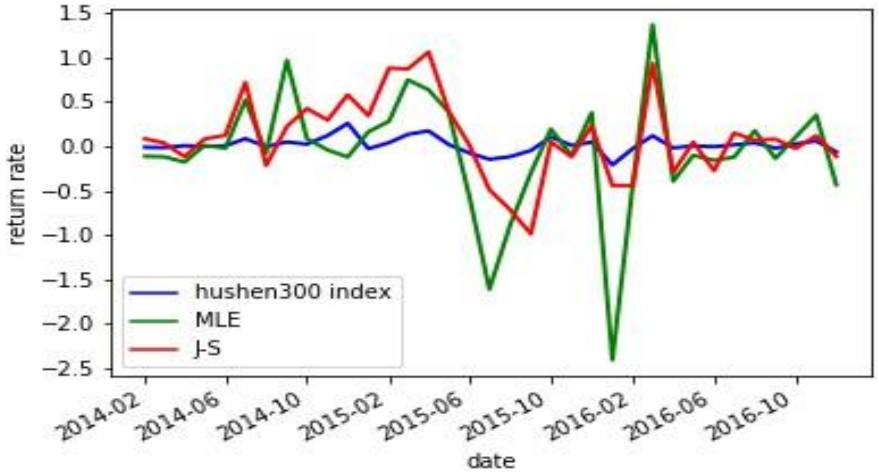
2014-11	na	na	na	na	na
2014-12	11.03	9.4	20.15	12.97	7.6
2015-01	11.6061	10.03851	22.0155	13.00061	7.818171
2015-02	12.68907	11.03959	21.50137	12.66647	8.166711
2015-03	12.8818	11.44942	22.27942	12.65678	8.611258
2015-04	13.99132	7.486511	5.467964	7.358755	15.3926

The following is a small tablet for MLE (of course, the complete data table is much bigger):

	000001.SZ	000002.SZ	000008.SZ	000009.SZ	000060.SZ
	平安银行	万科 A	神州高铁	中国宝安	中金岭南
2014-01	15.142	8.536667	8.652	9.994667	10.07333
2014-02	14.47375	9.35	9.751875	9.00125	7.8125
2014-03	14.47375	9.35	9.751875	9.00125	7.8125
2014-04	14.27706	9.195294	9.743529	9.129412	7.702941
2014-05	14.90813	8.464375	8.646875	9.955625	9.788125
2014-06	13.92737	9.068947	9.706316	9.252632	7.481053
2014-07	13.92737	9.068947	9.706316	9.252632	7.481053
2014-08	13.7265	9.029	9.641	9.3005	7.447
2014-09	13.59048	9.070476	9.581905	9.261429	7.491905
2014-10	na	na	na	na	na
2014-11	na	na	na	na	na
2014-12	11.03	9.4	20.15	12.97	7.6
2015-01	11.735	10.15	22.26	13.145	7.905

Obviously, the MLE is a bit larger than J-S as expected since J-S is a shrinkage of MLE. Then, based on the complete table and strategy, we make our trading decisions and form our portfolio for each month over and over.

The following is the comparisons among return rate for MLE and J-S and hushen300 index.



6 conclusion and further work

From the comparison between MLE, J-S and heshen300 index, it is easy to find out that the performance of J-S is better than that of MLE since when the strategy gets profits, J-S scheme makes a bit higher profit than MLE scheme and when the strategy suffers from loss, J-S scheme suffers a bit smaller loss than MLE scheme (at least for most of periods, the story makes sense). The underlying reason is as noted on the very beginning; J-S has smaller mean square error than MLE at cost of a bit higher bias. In my strategy, a smaller mean square error means better approximation for real price mean and better prediction for benchmark used for stock selection. On the other hand, the strategy performs not bad, at least, more than half of time, it beats the market in comparison with hushen300 index return rate. But it is worth noting that the degree of loss is much bigger than that of the profit in a specific.

There are some points worthwhile to watch out:

(1) perhaps the stock price does not follow normal distribution, in fact, the stochastic tools can model the stock price fluctuations better than just normal distribution. If it is so, the combination of stochastic analysis and J-S estimate can generate better results.
(2) The selection of macroeconomic indicator should be very careful. Some macroeconomic indicator is likely to have no business with stock performance and, on the contrary, good macroeconomic indicator has higher correlation with stock price and choosing such guys will lead to very good results.
(3) It is also likely to change the strategy after figuring out the macroeconomic regimes. For example, after we extract the data from these periods sharing the same macroeconomic performance, we can design up a scoring system as benchmark for stock selection based on the data, which perhaps produce more interesting results.
(4) All models are wrong; some just fits the real world better than the others. Never believe in what we have. Perhaps my strategy is absolute wrong, everything is just coincidence. But if my strategy can seize a bit of motion of real world, it is a not bad benchmark for us to some extent.

7 reference

James–Stein estimation problem for a multivariate normal random matrix and an improved estimator, by **XiaoqianLiu**[a],**LiangyuanLiu**[b] and **JianhuaHu**

An Introduction to Hidden Markov Models, by L. R. .Rabiner and B. H. Juang
.