International Journal of Economics and Finance; Vol. 6, No. 3; 2014
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

High Frequency and Dynamic Pairs Trading Based on Statistical Arbitrage Using a Two-Stage Correlation and Cointegration Approach

George J. Miao[1]

[1] Bluestar Investment Group, LLC, Cupertino, California, USA

Correspondence: George J. Miao, Bluestar Investment Group, LLC, P.O. Box 1042, Cupertino, CA 95015, USA. E-mail: george.miao@bluestarinvestgroup.com

Received: December 28, 2013 Accepted: January 13, 2014 Online Published: February 25, 2014
doi:10.5539/ijef.v6n3p96 URL: http://dx.doi.org/10.5539/ijef.v6n3p96

Abstract

In this paper, a high frequency and dynamic pairs trading system is proposed, based on a market-neutral statistical arbitrage strategy using a two-stage correlation and cointegration approach. The proposed pairs trading system was applied to equity trading in U.S. equity markets in any type of market cycle condition to capture statistical mispricing between the prices of each stock pair based on its residuals and to model the stock pairs naturally as a mean-reversion process. The proposed pairs trading system was tested for out-of-sample testing periods with high frequency stock data from 2012 and 2013. Our trading strategy yields cumulative returns up to 56.58% for portfolios of stock pairs, well exceeding the S&P 500 index performance by 34.35% over a 12-month trading period. The proposed trading strategy achieved a monthly 2.67 Sharpe ratio and an annual 9.25 Sharpe ratio. Furthermore, the proposed pairs trading system performed well during the two months in which the S&P 500 index had negative returns. Thus, the trading system might be especially more profitable at times when the U.S. stock market performed poorly. Therefore, the performance returns of the proposed pairs trading system were relatively market-neutral and were positive regardless of the performance of the S&P 500 index.

Keywords: cointegration, correlation, high frequency trading (HFT), long/short strategy, market-neutral, pairs trading, Sharpe ratio, statistical arbitrage

1. Introduction

High frequency trading (HFT), which is a type of algorithmic and quantitative trading, is characterized by short holding periods, specifically the use of sophisticated and powerful computing methods to trade securities rapidly. HFT has positions in equities, options, currencies, and all other financial instruments that possess electronic trading capability, aiming to capture small profits and/or fractions of a cent of profit on every short-term trade (Cartea & Penalva, 2012; Aldridge, 2010). It may have a potential Sharpe ratio, which is a ratio developed by Nobel laureate William F. Sharpe to measure risk-adjusted performance (Sharpe, 1975; 1994); this ratio could be hundreds of times higher than the Sharpe ratio of traditional buy-and-hold trading strategies (Aldridge, 2009).

Statistical arbitrage is a situation where there is a statistical mispricing of one or more assets based on the expected values of these assets or financial instruments, including stocks, derivatives, currencies, and bonds, etc. In other words, statistical arbitrage assumes the statistical mispricing of price relationships, which are true in expectation and in the long run when repeating a trading strategy (Pole, 2007). This is to say that when a profit situation takes place from pricing inefficiencies between securities, traders can identify the statistical arbitrage situation through mathematical and/or quantitative models. Statistical arbitrage depends heavily on the ability of market prices to return to a historical or predicted mean.

In the hedge fund industry, most hedge funds use statistical arbitrage, such as a market neutral strategy, long and/or short strategies, and so on (Wilson, 2010). Statistical arbitrage is referred to as a highly technical short-term mean-reversion strategy, which involves large numbers of securities, short holding periods, substantial computational models, and trading (Lo, 2010). The idea is to make many bets with positive expected returns and to produce a low-volatility investment strategy (Avellaneda & Lee, 2010), thereby taking advantage of diversification across assets.

Pairs trading is one of the most common strategies of statistical arbitrage and has been widely used by professional traders, institutional investors, and hedge fund managers. Pairs trading is a trading strategy, which takes advantage of market inefficiencies based on a pair of stocks. The perception is to identify two stocks that move together and to take long and short positions simultaneously when they diverge abnormally. Thus, it is expected that the prices of the two stocks will converge to a mean in the future (Perlin, 2009; Elliot, Hoek, & Malcolm, 2005; Caldeira & Moura, 2013). Pairs trading is one of the early quantitative methods of trading used at Wall Street that dates back to the 1980's (Vidyamurthy, 2004). Today, it continues to remain an important statistical arbitrage strategy used by hedge funds.

According to quantitative models, pairs trading requires a driving mechanism for mean-reversions using a statistical arbitrage strategy. If two stock prices were truly random, pairs trading would not work well. The Law of One Price (LOP) states that two stocks with the same payoff in every state of nature must have the same current value (Caldeira & Moura, 2013; Gatev, Goetzmann, & Rouwenhorst, 2006). Thus, two stock prices spread between close substitute assets should have a stable, long-term equilibrium price over time. This is consistent with the view that the profits are a compensation of statistical arbitrage according to the LOP.

Pairs trading can also be considered as a market-neutral trading strategy that matches a long position and a short position in a pair of highly correlated financial instruments. Like the process of statistical arbitrage, the profit value of pairs trading is derived from the difference of the price changes of the two instruments, rather than from the direction in which each instrument moves. This trading can be used to gain profit during a variety of market conditions, including periods when the equity market goes up, down, or sideways, along with low or high volatilities.

The market-neutral trading strategy often provides a hedge against market risk because one long position is taken in conjunction with another short position to reduce directional exposure. The market-neutral trading strategy is neither risk-neutral nor risk-free. As can be expected, the risks are different from those associated with market directional trading. Thus, the market-neutral trading strategy provides alternative and uncorrelated profits of return with market directions.

Recently, advances in wired and wireless high-speed wideband connections (Miao, 2007) and powerful computing methods have utilized HFT in conjunction with market-neutral statistical arbitrage strategies, such as pairs trading. Holding periods have significantly decreased from weeks to days, hours to minutes, or even shorter time periods, increasing the frequency of profit returns.

Correlation and cointegration in statistical arbitrage are related, but they highlight different concepts. High correlation in assets does not necessarily imply high cointegration in prices. Correlation reflects co-movements in assets, but it is usually unstable over time. High correlation alone is not sufficient enough to ensure the long-term performance of hedges. Correlations based on hedge strategies commonly require frequent rebalance (Alexander, 1999). On the other hand, cointegration measures long-term co-movements in prices even through a period when correlation appears low. Therefore, cointegration based on hedge strategies may be more effective in long-term running and short-term dynamic trends.

Thus, in this paper, a high frequency and dynamic pairs trading system is proposed, based on a market-neutral statistical arbitrage strategy using a two-stage correlation and cointegration approach. The proposed pairs trading system was applied to equity trading and was able to capture statistical mispricing between the prices of stock pairs, using regression residuals, and to model them as natural mean-reversion processes with a short holding period in the U.S. equities market under any market cycle conditions.

2. High Frequency Stock Data

A high frequency equity database contained 2,100 preselected individual stocks, in which each stock belonged to the NYSE and/or NASDAQ in the U.S. equities market. Each high frequency stock datum consisted of 15-minute stock prices, which included open, high, low, and close prices with a volume, and were referred to as open-high-low-close (OHLC) stock prices. These stocks had trading dates ranging from May 1, 2012 to the present, where new intraday data were updated automatically on a daily basis. Each high frequency stock datum was separated into different sectors. In this paper, 177 oil and gas stocks from the energy sector were selected for the high frequency and dynamic pairs trading system, based on the market-neutral statistical arbitrage strategy using the two-stage correlation and cointegration approach.

3. Algorithms, Modeling and Methods

Correlation is a statistical term derived from linear regression analysis, which describes the strength of a relationship between two variables. The central idea of pairs trading is to see if two stocks are highly correlated. Then, any changes in correlation may be followed by mean-reversion to the trend of stock pairs, thereby creating a profit opportunity.

On the other hand, cointegration is an attempt to parameterize pairs trading strategies, which explore the possibility of a statistical feature, where two stocks are cointegrated. That is, two stocks can be linearly combined to produce a stationary time series. Cointegration is a powerful tool, which allows the establishment of a dynamic model of two non-stationary time-series stocks.

Pairs trading attempts to identify a relationship between two stocks, determine the direction of their relationship, and execute long and short positions based on the statistical data presented simultaneously. Selecting a good stock pair for trading becomes the most important stage of mean-reversion of the market-neutral statistical arbitrage strategy.

Thus, in this section, the applicability of the two-stage correlation and cointegration approach for the high frequency and dynamic pairs trading system, based on the mean-reversion of the market-neutral statistical arbitrage strategy, is discussed in detail in terms of algorithms, quantitative models, and pairs trading strategies and methods.

3.1 Pairs Trading Using Correlation

Correlation measures the relationship between two stocks that have price trends. They tend to move together, and thus are correlated. By identifying two highly correlated stocks, we can look for periods of divergence, figure out why two stock prices are separating, and attempt to take profit through convergence, which is a mean-reversion process.

In selecting each of the stock pairs, all of the stocks belonging to the same energy sector were considered. Competing stocks within the same sector made natural potential stock pairs. This was due to similar market risks, exposed to all of the stocks in the same sector. Suitable stock pairs were found to have commonalities with good liquidity, and furthermore could be sold by using short-sell.

In order to identify matched stock pairs, all of the stocks from the same sector were screened by calculating correlation coefficients and/or using a minimum-distance criterion, which was the sum of squared deviations between two normalized stock prices. A matched stock pair was chosen by determining whether or not it had a high correlation coefficient and/or small minimum distance between the stock pairs.

Considering two stocks A and B, a correlation coefficient between the stocks was a statistic that not only provided a measure of how the two stocks A and B were associated in a sample but also its properties, which closely related them to a straight-line regression. The correlation coefficient ρ of stock A and stock B was obtained by (Kleinbaum, Kupper, & Muller, 1988; Miao & Clements, 2002),

$$\rho = \frac{\sum_i^N (A_i - \bar{A})(B_i - \bar{B})}{\left[\sum_i^N (A_i - \bar{A})^2 \sum_i^N (B_i - \bar{B})^2\right]^{1/2}} \tag{1}$$

where $\bar{A}$ and $\bar{B}$ were the mean prices of stocks A and B, respectively, and their formula was given by

$$\bar{A} = \frac{1}{N}\sum_i^N A_i \tag{2}$$

and

$$\bar{B} = \frac{1}{N}\sum_i^N B_i \tag{3}$$

where N denoted a stock trading data range and ρ was a dimensionless quantity, which was in the range of $-1 \leq \rho \leq 1$. In other words, ρ was independent of the units of stocks A and B. The more positive ρ was, the more positive the association of stocks A and B was. This meant that stocks A and B were highly matched.

The minimum-distance criterion, which could be considered as an alternative method of correlation calculation as the sum of the squared deviations between the normalized prices of stocks A and B, was obtained by:

$$\varepsilon = \sum_i^N (A_i - B_i)^2 \tag{4}$$

where $\varepsilon \geq 0$ usually. The smaller the value ε was, the more similar stocks A and B were. This meant that stocks A and B were a highly matched pair.

Thus, both the correlation coefficient and the minimum-distance criterion could be used as criteria for ranking and selecting stock pairs. Stock pairs with the highest correlation coefficients or the smallest minimum-distance values were chosen. In this paper, the correlation coefficients, criteria for pre-selection, were used for selecting potential stock pairs.

However, pairs trading based on a correlation approach alone would have a disadvantage of instabilities between prices of a stock pair over time. Correlation coefficients do not necessarily imply mean-reversion between the prices of the two stock pairs. This is because the correlation approach is sensitive to small time deviations (Harris, 1995; Lin, McCrae, & Gulati, 2006), especially in high frequency and dynamic pairs trading.

In order to overcome the above issue in the correlation approach for pairs trading, a cointegration approach was further used as the second-step of the selection process for the paired stocks, which were first selected by using the correlation coefficients for pairs trading.

3.2 Pairs Trading Using Cointegration

The cointegration concept, an innovative mathematical model in econometrics developed by Nobel laureates Engle and Granger (Engle & Granger, 1987), established much interest among economists in the last decade. Cointegration states that, in some instances, despite two given non-stationary time series, a specific linear combination of the two time series is actually stationary. In other words, the two time series move together in a lockstep fashion.

The definition of cointegration is the following: assume that x_t and y_t are two time series that were non-stationary. If there existed a parameter γ such that the following equation

$$z_t = y_t - \gamma x_t \tag{5}$$

was a stationary process, then x_t and y_t would be cointegrated. This path-breaking process emerged as a powerful tool for investigating common asset trends in multivariate time series. Cointegration provided a sound methodology for modeling both the long-term equilibrium and the short-term dynamic trends of the time-series samples.

3.2.1 Stationary Process

In time-series analyses and applications, the statistics or ensemble averages of a random process were often independent of time. It was commonly assumed that a time series signal had first-order and second-order probability density functions, which were independent of time. These conditions were sometimes referred to as stationary or statistical time-invariance.

A stationary process, or stationary for short in this paper, was essentially a stochastic process, in which its joint probability distribution did not change when it was shifted in time (Miao, 2007; Miao & Clements, 2002). Consequently, its corresponding mean and variance also did not change over time. Likewise, the mean and variance of a random process did not follow trends.

The process was said to be a first-order stationary if the first-order density function of a random process or a time series signal was independent of time. For the first-order stationary, the first-order statistics were invariant to a time shift of the process. That is, the mean of the random process was a constant,

$$m_t = m \tag{6}$$

and the variance was also a constant,

$$\sigma_t^2 = \sigma^2 \tag{7}$$

Similarly, if a second-order joint density function depended only on the difference, $t_2 - t_1$, rather than on individual times t_1 and t_2, then the process was said to be a second-order stationary. For the second-order stationary, the mean and variance were constants as well. In addition, the correlation between the random variables x_{t1} and x_{t2} depended only on the time difference,

$$R_{t2,\ t1} = R_{t2-t1} \tag{8}$$

where $R_{t2,\ t1} = E\{x_{t1}\ x_{t2}^*\}$ was an autocorrelation function (Miao, 2007; Miao & Clements, 2002) and the time difference, t_2-t_1, was referred to as the lag.

For the high frequency and dynamic pairs trading system, a recursive formula for frequently updating statistics was used to quickly calculate the mean and variance. A recursive mean was obtained by

$$m_{t+1} = \frac{1}{t+1}(m_t + x_{t+1}) \tag{9}$$

where x_{t+1} was the (t+1)*th* sample, and a recursive variance was computed by

$$\sigma_{t+1}^2 = \left(1 - \frac{1}{t}\right)\sigma_t^2 + (t+1)\left(m_{t+1} - m_t\right)^2 \tag{10}$$

Stationary was an important concept in time-series analysis and applications of stock pairs trading, where the raw stock data were often transformed to become stationary. It was an especially important concept tool in the application of the cointegration approach, as discussed in the next section in detail.

3.2.2 Cointegration Approach

Let P_t^A and P_t^B be the prices of two stocks A and B, respectively. If it was assumed that $\{P_t^A,\ P_t^B\}$ were individually non-stationary, there existed the parameter γsuch that the following equation was a stationary process

$$P_t^A - \gamma P_t^B = \mu + \varepsilon_t, \tag{11}$$

where μ was a mean of the cointegration model. ε_t was a stationary, mean-reverting process and was referred to as a cointegration residual, a regression residual, or a residual for short. The parameter γ was known as a *cointegration coefficient*. Equation (11) represented a model of a cointegrated pair for stocks A and B.

Also note that Equation (11) could be used in the logarithm for pair trading as well. In this case, it would only be valid to represent logarithm prices when these logarithm prices were cointegrated and their corresponding residuals were stationary.

The cointegration process determined the cointegration coefficient γ, and the long-term equilibrium relationship between stocks A and B determined the mean μ of the cointegration model. Thus, a quantity Q of profit or loss per trade of an investment or in pairs trading could be estimated as follows:

$$\begin{aligned} Q &= (P_t^A - P_{t+1}^A) - \gamma(P_t^B - P_{t+1}^B) \\ &= (P_t^A - \gamma P_t^B) - (P_{t+1}^A - \gamma P_{t+1}^B) \\ &= (\mu + \varepsilon_t) - (\mu + \varepsilon_{t+1}) \\ &= \varepsilon_t - \varepsilon_{t+1}. \end{aligned} \tag{12}$$

Equation (12) created three possible trading results in terms of profits or losses for each stock pairs trading process. As can be seen, if the term $(\varepsilon_t - \varepsilon_{t+1}) > 0$, the pairs trading made a positive profit. If the term $(\varepsilon_t - \varepsilon_{t+1}) = 0$, the pairs trading was a breakeven. If the term $(\varepsilon_t - \varepsilon_{t+1}) < 0$, then the pairs trading produced a negative return. Therefore, it is important to understand how the regression residual ε_t could be used for each pairs trading.

3.2.3 Cointegration Verification

In the Engle-Granger method (Engle & Granger, 1987), we first set up a cointegration regression between stocks A and B as stated in Equation (11), and then estimated the regression parameters μ and γ using an ordinary least squares (OLS) method. Subsequently, we tested the regression residual ε_t to determine whether or not it was stationary. If the regression residual ε_t was stationary, then the two stock prices $\{P_t^A,\ P_t^B\}$ were said to be cointegrated.

There existed a number of different stationary tests to verify the regression cointegration (Johansen, 1988; Cochrance, 1991; Wang & Yau, 1994). The most popular stationary test in the area of cointegration, the Augmented Dickey Fuller (ADF) test (Dickey & Fuller, 1979), was used on the regression residual ε_t to determine whether it had a unit root.

Testing for the presence of the unit root in the regression residual ε_t using the ADF test was given by

$$\Delta Z_t = \alpha + \beta t + \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + u_t \tag{13}$$

where α was a constant, β was the coefficient on a time trend, p was the lag order of the autoregressive process, and u_t was an error term and serially uncorrelated. If both parameters $\alpha = 0$ and $\beta = 0$, Equation (13) modeled a random walk. If $\alpha \neq 0$ and $\beta = 0$, Equation (13) modeled a random walk with a drift. To estimate all of the parameters in Equation (13), OLS was used.

The number of lag order p in Equation (13) was usually unknown and therefore had to be estimated. To determine the number of lag order p, the information criteria for lag order selection was used, such as the Akaike information criterion (AIC) (Akaike, 1992), Schwartz information criterion (SIC) (Schwarz, 1978), Hannan-Quinn criterion (HQC) (Hannan & Quinn, 1979), final prediction error (FPE) (Akaike, 1969), and Bayesian information criterion (BIC) (Akaike, 1979; Liew, 2004). The method for estimating the number of lag order p was to minimize one of the following criteria:

$$AIC = \ln(\hat{\sigma}_p^2) + \frac{2p}{T} \tag{14}$$

$$SIC = \ln(\hat{\sigma}_p^2) + \frac{p\ln(T)}{T} \tag{15}$$

$$HQC = \ln(\hat{\sigma}_p^2) + \frac{2p\ln[\ln(T)]}{T} \tag{16}$$

$$FPE = \hat{\sigma}_p^2 (T-p)^{-1}(T+p) \tag{17}$$

and

$$BIC = (\mathrm{T}-\mathrm{p})\ln\left(\frac{T\hat{\sigma}_p^2}{T-p}\right) + T\left[1+\ln\left(\sqrt{2\pi}\right)\right] + p\ln\left[\frac{\sum_{t=1}^{T}(\Delta Z_t)^2 - T\hat{\sigma}_p^2}{p}\right] \tag{18}$$

where T was the sample size, and the estimation of the error variance $\hat{\sigma}_p^2$ was given by

$$\hat{\sigma}_p^2 = \frac{\sum_{t=p}^{T} \widehat{u_t}^2}{T-p-1} \tag{19}$$

where u_t was the error term in Equation (13). Thus, using one of these criteria, called $C_T[p]$, in Equations (14), (15), (16), (17) or (18), the number of lag order p could be estimated according to the formula as follows:

$$\hat{p} = argmin_{p \leq p_{max}} \{C_T[p]\}. \tag{20}$$

The p_{max} in Equation (20) was the maximum lag order, which could be determined by using a rule of thumb (Ng & Perron, 1995):

$$p_{max} = \left\lfloor 12 \cdot \left(\frac{T}{100}\right)^{1/4} \right\rfloor \tag{21}$$

where $\lfloor \cdot \rfloor$ denoted an integer operation. As can be seen, this choice allowed the maximum lag order p_{max} to grow with the sample size of T.

The unit root test for the regression residual ε_t using the ADF test was then carried out under the null hypothesis H$_0$: $\gamma = 0$ versus the alternative hypothesis H$_1$: $\gamma < 0$. A statistical value of the ADF test was obtained by

$$ADF\ test = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \tag{22}$$

where $\hat{\gamma}$ and $SE(\hat{\gamma})$ were the cointegration coefficient and standard errors of the OLS estimate, respectively. The standard errors $SE(\hat{\gamma})$ could be computed by

$$SE(\hat{\gamma}) = \sqrt{\frac{\sum_{i=1}^{n}\left(\Delta Z_{ti} - \widehat{\Delta Z}_t\right)^2}{(n-2)\sum_{i=1}^{n}\left(Z_{(t-1)i} - \overline{Z_{(t-1)\iota}}\right)^2}}\ . \tag{23}$$

The test result in Equation (22) was compared with the critical value of the ADF test. If the test statistic in Equation (22) was less than the critical value, then the null hypothesis H$_0$: $\gamma = 0$ was rejected. This meant that no unit root was present, and the regression residual ε_t in Equation (11) was thereby stationary. Thus, the two stock prices $\{P_t^A,\ P_t^B\}$ were cointegrated.

3.3 Pairs Trading Strategies

The pairs trading strategies, proposed in this paper, were referred to as market-neutral statistical arbitrage strategies using trading signals based on the regression residual ε_t in Equation (11) and were modeled as a mean-reverting process, and not the prices themselves.

In this paper, in order to select potential stocks for pairs trading, the two-stage correlation and cointegration approach was used. The first step of the pairs trading strategy was to identify potential stock pairs from the same sector, where the stock pairs were selected with correlation coefficients of at least 0.90 by using the correlation approach stated in Equation (1). The first step was also referred to as pre-selection of the stock pairs. The second step of the pairs trading strategy was to check to see if the pre-selected stock pairs were integrated in the same order, and were cointegrated according to Equation (11). If the null hypothesis H$_0$: $\gamma = 0$ was rejected, then no unit root was present. That is, the regression residual ε_t was stationary.

To test the cointegration, the Engle and Granger's approach was adapted by using the ADF test statistics based on Equations (13), (22), and (23). The cointegration tests were completed on all potential stock pairs that had been initially pre-selected based on the correlation coefficients. Selections of the stock pairs for pairs trading had to pass the cointegration test by using the ADF test statistics.

The third step of the pairs trading strategy was to rank all of the stock pairs based on the cointegration test values. The smaller the cointegration test value was, the higher the rank the stock pair was assigned to. Final selection of the stock pairs from the top rank was used for the out-of-sample testing periods.

The final step of the pairs trading strategy was to define a couple trading rules. By simultaneously taking both long and short positions for pairs trading, we must be able to determine when we should open and when we should close the pairs trading based on quantitative definitions. To open a pairs trading, the regression residual ε_t in Equation (11) must cross over and down the positive δ standard deviations above the mean or cross down and over the negative δ standard deviations below the mean, where δ was a positive value. When the regression residual ε_t in Equation (11) returned to the mean, the pairs trading was closed. Furthermore, in order to prevent the loss of too much trading capital on a single pairs trading, a stop-loss was used to close the pairs trading when the regression residual ε_t hit 2δ positive or negative standard deviations.

The composition of the portfolio for the pairs trading was not rebalanced even when stock prices moved and/or pairs trading positions might no longer be market-neutral after opening. However, only two actions of the pairs trading were used: to open a new position or to close the previously opened position with the total liquidation simultaneously.

3.4 Performance Evaluation

In order to evaluate the performance of the out-of-sample testing periods for the high frequency and dynamic pairs trading system based on the market-neutral statistical arbitrage strategy, a Sharpe ratio (Sharpe, 1975; 1994) was used to measure a risk-adjusted portfolio performance. The Sharpe ratio, also known as the Sharpe index and the Sharpe measure, measured the excess return per unit of derivation for an investment asset, which was the risk-adjusted performance. The Sharpe ratio (SR) formula was given by

$$SR = \sqrt{K}\,\frac{\overline{r_p} - r_f}{\sigma_p} \tag{24}$$

where $\overline{r_p}$ was an expected portfolio return, r_f was a risk free rate, σ_p was a portfolio standard deviation, and *K* was a constant. A 3-month U.S. Treasury Bill Rate at 0.08% was used as the risk free rate r_f for 2012 and 2013 (U.S. Treasury Bill Rate, 2013). To calculate an annual Sharpe ratio, the constant *K* was set to a different value, depending on whether the portfolio returns were hourly, daily, weekly, monthly, quarterly, or yearly. For the hourly returns, it was set as *K* = 1,638 for the annual Sharpe ratio. For the daily returns, *K* = 252; for the weekly returns, *K* = 50. For the monthly returns, *K* = 12; for the quarterly returns, *K* = 4. Finally, for the yearly returns, *K* = 1.

Equation (24) indicated whether the returns for a portfolio were due to smart investment decisions or excess risk. For the performance of the out-of-sample testing periods, the greater a portfolio's Sharpe ratio was, the better its risk-adjusted performance had been.

4. Simulation Results

During simulations of pairs trading, 177 stocks, which were all related to energy companies including oil and gas firms that were traded in public in the NYSE and/or NASDAQ markets, were collected to form a sector dataset. This dataset was referred to as energy sector. It contained all of the stocks with the 15-minute OHLC stock data along with volumes, which were split and dividend adjusted, from the trading dates ranging from May 1, 2012 to the present. All of the 15-minute OHLC stock data were initially divided into training (in-sample) and testing (out-of- sample) datasets. This method of separating a dataset into training and testing periods was referred to as the *holdout* method in statistical classification (Miao & Clements, 2002). A training period was preselected, in which the measures and parameters of the cointegration model in Equation (11) were estimated and computed. Immediately after the training period, a testing period followed, where the estimated model parameters based on the in-sample training period were used to test the performance of the pairs trading during the out-of-sample testing period.

4.1 A Dynamic Rolling Window

In the in-sample training period of the pairs trading, each of the training data contained a 3-month period, which was a dynamic rolling window size, where there were approximately 66 trading days including a total of 1,716 15-minute OHLC for each of the stocks; each of the testing data contained a 1-month period that immediately followed the previous 3-month training data period, where there were approximately 22 trading days including 572 15-minute OHLC for each of the testing stocks. After completing the first testing process, the dynamic rolling window was automatically shifted 1 month ahead for the next training and testing periods. This dynamic shifting process was repeated for each stock pairs trading until all of the 15-minute OHLC stock data were used completely.

4.2 In-Sample Training

During the in-sample training period, there was a total of 15,576 possible combinations of stock pairs using the 177 stocks within the energy sector. The first step of using the correlation approach was performed on all possible stock combinations, where each of the stock pairs was pre-selected if its correlation coefficient was equal to or greater than 0.90. The second step involving the cointegration approach was tested on all of the pre-selected stock pairs, which were already selected based on the results from the correlation approach. A subset of the stock pairs was further selected if the test value of the cointegration approach was equal to or less than -3.34. This indicated that the test statistics of the cointegration approach was at a 95% criteria level. Of the initial 15,576 possible stock pairs, an average of 53 stock pairs that passed both requirements of the correlation approach and the cointegration test statistics was finally obtained during the in-sample training period.

The selected stock pairs were subsequently ranked based on the test statistics of the cointegration approach. The smaller the test value of the cointegration approach was, the higher the rank that the stock pair was given. Only the top 10 stock pairs with the highest ranks were selected from each of the 3-month in-sample training periods; the following 1-month testing period of pairs trading was then carried out for evaluating the cointegration model of pairs trading performance. Once this first testing period was finished, the dynamic rolling window was initiated to shift it 1 month ahead to form the second training and testing periods. All of the stock pairs were reselected according to the requirements of both the correlation and the test statistics of the cointegration approaches. Then all of the model parameters in Equation (11) were re-estimated and recomputed to be used for the second testing period. This procedure was continued in an overlaid and dynamic rolling window fashion to the end of all of the 15-minute OHLC stock data.

Table 1 shows the details of the top 10 ranked stock pairs, obtained during the first in-sample training period dating from May 1, 2012 to July 31, 2012 and each was associated with its correlation coefficient, minimum distance, test value of its cointegration, and rank. These stock pairs were used for the high frequency and dynamic pairs trading system based on the market-neutral statistical arbitrage strategy. Note that out of the 15,576 possible stock pairs, an average of only 53 stock pairs had both the correlation coefficient of at least 0.90 and the cointegration test value of less than -3.34. Of those stock pairs, the top 10 stock pairs with the highest ranks were finally selected to be used for the out-of-sample testing period. Although all of the top 10 ranked stock pairs presented positive returns during the in-sample training period, there was no guarantee that all top 10 of the stock pairs would produce positive returns of the pairs trading during the out-of-sample testing period.

Table 1. A typical list of the top 10 ranked stock pairs obtained during the first in-sample training period (from May 1, 2012 to July 31, 2012)

Stock Pairs	Correlation Coefficient	Minimum Distance	Cointegration Test Value	Rank of Stock Pair
ESV/NE	0.9436	0.6712	-4.5454	1
ETP/MWE	0.9120	0.1354	-4.2033	2
CLMT/RIG	0.9218	0.1141	-4.1938	3
RDC/SGY	0.9226	0.2946	-4.0770	4
ENB/SDRL	0.9245	0.1241	-4.0511	5
EOG/WMB	0.9598	0.8773	-3.7894	6
AREX/WLL	0.9530	0.0375	-3.6805	7
SEMG/VLO	0.9031	0.2712	-3.6805	8
OAS/SPN	0.9439	0.0206	-3.4807	9
CNQ/PXD	0.9666	0.0007	-3.4405	10

Figure 1 shows a typical intraday chart of the stock pair (CLMT/RIG) with normalized prices in the 15-minute time duration, during the first in-sample training period from May 1, 2012 to July 31, 2012. This stock pair had a correlation coefficient of 0.9218 and cointegration test value of -4.1938, which was ranked third, as listed in Table 1. As can be seen, by identifying correlated and cointegrated stock pairs as shown in Figure 1, periods of divergence could be found where the prices of the stock pairs were separating in statistics. Therefore, we attempted to profit through convergence by using mean-reversion of the pairs trading based on the market-neutral statistical arbitrage strategy.

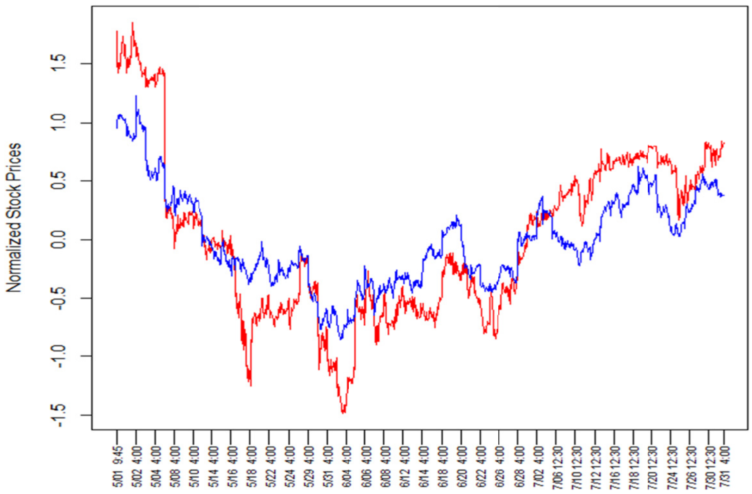

Figure 1. Normalized prices of the stock pair (CLMT/RIG) in a 15-minute time duration, in which the red curve was the stock (CLMT) and the blue curve was the stock (RIG)

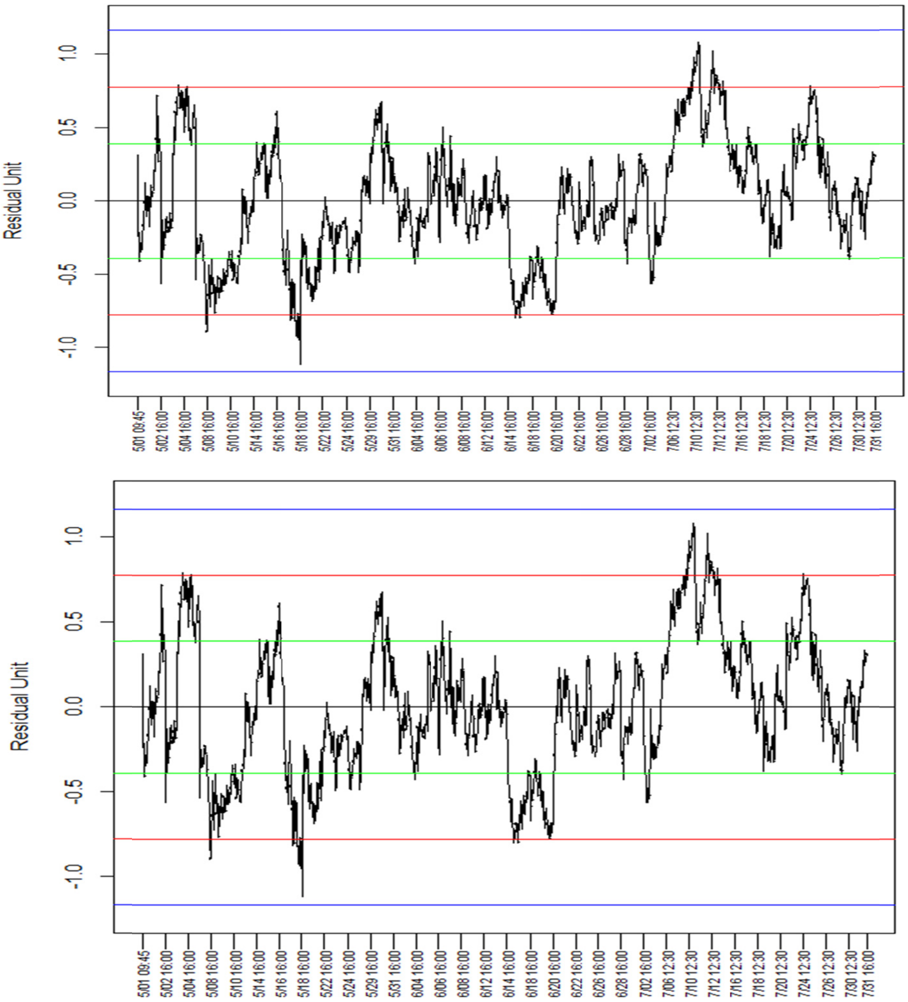

Figure 2. A corresponding residual chart of the stock pair (CLMT/RIG), in which the black line was the zero-mean, the green lines were ±1 standard deviations, the red lines were ±2 standard deviations, and the blue lines were ±3 standard deviations

The corresponding residual chart of the stock pair (CLMT/RIG) is shown in Figure 2, in which the black line was the zero-mean; the green, red, and blue lines represented ± 1, ± 2, and ± 3 standard deviations, respectively. In this figure, there were 3 trading opportunities in the area spanning from the 2 positive standard deviations (red line) down to the zero-mean (black line) and 3 trading opportunities in the area spanning from the 2 negative standard deviations (red line) up to the zero-mean (black line). These trading opportunities reinforced the mean-reversion characteristic fashions. Thereby, the residual chart of the stock pair as shown in Figure 2 had a desirable characteristic for the mean-reversion of the pairs trading based on the market-neutral statistical arbitrage strategy.

4.3 Out-of-Sample Testing

During the out-of-sample testing periods, the high frequency and dynamic pairs trading system based on the market-neutral statistical arbitrage strategy opened long and short positions simultaneously when the regression residual ε_t in Equation (11) passed through either the positive or negative 2 standard deviations line twice. The trading system also closed both positions at the same time when the regression residual ε_t reached the zero-mean. A stop-loss was also employed when the regression residual ε_t reached ± 4 standard deviations.

Table 2 shows the summarized results of the proposed high frequency and dynamic pair trading system based on the market-neutral statistical arbitrage strategy, obtained from all of the out-of-sample testing periods using the top 10 ranked stock pairs ranging from August 2012 to July 2013. There was a total of 169 trades, including 149 winning trades and 20 loss trades. The ratio of the wining trades divided by the loss trades was 7.45. The percentages of the winning and loss trades were 88.17% and 11.83%, respectively. The average percentage of a winning trade was 3.22%, while the average percentage of a loss trade was -1.88%. The average percentages were calculated under an assumption of zero transaction costs. Even if transaction costs were considered, such as a fixed standard transaction fee of $7 (U.S. dollars) per trade for an unlimited number of shares or a transaction fee of $0.002 per share per trade, the calculation of the winning and loss trades would not be affected significantly. The inclusion of transaction costs in calculating the trades would be negligible if a relatively larger amount of capital was used for each of the pairs trading.

In Table 2, the absolute ratio of reward divided by risk per trade was 1.71. The maximum and minimum percentages of monthly returns were 5.88% and 1.61%, respectively. The average of the monthly returns was 3.82%, with a standard deviation of 1.40%. Thus, the proposed pairs trading system achieved a monthly 2.67 Sharpe ratio and annual 9.25 Sharpe ratio for the out-of-sample testing periods from August 2012 to July 2013.

Table 2. A summary of performance results of the proposed pairs trading system during the out-of-sample test periods (from August 2012 to July 2013)

Pairs Trading Portfolio (All Trading Results)	Top 10 Ranked Stock Pairs	Pairs Trading Portfolio (Monthly Return Results)	Top 10 Ranked Stock Pairs
Number of Winning Trades	149	Mean	3.82%
Number of Loss Trades	20	Median	3.97%
Total Number of Trades	169	Minimum	1.61%
Winning/Loss Ratio	7.45	Maximum	5.88%
Percentage of Winning Trades	88.17%	Standard Deviation	1.40%
Percentage of Loss Trades	11.83%	Skewness	-0.0468
Average Percentage of Winning Trades	3.22%	Kurtosis	-0.7904
Average Percentage of Loss Trades	-1.88%	Monthly Sharpe Ratio	2.67
Reward/Risk Ratio	1.71	Annual Sharpe Ratio	9.25

Table 3 shows the comparative monthly and cumulative returns of profits or losses during the out-of-sample testing periods using the top 10 ranked stock pairs, which were selected during each of the in-sample training periods, along with those from tracking the S&P 500 index performance in the NYSE equity market. The cumulative return of the S&P 500 index performance was 22.23% during the period from August 2012 to July 2013, while the proposed high frequency and dynamic pairs trading system, based on the market-neutral statistical arbitrage strategy using the two-stage correlation and cointegration approach, achieved 56.58%. Compared to the S&P 500 index performance, the proposed high frequency and dynamic (HFD) pairs trading system performed 34.35% more in terms of cumulative returns. In addition, the S&P 500 index had negative returns during the two months, October 2012 and June 2013, while the proposed pairs trading system showed positive returns for all of the months during 2012 and 2013.

Table 3. A summary of performance results of the monthly and cumulative returns of profits or losses based on the proposed pairs trading system and the S&P 500 index respectively

	Monthly Returns Performance		Cumulative Returns Performance	
Date Periods	S&P 500 Index	HFD Pair Trading	S&P 500 Index	HFD Pair Trading
2012				
August	1.98%	1.88%	1.98%	1.88%
September	2.42%	4.45%	4.45%	6.41%
October	-1.98%	4.31%	2.38%	11.00%
November	0.28%	5.88%	2.67%	17.53%
December	0.71%	3.97%	3.40%	22.19%
2013				
January	5.04%	1.61%	8.61%	24.16%
February	1.11%	4.93%	9.81%	30.28%
March	3.60%	3.36%	13.77%	34.66%
April	1.81%	5.87%	15.82%	42.56%
May	2.08%	3.97%	18.23%	48.22%
June	-1.50%	2.66%	16.46%	52.17%
July	4.95%	2.90%	22.23%	56.58%

Figure 3 shows a line chart comparing the performances of the proposed high frequency and dynamic pairs trading system using the market-neutral statistical arbitrage strategy and the S&P 500 index based on cumulative returns of profits or losses for the top 10 ranked stock pairs. In this figure, the red line represented the proposed pairs trading system performance, and the blue line represented the S&P 500 index performance. The proposed pairs trading system performance well exceeded the S&P 500 index performance by 34.35% after the end of the 12-month trading period from August 2012 to July 2013. In addition, the proposed pairs trading system performed well during the two difficult months (October 2012 and June 2013) when the S&P 500 index had negative returns. In other words, the proposed pairs trading system might be more profitable in times when the U.S. stock market performed poorly. Furthermore, a correlation measure of the monthly returns between the proposed pairs trading system and the S&P 500 index performance was about -0.38. As can be expected, there was a low negative correlation between their performances in which the returns of the proposed pairs trading system performed positively regardless of the performance of the S&P 500 index.

Figure 4 shows a line chart comparing the performances of the proposed pairs trading system (red line) and the S&P 500 index (blue line) based on their monthly returns. The performance returns of the proposed pairs trading system were relatively market-neutral and independent of the performance returns of the S&P 500 index. Likewise, the performances of the proposed pairs trading system and the S&P 500 index did not correlate; the proposed pairs trading system performed more positively than the S&P 500 index.

Additionally, a bar chart comparing the performances of the proposed pairs trading system (red) and the S&P 500 index (blue) based on monthly returns is shown in Figure 5. As can be seen, all of the performance returns of the proposed pairs trading system were positive returns, while the S&P 500 index had two negative returns in October and June during the 12-month trading period. The proposed pairs trading system generated an average monthly return of 3.82% in profits while the S&P 500 index produced an average monthly return of 1.71% in profits. Thus, the proposed pairs trading system outperformed the S&P 500 index by 123% in monthly returns.

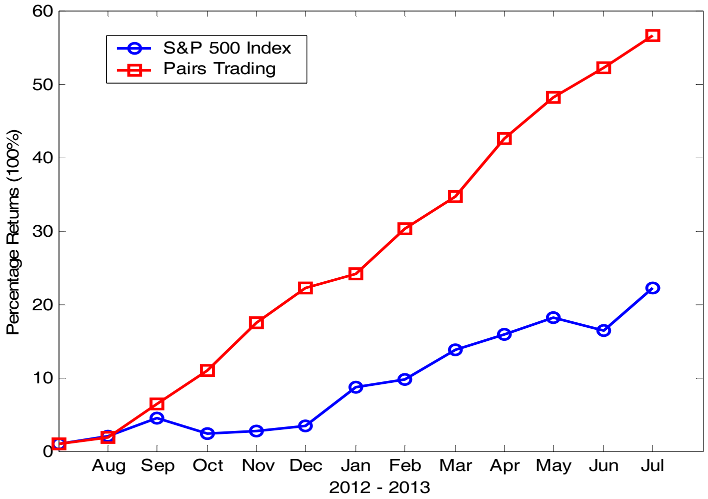

Figure 3. A line chart comparing the performances of the proposed pairs trading system using the market-neutral statistical arbitrage strategy (red line) and the S&P 500 index (blue line) based on their cumulative returns

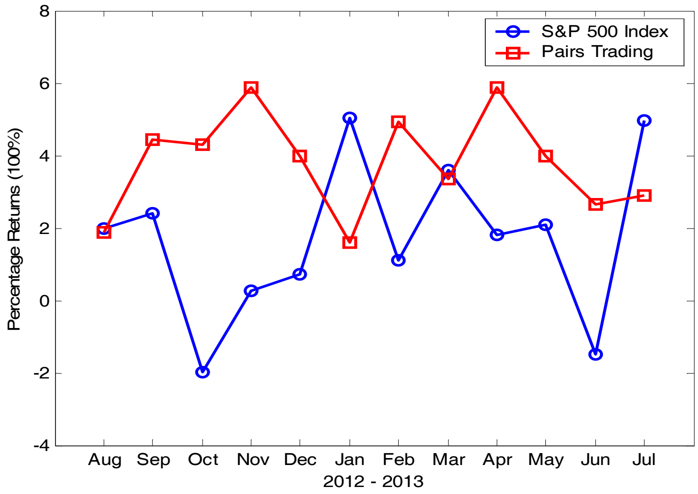

Figure 4. A line chart comparing the performances of the proposed pairs trading system (red line) and the S&P 500 index (blue line) based on their monthly returns

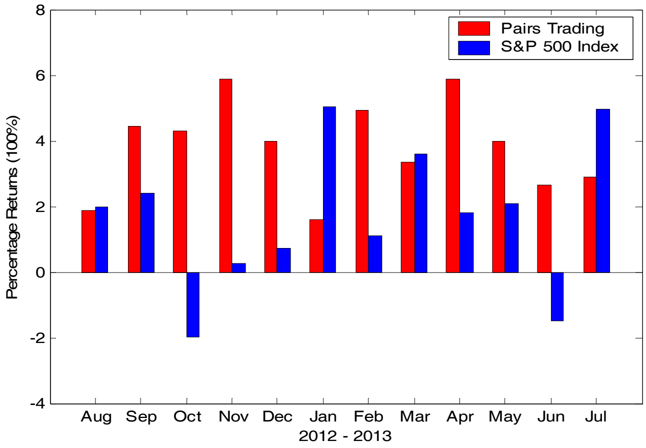

Figure 5. A bar chart comparing the performances of the proposed pairs trading system (red) and the S&P 500 index (blue) based on their monthly returns

5. Conclusion

In this paper, a high frequency and dynamic pair trading system was developed based on a market-neutral statistical arbitrage strategy using the two-stage correlation and cointegration approach. All of the stocks that were used for the proposed pairs trading system belonged to the same energy sector in the NYSE and/or NASDAQ equity markets. The composition of the portfolio of the proposed pairs trading system had vanishing portfolio betas. Thus, it was uncorrelated with the market factors that drove the portfolio returns. The trading strategy of the proposed pairs trading system was implemented using the two-stage correlation and cointegration approach, which explored the mean-reversion of the top 10 ranked stock pairs from each of the in-sample training periods.

Correlation measures were first applied to all possible stock pair combinations. The pre-selected stock pairs must have had correlation coefficients that were at least 0.9. Then, the cointegration tests were applied on all of the pre-selected stock pairs in order to identify stock pairs that shared long-term equilibrium relationships. The selected stock pairs should have had test values of cointegration that were less than -3.34, which indicated that the test statistics of the cointegration approach were at the 95% criteria level. Out of 15,576 possible stock pairs, an average of 53 correlated and cointegrated stock pairs from each of the in-sample training periods were obtained. Subsequently, all of the stock pairs were ranked based on the cointegration test values obtained from the in-sample training periods. From there, a portfolio containing the top 10 ranked stock pairs that displayed the lowest test values of cointegration in the in-sample training periods was finally compiled to be traded for the out-of-sample testing periods.

The cumulative net profit of the proposed pairs trading system during the 12 months of the out-of-sample test periods was 56.58%, with an average monthly return of 3.82% and a standard deviation of 1.40%. This resulted in a 2.73 monthly Sharpe ratio and 9.45 annual Sharpe ratio for the out-of-sample test periods. Furthermore, the proposed pairs trading system showed relatively low levels of volatility and no significant correlation with the S&P 500 index performance in the NYSE equity market, thereby confirming its market neutrality. In addition, the test results during the out-of-sample test periods were attractive when compared to other strategies employed by hedge funds and professionals (Lo, 2010; Jaeger, 2003). From these results, the proposed high frequency and dynamic pairs trading system, produced by the market-neutral statistical arbitrage strategy using the two-stage correlation and cointegration approach, reinforced the use of correlation and cointegration as important integration tools in quantitative trading, risk control, and money management.

In future research, Kalman filtering techniques (Miao & Clements, 2002; Grewal & Andrews, 2008) would be used for estimating and identifying the adaptive stability of the parameters of the cointegration model in real-time mode. Therefore, there would be further enhancements of profitability and mitigation of risks for pairs trading based on the market-neutral statistical arbitrage strategy.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21*, 243–247. http://dx.doi.org/10.1007/BF02532251

Akaike, H. (1979). A bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika, 66*, 237–242. http://dx.doi.org/10.2307/2335654

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (Vol. 1, pp. 610–624). London, England: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4612-0919-5_38

Aldridge, I. (2009). *How profitable are high-frequency strategies?* FINalternatives: Hedge Fund & Private Equity News. Retrieved from http://www.finalternatives.com/node/9271

Aldridge, I. (2010). *High-frequency trading: A practical guide to algorithmic strategies and trading system*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Alexander, C. (1999). Optimal hedging using cointegration. *Philosophical Transactions of the Royal Society, Series A, 357*, 2039–2058. http://dx.doi.org/10.1098/rsta.1999.0416

Avellaneda, M., & Lee, J. H. (2010). Statistical arbitrage in the U.S. equities market. *Quantitative Finance, 10*, 1–22. http://dx.doi.org/10.2139/ssrn.1153505

Caldeira, J. F., & Moura, G. V. (2013). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *Revista Brasileira de Financas (Online), Rio de Janeiro, Brazil, 11*(1), 49–80. http://dx.doi.org/10.2139/ssrn.2196391

Cartea, A., & Penalva, J. (2012). Where is the value in high frequency trading? *Quarterly Journal of Finance, 2*(3), 1–46. http://dx.doi.org/10.1142/S2010139212500140

Cochrance, J. H. (1991). A critique of the application of unit root tests. *Journal of Economic Dynamics and Control, 15*, 275–284. http://dx.doi.org/10.1016/0165-1889(91)90013-Q

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431. http://dx.doi.org/10.2307/2286348

Elliot, R. J., Hoek, J. V. D., & Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance, 5*(3), 271–276. http://dx.doi.org/10.1080/14697680500149370

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica, 55*(2), 251–276. http://dx.doi.org/10.2307/1913236

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies, 19*(3), 797–827. http://dx.doi.org/10.1093/rfs/hhj020

Grewal, M. S., & Andrews, A. P. (2008). *Kalman filtering, theory and practice using MATLAB* (3rd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/9780470377819

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, B*(41), 190–195.

Harris, R. I. D. (1995). *Using cointegration analysis in econometric modeling*. London, England: Prentice Hall.

Jaeger, R. A. (2003). *All about hedge funds: The easy way to get started*. New York: McGraw-Hill Companies, Inc.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control, 12*, 231–254. http://dx.doi.org/10.1016/0165-1889(88)90041-3

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods* (2nd ed.). Boston, Massachusetts: PWS-KENT Publishing Company.

Liew, V. K. S. (2004). Which lag length selection criteria should we employ? *Economics Bulletin, 3*(33), 1–9.

Lin, Y. X., McCrae, M., & Gulati, C. (2006). Loss protection in Pairs trading through minimum profit bounds: A cointegration approach. *Journal of Applied Mathematics and Decision Sciences*, 1–14. http://dx.doi.org/10.1155/JAMDS/2006/73803

Lo, A. W. (2010). *Hedge funds: An analytic perspective*. Princeton, New Jersey: Princeton University Press.

Miao, G. J. (2007). *Signal processing in digital communications*. Boston, Massachusetts: Artech House, Inc.

Miao, G. J., & Clements, M. A. (2002). *Digital signal processing and statistical classification*. Boston, Massachusetts: Artech House, Inc.

Ng, S., & Perron, P. (1995). Unit root tests in ARMA Models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association, 9*(429), 268–281. http://dx.doi.org/10.2307/2291151

Perlin, M. S. (2009). Evaluation of Pairs-trading strategy at the Brazilian financial market. *Journal of Derivatives & Hedge Funds, 15*(2), 122–136. http://dx.doi.org/10.1057/jdhf.2009.4

Pole, A. (2007). *Statistical arbitrage: Algorithmic trading insights and techniques*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Schwarz, G. (1978). Estimating the dimension of a Model. *The Annals of Statistics, 6*(2), 461–464. http://dx.doi.org/10.1214/aos/1176344136

Sharpe, W. F. (1975). Adjusting for risk in portfolio performance measurement. *Journal of Portfolio Management*, 29–34. http://dx.doi.org/10.3905/jpm.1975.408513

Sharpe, W. F. (1994). The sharpe ratio. *Journal of Portfolio Management, 21*(1), 49–58. http://dx.doi.org/10.3905/jpm.1994.409501

U.S. Treasury Bill Rate. (2013). *3-month historical data* (January 2012–December 2013). Retrieved from http://ycharts.com/indicators/3_month_t_bill

Vidyamurthy, G. (2004). *Pairs trading: Quantitative methods and analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Wang, G. H. K., & Yau, J. (1994). A time series approach to testing for market linkage: Unit root and cointegration tests. *Journal of Futures Markets, 14*(4), 457–474. http://dx.doi.org/10.1002/fut.3990140407

Wilson, R. C. (2010). *The hedge fund book: A training manual for professionals and capital-raising executives*. Hoboken, New Jersey: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/9781118266465

Copyrights