Nonparametric Pairs Trading with International ETFs

Tomoya Greve
jangreve4@toki.waseda.jp
School of Political Science and Economics, Waseda University

Supervisor: Prof. Hiroshi Saigo
saigo@waseda.jp
School of Political Science and Economics, Waseda University

Supervisor:Associate Prof. Kenichiro Tamaki
tamaki@waseda.jp
School of Political Science and Economics, Waseda University

1 Summary

The objective of this thesis is to propose a new method for algorithmic pairs trading based on nonparametric methodologies such as Random forest and Bootstrapping. The backtest results of this algorithm are compared with that of the conventional linear regression based algorithm detailed in Chan (2013) [2]. We concluded that my algorithm is less parameter dependent in overall performance, thus making it easier for individual and small capital investors to adopt this strategy which used to be considered only viable for institutional investors and hedge funds.

2 Methodology

There are three major methods to implement pairs trading: the distance method used in Gatev et al (1999) [4], the stochastic spread approach proposed by Elliott et al (2005) [3] and the cointegration method discussed in the textbook by Chan (2013) [2]. Pairs trading in this paper fits into the third category.

The cointegration method has three steps: step1 is identification of securities to be used in pairs trading, step 2 is the spread series creation and step3 is formation of trading rule. In this paper, we made modifications to step2 and step3 by replacing parametric methods with nonparametric ones and added a mathematical transformation as well. The same rule applied in Chan (2013) [2] will be used in step 1.

Firstly, we show in this paper how step 2 can be modified in order to create more predictable spread series. The Conventional algorithm uses residuals generated by OLS as spread series. The rational behind this choice is that residuals series from the regression of cointegrated time series are expected to posses stationarity and mean-reversion. However, in practice, generating a residual series possessing optimal level these characteristics is entirely dependent on lookback period (the amount of past data to be included in regression). This is because OLS is a global optimizer and therefore inclusion of old data which has little predictive power in current price is still influencial in prediction.A measurement such as the half-life of mean reversion discussed in Chan (2013) [2] can be utilized in finding an optimal lookback length. However, the half-life can only be estimated by using the past data and optimal length of lookback is according to Chan (2013) [2] a small multiple of this value, which is quite speculative.

To ameliorate these concerns, we replaced linear regression with Random Forest regression theorized by Breiman(2001) [1] in order to make spread series more robust to the change in lookback period. Of all the nonlinear regressions, the reason why we choose Random Forest is because it's virtually parameterless. Therefore, Random Forest is no more susceptible to data snooping bias than linear regression and yet manages to create spread series that are more robust against a change in lookback because of it being a local optimizer. Details about how Random Forest is able to bring such changes and why it can be considered parameterless are provided in the paper.

Secondly, in step 3, we replaced commonly used Bollinger Band with Stationary Bootstrapped mean intervals to detect divergence in spread series. As with linear regression, Bollinger Band is also sensitive to a change in lookback period. Therefore, a resampling method such as Bootstrap will be considered a more robust alternative. Moreover, in last model, we made a nonlinear transformation to spread series in order to amplify the price change when it's distant from mean and condense the value otherwise. This transformation cut the number of unprofitable transactions and also reduced holding times dramatically, thus making the strategy less prone to market exposure.

3 Results

The table below shows one example of backtest results conducted in this paper. Parametric model is representing the conventional model while the Nonparametric model is what this thesis is proposing. Three set of values shown below these headers are the total returns, Sharpe ratio and β. In this example, pairs trading is conducted with EWD and EWU from 2014/01/01 to 2016/07/01. The half-life of mean reversion of this pair estimated from the past data is 10 days while in-sample estimate is 12 days. Therefore, the optimal look-back period would be approximately in a range from 20 to 40. To check that algorithm would work well with extreme parameters we also tested with lookback of 100, 200 days.Three different thresholds are used to signal trade entry: in nonparametric case 99%, 95% and 90% bootstrap confidence intervals are used while in parametric case z-value of Bollinger band strategy with threshold of 1.5,1.0 and 0.75 are used instead. More details about the settings such as commission fees and slippage is provided in the paper.

We can clearly observe that Nonparametric model is more consistent in all measurements. Parametric model on the other hand did outperform nonparametric one in some rare occasion, but it's overall inconsistency of returns makes it difficult to tune accordingly. Especially its sensitivity to the threshold is problematic since there are no concrete rule to set this value. Institutional traders which can try with many different parameters may be able to set parameters accordingly, but for most individual and small capital investors, parameter sensitivity of conventional algorithm is unacceptable. My nonparametric algorithm with its consistency exemplified in this table would have a significant appeal to those players.

Lookback	Threshold	Parametric	Nonparametric
20	0.75/90%	17.7%/1.18/0.02	13.7%/1.25/0.00
	1.0/95%	12.2%/0.97/-0.00	14.5%/1.35/0.00
	1.5/99%	9.8%/1.16/0.00	16.8%/1.50/0.00
30	0.75/90%	12.2%/0.86/-0.00	14.5%/1.27/-0.01
	1.0/95%	13.9%/1.09/-0.02	13.5%/1.22/-0.01
	1.5/99%	0.2%/0.04/0.01	16.4%/1.45/-0.00
40	0.75/90%	14.9%/1.05/-0.00	10.3%/1.00/0.00
	1.0/95%	15.1%/1.18/-0.01	12.9%/1.27/0.00
	1.5/99%	2.3%/0.30/0.00	13.6%/1.33/0.00
100	0.75/90%	14.8%/1.15/0.01	13.1%/1.28/-0.01
	1.0/95%	15.9%/1.37/0.01	16.7%/1.61/-0.01
	1.5/99%	4.0%/0.54/0.00	17.7%/1.64/-0.01
200	0.75/90%	15.3%/1.24/0.01	17.2%/1.59/-0.01
	1.0/95%	18.0%/1.64/0.01	18.4%/1.71/-0.01
	1.5/99%	4.9%/0.64/0.01	19.4%/1.80/-0.01

Table 1: Backtest results with EWD and EWU with different parameters

4 References

[1] Leo. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Ernie." "Chan. *"Algorithmic trading: winning strategies and their rationale."*. "John Wiley Sons", 2013.

[3] Robert J. Elliott, John Van Der Hoek*, and William P. Malcolm. Pairs trading. *Quantitative Finance*, 5(3):271–276, 2005.

[4] Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.