Proceedings of Business and Economic Studies, 2021, Volume 4, Issue 5
http://ojs.bbwpublisher.com/index.php/PBES
ISSN Online: 2209-265X
ISSN Print: 2209-2641

Pairs Trading Strategy for A and H Shares Based on Kalman-HMM Approach

Ming Zang*

Shanghai University of International Business and Economics, Shanghai 201600, China

***Corresponding author:** Ming Zang, 18805277670@163.com

Abstract: Pairs trading is a statistical arbitrage strategy that takes advantage of unbalanced financial markets. A common difficulty for quantitative trading participants is the detection of market institutional changes in financial markets. In order to solve this issue, the hidden Markov model (HMM) is applied for status detection. The research objective is to use Kalman filter to predict and the hidden Markov model (HMM) to identify state transitions on the basis of screening transaction pairs with obvious co-integration relationship. This research would prove the profitability of the strategy and the ability to resist risk through the combination of these two methods with real data. The empirical results showed that compared with the traditional cointegration strategy, the holding yield increased from 1.6% to 16.2% and the maximum pullback reduced to 0.02%. Further research is required to improve trading rules.

Keywords: Pairs trading; Kalman filtering; State transition; Hidden Markov model (HMM); Cointegration relationship

Publication date: October 2021; ***Online publication:*** October 29, 2021

1. Introduction

Pairs trading, which was first proposed by Morgan Stanley's quantification team in mid-1980s, is a statistical arbitrage strategy. Its core idea is first, building a highly relevant asset pair with iso-varying prices from historical data; second, assuming that the price difference of the selected asset pair is subjected to the mean reply (mean-reverting) characteristic [1]; finally, based on the extent of the future price difference from the historical average, short the middle assets to the overvalued assets, assume the undervalued assets, wait for the price difference to respond to the historical average level while settling the position at the same time and for investors to achieve positive returns.

Foreign research on this strategy first appeared in a literature by Gatev and other researchers [2]. The development process of statistical arbitrage strategy is mainly divided into four major mainstream strategies: distance method, cointegration method, stochastic spread method, and machine learning methods. Gatev, Goetzmann, and Rouwenhorst [3] used the minimum distance method to find matching stocks. They selected stocks with the minimum distance by calculating the distance between two stocks standardized price sequences to build the portfolio. Huck performed an empirical analysis for Gatev's classical theory and found that the assets obtained by cointegration method have more convergence [4]. Based on the cointegration theory, Vidyamurthy constructed a Pearson correlation coefficient based on the common factor yield to measure the absolute value distance between stocks [5]. Jurek and Yang [6] applied the stochastic control theory and dynamically configured arbitrage models as well as risk-free assets. The machine learning method was introduced in pairs trading by Huck [7]. However, the above methods have certain issues such as the inability to effectively capture market changes, neglecting asset pairs with stable relations, and the appearing data are over-fitting.

After a sharp pullback in 2020, there were two unprecedented rising trading restrictions; hence, it became more important than ever to use a strong risk management system to detect frequent behavioral changes in financial markets. This shift is often due to changes in government policy, the regulatory environment, and other macroeconomic impacts. Therefore, it is necessary to find ways to effectively detect these state transitions. The hidden Markov model can well solve this issue by reasoning about the hidden state through indirect noise observation related to the process. Kalman filter is also combined with this model in this study, using its prediction error as a threshold to set the trading strategy and reduce the risk. Previous studies were lacking in the stock selection process; thus, this study added two parts in the stock selection process; the first is the selection of high correlation contract pairs in the contract pool while the second is to select contract pairs with significant cointegration relationship for matching transactions.

2. Methods

2.1. Kalman filter

The Kalman filter can be used to infer information about the hidden state. In a linear space model, the state X_k at k is a linear combination of its prior state and the process noise at k-1, defining the following state transfer equation:

$$X_k = AX_{k-1} + BU_k + w_k \tag{1}$$

Here, the matrix A is a state transfer matrix (known coefficients), a transformation matrix acting on the k-1 moment state X_{k-1}, defining the transformation of the state vector and is dynamically updated over time. B and U_k are the system parameters, and the process noise w_k is a time-dependent multivariate noise. Since the state is unobservable, it needs to be inferred by defining the values of the actual observables. The equation of state observation is defined as follows:

$$Z_k = HX_k + v_k \tag{2}$$

The observation Z_k is a linear combination of the current real state X_k and observed noise v_k. H is the observation matrix, and the action is to map the implied real state space to the observation space, the observation noise represented by v_k. Then, it needs to be updated with an algorithm.

$$< \mathrm{X}_{\mathrm{k|k-1}} >= \mathrm{A} < \mathrm{X}_{\mathrm{k-1}} > + \mathrm{BU}_{\mathrm{k}} \tag{3}$$

$$P_{k|k-1} = AP_{k-1}A^T + Q_k \tag{4}$$

$< \mathrm{X}_{\mathrm{k-1}} >$ is the estimated state at k-1 time, and $< \mathrm{X}_{\mathrm{k|k-1}} >$ is the predicted value of predicting k time estimation state based on k-1 time estimation state; $P_{k|k-1}$ is the posterior estimation error covariance matrix, which predicts the estimated value at k-1 time first, and then calculates the posterior estimation error covariance matrix used to measure the prediction accuracy.

Finally, in order to verify the correct estimate of the k time state at k-1 time, it needs to be measured by the error between the actual measurement and consider the error compensation. The complete iterative update process is clarified in a study [8].

2.2. Hidden Markov state transfer model

Hidden Markov state transfer models are special state-dependent models of mixed distribution. In such models, there are potential states and probability transitions between them, but they cannot be observed directly; instead, these potential states affect the observations. The core idea of the hidden Markov model is that there are different states behind the price (or yield) of the financial asset, and the state determines the price (yield), which in turn is hidden and requires inference (decoded).

Hidden Markov models tend to remain for a particular state; then, suddenly jump to a new state and repeat the same behavior as market regimes do not change too frequently under long observations, where they most likely last when it changes.

The hidden Markov model is defined in the form of a joint probability distribution as follows:

$$P(X_1^T = x_1^T, C_1^T = c_1^T) = P(x_1^T, c_1^T) = P(c_1^T)P(x_1^T|c_1^T)$$

$$=P(c_1)\prod_{t=2}^{T} p(c_t|c_{t-1})\prod_{t=1}^{T} p(x_t|c_t) \quad (5)$$

With formula (5), it is known that the joint probability of the hidden state and the observation result is equal to the probability of seeing the hidden state multiplied by the probability of the observation result (conditioned on the state). The observations cannot affect the state, but the hidden states do affect the observations indirectly. **Figure 1** intuitively shows the dynamic process of the state c_t and how it dynamically affects the observation x_t.

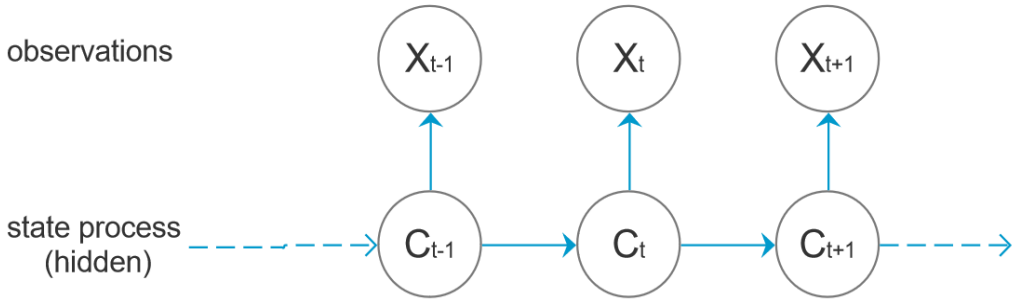

Figure 1. Hidden Markov model

The final yet an important problem in applying the hidden Markov model is the global decoding of finding out the optimal sequence of states $\{s_1, s_2, s_3, ..., s_t\}$ which can best explain the observation sequence $\{x_1, x_2, x_3, ..., x_t\}$. In this case, the Viterbi algorithm can be used [9]. It is an algorithm to compute the optimal state sequence of chain structure graphical models which can be used to find the implied state sequences that are most likely to produce a sequence of observed events:

$$s_1, s_2, s_3, \ldots, s_t = argmax_{c_1, c_2, \ldots, c_t} P(X_1^T = x_1^T, C_1^T = c_1^T) \quad (6)$$

The complete introduction of the hidden Markov model can be found in a literature [10].

2.3. Cointegration

HMM is a probabilistic model about timing. Models about time series are required to be stationary by the analyzed sequence; the modeled must be stationary and non-white noise data. Therefore, before performing the cointegration test, the data should be verified if it is a first-order single integer; that is, the first-order difference after the original non-stationary data becomes stationary.

First, $\{\boldsymbol{x_t}\}$ is defined as a time series, if $t_1, t_2, ..., t_n \in T$, for any positive integer τ :

$$F_X(x_{t_1}, x_{t_2}, \ldots, x_{t_n}) = F_X(x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_n+\tau}) \tag{7}$$

When the joint distribution of x_{t_1}, x_{t_2},..., with the joint distribution of $x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_n+\tau}$.., are identical, then the time series $\{X_t\}$ is stationary. In order to detect whether a time series is stationary, a statistical hypothesis test for the existence of a unit root in a time series sample needs to be established. Using the unit root test (ADF) by Shumway and Stoffer [11], the stationarity of the time series can be tested. ADF testing is applied to the model as follows:

$$\Delta x_t = \alpha + \beta t + \gamma x_{t-1} + \sum_{I=1}^{k} \delta_i \Delta x_{t-i} + \varepsilon_t \tag{8}$$

Where α is a constant, β is a coefficient on a time trend. When α and β are both 0, the random walk process corresponds, and when only β is 0, that with drift terms. The unit root test is then performed based on the zero assumption of $\gamma = 0$. In comparison with the DF (degrees of freedom) distribution critical value table test, if the calculated statistical value is less than the critical value, the zero assumption of $\gamma = 0$ is rejected, and there is no unit root present.

3. Transaction framework

From previous research, it is known that A + H shares have the mean response characteristics, thus arbitrage opportunities, but with the single use of the average response characteristics in past research and the lacking in the stock pair selection process, so this study added two parts in the stock selection process; one of it is to select the contract pair in the contract pool, and the other is to select the contract pair with significant cointegration relationship for pairs trading.

The final contract price is y_t and x_t, where they have the following relationship:

$$y_t = \theta_t x_t + v_t \tag{9}$$

θ_t represents the vector of the intercept and slope values in the linear regression between the two stocks at t time, and the dynamic hedging ratio is defined by a component of the hidden state vector θ_t. Kalman filter is used to dynamically track the hedge ratio between the two to keep the spread stable, and it also helps calculate e_t and Q_t, where e_t represents the prediction error or residual error of the predicted value at t, while Q_t represents the variance of the predicted value at time t, and $\sqrt{Q_t}$ is the standard deviation of the prediction.

The dynamic "spread" between contract pairs is the concerned time series that is assumed more or short. When interruption suddenly occurs, it is necessary to determine when the difference can return to the expected value, so a certain threshold needs to be set to determine the range where the difference exceeds the expected value. Here, consider the standard difference $\sqrt{Q_t}$ of the predicted value to measure, use this multiple as the boundary, and set a certain stop-loss rate for policy improvement.

The hidden Markov model is trained on the selected training set of the two contract pairs. The trained model is then used to predict each hidden state of each asset, where the state indexes are 0 and 1.
Based on the above, the policy has been implemented as follows:

(1) seek for optimal stock pairs for A and B (selects contract pairs with significant cointegration among contracts with high correlation);

(2) use Kalman filter to dynamically estimate asset A prices based on the previous day's B observations;

(3) use the difference between the Kalman estimate and actual value of stock A to measure the gap between stock A and stock B deviation from expectation;

(4) short high price targets when e_t is above the upper bound (given threshold) and reverse when e_t is below the lower bound; if the income of the current transaction is negative and its absolute value is less than the loss of the total income, the closing loss;

(5) check the institutional status of asset A and asset B, and if they are both in institutional status 0, the above steps are performed; otherwise, wait for a new opportunity;

(6) stop assuming more and short when restored to its expected value.

4. Data and results

4.1. Selection of stock pairs

The idea behind matching transactions is two assets with similar price trends and similar potential factors affecting their price changes. A + H shares meet such characteristics; however, not every A + H share is the best stock pair, and not every pair meets the requirements of the trading strategy. Hence, the A + H shares had to be screened.

More than 90 A+H shares were obtained from January 1, 2019, to December 31, 2019. These shares were from the China Stock Market & Accounting Research (CSMAR) database corresponding to the stocks listed on the Shanghai and Hong Kong Stock Exchange, with high frequency trading data of 9:31-11:10; 13:00-15:00, 244 trading days, and about 58,560 data; thus, about 5.44 million data with more than 90 stock companies.

Each stock data is as follows, with paired transactions based on this data (**Table 1**):

Table 1. Stock data information

	date	stamp	open	high	low	close	volume	turnover
0	2019/01/10	09:31	17.15	17.15	16.98	17.05	318148.0	543968452.0
1	2019/01/10	09:32	17.02	17.03	16.95	16.95	92009.0	156349376.0
2	2019/01/10	09:33	16.94	16.94	16.83	16.90	86791.0	146368960.0
3	2019/01/10	09:34	16.89	16.95	16.88	16.95	63378.0	107233728.0

The closing prices for more than 90 stocks were used for stock pair selection (**Table 2**).

Table 2. Stock pool established at the closing price

	00038	00107	00168	00177	00187	00317	00323	00338	00358	00386
0	1.82	2.40	31.60	10.92	1.31	5.11	3.41	3.43	9.14	5.56
1	1.82	2.40	31.60	10.92	1.31	5.11	3.41	3.42	9.23	5.59
2	1.82	2.40	31.60	10.92	1.31	5.11	3.41	3.44	9.23	5.59
3	1.82	2.40	31.60	10.92	1.31	5.11	3.41	3.44	9.23	5.60
4	1.82	2.40	31.60	10.92	1.31	5.11	3.41	3.44	9.23	5.60
...	...	...	...	...	...	...	...	...	...	...
58555	1.75	2.43	52.05	10.72	1.49	NaN	3.14	2.30	10.78	4.67
58556	1.75	2.43	52.05	10.72	1.49	NaN	3.14	2.30	10.78	4.67
58557	1.75	2.43	52.05	10.72	1.49	NaN	3.14	2.30	10.78	4.67
58558	1.75	2.43	52.05	10.72	1.49	NaN	3.14	2.30	10.78	4.67
58559	1.75	2.43	52.05	10.72	1.49	NaN	3.14	2.30	10.78	4.67

58560 rows × 186 columns

First, the correlation between the contract target in China was calculated, and several pairs of contracts with high correlation were selected. The threshold was set to 0.97; thus, contract pairs with correlation more than 0.97 were selected with 8 pairs meeting the condition (**Table 3**).

Table 3. Contract pairs with a correlation greater than 0.97

	0	1	2	3	4	5	6	7
cont1	00338	00338	00386	00390	00991	sh600688	sh600688	sh601390
cont2	03328	sh601857	01800	01766	01618	sh601857	sh601991	sh601766
cor_val	0.971603	0.979286	0.977742	0.986266	0.970211	0.971684	0.979624	0.984963

The most relevant pair, 00390 and 01766, was chosen for the trading strategy study. They represent the Bank of Communications (H-share Financial Industry) and CRRC Limited (H-share Industry), respectively. The hidden Markov model was trained from January 1, 2019, to April 31, 2019.

4.2. Empirical analysis

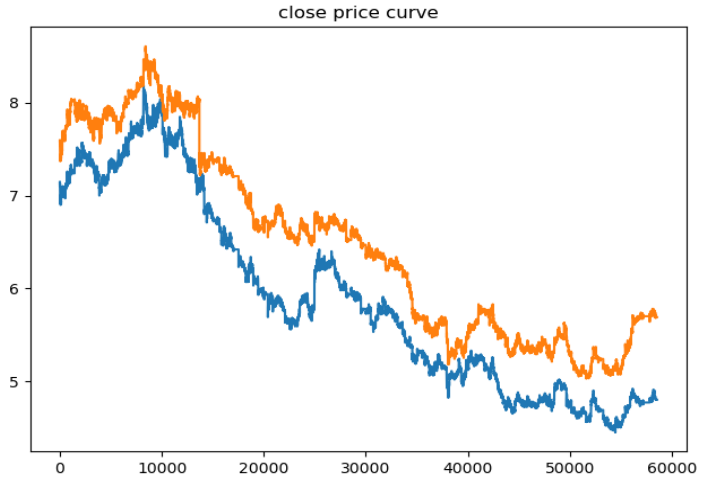

Figure 2. Closing price figure

Figure 2 shows that the closing price of stocks 00390 and 01766 share the same trend; there is arbitrage in the short-term timestamp.

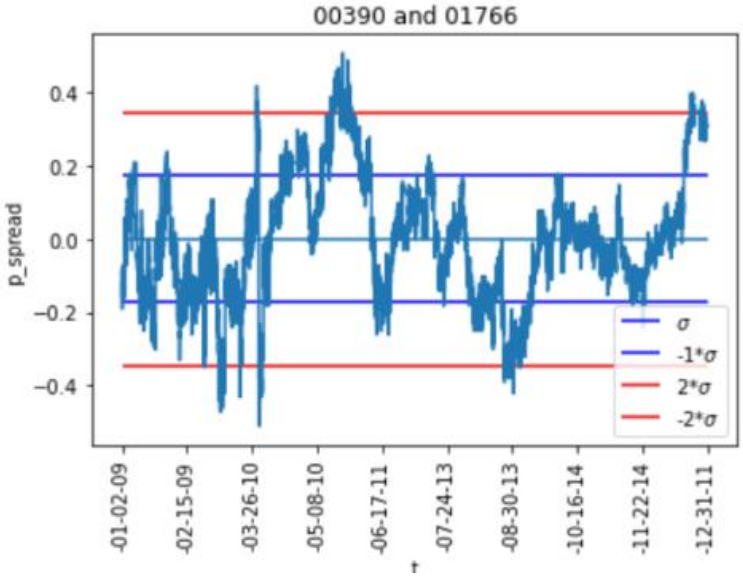

Figure 3. 00390 and 01766

Figure 3 shows the spread curve of the two fluctuating around 0, and $1.5\sqrt{Q_t}$ was set as the threshold for the implementation of the trading strategy.

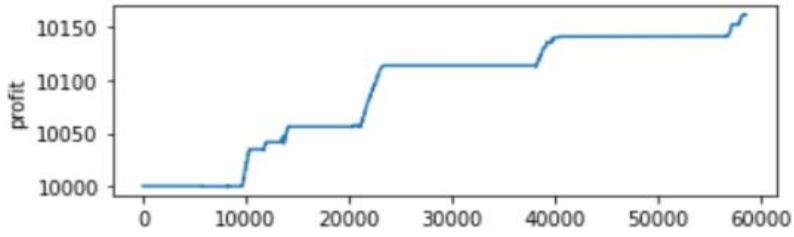

Figure 4. Strategy based on cointegration

The implementation of pairing transactions based on the cointegration theory is shown in **Figure 4**. According to the income chart, the transaction return remained unchanged for a long period of time, with the income of 1.618%, a total of 1320 transactions, an income loss ratio of 6.13, and a maximum pullback of 0.066%.

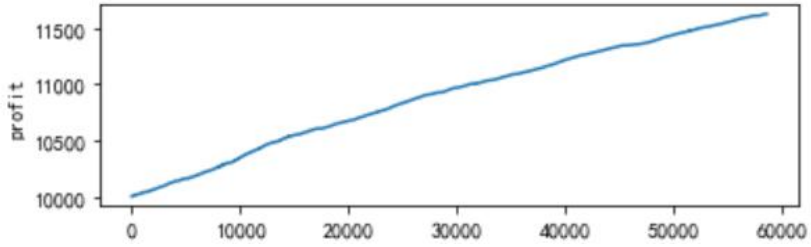

Figure 5. Strategy based on Kalman-HMM

Figure 5 shows the implementation of Kalman-HMM trading strategy. Based on Kalman-HMM pairs trading research, $1.5\sqrt{Q_t}$ was set as the threshold. From the figure, continuous income can be realized for long with 13,638 transactions, 16.238%, income loss ratio of 10.68, and maximum pullback of 0.0199%. Based on **Figure 4** and **Figure 5**, compared with the cointegration theory, the revenue reached 16 times with Kalman-HMM strategy and the maximum pullback reduced to 0.0199%. By comparing these two strategies, Kalman-HMM strategy provides a high-yield and low-risk method for the implementation of paired transactions.

5. Conclusion

The proposed Kalman-HMM approach in pairs trading strategy achieved good results based on screening contract pairs with distinct co-consolidation relations. This strategy of combining the Kalman filter with the hidden Markov model does not only help to achieve ideal return effects, but also significant changes in market fluctuations and the operating conditions of companies.

This strategy can also take into consideration of parameter optimization in order to achieve a better prediction effect by optimizing the parameters of the Kalman filter and reducing the impact of noise on the pairing transaction research, starting with the system and measurement noise, the noise reduction processing, and the restoration of real stock price process.

Disclosure statement

The author declares that there is no conflict of interest.

References

[1] Huang XW, Yu M, Pi DY, 2015, Pairs Trading Strategies and Financial Market Efficiency based on O-U process. Management Review, 27(1): 3-11.

[2] Gatev E, Goetzmann WN, Rouwenhorst KG, 1999, Pairs Trading: Performance of a Relative-Value Arbitrage Rule. Working paper, Yale School of Management's International Center for Finance.

[3] Gatev E, Goetzmann WN, Rouwenhorst KG, 2006, Pairs Trading: Performance of a Relative-Value Arbitrage Rule. The Review of Financial Studies, 19(3): 797-827.

[4] Huck N, Afawubo K, 2015, Pairs Trading and Selection Methods: Is Cointegration Superior?. Applied Economics, 47(6): 599-613.

[5] Vidyamurthy G, 2004, Pairs Trading: Quantitative Methods and Analysis, John Wiley & Sons, Hoboken.

[6] Jurek JW, Yang H, 2007, Dynamic Portfolio Selection in Arbitrage. Working paper, Harvard University.

[7] Huck N, 2010, Pairs Trading and Outranking: The Multi-Step-Ahead Forecasting Case. European Journal of Operational Research, 207(3): 1702-1716.

[8] Jwo DJ, Cho TS, 2007, A Practical Note on Evaluating Kalman Filter Performance Optimality and Degradation. Applied Mathematics and Computation, 193(2): 482-505.

[9] Forney GD, 1973, The Viterbi Algorithm, in Proceedings of the IEEE 61(3), 268-278.

[10] Li H, 2012, Statistical Learning Method, Tsinghua University Press, Beijing.

[11] Shumway RH, Stoffer DS, 2000, Time Series Analysis and Its Applications. Studies in Informatics and Control, 9(4): 375-376.