Machine Learning in Pairs Trading Strategies

Yuxing Chen (Joseph)
Department of Statistics
Stanford University
Email: josephc5@stanford.edu

Weiluo Ren (David)
Department of Mathematics
Stanford University
Email: weiluo@stanford.edu

Xiaoxiong Lu
Department of Electrical Engineering
Stanford University
Email: lxx@stanford.edu

Keywords: pairs trading, mean reverting, Ornstein-Uhlenbeck process, portfolio rebalancing, Kalman filter, Kalman smoother, EM

1.Introduction

Pairs trading consists of long position in one financial product and short position in another product and we focus the form of statistical arbitrage instead of trend following; these strategies are market neutral and have low risk.

Choose two securities 1, 2 and denote their prices as S_1, S_2 . Then the spread is $S_1 - \beta S_2$, where β is a carefully chosen constant depending on time. The simplest case is that $\beta = 1$; the spread becomes simply difference between two prices.

We assume that the spread is a mean reverting process, meaning if deviations of spread from its mean occur, this deviation will eventually vanish. Then when deviations arise, we long the relatively cheap securities and short sell the relatively expensive securities and then wait for the spread will go back to its mean level to make profit. This is the basic idea behind many pairs trading strategies including ours.

The question now becomes how to model the mean-reverting process of spread so that entering and exiting trading signal can be developed from that model. In this paper, Ornstein-Uhlenbeck process is used as the underlying model of spread:

$$dX(t) = \theta(\mu - X(t))dt + \sigma dW(t) \tag{1.1}$$

where $X(t)$ is the spread at time t, θ measures the speed of $X(t)$ returning to its mean level μ , and σ is the volatility of spread.

In this project, two approaches are applied. One is starting from difference of daily returns instead of spread of prices, and integrating this process and using a linear regression to estimate coefficients θ, μ, σ . Another one is assuming a spread model which is a latent O-U process plus some noise and building signals based on prediction generated from Kalman filter; E-M algorithm modified for Kalman smoother/filter is applied to estimate coefficients in the spread model.

In Section 2 and 3, models and algorithms are given in a backwards order first, starting from models and then introducing algorithms in order to estimate parameters in models. Very brief summaries of real procedures are given in later part of Section 2 and 3, showing the order of how algorithms should be implements.

2.Portfolio Rebalancing & Linear Regression Approach

The advantage of this approach is simplicity: linear model is convenient to be interpreted and if anything goes wrong, it is easy to spot the source of problem.

2.1 Assumptions and Portfolio Rebalancing

We assume that the daily returns of two financial products satisfy the following stochastic differential equation:

$$\frac{dS_1(t)}{S_1(t)} = \alpha dt + \beta \frac{dS_2(t)}{S_2(t)} + dX(t) \tag{2.1}$$

The drift term α is the trend of spread of daily returns; β is a constant which cannot change much along time; $X(t)$ is a mean reverting process. In practice, this equation says that if we long \$1 securities one and short selling \$ β securities two, the daily return of our portfolio should be mean reverting given the condition that $|\alpha| \ll$ the magnitude of fluctuation of $X(t)$, which is usually the case.

From the above explanation, we can see why β cannot change much along time. Since $\left(1/(1+\beta), \beta/(1+\beta)\right)$ are weights within our portfolio, if β changes frequently and in a large magnitude, meaning that portfolio needs rebalancing frequently and weights change much. Then the profit cannot cover the cost of rebalancing. It is better to run a regression to find β at t_0 and keep β as a constant for a short period of time e.g. 5 or 10 days, and check whether $X(t)$ from $dX(t) = \frac{dS_1(t)}{S_1(t)} - \beta \frac{dS_2(t)}{S_2(t)}$ has mean-reverting property.

2.2 The O-U Model of Spread

As stated in Section 1, we use O-U process (1.1) to model the dynamic of the spread $X(t)$, thus we have

$$dX(t) = \theta(\mu - X(t))dt + \sigma dW(t) \qquad (2.2)$$

Integrating the above equation we have [1]

$$X(t_0 + \Delta t) = e^{-\theta\Delta t} X(t_0) + (1 - e^{-\theta\Delta t})\mu + \sigma A(t_0, \Delta t)$$

$$A = \int_{t_0}^{t_0+\Delta t} e^{-\theta(t_0+\Delta t - s)} dW(s) \qquad (2.3)$$

Now let Δt tend to infinity, the equilibrium probability distribution of $X(t)$ is normal with

$$E[X(t)] = \mu \quad \text{and} \quad Var[X(t)] = \frac{\sigma^2}{2\theta} \qquad (2.4)$$

With (2.3) and (2.4), we are able to estimate the parameters in the O-U process.

2.3 Linear Regression for Estimating Weights and Parameters

Let us denote

$$R_t^1 = \frac{S_t^1 - S_{t-1}^1}{S_{t-1}^1}, \; t \ge 1$$

Run a linear regression of R_t^1 against R_t^2 on a moving window with length 60.

$$R_t^1 = \hat{\beta}_0 + \hat{\beta} R_t^2 + \varepsilon_t, \quad t = t_1, \ldots, t_{60}$$

and note that $\left(1/(\hat{\beta}+1), \hat{\beta}/(\hat{\beta}+1)\right)$ is the weight of portfolio and also that we may run the above regression every 5 days as indicated in Section 2.1.

Then we use the sum of residuals to obtain the discrete version of X_k

$$X_k = \sum_{j=t_1}^{k} \varepsilon_j, \; k = t_1, \ldots, t_{60}$$

Then use these X_k and linear regression again to estimate parameters θ, μ, σ as below.

$$X_{k+1} = a + bX_k + \zeta_{k+1}, \; k = t_1, \ldots, t_{60}$$

By (2.3) we have

$$a = (1 - e^{-\theta\Delta t})\mu$$

$$b = e^{-\theta\Delta t}$$

$$Var(\zeta) = \sigma^2 \frac{1 - e^{-2\theta\Delta t}}{2\theta}$$

Thus

$$\theta = -\log(b) \times 252$$

$$m = a/(1-b)$$

$$\sigma = \sqrt{\frac{Var(\zeta) \cdot 2\theta}{1 - b^2}}$$

Moreover, we are able to get the equilibrium standard deviation from (3.4) now.

$$\sigma_{eq} := \sqrt{Var(X(t))} = \frac{\sigma^2}{2\theta} \qquad (2.5)$$

At this stage, we can use the standardized version of $X(t)$, called Z-score as trading signal. This factor measures how far $X(t)$ deviates from its mean level and is a valid measure across all securities since it is dimensionless. More details of signal will be given later.

2.4 Summary of the Procedure

A summary of the whole procedure will be given and it displays the order within the implementation of the trading strategy.

First run a linear regression on daily returns on a moving window to get ε_t and new weight $\hat{\beta}$ (performing rebalancing if necessary); then use ε_t to sum to discrete version of X_k and run another regression of X_k to obtain parameters θ, μ, σ in the O-U process and z-score as trading signal.

The buying and selling rules are

buy to open if $s_i < -s_{bo}$

sell to open if $s_i > +s_{so}$

close long position if $s_i > -s_{sc}$

close short position if $s_i < +s_{bc}$

2.5 Back-Test Results

We use closing price (daily data) of two chosen future contracts in China future market. The daily return plot is shown as below.

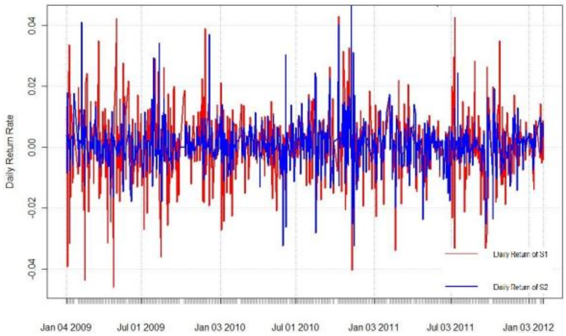

Figure 2.1 Daily returns of two securities

The plot of $\hat{\beta}$ (updated every five trading days) is

Figure 2.2 $\hat{\beta}$ which indicates the weight of portfolio

The plot of cumulative profit and summary are given; we consider the transaction cost of buying/selling plus slippage is 20 basis points.

Figure 2.3 Cumulative profit

Annualized Return Rate	Volatility	Sharpe	Maximal Drawdown
8.10%	7.59%	1.14	5.57%

3.Kalman Filter and EM Algorithm Approach

3.1 The Spread Model

The State Process

We studied $\{y_k\}$, the simplest spread $S_1 - S_2$, with the assumption that it is a noisy observation of a latent mean-reverting state process $\{x_k\}$.

Here τ stands for one day and x_k is some variable at time $t_k = k\tau$ for $k = 0,1,2\cdots$, satisfying the following mean-reverting dynamic (discrete version of OU process).

$$x_{k+1} - x_k = (a - bx_k)\tau + \sigma\sqrt{\tau}\varepsilon_{k+1} \qquad (3.1)$$

where $\sigma \geq 0,\ b \geq 0,\ a \in \Re$, and $\{\varepsilon_k\}$ is i.i.d Gaussian $N(0,1)$. Thus, ε_{k+1} in the above equation is independent of all x_k. And the process mean reverts to $\mu = a/b$ with "speed" b. Therefore, we can rewrite (3.1) as follows

$$x_{k+1} = A + Bx_k + C\varepsilon_{k+1} \qquad (3.2)$$

The state x_k is hidden, which needs observations to estimate.

The Observation Process

We assume the spread process $\{y_k\}$ is the observation of $\{x_k\}$ with Gaussian noise,

$$y_k = x_k + D\omega_k \qquad (3.3)$$

where $\{\omega_k\}$ are also i.i.d Gaussian $N(0,1)$ and independent of $\{\varepsilon_k\}$.

The Trading Signal

Here we define

$$\hat{x}_{k|l} = E[x_k \mid F_l] \qquad (3.4)$$

$$\Sigma_{k|l} \equiv E[(x_k - \hat{x}_k)^2 \mid F_l] \qquad (3.5)$$

$$\hat{x}_k \equiv \hat{x}_{k|k} \qquad (3.6)$$

The conditional expectation given observed information F_l, either from an expanding window or a moving window.

If $y_k > \hat{x}_{k|k-1}$ + transaction cost + premium, here the premium is the profit that we want to ensure when enter a position, then the spread is regarded as too large, meaning the securities 1 if relatively expensive than securities 2; we would take a long position in the spread portfolio (short selling one unit 1 and longing one unit of product 2), expecting that the spread will shrink eventually.

Similarly, if $y_k < \hat{x}_{k|k-1} -$ transaction cost - premium, then the spread is lower than the expectation significantly; we would take a short position in the spread portfolio.

We close positions when $|y_k - \hat{x}_{k|k-1}| \leq \sigma$, where σ is a predetermined threshold.

3.2 Kalman Filter

Equiped with state equation (3.1) and observation equation (3.2), our next step is to estimate the hidden process.

In order to estimate $\hat{x}_{t+1|t}$, the prediction of the next day at time t, we will start from time 0 and the initialization $\hat{x}_0 = y_0$ and $\Sigma_{0|0} = D^2$; recall the definition (3.4) to (3.6).Then perform the following procedure iteratively until $k = t$.

In the prediction step, we would compute the "prediction"

$$\hat{x}_{k+1|k} = A + B\hat{x}_{k|k} \qquad (3.7)$$

$$\Sigma_{k+1|k} = B^2\Sigma_{k|k} + C^2 \qquad (3.8)$$

The optimal Kalman gain is

$$K_{k+1} = \Sigma_{k+1|k} / (\Sigma_{k+1|k} + D^2) \qquad (3.9)$$

Then we can compute our estimation with the new observation y_{k+1} in the following update step:

$$\hat{x}_{k+1} = \hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}[y_{k+1} - \hat{x}_{k+1|k}] \qquad (3.10)$$

$$R_{k+1} = \Sigma_{k+1|k+1} = D^2 K_{k+1} = \Sigma_{k+1|k} - K_{k+1}\Sigma_{k+1|k} \qquad (3.11)$$

Repeat the above process to obtain $\hat{x}_{t|t}$ and $\hat{x}_{t+1|t}$.

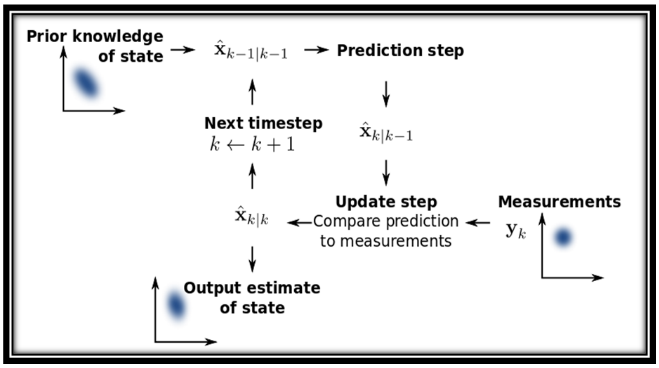

Figure 3.1 Kalman filter

The above algorithm is based on knowing $\Theta = (A, B, C^2, D^2)$, but Θ is unknown now. We need an extra training set to estimate the parameters A, B, C and D before using Kalman filter and these parameters are estimated via EM algorithm.

3.3 Estimation using EM Algorithm

Before predicting $\hat{x}_{k+1|k}$ We now use the EM algorithm to estimate $\Theta = (A, B, C^2, D^2)$ based on the observations $y_0, y_1, \ldots, y_k$.

Recall the general form of EM algorithm (P_Θ is a probability measure with parameter Θ):

Step 1 (E-step)

$$\text{Compute} \quad E_{\Theta_j}\left[\log \frac{dP_\Theta}{dP_{\Theta_j}} \,|\, F_N\right]$$

Step 2 (M-step)

Update Θ_j to Θ_{j+1} by maximizing conditional expectation.

$$\Theta_{j+1} = \arg\max_{\Theta} E_{\Theta_j}\left[\log \frac{dP_\Theta}{dP_{\Theta_j}} \,|\, F_N\right]$$

Here we are going to use Shumwau and Stoffer [6] smoother to implement EM algorithm. We define smoothers for $k < N$ (note that they have the same form of definition for filtering when $k \geq N$):

$$\hat{x}_{k|N} = E[x_k \,|\, F_N] \qquad (3.14)$$

$$\Sigma_{k|N} = E[(x_k - \hat{x}_{k|N})^2 \,|\, F_N] = E[(x_k - \hat{x}_{k|N})^2] \qquad (3.15)$$

$$\Sigma_{k-1,k|N} = E[(x_k - \hat{x}_{k|N})(x_{k-1} - \hat{x}_{k-1|N})] \qquad (3.16)$$

They can be computed backwards, meaning that $\hat{x}_{k|N}$, $\Sigma_{k|N}$, and $\Sigma_{k-1,k|N}$ can be obtained from $\hat{x}_{k+1|N}$, $\Sigma_{k+1|N}$, and $\Sigma_{k,k+1|N}$ [5].

Let us use first few steps to illustrate how EM algorithm works in the form of Kalman smoother/filter.

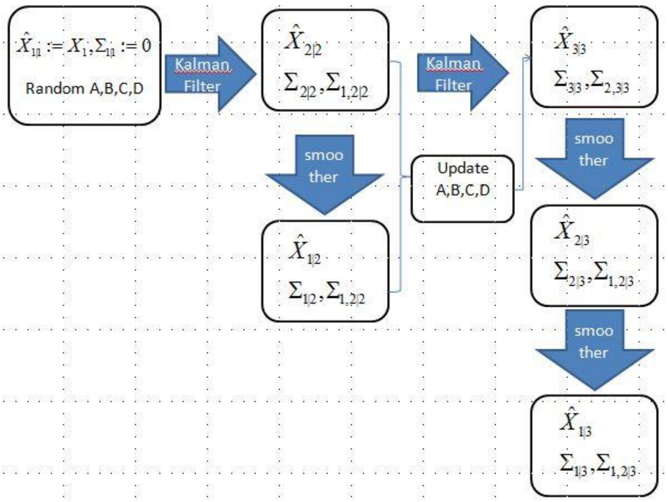

Figure 3.2 EM in the form of smoother/filter

From the above figure, the operation rules with blue arrows indicating smoother and filter are known; so now we go to the rule of updating A, B, C, D using $\hat{X}_{k|N}$, $\Sigma_{k|N}$, and $\Sigma_{k-1,k|N}$.

Given $\hat{\Theta}_j = (A, B, C^2, D^2)$ and initial values for the Kalman Filter, the update $\hat{\Theta}_{j+1} = (\hat{A}, \hat{B}, \hat{C}^2, \hat{D}^2)$ are calculate as follows:

$$\hat{A} = \frac{\alpha\gamma - \delta\beta}{N\alpha - \delta^2} \qquad (3.17)$$

$$\hat{B} = \frac{N\beta - \gamma\delta}{N\alpha - \delta^2} \qquad (3.17)$$

$$\hat{C}^2 = \frac{1}{N}\sum_{k=1}^{N}[(x_k - \hat{A} - \hat{B}x_{k-1})^2 \,|\, F_N] \qquad (3.19)$$

$$\hat{D}^2 = \frac{1}{N+1}\sum_{k=0}^{N}[(y_k - x_k)^2 \,|\, F_N] \qquad (3.20)$$

where

$$\alpha = \sum_{k=1}^{N} E[\hat{x}_{k-1}^2 \,|\, F_N] = \sum_{k=1}^{N}[\Sigma_{k-1|N} + \hat{x}_{k-1|N}^2]$$

$$\beta = \sum_{k=1}^{N} E[x_{k-1}x_k \mid F_N] = \sum_{k=1}^{N} [\Sigma_{k-1,k|N} + \hat{x}_{k-1|N}\hat{x}_{k|N}]$$

$$\gamma = \sum_{k=1}^{N} \hat{x}_{k|N}$$

$$\delta = \sum_{k=1}^{N} \hat{x}_{k-1|N} = \gamma - \hat{x}_{N|N} + \hat{x}_{0|N}$$

3.4 Summary of Algorithms

As shown in Figure 3.2, we start from

$$\hat{X}_{1|1} := X_1, \Sigma_{1|1} := 0$$

and continuing updating. After $k \geq 300$, we use the prediction $\hat{x}_{k+1|k}$ as trading signal.

Besides, the EM algorithm in [4] is implementated using expanding window as above which would cause heavy computational burden. So we also implement the EM using moving window for updating as well.

3.5 Result of Implementation

We pair two future contracts of agricultural products in China market and estimate the process of spread.

With Kalman Filter and EM Algorithm, we can predict the hidden state $\hat{x}_{k+1|k}$ using an expanding window, plotted in Figure 3.3.

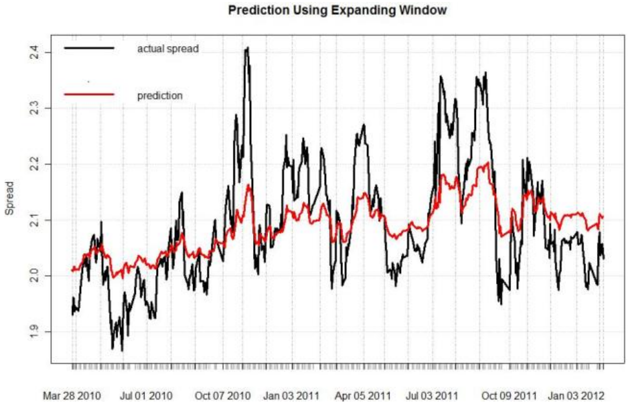

Figure 3.3 Actual Spread vs Prediction

As shown in Figure 3.3, the actual spread y_k is osillating around our prediction. When y_k deviates from $\hat{x}_{k|k-1}$ significantly, we would expect that the spread will shrink eventually and we will take advantage of this to make profit.

Figure 3.4 is the prediction made from EM on a moving window. Comparison of cumulative profits of using two windows is given as well (Figure 3.5).

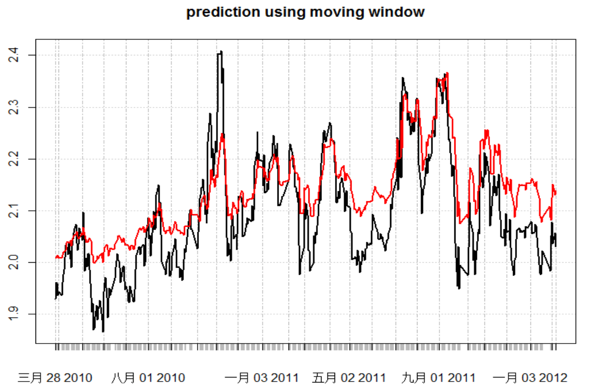

Figure 3.4 Actual Spread vs Prediction

Figure 3.5 Cumulative Profits

4.Conclusion and Analysis

We are betting on meaning reversion of spread, thus it is necessary to check whether the spread between two securities has such property in historical data. If the spread has obvious upward or downward trend, then loss may incurr; setting a loss cutting limit can be implemented.

It seems that the smoother/filter approach has better returns (10.41% and 8.79%) compared with the linear regression approach. The cost of rebalancing in the latter one may be the reason. We have to realize that linear regression is simple and it is easy to be interpreted while smoother/filter approach is a black box.

Reference

[1] Statistical Arbitrage in the U.S. Equities Market, Marco Avellaneda and Jeong-Hyun Lee.

[2] Sanford Weisberg, Applied Linear Regression, third edition, John Wiley & Sons, Inc.

[3] David Serbin, Trend following signal confirmation using non-price indicators.

[4] Robert J. Elliott, John van der Hoek and William P. Malcolm, "Pairs Trading," Quantitative Finance, Vol. 5(3), pp. 271-276.

[5] Elliott, R.J., L. Aggoun and J.B. Moore, Hidden Markov Models. (Springer Verlag 1995)

[6] Shumway, R.H. and D.S. Stoffer. An Approach to Time Series Smoothing and Fore-casting using the EM Algorithm. Journal of Time Series 3 [4] (1982), 253-264.