Pair Trading Models

Data Analytics Project

1st Hritik Shanbhag
PES University
Bangalore, India
Hritik.24shanbhag@gmail.com

2nd Varun Seshu
PES University
Bangalore, India
seshuvarun@gmail.com

3rd Manas V Shetty
PES University
Bangalore, India
manasvshetty@gmail.com

4th Shashwath S Kumar
PES University
Bangalore, India
shashwath457@gmail.com

***Abstract*—This project endeavours to simulate pair trading for correlated and co-integrated pairs and evaluate various machine learning models on time series stock data and see which model best predicts future trades on a stationary pair of stocks.**

***Index Terms*—Pair Trading, Data Science, Machine learning**

I. Introduction

Pairs trading is simple: find a pair of stocks which track each other, and when they diverge, buy the lower one and short the high. If they converge again, you pocket the difference.

The simplicity of the idea of pairs trading is probably the reason for it being so popular among the Wall Street's Experts .It was explored in the 1980s by a group of applied mathematicians and scientists led by Nunzio Tartaglia.Many high profile quantitative firms, such as D.E. Shaw and Long Term Capital have also tried this strategy with small tweaks in the algorithm.

Pairs Trading can be called a Mean Reversion Strategy where we one bets that the prices will eventually revert to their historical trends.The most basic or the first step is to find pairs that are highly cointegrated or correlation.The best way to do this is to start with stocks that might be cointegrated(probably based on some common properties maybe?) and perform a statistical test. Usually the businesses that are in the same industry or sub-sector show good correlation, but this is not always true.

After finding a pair we can expect the ratio or difference in prices (also called the spread) of these two pairs to remain constant with time. However,sometimes there might be a divergence in the spread between these two pairs caused by temporary supply or demand changes, large buy or sell orders for one stock, reaction for important news about one of the companies etc. In this scenario,if one stock moves up while the other moves down relative to each other.If you expect this divergence to revert back to normal with time, you can make a pairs trade.

When we observe a temporary divergence,the basic idea is to sell the outperforming stock i.e the stock that went high and to buy the underperforming stock-the stock that went down and assume that the spread between the two stocks would eventually converge by either the outperforming stock moving back down or underperforming stock moving back up or both — whatever the scenario we will make the money. However if both the stocks move up or move down together without changing the spread between them, no money is lost or gained.Therefore pairs trading is a market neutral trading strategy enabling traders to profit from any of the market conditions: uptrend, downtrend, or sideways movement.

In this project we explore and attempt to answer these questions using data collected over the past 20 years(240 months).We scraped data from Yahoo Finance, which provided us with open and close prices at each day and other meta data regarding the specific company.We first preprocess our data and find 4 stock pairs and the periods in which they are highly correlated and co-integrated. We then build our model after deciding the time period for prediction. We use 3 models per stock, namely -Linear Regression, ARIMA and LSTM to get predictions for our chosen stocks. We then go about choosing the best model by calculating our returns and compare the real data and the predicted data.

II. Literature in this field

There have been numerous ways of approaching pair trading strategy: In a past study, Huang et al. [1] used moving averages and Bollinger bands [2] to develop a trading system. Genetic Algorithms were used to predict the sales by encoding the moving averages and Bollinger bands in binary format as a chromosome.

The final output is a set of model parameters (optimized by the GA) that prescribes the pairs-trading and timing models. The limitation to this model is that genetic algorithms do not scale well with complexity. Finding the optimal solution often requires very expensive fitness function evaluations. Since we are computing these predictions for a very large dataset of sales, GA is not very feasible for our model.

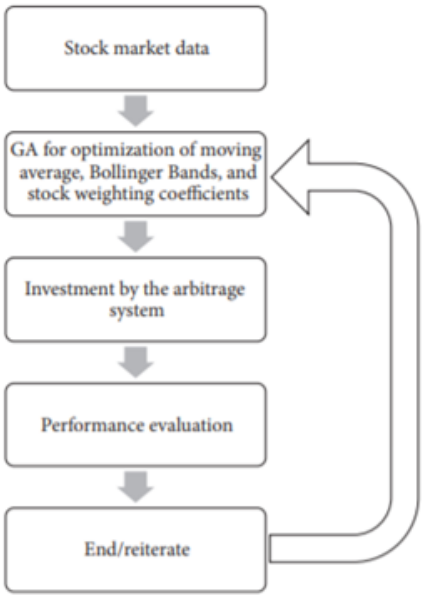

Fig. 1. Example of a figure caption.

Another study, jiayu wu et al. [3] used a spread model and a modified version of pair trading model with the following equation for Spread:

$$\frac{dA_t}{A_t} = \alpha dt + \beta \frac{dB_t}{B_t} + dX_t$$

Fig. 2. Example of a figure caption.

Here there are two securities A and B. The above equation shows both the securities are linearly related.Here Beta here helps solve the normalization problem since the two stock prices that are (A and B) may not fluctuate in the same range.Also Beta may change over time because of change of some intrinsic characteristics of either company or change of overall stock market regime.

In the modified version of pair trading model, We have selected 3 good technical indicators:-

i) Simple moving average

ii) weighted moving average

iii) Relative strength index

The performance of the above models can be improved if we design an update rule to update Beta for every n-period of time.

III. Project Workflow

This is a broad outline of the steps taken in this project.

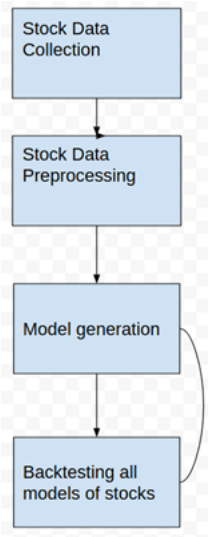

Fig. 3. Example of a figure caption.

In Stock data collection, we collect upto last 20 years of Candlestick data for all stocks actively trading in the stock market.

For stock data preprocessing, we first clean the data we have and then we generate stock pairs. We filter to only include pairs that fit our criteria of correlation and cointegration. After we have correlated and cointegrated pairs, we choose any 4 best pairs and generate orders using our pair trading algorithm (mentioned later in the report) over the period where they are correlated and cointegrated.

We then generate 3 models (Linear Regression, ARIMA, LSTM) for each of the 8 chosen stocks (4 pairs and 24 models in total) to predict the stock data for the correlated period. We give each of the models 1 year's worth of training data and make them predict values for the period where the stocks are correlated and cointegrated.

We then backtest all the orders produced during preprocessing on both actual data and predicted data for each stock. During backtesting we decide how many shares to buy in each trade and the profit made per trade on actual data and predicted data of the models. We then compare the profits generated by the models and decide which model is the best model for each stock.

IV. Trading Methodology

In trading methodology, we outline the parameters we use to generate trades and perform the experiment.

Assumptions based on pair trading:

1) Pair trading is a mean reverting strategy, where we make a bet expecting the spread between the pair

of stocks to revert back to the mean spread from its current position. Thus, when the spread increases, we bet that it decreases back and reverts to the mean. Vice Versa when the Spread decreases.

2) Thus we take a Short position when the Spread increases beyond a threshold and a Long position when the Spread decreases beyond a threshold.

Criteria to place orders and trade:

1) Capital: The Capital we are using to perform the trades per stock is Rs. 10,00,000.

2) Risk:We are allowing a maximum risk of 2% of the total capital per trade, i.e, Rs. 20000.

3) Finding pairs: For each sector in the market, we create pairs of stocks and find 90 day periods where they are correlated and cointegrated. We choose stock pairs with correlation coefficient r $>$ 0.85 and with p-value $<$ 0.05 in the cointegration test to see if the correlation calculated is statistically significant.We then choose the 4 best stock pairs according to correlation and cointegration value for the rest of the experiment.

4) Spread: We then calculate the spread (Refer to Appendix) for the chosen stock pairs.

5) z-score: We calculate the z-score of the spreads, using which we generate orders to place based on thresholds below.

6) Open position: We open a Short position on the pair if the z-score of the spread $>$ 1.5, i.e, Short position on stock 1 and Long position on stock 2.We open a Long position on the pair if the z-score of the spread $<$ -1.5, i.e, Long position on stock 1 and Short position on stock 2.

7) Close Position: We close a position on the basis of 2 criteria.

8) Take Profit: We consolidate our position in order to take the profits.

9) Stop Loss: We consolidate our position in order to cut losses in our trade.

10) For a Long position: Take profit criteria - z-score $\geq$ 0. Stop Loss criteria - z-score $\geq$ 3.

11) For a Short position: Take profit criteria - z-score $\leq$ 0. Stop Loss criteria - z-score $\leq$ -3.

V. Data Collection

We have scraped Indian stock market daily candlestick data for a period of 20 years on all actively trading stocks from the Yahoo Finance API.

We have collected data for 4300 stocks, each with Columns:

1) Date - Date in format yyyy-mm-dd
2) High - High price of the day
3) Low - Low price of the day
4) Open - Open price of the day
5) Close - Close price of the day
6) Adjusted Close - Close price adjusted for stock splits
7) Volume - Number of trades of the stock in a day

We also have a Sector-wise stock ticker csv file which helps us to find pairs of companies within the same sector.

Lastly, we have 2 csvs, NSETICKERS.csv and BSETICKERS.csv, which consists of all actively traded stocks in the Stock exchanges. We use this data to automate collection of stocks historical data using the Yahoo Finance API.

VI. Data Preprocessing

1. Cleaning Data

We are dropping rows of the datasets which have missing data. We are not Interpolating the missing data as filling in the data for missing days may lead to discrepancies and might affect the mean reverting nature of larger stocks and increase volatility.

We then delete datasets which have data for less than 3 years as we would not have significant correlations and very less data would be available for training and testing models. We then crop the remaining datasets to be between the years 2017-2020 as to reduce compute time for the models and have enough data to find significant correlations. We have chosen not to take 2020 stock data as due to the pandemic, the volatility of the stocks have increased and this may lead to skewed overall results.

We then added the company names and Exchange as columns to the dataset for easier identification later.

We then group the stock tickers by their sectors to make it easier to identify correlated and co-integrated pairs.

2. Generate pairs

For Identifying correlated and cointegrated pairs, we follow the algorithm:

For each sector: For each stock in sector: For each 90 day window in stock data: Calculate correlation on close

price Calculate p-value of cointegration test on close price If correlation ¿ 0.85 and p-value ¡ 0.05 Record pair data and time period Store all recorded pair data in csvs.

3. Calculate spread and z-score

For the pairs that we have stored, calculate the Spread (Formula in Appendix) and z-score.

4. Generate trade orders

For each pair, we go through the z-score of the spread calculated and open or close positions depending on the criteria mentioned in the Trading Methodology.

We then label each day as:

1) LONG - Denotes buy the first stock and sell the second.
2) SHORT - Denotes sell the first stock and buy the second.
3) FLAT - Denotes no order to be placed on that day.
4) GET_OUT_OF_POSITION - denotes to cash in on all previous orders on that date and have no outstanding LONG or SHORT positions as of that date.

For the pair BAJAJ-AUTO and HEROMOTOCO:

1. Close prices movement

Fig. 4. Example of a figure caption.

- Blue - Heromotoco
- Orange - Bajaj-Auto

2. Calculated correlation matrix

	HEROMOTOCO_Close	BAJAJ-AUTO_Close
HEROMOTOCO_Close	1.000000	0.899256
BAJAJ-AUTO_Close	0.899256	1.000000

Fig. 5. Example of a figure caption.

3. P - value

Calculated p-value = 0.024984061040578365

4. Z-score of spread visualization

Fig. 6. Example of a figure caption.

- Blue - z-score of spread
- Black line - mean
- Red Lines - +1 and -1 standard deviations above and below the mean

5. Orders generated Visualization

Fig. 7. Example of a figure caption.

- Blue - z-score of spread
- Black horizontal line - mean
- Red Lines - +1 and -1 standard deviations above and below the mean
- Red Lines - Short position on pair
- Green Lines - Long position on pair
- Black Vertical Lines - Close all previous positions on pair

VII. Models

There are many models that can be used for a pair trading system. Given the market timing and pairs-trading models, the performance of a trading system can be enhanced with suitable training data for each model.

We train our model with a sample of one year prior to the pair generated with the test set being the 3 month time period after the training set. We are making predictions on the set which yielded high correlation and low cointegration compared.

We are using 3 different models to test our pair trading strategy:

1. Linear Regression: Linear regression [7] attempts to model the relationship between two variables by fitting a linear equation to observed data.

The model is trained using the date and closing price of the stock on that day. It attempts to predict the closing price on a day based on the date given to the model.

The results we got for one out of our 8 tests in linear regression for the company

Bajaj-auto: Training set being from sept 2017 to sept 2018. Testing set being from sept 2018 to dec 2018

```
Mean Absolute Error: 117.8660153228754
Mean Squared Error: 18859.49259140318
Root Mean Squared Error: 137.3298678052345
R2 Score: -0.6364630068913395
```

Fig. 8. Example of a figure caption.

Visualization of training vs test data for the same:
X-axis -> Date
Y-axis -> Closing price

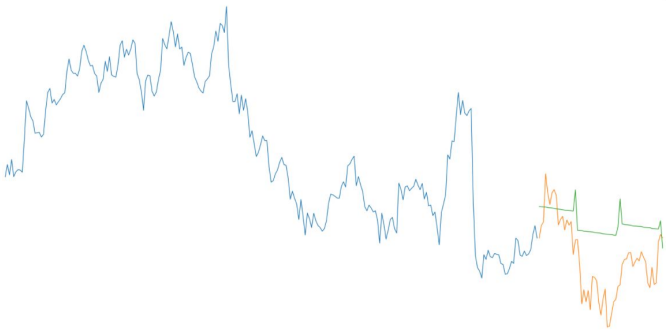

Fig. 9. Example of a figure caption.

- Training data in blue
- Test data in orange
- Predictions in green

There are obvious caveats when performing linear regression on a time-series data. Though it is simple and inexpensive, we are assuming that the data has no autocorrelation and this leads to inconsistency and confidence intervals being unreliable when our assumption proves to be false.

2. ARIMA

ARIMA [8] (autoregressive integrated moving average) is a model that uses time series data to either better understand the data set or to predict future trends. It uses both the concepts of auto-regression and moving averages.

The model is trained using the close price for the period of one year and predicts the close price for the next 3 months.

The results we got for one out of our 8 tests in ARIMA for the company Bajaj-auto:

```
Mean Absolute Error: 121.81543019473166
Mean Squared Error: 22522.468849934037
Root Mean Squared Error: 150.07487747765794
```

Fig. 10. Example of a figure caption.

Visualization of training vs test data for the same:
X-axis -> Date
Y-axis -> Closing price

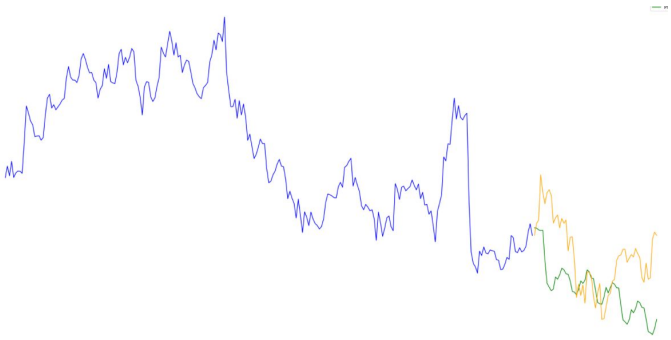

Fig. 11. Example of a figure caption.

- Training data in blue
- Test data in orange
- Predictions in green

3. LSTM

Long Short-Term Memory [8] (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately.

This architecture makes more effective use of model parameters than the others considered, converges quickly, and

outperforms a deep feed forward neural network having an order of magnitude more parameters.

The model is trained using only the close price for the period of one year after downscaling it to the range 0-1. This model is then used to predict the closing price for the next 3 months.

The results we got for one out of our 8 tests in LSTM for the company Bajaj-auto:

```
Mean Absolute Error: 72.42936363998724
Mean Squared Error: 7554.010239179037
Root Mean Squared Error: 86.91380925479585
```

Fig. 12. Example of a figure caption.

Visualization of training vs test data for the same:
X-axis -¿ Date
Y-axis -¿ Closing price

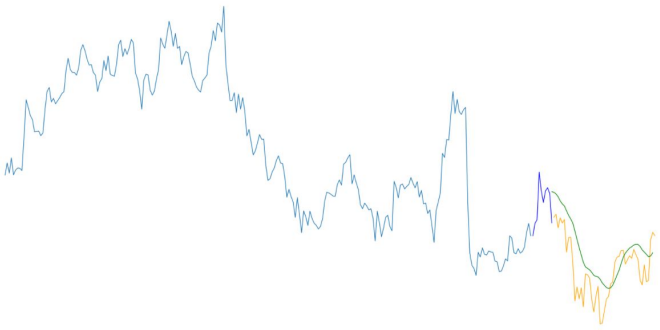

Fig. 13. Example of a figure caption.

- Training data in blue
- Test data in orange
- Predictions in green

Visualization of the predictions by the different models for Bajaj-auto:

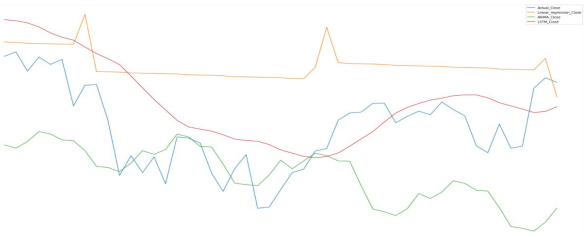

Fig. 14. Example of a figure caption.

- x-axis - Date
- y-axis - Price in rupees
- Blue line - Actual data
- Orange line - Linear Regression
- Red line - LSTM predictions
- Green line - ARIMA predictions

VIII. Backtesting

Backtesting is done to get statistics on our model to gauge the effectiveness of the strategy. It is a key component of effective trading system development. It is accomplished by reconstructing, with historical data, trades that would have occurred in the past using predictions given by a strategy.

We have used 3 different strategies for each stock out of the 8 stocks (Linear Regression, ARIMA, LSTM) and we are backtesting them to measure which strategy works better.

Using capital and risk as mentioned in Trading Methodology above, we evaluate the orders for each of the models and compare it to the actual order.

Evaluating the orders for one of our 8 stocks, Bajaj-auto, with the 5 orders in our test set, we get the profits as:

Actual_profit	Linear_regression_profit	ARIMA_profit	LSTM_profit
510.299316	15.742308	-64.283509	250.8926
0.000000	0.000000	0.000000	0.0000
-652.049316	-140.370906	-57.452599	-648.7824
0.000000	0.000000	0.000000	0.0000
-319.200684	-7.496295	-300.210839	-92.0199

Fig. 15. Example of a figure caption.

Visualizing the profits evaluated for the different models for the same:

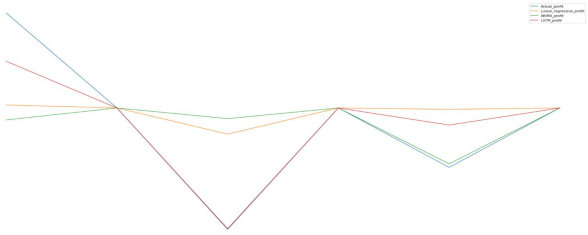

Fig. 16. Example of a figure caption.

- x-axis - Date
- y-axis - Profit
- Blue line - Actual profit
- Orange line - Linear Regression predicted profit
- Green line - ARIMA predicted profit
- Red line - LSTM predicted profit

IX. Conclusion

In this section we examine the performance of our proposed method for pair-trading systems. We used 8 different companies to make 4 different pairs to test our models and find the best model for each company. We used 3 different models for each company totalling 24 models. The best model being the one which predicts the stock price of the day closest to the actual price.

Results found:

Company	Best Model	RMSE
Bajaj-Auto	LSTM	86.913809
Heromotoco	LSTM	150.025906
BEML	LSTM	59.941932
ESCORTS	LSTM	60.090202
TATAMOTORS	LSTM	19.74632
TWL	LSTM	7.82924
ASHOKLEY	Linear Regression	6.878291
TATAMTRDVR	LSTM	11.52505

We observe that on majority LSTM performs the best with very accurate predictions. Refer to the appendix for data on all companies.

References

[1] Huang, Chien-Feng Hsu, Chi-Jen Chen, Chi-Chung Chang, Bao Li, Chen-An. (2015). An Intelligent Model for Pairs Trading Using Genetic Algorithms. Computational intelligence and neuroscience. 2015. 939606. 10.1155/2015/939606.

[2] Murphy, Technical Analysis of Financial Markets, New York Institute of Finance, New York, NY, USA, 1999.

[3] A Pairs Trading Strategy for GOOG/GOOGL Using Machine Learning Jiayu Wu. December 9, 2015

[4] Yuxing Chen, Weiluo Ren, Xiaoxiong Lu. Machine Learning in Pairs Trading Strategies.

[5] Xing Fu, Avinash Patra. Machine Learning in Statistical Arbitrage.

[6] Kumari, Khushbu Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.

[7] Badra, Niveen Sabry, M Abdel-Latif, Hatem. (2007). Comparison Between Regression and Arima Models in Forecasting Traffic Volume. 1. 126-136.

[8] Hasim Sak, Andrew Senior, Francoise Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling

X. Appendix

1. z-score

It shows the distance of a data point from the mean (normalized to 0).

z = (x - mu) / sigma

Where
x is a data point
mu is the mean
sigma is the standard deviation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

2. Correlation

Where r is the correlation coefficient
xi - data point of x variable
yi - data point of y variable
xbar - mean of x variable
ybar - mean of y variable

3. Co-integration test and p value

We use the augmented Engle-Granger two-step cointegration test for finding if the close prices of 2 stocks are co-integrated.

4. Spread

Spread refers to the difference of prices between any 2 stocks in a pair.

Spread = A - B

Where A is the price of stock A
Where B is the closing price of sock B

5.Closing Price

The final price at which a stock trades before the stock market closes for the day.

6. Price Ratio

The ratio between the prices of any 2 stocks.

PR = A / B

Where A is the price of stock A
Where B is the closing price of sock B

7. Long position on a stock

Buying a stock in hopes of it rising in the future.

8. Short position on pair

Selling a stock in hopes of it falling in the future.

9. Flat position on pair

Taking no position on a stock.

10. Capital

The amount of money available to trade.

13. Risk

The maximum amount of money to spend, or the maximum amount of money willing to be lost on a single trade. Usually 2% of capital.

14. Open position

Taking a position on a stock, long or short.

15. Close position

Consolidation of all open positions on a stock to either take profit or stop loss.

16. Take profit

Criteria to be met to leave a profitable position take profit.

17. Stop loss

Criteria to be met to leave an unprofitable position and take loss.