ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 10, Issue 7, July 2023

Pairs Trading Management Problem: Linear Regression vs. Neural Network Frameworks

R. Sivasamy

Statistics Department, University of Botswana
sivasamyr@ub.ac.bw

Abstract— This paper chooses two correlated (Y(t), X(t)) stock prices observed at t = 1, 2..., N and develops algorithms to locate best profitable positions for future trading. We divide the past N1 pairs as training set, say TR and the remaining (N-N1) pairs as future series, say TE, to be used for testing tasks and fixing a trading strategy. Assuming the response variables as Y=Y(t) and the predictor as X=X(t) in the paired dataset TR. We fit a simple linear model (LM) of the response variable Y of TR with the predictor X using the statistical command "lm()" and then a non-linear model (NLM) using the command "neuralnet()' of the neural net package in R. The stationary property of the residuals is then checked by the "Augmented Dickey-Fuller" (ADF) test. We then develop a trading strategy to maximize profits over trading period of TE and thus a positional chart to show profitable positions using co-integrated stationary spread obtained from TE based on the estimated results by both the LM and NLM cases. Only real data sets are used for illustration and optimal performance is determined. Finally, the total profit of pairs trading is calculated on a case-by-case basis and compared. Finally, the NLM is shown to perform better than the LM approach.

Index Terms— correlated stocks, linear model, non-linear model, trading strategy, total profit

I. INTRODUCTION

The success of pair trading depends on hash time series modeling and forecasting. Being able to predict the "direction" of this spread is key. Several methodologies have been used from an econometric perspective to model mean reversion (stationarity, cointegration, etc.).

The framework for the discussed cointegration concept is as follows: "Choose two cointegrated stock price series, then open a long/short position when the stock deviates from its long-term equilibrium, and finally close the position after convergence or at the end of the trading period. Some of the motivating factors for this article include:

- Trades are initiated when stock price deviates from balance or equilibrium state.
- Pairs trading is a popular dollar-neutral trading strategy.
- Selection of pairs trading following the criterion of stationarity leads to poor performance, but cointegration gives high, stable, and robust returns.

Any two financial stock prices "X and Y" behave unpredictably and neither form a stationary process. Each series can be tested for non-stationarity by unit root verification using the popular ADF test.. Testing of cointegration or stationarity was developed by a seminal paper due to Engle & Granger [1] in two steps:

- The first step verifies the individual series are indeed integrated of order *I(1)* (being non-stationary). Then we regress one on the other using ordinary least square (OLS) and check if the residual series is integrated of order *I(0)* (suggesting stationarity).
- *Secondly if the residuals are stationary, then we can extract the coefficients from the OLS model and use those to form a stationary linear combination of the two-time series.*

A. Stationarity and Cointegration

Although correlation and co-integration are theoretically similar, they are not the same. A management method that can increase the probability of winning is "pair(s) trading with integration". A stationary series is a series whose changes are not functions of time. We can create a portfolio of two price series so that their linear combination is fixed, i.e., stationery. If the linear combination of two price series is stationary, then the individual price series are said to be cointegrated. Mathematical and statistical modelling aspects of pair trading and cointegration are discussed by Vidyamurthy [2], Lin et al. [3], Chiu and Wong [4] and Galenko, Popova and Popova, [5].

The stable relationships implied by economic theories are referred to as long-run equilibrium or stationary relationships. According to Elliott et al. [6], the spread between two financial assets can be modelled as a mean reversion process.

B. Problem Statement and Methodology

This paper considers a pairs trading problem on two correlated (Y(t), X(t)) stock prices observed at t = 1, 2..., N. The main objective is to obtain a best trading strategy for maximizing the total profit based on a co-integrated relationship between the two observed datasets on X(t) and Y(t).

We divide the past N_1 pairs as training set, say TR and the remaining $(N-N_1)$ pairs as future series, say TE, to be used for testing tasks and fixing a trading strategy. Assume that the response variable is Y= log(Y(t)) and the predictor is X=log(X(t)) which belong to the paired dataset TR.

1) *Linear regression of Y on X*

As a first step, we fit a simple linear regression of Y on X model (LM) given in (1):

$$Y = \alpha + \beta X + e; \quad e \sim N(0, \sigma 2) \qquad (1)$$

After the execution of the statistical command "lm(Y ~X, data=cbind(Y,X))" in R, we can extract the estimate of 'α' as 'a' and 'β' as 'b' that are involved in (1). Thus, the estimate of the observed response variable Y=Y_o is given by Y_e = a + b X. Thus, the spread measure is given by,

$$Spread = Yo - Ye \qquad (2)$$

In addition, a non-linear model (NLM) is also fitted using the command "neuralnet()' of the neural net package in R. The stationary property of the residuals (or the spread) of each model is then checked by the "Augmented Dickey-Fuller" (ADF) test.

Second, we develop a trading strategy to maximize profit during the TE trading period. Therefore, the position chart is designed to show profitable positions using the co-integrated spread process obtained by TE based on the estimated results of both the LM and NLM cases.

2) *Positional Chart for Trading*

Let the mean and standard deviation of the underlying spread series, say R=r(t), be denoted by "m" and "s" respectively. The focus is on the "sell, buy or no trade" options of a trader, which requires (i) sell stock 1 for $1 and buy stock 2 for $1 if R just passes 'm + k s' (i.e., r > m+ k s) is called an "open position". The value of 'k' is arbitrary but positive. Wait for the string "R" to fall and cross only "m", at that point close the position and buy stock 1 at $1 and sell stock 2 at $1, (ii) sell stock 2 at $1 and buy stock 1 at $1, when 'r' exceeds only r > (m - k s), is called the "open position". Wait for string "r" to rise and cross "m", at that point close the position and buy stock 2 at $1 and sell stock 1 at $1 and (iii) do not trade another time.

The upper threshold level (UTL), the middle Level ML) and the lower threshold level (LTL) for the positional chart are defined by:

$$UTL = m + k\,s, ML = m,$$
$$and\ LTL = m - ks \qquad (3)$$

Finally, the total profit of pairs trading is calculated on a case-by-case basis and compared. As expected, the NLM is shown to perform better than the LM approach.

3) *Selection of securities*

Every investing management must choose a few securities of interest and then look for potential combinations among the chosen securities, as stated in [7], and finally the management can decide a proper pair of equities for pairs trading based on co-integration.

Sivasamy and Omolo [8] constructed a positionl cahart for a trader to know profitable time points for buying and selling activities using two correlated moving average processes.

For a specific portfolio of Chinese equities, Xiang, and He [9] explained ideas on how the factors chosen to explain the cross-section of stock returns based on pairs trading method. Han et al. [10] created a long-short portfolio for pairs trading using an unsupervised learning that was applied to the US stock market from January 1980 to December 2020. Additionally, Sabino da Silva et al. [11] provided an alternative pairs trading method for calculating a mispricing index in a novel manner using an optimal linear combination of copulas, by looking at the daily returns of the S&P 500 stocks between 1990 and 2015.

The following sections are organized as follows: Section II explains how four stock prices are randomly collected and how a pair of stocks are selected to develop a trading strategy using a position chart. In addition, we process numerical representations to support theoretical tools to facilitate their use. Section III presents the formal conclusion.

II. DATA COLLECTION AND ANALYSIS

For our illustration, we choose four securities, say Y1, Y2, Y3, and Y4 and collect prices from the Yahoo website using the following R programming commands.

```
Y1=getSymbols("WFC",start="2023-01-01",
                    auto.assign=F)
Y2= getSymbols("XOM",start="2023-01-01",
                    auto.assign=F)
Y3=getSymbols("GE",start="2023-01-01",
                    auto.assign=F)
Y4=getSymbols("JNJ",start="2023-01-01",
                    auto.assign=F)
```

It should be noted that we have stored the prices in the Y1 matrix with the column names "WFC.Open, WFC.High, WFC.Low, WFC.Close, WFC.Volume, WFC.Adjusted". A similar registration was made with other matrices Y2, Y3 and Y4, each row length is N=4152 and column length is 6.

The proposed analysis considers only the Closing prices as Z_1= WFC.Close, Z_2= XOM.Close, Z_3=GE.Close, and Z_4= JNJ.Close. The correlation matrix of the vector V=(Z_1, Z_2, Z_3, Z_4) is reported in Table 1:

Table 1: Correlation co-efficients

	z1	z2	z3	z4
z1	1.00	0.35	0.14	0.56
z2	0.35	1.00	0.45	-0.09
z3	0.14	0.45	1.00	-0.46
z4	0.56	-0.09	-0.46	1.00

We divide the vector V into two disjoint sub-vectors V_1 containing the past N=4000 items and V_2 which includes the recent prices of N_2=N-N_1=152 items. The graphical representation of the vector V_1 is drawn in Figure 1.

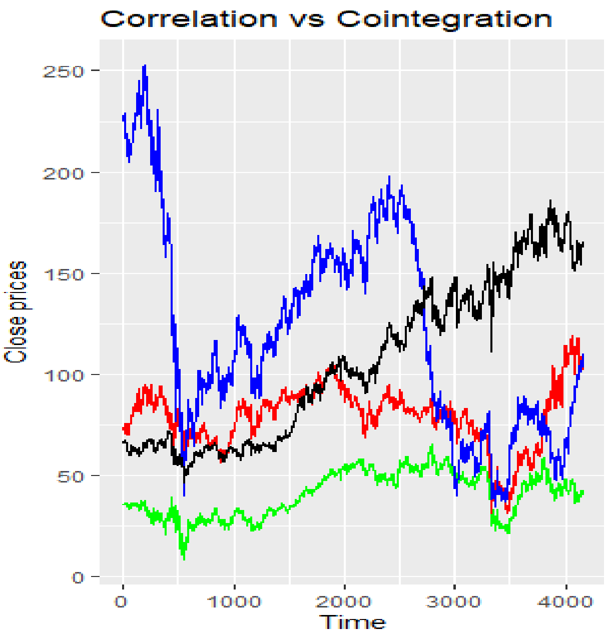


Figure 1: Graph showing the changes of the four types of closing prices Z_1(green), Z_2(red), Z_3(blue) and Z_4(black) observed for 4000 days.

A simple examination of the correlation matrix in Table 1 allows us to notice that the variables Z1 and Z2 are positively correlated. Further inspection of the graphs shown in Figure 1 confirms that the variables Z1 and Z2 get co-integrated since the two curves move in the same direction. So, we consider the pair (Z_1, Z_2) for further pairs trading investigations.

Figure 2 presents the behavior of pairs (Z1, Z2) as days change from 1 to 4000 days (left panel) and from 4001 to 4152 days (right panel).

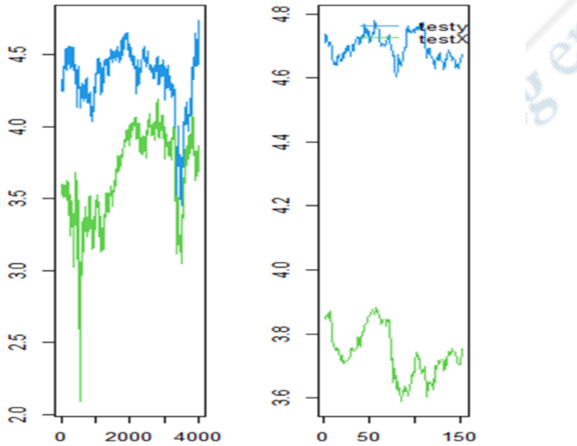


Figure 2: Graphs of Z_1 and Z_2 for TR (left) and TE (right) sets.

A. LM fitting and its predictions

Further analysis is carried out in two stages. The first stage fits the linear regression of Z_2 on Z_1 based on those 4000 data points to obtain the values of the intercept "a" and the slope co-efficient "b". Thus, the outcomes from the execution of the commands are summarized:

```
data <- ts.union(z2, z1)
z2.eq <- lm(z2 ~ z1, data = data)
summary(z2.eq)
"z2.eq= 3.384103 + 0.261094 z1"
Spread= z2-z2.eq
```

Let the mean metric of the spread series be 'm' (mist be zero) and its standard deviation be "s". Thus, for any arbitrary value assigned to k>0, one can compute the UTL, ML and the LTL values that were stated by (3).

We consider the dataset on ts.union(z_2, z_1) belonging to TE, each of which has the recent observation of length 152. The corresponding prediction series on spread, say spreadP, is then calculated:

z_2.P= 3.384103 + 0.261094 z_1
SpreadP= z_2-z_2.P

Figure 3 shows a graph showing the variation of SpreadP during these 152 days and profitable positions depending on the UTL, ML, and LTL lines. R code is then used to calculate the total profit and the details are shown in Table 1.

The total profit from the LM source is $1.008 per dollar invested.

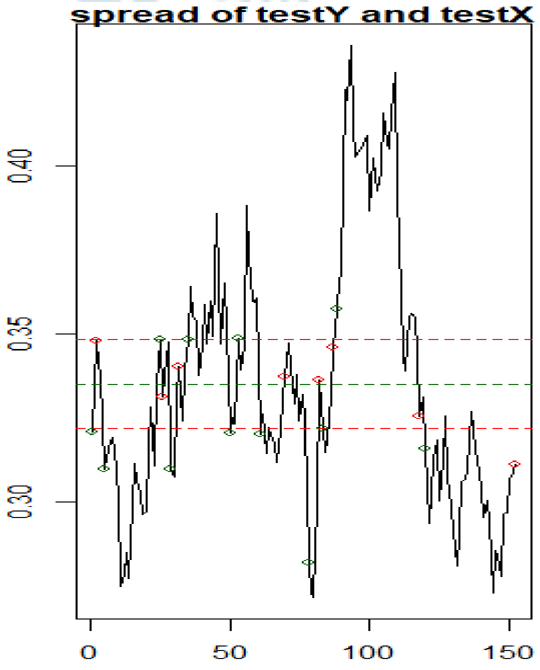


Figure 3: Position chart showing 152 days of data points and profitable positions to buy and sell stocks when k = 0.35

If you read Table 2 and inspect Figure 3 simultaneously, you can see 12 profitable positions with a total profit of $1.008, so the winning amount is $0.008 if you invest $1.

B. NLM fitting and its prediction.

The goal here is to check if the neural network (NN) method can provide high performance compared to the linear models discussed in the previous section. Using the same data sets to train the network with 4000 values and test the network, the NN model was fitted and thus predicted 152 values as before. The corresponding diagram is in Figure 4.

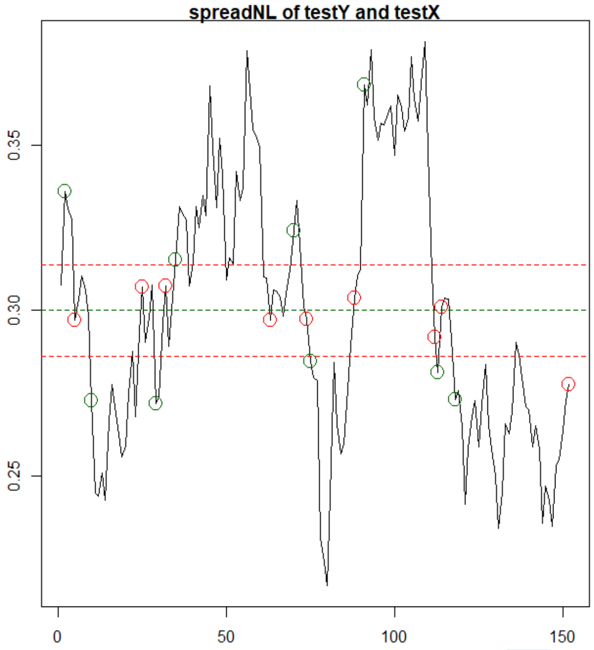


Figure 4. Position plot showing nine feasible positions based on NN where NLM fits to 152 data points for k = 0.35

The total profit is then calculated by NLM and adjusted by the NN method: the profitable positions, related profit, and average profit for both LM and NLM sources are shown in Table 2.

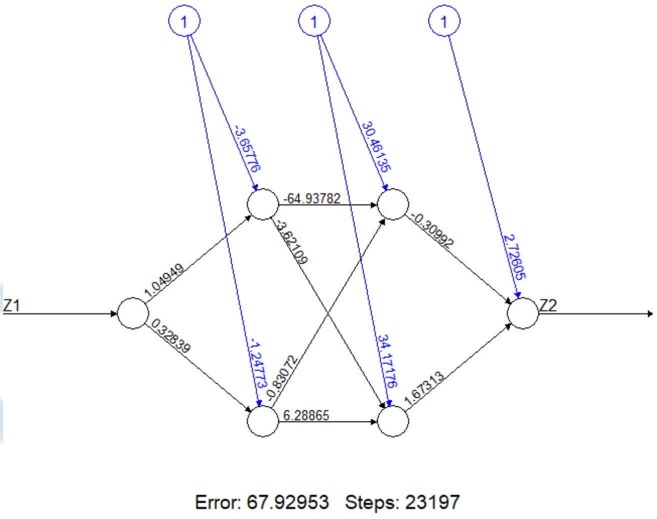
Table 2: Profitable positions and the expected profit(s)					
LM source			NLM source		
No.	Positions	Profit per $	No	Pos	Profit
1	1,2	1.011	1	2,5	0.983
2	5,25	1.086	2	10,25	1.054
3	25,26	1.009	3	29,32	1.004
4	29,32	1.004	4	35,63	0.949
5	35,50	0.940	5	69,74	1.056
6	50,53	1.005	6	75,88	0.943
7	53,61	0.992	7	91,112	1.052
8	61,70	1.015	8	113,115	1.014
9	78,82	0.969	9	118,152	1.063
10	83,87	0.988		**Mean**	**1.013**
11	88,118	1.021			
12	120,152	1.052			
	Average	**1.008**			

Average of NLM source is $ 1.013 i.e., profit is $0.013 per dollar invested which is higher than the profit earned by LM approach.

The following command is used to fit the non-linear model using the training data sets:

```
nnNL<- neuralnet(Z2~ Z1,
            data = cbind(Z2,Z1), hidden = c(2,2))

plot(nnNL)
```



The prediction task is computed using the command: Z2.E=compute (nnNL, Z1). Thus, the non-linear spread is (Z2-Z2.E).

III. CONCLUSION

Statistical aspects of both correlated and integrated stock price pairs are applied in the secondary data on the closing prices of two stocks Z_1= WFC.Close, Z_2= XOM.Close, that were collected over 4152 days. A training sample size of the past 4000 days was used to fit the LM, i.e.,

z_2.eq= 3.384103 + 0.261094 z_1

The corresponding spread is obtained, which is used to visualize the position graph in Figure 3. This LM approach allows 12 profitable trades out of a total of 152 days. This results in a total profit of $1.008. A similar experiment was performed by fitting NLM using past data sets of size 4000 and evaluating the spread for Z1 and Z2 closing rates for the last 152 days under the NN method. During these 152 days, profitable positions were also found in 9 days with a total profit of $1.013 in the analysis of the report.

A simple comparison of Table 1 results with such total profits calculated by LM and NLM shows that NLM produces higher profits than the LM approach.

REFERENCES

[1] R. F. Engle and C. W. J. Granger, "Co-integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, pp. 55, 251-276, 1987.

[2] G. Vidyamurthy, Pairs Trading, Quantitative Methods and Analysis, Hoboken, NJ: John Wiley, 2004.

[3] Y.-X. Lin, M. Michael and G. Chandra, "Loss protection in pairs trading through minimum profit bounds: a cointegration approach," *Journal of Applied Mathematics and Decision Sciences,* vol. 1, no. 1, pp. 1-14, 2006.

[4] M. C. Chiu and H. Y. Wong, "Mean–variance portfolio selection of cointegrated assets," *Journal of Economic Dynamics and Control,* vol. 35, p. 1369–85, 2011.

[5] A. Galenko, E. Popova and I. Popova, "Trading in the presence of cointegration," *Journal of Alternative Investments,* vol. 15, pp. 85-97, 2012.

[6] R. Elliott, J. Van Der Hoek and W. Malcolm, " Pairs Trading," *Quantitative Finance* , vol. 5, pp. 271-276, 2005.

[7] J. Caldeira and G. Moura, "Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy," *Rev. Bras. Financ¸ as (Online), Rio de Janeiro,* vol. 11, no. 1, pp. 49-80, 2013.

[8] R. Sivasamy and B. Omolo, "Inteligent Computing and Mathematical Modelling," *Easy Chair, online, June 20-21,* pp. 1-6, 2021.

[9] Y. Xiang and J. He, "Pairs trading and asset pricing," *Pacific-Basin Finance Journal* , vol. 72, pp. 1-20, 2022.

[10] C. Han, Z. He and A. Toh, "Pairs trading via unsupervised learning," *European Journal of Operational Research,* vol. 307, p. 929–947, 2023.

[11] F. Sabino da Silva, F. Ziegelmann and J. Caldeira, "A pairs trading strategy based on mixed copulas," *Quarterly Review of Economics and Finance,* vol. 87, p. 16–34, 2023.