POLITECNICO DI MILANO

Faculty of Industrial and Information Engineering

Master of Science
in Mathematical Engineering (Quantitative Finance)

Integration of a Deep Learning Algorithm and Network Approach to Portfolio Optimization and Reversal Pairs-Trading

Supervisor: Prof. Andrea Flori

Master's dissertation of:
Amir Moradibaad
Student ID number: 900263

Academic Year 2020 - 2021

1. Abstract

Everyday financial institutions, fund managers, brokers and individual investors put millions of dollars in various sectors all over the world. Therefore, a systemized approach is crucial to properly select securities. It comes into a great prominence as it helps to monitor and manage assets by generating profit in market and minimize losses. Consequently, introducing an efficient and feasible approach is vital to asset managers, which is qualified to accurately bring effectual investment strategy for financial needs and to implement appropriate regulatory policies. Currently, there are numerous researches and approaches available, from old financial-statistical security simulations to new combined approaches developed by computer based techniques like machine learning. However, each of these approaches has some limitations like uncertainties in input data and incomplete strategies to consider set of securities within a financial market that could cause less profitable portfolios.

In this thesis, I integrated a pairs-trading strategies coupling Deep Learning algorithm approach to better foresee securities behavior in financial market under investigation. I assumed that these integrations allow us to deal with uncertainties in input data, beside capturing intercorrelation amongst securities in the specified financial context to detect and predict each stock's peer. This accommodates the possibility to better discover peers in mentioned financial market and provide a baseline to find foremost security selection criteria to maximize portfolio return alongside minimizing loss with aims of pairs-trading strategy.

2. Abstract (Italian)

Istituzioni finanziarie, singoli investitori, gestori di fondi e broker investono milioni in vari settori in tutto il mondo. Pertanto, un approccio sistematico diventa fondamentale per selezionare i titoli. Questo acquista grande risalto poiché aiuta a monitorare e gestire le risorse, generando profitti e riducendo al minimo le perdite. Di conseguenza, l'introduzione di un approccio efficiente è vitale per i gestori patrimoniali, qualificati per elaborare con precisione una strategia di investimento efficace per le esigenze finanziarie e per implementare appropriate politiche normative. Ad oggi sono disponibili numerosi approci, dalle vecchie simulazioni di financial-statistical security ai nuovi approcci combinati sviluppati da tecniche informatiche come l'apprendimento automatico. Tuttavia, ciascuno di questi approcci presenta alcune limitazioni, ad exemplum possibili incertezze nei dati di input ed eventuali strategie incomplete, con l'obbiettivo di considerare un insieme di titoli, all'interno di un mercato finanziario, che potrebbero portare a portafogli meno redditizi.

In questa tesi, abbiamo integrato due strategie di pairs-trading con approcci di noise clearing e di Deep Learning, al fine di prevedere il comportamento dei titoli nel mercato finanziario in oggetto. Nella nostra ipotesi, queste integrazioni ci permettono di affrontare le possibili incertezze nei dati di input, oltre a catturare l'intercorrelazione tra i titoli nella rete finanziaria considerata, per rilevare e prevedere il peer di ciascun titolo. Questo offre la possibilità di indagare a fondo i peer nel mercato finanziario, e fornire una linea guida per individuare i principali criteri di selezione dei titoli, al fine di massimizzare il rendimento del portafoglio e, contemporaneamente, ridurre al minimo le perdite, con gli obiettivi della strategia pairs-trading.

Table of Contents

List of Figures

List of Tables

2. Introduction

Researchers and major institutional investors at large banks and desks, always were seeking for a market neutral trading strategy to overcome the overall effects of the market movements toward their profiting strategy. Since it is difficult to predict short to medium market movements, it became necessary to take advantage of simple, steady and profitable strategy with low-risk position. This led investors to build statistical arbitrage and convergence trading strategies. Among which, pairs-trading that first introduced by group of quantitative analysts working at Morgan Stanly in 1980s. Ever since, many attempts conducted to prove the profitability of this market neutral strategy, while predictability of future stock prices seemed to be impossible as a consequence of a well-known random walk theory around price predictions. Gatev [52], in a research paper achieved 12 percent return within six month trading period, by training over a large amount of data between 1967 and 1997.

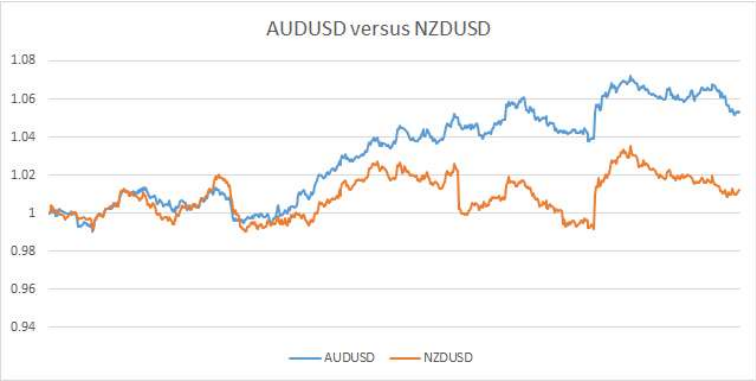


Figure 2.1. Representation of divergence from spread

The main idea behind pairs trading strategy is to seek 2 securities or funds with similar characteristics, where they follow an equilibrium spread, while currently experiencing a price that deviates from their historical equilibrium (Figure 2.1.). Pairs-trading is not known as a risk free strategy, specially there is a risk that two co-moving securities begin to drift apart from

the expected historical spread instead of converging. To overcome this difficulty, it needs an efficient methodology to be developed.

With the aim of rapid advances achieved in recent years in software and hardware technologies and public availability of financial data, stepping toward a more efficient and ambitious model to assess the current available pairs-trading strategies, became feasible. In financial data market that contains many features and includes huge amount of data with their instant characteristics, a fast comprehensive methodology to readily understand their structure is crucial. Fast computer generation systems, especially in pairs-trading approach enabled researchers to develop algorithms and frameworks to uncover and establish the hidden structure laid behind. Nevertheless, Artificial Neural Network and Deep Learning as modern statistical techniques based on computer computational power, constructed a new era in asset management and portfolio construction strategies.

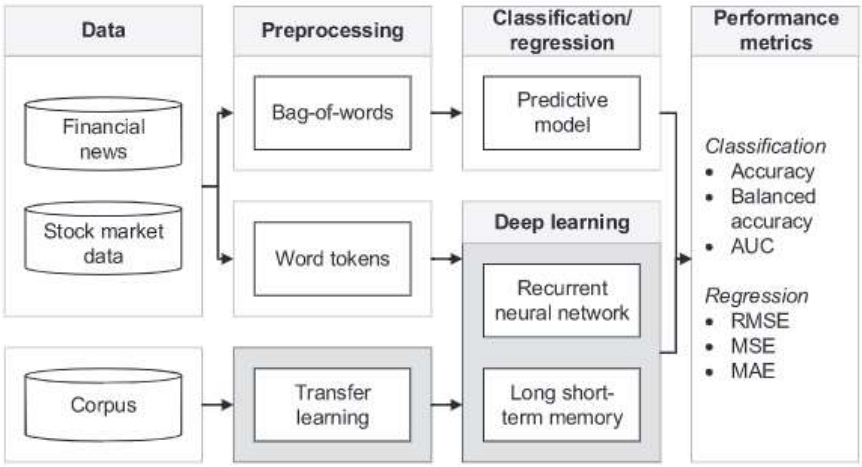


Figure 2.2. Research framework evaluating the performance gains from deep learning

This research aims to integrate the pairs-trading approach coupled with machine learning-based LSTM method to take advantage of its ability to capture hidden financial context within S&P500 stock data. The framework of this methodology splits into two main steps: First detecting possible candidate of co-moving stocks by correlating pre-selection and then cointegration tests, and second, to consider different sorting criteria within each group to detect any possible divergence from group spread to detect underperforming and overperforming

securities. In more details, the goal of this thesis is to investigate the effect of correlation pre-selection in co-moving group construction and also study and compare the performance of LSTM technique with respect to price and return based pairs-trading approaches. Another objective of the thesis is to investigate and study the effect of COVID-19 and after COVID-19 over different pairs-trading strategies.

The rest of thesis organized as follow: Section 3 demonstrates a review of the research and studies done using pairs trading based on different methods starting from Markowitz framework to reversal, CAPM and machine learning based strategies; Section 4 defines the framework and basic concepts of the methodology to be followed in this work by introducing rationality behind pairs-trading and different concepts of correlation and cointegration logics to deep learning methods, and more specifically, LSTM structure and elements to be constructed and designed for my study in the next chapters; in section 5, S&P 500 stocks as my case study has been examined within my methodology with all detailed steps; in section 6, I conclude the thesis in more details.

3. Background

In this section, I am going to briefly review some of the researches have been done in portfolio optimization and asset management approaches focusing on reversal pairs-trading strategy. Starting from description of modern portfolio theory developed by Markowitz, briefly reviewing pairs-trading investment strategies, based on reversal effect, sorting criterion based on CAPM and factor models with specific metrics defined on price and return measures, and finally description of statistical machine learning frameworks and how to combine them with other financial techniques to apply in financial markets. The chapter is organized as follows:

- Modern portfolio theory, its benefits and results are provided
- Basics of reversal effect and Pairs-trading investment strategy in portfolio construction
- CAPM and consequently factor model theories
- Clustering and correlation-cointegration based pairs-trading approach
- Machine learning approach for portfolio optimization coupling with pairs-trading strategy

I will briefly describe main concepts and nominal works at each point and will go into more accurate and deeper details in “Methodology” section to be used for building my theory.

3.1. Modern Portfolio Theory

Modern portfolio theory foundation is widely known to be built on Markowitz seminal work [1]. Where in Markowitz framework investors endeavor to achieve maximized future returns. However, the hypothesis of maximized anticipated return is not sufficient and efficient condition to consider, since it implies that investors will pull all his fund in the security with most expected return, disregarding the portfolio risk and market dynamic. However it is not straightforward how to choose the best portfolio considering that the dynamic of market is

variating fast enough that could turn an anti-correlated set of stocks to correlated ones. Instead, Markowitz discussed also variance, proposing a tradeoff between maximizing expected return and minimizing variance.

Performance of Markowitz framework for out-of-sample cases portfolio is not reliable. First reason of this nonpromising behavior on out-of-sample portfolios, is related to poor performance on estimating expected returns [2] as time series of realized stock returns data is erroneous. Even if expected return is merely constant for all time, it would need a huge historical data to calculate adequate estimation. Second reason of poor performance of Markowitz framework for out-of-sample portfolios is error of estimation on covariance matrices. It was shown by a simulation approach using the Monte Carlo experiment [3], that because of large sampling error, the portfolio selected by applying Markowitz framework is not more efficient than an equally weighted portfolio. Some techniques being developed to overcome these limitations by researchers [4].

Some studies also designed to enclose this hurdle of erroneous estimations by investigating on conditions under which mean variance portfolio expected to exhibit reliable results even in presence of estimation risk [5] by comparing different portfolio models with equally weighted portfolio, suggesting complementing methods to traditional Bayesian statistics by exploiting empirical regularities would lead to promising direction to pursue.

Another challenge arise from Markowitz theory is related to computational difficulties associated with solving a large-scale quadratic programming problem with a dense covariance matrix in large-scale portfolio. Development in computer science and it's computational power facilitated scientist for analyzing more complex portfolios with bigger and intricated combination of financial instruments. Konno[6] demonstrated that L1 risk(mean absolute deviation risk) model, which is a special case of piecewise linear risk model, can remove the above mentioned difficulties. He considered optimization problem consisting of more than 1000 stocks, implementing L1 risk model lead to a linear program instead of a quadratic one while the result is quite similar to that of Markowitz's model.

Capital Asset Pricing Model (CAPM), developed by Sharpe [7] over Markowitz model is a well-known and essential model in asset management world. CAPM evaluate the tradeoff between asset risk and its return by measuring covariance of an individual asset with respect to the whole market [8]. CAPM is a methodology for translating risk into estimated expected

return. Moreover, This model considers expected return for each asset to be calculated as linear relation between risk free rate, beta value of the asset and Expected return of market as below formula:

$$E(r_i) = R_f + \beta_i \left(E(r_m - R_f)\right) \qquad (3.3.1)$$

we can interpret beta as a scale for sensitivity for each asset's return with respect to variations in the market return.

Although Value at Risk (VaR) as a statistical measure that quantifies potential financial losses, is not straight to be applied in portfolio optimization and construction. Since it lacks convexity and subadditivity properties, makes it difficult to optimize when calculating from scenarios. New approaches for huge portfolios introduced [9] for calculating simultaneously VaR and CVaR (conditional value at risk) with possibility to bee combined with other analytical techniques. This approach uses linear programming and non-smooth optimization algorithms for collecting the best potential portfolios. In case securities return are skewed or investor's utility function, is more risk averse than the one implied by mean-variance (MV) analysis, problem involves a large number of decision variables, minimax approach [10] can be a feasible solution. In minimax instead of considering variance as measure of risk, it considers to minimize the maximum loss over historical data for a specific level of return. Minimax approach is a linear optimization problem and it's results are comparable to those chosen by mean-variance rule.

All these mentioned researches are based on modern portfolio theory developed by Markowitz in which consider the first and second moment of the returns distribution. Some paucity of this framework are related to high dimensional historical data, out-of-sample examples, erroneous estimation of expected return or covariance matrix and etc. Engaging different approaches and techniques that choose the most efficient portfolio, could be a challenging and vital.

3.2. Reversal effect and Pairs-trading investment strategy

For many years researchers and investors have been examining to explore how historical data can be used for future stock's price prediction. These prediction can be used by practitioners to investigate future returns and select properly their portfolio. On one hand, the random walk theory asserts that future stock prices are memoryless in the sense that future stock prices are independent and identically distributed random walks. With acceptance of random walk behavior the past prices are not useful anymore for future price predictions.

However, random walk theory involves two hypothesis: 1. Successive stock price changes are independent 2. The prices follow some probability distribution. The independence assumption states that the price series is not sufficient to gain reliable prediction of the future prices and to achieve greater expected return. While second hypothesis states that, the process generating the prices should follow a distribution which doesn't need to be specified exactly[11].

However, contrarian strategy also became an academic debate. Buying past losers and selling past winners famed as a beneficial strategy which directly oppose random walk theory and unpredictability of future prices. If stock price overreact or underreact to the public information or personal insights through the market, then profitable trading strategies based on historical prices will become relevant. Furthermore, there was also additional evidence indicating that the profitability of the relative strength are not due to their systematic risk[2]. In [12], Jegaeesh and Titman, investigated the portfolios constituting past losers-winners and the abnormal return of earning announcement were examined. As a result, within this portfolio they realized a significantly higher return in the 36 month following the formation date.

The link between past prices and contrarian strategy could be associated with the liquidity provision effects caused by active institutions in a way that holding poorly performed stock in past may affect their fund and facing bigger outflow compelling to sell them. Another reason could be due to window dressing strategy through which portfolio managers try to improve the appearance by selling stocks with large loss and purchase in well performing stocks to be presented to clients [13]. This non-informational demands for immediacy inject extra pressure to stock prices and consequently affects liquidity. By decreasing liquidity, practitioners face are not able to easily sell or buy sufficient quantities, faster and above all higher transactional

cost. In other words, the largest short term price reversal occurs in high volume – low liquidity stocks[14].

Two other factors that cause past prices to affect contrarian strategy was mentioned by Cheng et al [13] is related to fire sales and institutional reaction to past prices. Fire sales refer to selling stock at low price with high discount which in financial market called when securities are trading extremely below their intrinsic values and causing an excessive price decline with a possible subsequent rebound. While there is a behavioral model based on imperfect investor rationality that cause investors to overestimate precision of their individual information. Consequently will lead to an overreaction that will cause a momentum in security prices. This model states that prices will be reversed smoothly by public signals and rebound to fundamentals.

However a vital question that could be asked is whether stock market prices following the random walks? With positive answer, how could one expect to predict future prices using given historical data? Many researchers debated that, one could consider behavioral study for short term returns as an anomaly to CAPM model, since it cannot be explained completely by CAPM model. While Lo [16] tested random walk hypothesis for weekly stock market returns and showed that stock returns contains predictable signal component. Stambaugh [15], tried to construct proxy variables for both bond and stocks level of prices to demonstrate its direct relation to predicting risk premium. Also, he studied seasonality effect importance for predicting expected return of each asset price level. Lo [16] test was a simple specification based on variance estimators and concluded that random walk model is not convincing enough to explain stochastic behavior of weekly returns mainly for stock with smaller capitalization. Consequently he rejected random walk model for weekly stock returns by his volatility based test.

All of the above mentioned researches and investigations on different models and theories, all of which believe and discuss random walk model are considering two main hypothesis stating that series of stock prices are independent and the prices are following some probability distribution. Although base on random walk model, stock prices are unpredictable, but there is some researches illustrate some kind of short term predictability in future prices. Predictability could be motivated by three reason:

1. Influence of liquidity provision

2. Investors overreaction caused by individual cognitive to personal information

3. Institutional and personal fire sales

Empirical evidences and studies around predictability of short-term return and price momentum, proves the feasibility of price prediction and return approximation based on historical stock price series using various selection criteria developed by researchers [17],[18]. However I rely on price prediction feasibility assumption and its study is out of the scope of this thesis. I have decided to focus on testing different criteria for portfolio selection to optimize and maximize the price and consequently the return.

3.3. CAPM and Factor Models

Academic financial researches are based on the rationality of the market, which means all the investors have homogenous expectations about the expected return and covariance of the different returns. An essential model that built over Markowitz model, is Capital Asset Pricing Model (CAPM), developed by Sharpe [7] which is well known as the birth of asset pricing model. As stated before, Markowitz model is based on the mean-variance model. Suppositionally, rational investor always choose the mean-variance efficient portfolio:

1. At a given level of risk, choose the portfolio with maximum expected return

2. Given expected return, choose the portfolio with minimum risk

Furthermore, Linter [20] added two key assumption to Markowitz model. He augmented assumptions of “complete agreement” and “borrowing-lending at risk-free rate” in which investors agree about the expected return and covariance of the different returns and are allowed to borrow or lend money identically in risk-free rate on equal terms.

In figure 3.3.1 Fama[19], characterize the portfolio opportunity and efficient frontier to clarify the CAPM. Considering the portfolios A and B comprising of only risky assets , with expected return $E(r_i)$ and variance $Var(r_i)$, the mean-variance criteria states that portfolio A is preferred to portfolio B if:

$$E(r_A) > E(r_B) \qquad (3.3.1)$$

And

$$Var(r_B) > Var(r_A) \qquad (3.3.2)$$

Set of portfolios that satisfy mean-variance criteria, are known as efficient portfolios and under rationality assumption each individual investor, would hold portfolio A instead of B since it contribute to higher return while being less risky.

In Figure 3.3.1, the vertical line is the expected return over the chosen portfolio. While the vertical line is the standard deviation of return as a measure of risk. Considering only a portfolio of risky assets, a-b-c curve is the represent the mean-variance levels tradeoff, and considering the rationality, the a-b part is the efficient portfolio. Considering also the risk-free chance r_f , the efficient frontier change to straight line tangent to min-variance efficient frontier at point T. Over this line all combination of risk-free and risky asset are possible and one could achieve higher return by borrowing at risk-free rate and invest in risky asset.

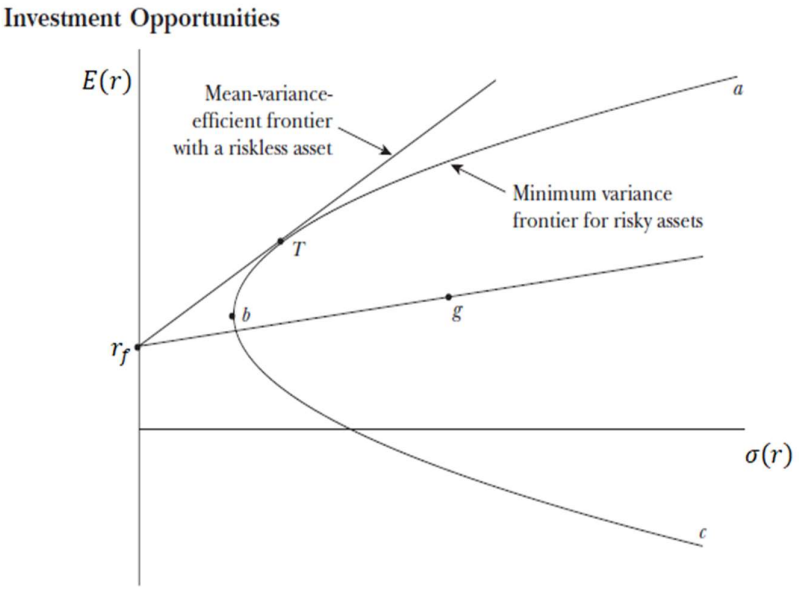


Figure 3.3.1: Investment Opportunities with efficient portfolios

At this stage one could ask how investors decision or preference will come to play. There is remarkable "Separation Theorem" [21] stating that the investment decision can be split into two stages:

- Find the efficient portfolio of risky assets
- Find the optimum fraction to invest in the efficient portfolio of risky assets and the risk-free asset.

the role of risk aversion is confined to the second stage and plays no role in the first stage since all investors hold the same portfolio of risky asset.

However CAPM is able to demonstrate the tradeoff between asset risk and its return by measuring covariance of an individual asset with respect to the whole market [8]. CAPM is a methodology for translating risk into estimated expected return. Moreover, This model considers expected return for each asset to be calculated as linear relation between risk free rate, beta value of the asse and Expected return of market as below formula:

$$(Sharpe - Linter\ CAPM) \quad E(r_i) = r_f + \beta_i \left(E(r_m - r_f)\right) \tag{3.3.3}$$

$$(Market\ Beta)\ \beta_i = \frac{cov(r_i, r_m)}{\sigma(r_m)^2} \tag{3.3.4}$$

We can interpret beta as a scale for sensitivity of each asset's return with respect to variations in the market return. However Douglas[22] and Linter[23] pioneered to use CAPM on individual security returns with focus on risk-return relevance, though their empirical results didn't meet expectations. Whereas they found the intercept had much larger value with respect to risk-free rate return even the coefficient of beta was statistically lower. Same problem was reported by Miller and Scholes [24], stating that the model does not fully describe the structure of security return.

Likewise Black[25] empirical test asserted that the expected excess return on an asset is not strictly proportional to its β. The test was done over all securities listed on New York Stock Exchange in the interval between 1926-1966 and rejected traditional form of the model. To overcome the above mentioned obstacle they used two factor model considering also portfolio return as 'Beta factor' as the second factor in CAPM. Basically their two factor model lies under the hypothesis of relaxation of risk-free borrowing and landing opportunity.

However, in many researches(Basu[26],Banz[27],Bhandari[28]) and empirical results that was made on CAPM afterward, there was evidences where expected return variation was not directly related to market beta in which they compared earnings-price ratios sorted portfolio, size sorted portfolio and debt-equity ratios with respect to CAPM predicted average return. Also Book-to-market equity ratio (the ration of the book value of a common stock to its market value) effect analyzed by Statman [29] and Rosenberg [30] and they found that high B/M leads to high average return that was not previously captured by CAPM betas. All these researches shows that ratios containing stock prices reveal essential information missed by CAPM betas.

Fama and French [31] used cross-section regression approach to study joint pattern of market beta, size, earning-price ration leverage and B/M equity of average stock returns and confirmed with empirical results over price sorted portfolios of NYSE, Amex and NASDAQ stocks for 1963-1990 that market beta has little information about average return. They affirmed that size and B/M obtain strong joint variation in average return for stock portfolios built to test risk factors. Also Fama stated that there are at least three explanatory stock-market factors of size, book-to-market and overall factor in expected stock return. Hence Fama and French proposed the following 3 factor model:

$$E(r_i) = r_f + \beta_{im}\Big(E\big(r_m - r_f\big)\Big) + \beta_{is}\big(E(SMB\ \ )\big) + \beta_{ih}\big(E(HML)\big) \qquad (3.3.5)$$

where SMB is the difference of small and big stocks portfolios, while HML is the delta return of high and low book-to-market stocks and betas are regression slopes.

Thereafter, Carhart [32] formed his 4-factor model using Fama and French's 3-factor model by adding one-year momentum versus contrarian stocks to describe portfolio returns as below:

$$E(r_i) = r_f + \beta_{im}\Big(E\big(r_m - r_f\big)\Big) + \beta_{is}\big(E(SMB\ \ )\big) + \beta_{ih}\big(E(HML)\big) + \beta_{iy}\big(E(PR1YR)\big) \qquad (3.3.6)$$

By considering the low correlation of SMB, HML and PR1YR with each other in zero investment portfolios, he noticed high variance for them. Summary statistics over portfolios of NYSE, Amex and NASDAQ is disposed to explained sizeable time-series variations and 3 factor of SMB, HML and PR1YR with high mean return account for most cross-sectional variation in mean return on stock portfolios. In analyzing performance of one year lagged returns, he considered monthly returns for ten equally weighted portfolios formed on 1st of January of each year. Buying last year top-decile and selling last year's bottom-decile yield

about 8 percent per year and he stated that with 4-factor model can explain short-term persistency to a good extent.

However some evidences exhibited that 3-factor is not explaining much of the variation in average return related to profitability and investment; two factors that was proved to be connected to average return. Fama and French[33] two factors of profitability and investment to the 3-factor model and demonstrated the relation between average return and new factors of profitability and investment through a model for market value based on discounted value of expected dividends per share. The 5-factor model of Fama and French is:

$$E(r_i) = r_f + \beta_{im}\left(E(r_m - r_f)\right) + \beta_{is}(E(SMB\ )) + \beta_{ih}(E(HML)) + \beta_{iw}(E(RMW)) + \beta_{ia}(E(CMA)) \qquad (3.3.6)$$

Where in this equation RMW is the difference between the returns on diversified portfolios of stocks with robust and weak profitability and CMA is difference of low and high investment firms. With 5-factor model they explained approximately between 71% and 94% of the cross-section variance of expected return of portfolios under investigation.

The aforementioned leading researchers showed the evolution of portfolio management theory that was built on Markowitz model, where researchers were seeking for privileged model that can describe properly the market expected return which imperative in asset pricing. Different technical, fundamental and macroeconomics factors were tested to relate risk and explain stock returns. Although there is no panacea financial model that can precisely predict the expected return, but to a good extent support perceiving the expected return.

3.4. Correlation-Cointegration Approach

Along Since the seminal paper of Mantegna[34], researchers studied structure of financial market and investigate their statistical properties. In mathematical finance, evolution of stock return form a time series which is unpredictable to up to a random process[34]. These random processes correlation and common economic factors mutating several stocks and at the same constitute the main discussion. Correlation and clustering as methods to study such financial structure are deemed repeatedly in financial literature, as well as pairs-like reversal strategies.

Correlation is a useful measure of linear codependence between securities and will extract estimation of co-movement between observed time series of two stock prices. However, stock prices are noisy, nondeterministic and ill-conditioned where signal to noise ratio are low and needs to be treated carefully to no affect metric construction. There has been developed techniques as in Lopez[35] to reduce the noise and increase the signal included in correlation matrix.

Perlin[36], used minimum squared distance to study the performance (profitability and risk) of pairs-trading approach in Brazilian stock market. Where he used normalized stock prices based on variance and mean, to better evaluate the minimum squared distance between historical prices of two stocks (Figure 3.4.1 depicts an example of this approach for two stock TNLP4 and TNLP3). Also he picks two stock as candidate if their minimum squared distance exceed a predefined threshhold d in prespecified time horizon h. At the end, his results shows an acceptable performance of pairs-trading in Brazilian stock market with one day period and distance between 1.5 and 2.

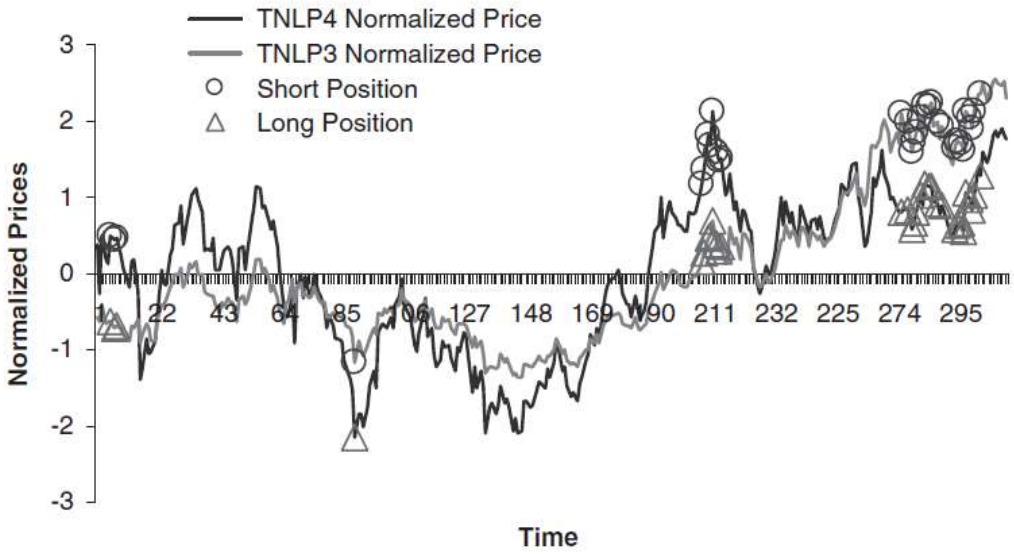


Figure 3.4.1. Minimum squared distance of TNLP4 and TNLP3

However, to calculate correlation between time series of two security, different choices exists. Wang [37] studied the performance of pairs-trading approach and its performance using two Maronna and Pearson correlation with three trade triggering mechanism with different permutation and combination of Maronna and Pearson correlation. He used correlation as main

ingredients of reversal approach and derived the Sharpe ratio and win-loss ratio performance over 61 stocks in NSYE using back-testing. As Pearson correlation is sensitive to outliers and Maronna is a more stable correlation measurement, he expected to perform better, but stability of Marrona cause the lag in identifying the correlation divergence and consequently slower to capture price changes. There are also some non-paramteric rank-based correlation coefficients like Kandall tau and Spearsman rho. In [38], significance and its relation for Pearson's correlation is compared to Spearsman's correlation, noting that Pearson's correaltion is based on linear association of two variables while Spearsman's correlation is a measure of monotone association.

Another approach is developed by Wen et al.[39], where they introduced a new pairs-trading approach which uses correlation and cointegration successively over Chinese Stock market. Hence, Pearson's correlation of daily stock prices being used to pre-select potential stock candidates within their strategy and cointegration being utilized to construct the weighted cointegration network. Network considered to be the weighted graph obtained from cointegraion matrix, where securities corresponding to nodes and weighted links are related to correlation coefficient matrix of returns[40]. After building the full graph (FG), they used MST(Figure 4.3.3.) to study network systematically with respect to overall market trend. The structure and edge evolution analysis shows that only a few number of linkages remains from one month period to the other and above all they showed once again proves cointegraion based pairs-trading as a market neutral strategy.

Another research [41] done investigating weighting criterias in portfolio construction in comparison to equally weighted portfolio. Where four different pairs-trading based on distance (squared distance of normalised prices), Spearman's correlation, Engle-Grangle cointegration and Hurst exponent used for pairs selection. Also new weightening approaches based on volatility, log-prices minimal distance, correlation of returns, cointegration of prices and Hurst exposure factor was introduced and compared ot equally weighted portoflio. The experimental result shows improvement in new proposed approaches for constructing weighted portfolios.

The aforementioned researches developed and proposed different approaches in reversal pairs-trading approach based on various criteria, including correlation and clustering coupled with complements in weight selection or misxed with other criteria. Correlation as a measurement of co-movement, is utilized to capture useful information to be used in pairs selection. Correlation can be used to study structure of network, build a metric to use in clustering

methods, or decouple with cointegration to pre select the candidate market elements to find pairs. Clustering is another powerful pair selection method that facilitate investigating formation of pairing groups over considered time horizon. It can display clearly the birth, death, merge and spliting of pairs group over time taking the chance of further study of overall market elements or improving pairs selection approach.

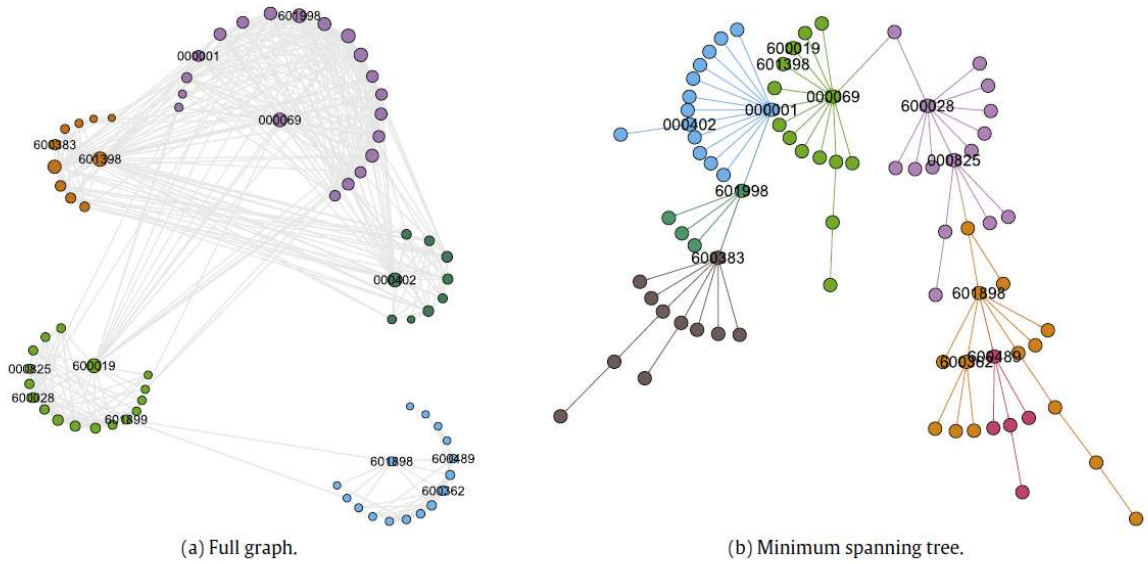


Figure 4.3.3. (a): Full graph of pre-selected 25 market element by correlation. (b) Minimum Spaning Tree to find the clusters

The aforementioned researches developed and proposed different approaches in reversal pairs-trading approach based on various criteria, including correlation and clustering coupled with complements in weight selection or misxed with other criteria. Correlation as a measurement of co-movement, is utilized to capture useful information to be used in pairs selection. Correlation can be used to study structure of network, build a metric to use in clustering methods, or decouple with cointegration to pre select the candidate market elements to find pairs. Clustering is another powerful pair selection method that facilitate investigating formation of pairing groups over considered time horizon. It can display clearly the birth, death, merge and spliting of pairs group over time taking the chance of further study of overall market elements or improving pairs selection approach.

3.5. Machine Learning Approach to Portfolio Optimization

Along with the developments in machine learning technology, and process automation capability of Machine Learning, captured great attention of scientists also in financial field. Coupling financial theories and data with emerging machine learning techniques, made building an accurate, efficient, consistent and powerful model feasible. Models that could be enlist for predicting , controlling and diagnosis of the financial data and systems.

Financial Machine Learning usage in asset management being categorized by researchers. Among those, Snow [42] divided financial machine learning research into four streams as in (Figure 3.4.1). First stream is related to predicting the future value of securities, while second one is related to predicting financial events like regime changes, corporate defaults, mergers and acquisitions. Third stream entails estimation and prediction of factors that are not indirect related factors of financial market like volatility, firm evaluation and credit rating. After having all these information in previous streams, in the last one, comprises the machine learning techniques to optimize constructed portfolio. Which I will discuss in more details the fourth stream.

There are wide range of machine learning techniques to be used in financial modeling consists of two categories of supervised and unsupervised method. Linear regression as the most used and simplest supervised-learning algorithm, being used in many researches for asset management and portfolio selection. It's a linear approach to find out the relation between a dependent variable and one or more independent variables. However linear models being developed with some regularization functions to overcome ill-posed linear operator problems to prevent overfitting. Among most famous regularization method, Lasso and ridge regression is widely used. Another regularization approach is befittingly known as performance-based regularization (PBR)[43]. In this method, the goal is to steer the solution toward having less performance estimation error. G-Y Ban[43] developed PBR model for Markowitz problem with considering performance-motivated regularization by constraining sample variance of estimated portfolio risk and return. Their result shows that PBR approach improves between 5%-10% in out-of-sample Sharpe ratio[44] with respect to other benchmarks.

Artificial Neural Network (ANN), a complex learning algorithm, being used for information processing, mainly inspired by biological nervous system, is a kind of supervised learning. This network learns through iterative process relating the input features to desired results and reducing the error in each step, similar to learning a task by a human through repetition and correction by a trainer. While ANN is used for different science branches, in [37] Artificial Neural Network being used to predict and forecast NASDAQ stock exchange rate. In this research several feed forward ANNs being trained using back propagation assessment, where tangent sigmoid (TANSIG) being used as transfer function and one step secant (OSS) back propagation approach. The result shows that a network with 20-40-20 neurons in hidden layers results in an optimized network with R-Squared value of 0.94.

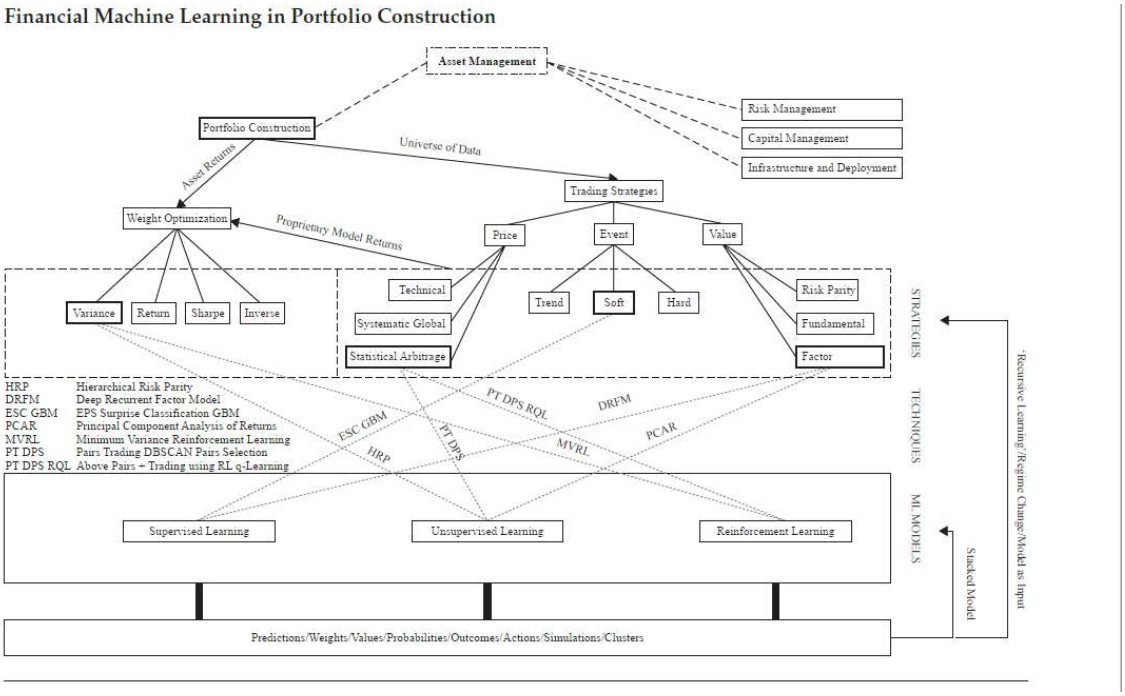


Figure 3.4.1: Financial Machine Learning in Portfolio Construction Stratification

Support Vector Machine (SVM), as a supervised learning method is widely used in finance. It can be a linear classification or non-linear kernel based, for both binary or multi-classification. In ANN if the number of parameters to be fix is large, which is a common situation in most of cases, it will lead to over-fitting problem. As alternative to avoid this limitation made by ANN,

SVM method had been developed[46] and became a popular tool used in various pattern recognition fields.

In a research SVM method coupled with particle swarm optimization(PSO)[47] was used to forecast financial time series for Shanghai stock and Dow Jones index[48]. In this paper, SVM method being compared in terms of accuracy and return of investment prediction with neural network and other soft computing proposed methods. The results shows that the proposed method, obtains a good performance with accuracy rate of 61.728 % on the Shanghai stock market index and 57.937 % on the Dow Jones index. Also In [47] Hegazy studied regularized SVM with least square (LS-SVM) for predicting future prices of daily stocks based on historical data and technical indicators . Note that in SVM method, parameters strongly influence performance.

Sometimes researchers confront curse of dimensionality which is being caused by high dimensional real word financial data in machine learning. In third stream consider the problem of credit scoring models, it's data structure usually experience high number of features. Some of these features may be irrelevant or redundant due to high intercorrelation and needs efficient criteria to select the optimal feature set possible and overcome drawbacks related to working with high dimensional data.

An interesting research done to optimize portfolio of Exchange Traded Funds (ETF) over a testing period from 2011 to the end of April 2020, which includes financial instabilities related to COVID-19 [49]. To form a portfolio, four ETFs of US market indices with high liquidity and small expense ratio being chosen: US total stock index(VTI), US aggregate bond index (AGG), US commodity index (DBC) and Volatility Index (VIX). The reason behind selecting ETFs instead of choosing individual assets is explained with the fact that these indices are generally uncorrelated or even negatively correlated, thereby help significantly with diversification requirements. The Long Short Term Memory (LSTM) method [50], an artificial recurrent neural network architecture used in the field of deep learning, is used to build the model. As a result of this study, comparing to a wide range of popular algorithms, this model delivers the best performance and a detailed study of the model performance during the crisis shows the rationality and practicability.

All these mentioned researches showed the importance of machine learning to explore efficiently the financial data and make data-driven predictions operating a model from a

training set of inputs and observations. Since the financial data to build a portfolio might come from different resources and be collected for specific usage, asset managers usually face very high dimensional data with lots of features. Some of these features are irrelevant or redundant to discussed problem. Also in most cases curse of dimensionality in data cause the paradigm of optimized portfolio to not be extracted easily. Machine learning methods not even helps to opt for relevant features, but also comes up with the best reliable model possible.

3.6. Background Summary

Considering all these approaches (modern portfolio theory, clustering, correlation-cointegration approaches and machine learning approach), their benefits and limitations, it is hard to favor one approach to others, since each of them being widely used. Markowitz modern portfolio theory introduced the basic idea and requirements in portfolio selection, where it considered first and second momentum to fulfil the diversification requirement beside maximizing expected return in portfolio selection. Markowitz framework considered to be the benchmark for future developments in asset management theory and portfolio selection. On the other hand, considering pairs-trading approach where it tries to partition financial securities into co-moving groups embedded in a correlation-cointegration based structure, introduce a promising strategy for short-term price and return predictions. Having insight through centrality of asset, help to choose the optimal weights under the Markowitz framework. Finally, implementing machine learning approaches in supervised and unsupervised learning classes had a profound impact in detecting the best feature possible among the large set of available ones and also recommended acceptable models to predict future behavior of prices or to analyze high dimensional data.

3.7. Background Conclusion

For analyzing the time series data of financial instruments in order to optimize the inquired portfolio, Markowitz theory brings necessary conditions that should be considered but lacks accuracy in out-of-sample examples. These inaccuracies could be caused by the erroneous in expected return or variance calculation. However, to solve this matter one could improve the model using CAPM to analyze the centrality of securities and enhance the selection process of a more robust portfolio. Furthermore, pairs-trading strategy as an examined method to seek out potential market-neutral profits. Researches show significantly better Sharpe ratio in out-of-sample examples comparing to the one given by Markowitz. The importance of pairs-trading approach, being specified by the fact that it facilitates, to some extent, future price prediction. Whereafter defining the co-moving securities using correlation-cointegration measures, deploying machine learning methodologies enables proper pairs selection.

4. Methodology

In this chapter, I go through the principal steps that are going to be pursued in the next chapters. Starting from introduction of cointegration methods and construction of co-moving groups, using two different cointegration tests based on an appropriate approach. Then I will present correlation as a measure of co-movement for securities and couple it with cointegration approach to recommend my pairs-trading method to be applied over the set of securities as asset picking strategy. Artificial neural network, specifically LSTM, is a powerful tool that will be used within my strategy as a sort of criteria within cointegration groups and further investigation over cointegrated peers and finally portfolio construction. This chapter is organized as follows:

- Cointegration groups
- correlation-cointegration approach
- Pairs trading approach
- Artificial neural network and LSTM

I will try to cover all the basics and essential concepts needed in this thesis. While for the other available techniques, I will provide the general idea with useful references to be followed by enthusiast readers.

4.1. Cointegration Groups

The idea of pairs trading is very simple and practical and, in sum, it consists of two steps. First to find two securities whose prices have historically moved together on average and monitor their spread. If the monitored spread between them diverges, I sell security which negatively diverges from spread and buy the one positively diverges. As these securities have the potential to rebound to the equilibrium, this strategy makes profit. The same approach can be extended to multi-securities pairs trading with trading a set of securities against the others by considering the spread between two sets. Krauss [50] classified pairs-trading approaches into five

categories of: 1- distance approach 2- cointegration approach 3- Time series approach 4- stochastic control approach and 5- other approaches. I will explain aforementioned methods briefly:

1. **Distance approach:** Distance approach which was promoted first with Gatev[52], and is the most noted researched approach. In this approach co-moving securities and groups can be formed through various distance metrics. Gatev considered sum of squared deviation between normalized prices as a metric for divergence and convergence valuation. Metrics and distances as the most intuitive and simple concepts, are essential property of this method, and one could easily take profit across different markets by applying distance approach.

2. **Cointegration approach:** In this approach there is two main step, first it should be tested if two series are cointegrated then if the deviation assure some deviation rule to be triggered. The most common method of cointegration in financial literature is ordinary least square. To generate trading signals in trading period, GGR threshold method can be used.

3. **Time series approach:** Time series approach assumes that groups of co-moving securities are already detected. Instead the focus is on time series analysis with different techniques to optimize trading signals for the co-moving sets of securities.

4. **Stochastic control approach:** As in time series approach, it is ignoring formation period and just trying to find the optimal portfolio in each leg of pairs trading. A seminal work in stochastic control approach to pairs-trading done by Mudchanatongsuk[53] where he proposed a portfolio model based on log-prices difference of a pair of stocks as Ornestein-Uhlenbeck process. Then he used a dynamic stochastic control approach for portfolio optimization that may have a closed form solution for parameters under maximum likelihood.

5. **Other approach:** All the other approaches like Principal Component Analysis or Neural Network approaches and less propounded techniques lie in this class.

I am going to consider Cointegration approach for pairs trading which is a powerful tool to retrieve co-moving groups within financial instruments. This approach is used by many

academic researchers in financial world, with proper and competent results with respect to the other available methods.

4.1.1. Cointegration Approach to build Cointegration Groups

Almost all the economic factors are non-stationary and for every equilibrium theories, a combination of non-stationary variables, needs to follow some stationarity otherwise any deviation from equilibrium will not be temporary. In other words, equilibrium is a kind of stationary point characterized by different forces and in case it wanders away, these forces will push it back toward equilibrium point. Since in my pairs trading analysis I need some kind of equilibrium situation to be satisfied in prices spread, I will clarify some stationarity definitions.

Definition: Considering a $k * 1$ vector y, it is said to be stationary of order b (shown as I(b)) if it only needs b difference to induce stationary. Then we say a $k * l$ vector y_t is stationary of order b, d (shown as CI(b,d)) if every component of y_t is I(b) and there exist a vector β such that $z_t = \beta' y_t$ which is I(b-d) and vector β is the cointegration vector.

An integration test proposed by Engle and Granger[54], where for a times series they used Augmented Dickey-Fuller test to show the stationarity of its components. Let a multivariate time series be defined as:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + e_t \qquad (4.1.1.1)$$

Where p is the number of lags and each univariate component of X_t defined as:

$$y_t = \theta y_{t-1} + \varepsilon_t \qquad (4.1.1.2.)$$

where ε_t is a white noise.

The null hypothesis in Engle-Granger cointegration test is if eigenvalues of estimated $\varphi = 1 - \varphi_1 - \varphi_2 - \cdots .. \varphi_p$ are not significantly different from zero. If y_t and z_t are two univariate component of X_t and their integration order is 1, then Eagle-Granger tests if y_t and z_t are cointegrated of order CI(1,1). To perform Engel-Granger cointegration test, I will use three straightforward steps:

1. First defines the order of integration for components y_t and z_t. If y_t and z_t have the same order of cointegration, the Augmented Dickey-Fuller test can be applied to check whether $\theta = 0$ for each y_t and z_t. If the test satisfies, then it would meant that components are stationary and their difference could be integrated of order zero.

2. If the test on step 1 is satisfactory, and components are of order one, it keeps the residual in following regressions:

$$y_t = a_0 + a_1 z_t + \varepsilon_{y,t} \qquad (4.1.1.3.)$$

$$z_t = b_0 + b_1 y_t + \varepsilon_{z,t} \qquad (4.1.1.4.)$$

3. Consider the null hypothesis of $|a_1| = 0 = |a_2|$ in the following regression over the residuals:

$$\Delta\varepsilon_{y,t} = a_1 \varepsilon_{y,t-1} + v_{y,t} \qquad (4.1.1.5.)$$

$$\Delta\varepsilon_{z,t} = a_2 \varepsilon_{z,t-1} + v_{z,t} \qquad (4.1.1.6.)$$

If we couldn't reject the null hypothesis, then we cannot reject the hypothesis that y_t and z_t are not cointegrated.

Another test that I am going to consider is Johansen cointegration test [55]. In calculation of maximum likelihood estimation for cointegration of vectors, Johansen used the null hypothesis $H_0: rank(\varphi) \leq r \; or \; \varphi = \alpha \beta'$ for any $r \leq p$ and if there is cointegration within variables then X_t is cointegrated with vector β. However, the estimation of the parameters α and β is impossible, but I can approximate the space spanned by matrix β as stated in following theorem by Johansen.

Theorem: Bilgili[56], The maximum likelihood estimate of the space spanned by β, is the space spanned by r canonical variates corresponding to the r largest squared canonical correlations between residuals of X_{t-p} and ΔX_t that are corrected for the effect of the lagged difference of the X process.

To calculate r largest canonical correlations, following steps may be used:

1. Regress both X_{t-p} and ΔX_t over $\Delta X_{t-1} + \Delta X_{t-2} + \cdots + \Delta X_{t-p+1}$ and keep the residuals v_t and w_t respectively

2. By another method introduced , I compute the squared of canonical correlation between two residuals v_t and w_t we calculated in step 1 as:

$$\gamma_1^2 > \gamma_2^2 > \cdots > \gamma_p^2 \qquad (4.1.1.7.)$$

3. Then I will test the number of non-zero eigenvalues using Trace test or Maximal eigenvalue test as:

$$\gamma \max(r, r+1) = -Tln\left(1 - \overline{\gamma_{r+1}^2}\right) \qquad (4.1.1.8.)$$

$$\gamma \operatorname{trace}(r) = -T \sum_{i=r+1}^{n} \ln\left(1 - \overline{\gamma_i^2}\right) \qquad (4.1.1.9.)$$

However, cointegration is a trick that makes regression works for non-stationary time series. Considering the aforementioned cointegration tests of Engle-Granger and Johansen, next step is to build cointegration groups of securities.

Assuming a set of securities $\{S_i\}$ under investigation, then I will go through following steps to build my cointegration groups:

- For each security $S_{i,t}$, at January of year t, consider the time series of adjusted close prices for past three years' period $[t-3, t-1]$.

- For each pair of securities $S_{i,t}$, $S_{j,t}$, use both Engle-Granger and Johansen cointegration test with significance level of 1% . If both tests show no evidence that two series are not cointegrated, I consider that two securities $S_{i,t}$, $S_{j,t}$ are cointegrated and I denote as $S_{i,t} \sim S_{j,t}$.

- For security i , I define the cointegrate groups as the set of securities having cointegration relation to security i:

$$CG_{i,t} = \{S_{j,t} : S_{i,t} \sim S_{j,t}\} \qquad (4.1.1.10.)$$

It should be noted that the cointegration relation is not symmetric, and consequently, an equivalence relation, as it may be in some cases, the security $S_{j,t}$ belongs to cointegration group $CG_{i,t}$ but security $S_{i,t}$ does not belong to cointegration group $CG_{j,t}$. This behavior is expected as a consequence of Engle-Granger and Johansen cointegration tests.

4.2. Correlation-Cointegration approach

In this thesis, I use correlation, as a measurement of two co-movement between time-series of two securities, to check the codependences of securities and I am going to use correlation as pre-selection criteria to retrieve possible candidates to be fed into cointegration approach. In more details, using correlation, I detect highly-correlated securities and build my cointegration groups starting with these set of securities.

4.2.1. Correlation Formation and Metric

With correlation coefficient one could get the linear dependence and relationship between two variables. There exists different correlation coefficients as Pearson, Marrona and Spearsman where each has some advantages and disadvantages. Based on the researches done, I decided to use Pearson's correlation coefficient. Considering securities X and Y with P^X and P^Y respectively as prices, the Pearson's correlation coefficient is defined as:

$$\rho_{XY} = \frac{\sum_{t=1}^{N}(P_t^X - \bar{P}^X)(P_t^Y - \bar{P}^Y)}{\sqrt{\sum_{t=1}^{N}(P_t^X - \bar{P}^X)^2(P_t^Y - \bar{P}^Y)^2}} \qquad (4.2.1.1.)$$

where $\bar{P}^X$ are the average of return over the observation period as:

$$\bar{P}^X = \frac{1}{N}\sum_{t=1}^{N} P_t^X \qquad (4.2..1.2.)$$

and ρ_{XY} is a measurement of linear co-movement of two securities.

However, the correlation is not satisfying the axioms of metric and cannot evaluate the distance of two time-series. Thus, I define a metric as a function of correlation in a way that the distance of two variables will be high (low) if the correlation is high (low). I will use the constructed metric in pre-selecting securities. To this end, for two securities X and Y, I define the distance as a function of their correlation by:

$$d(X,Y) = \sqrt{2(1-\rho_{XY})} \qquad (4.2.1.3.)$$

Distance $d(X,Y)$ defined above satisfy all the three metric axioms of: 1- identity of indiscernible, 2- symmetry and 3- triangle inequality.

4.2.2. Correlation Coupled with Cointegration

Within this approach, I first select highly-correlated stocks by applying Pearson's correlation coefficient, defined as measure of co-movement of securities. For each security, X and Y with prices P^X and P^Y, I consider three years' time window aligned to the time window used in cointegration section. At the beginning of each year, I look-back at past three years' daily prices and calculate the correlation between prices of the two stocks within this period, to select potential candidates. By employing cointegration theory of Engle-Granger and Johansen, I build my cointegration groups of securities to be fed into the pairs picking process.

Although both correlations and cointegrations, at some extent, explain the relation of the two time-series, but both are not synonymous in the sense that, it can happen that two time-series have high correlation but low cointegration and vice versa. By the fact that the correlation captures the linear codependence of two series and cointegration seeks stationarity, this approach is capable to reveal stock's time-series that tend to walk together randomly. Specifically cointegration captures the common trends beside long-term equilibriums and short-term deviations.

4.3. Pairs Trading Approach

Whereas each pair trading strategy comprises of two steps, first finding a pair of co-moving securities as potential candidates and second, to introduce a criteria to open or close the positions. For co-moving time-series, especially in case that they are cointegrated and have some form of stationarity, it is expected that if they wander away from equilibrium, they will rebound in long term. Thus, after finding co-moving groups, it is necessary to detect divergence from equilibrium. When a security underperforms with respect to the equilibrium, I take a long position and when it is overperforming I will open short position. With this strategy and rebounding fact, I will produce some profits.

4.3.1. Pairs Trading Based on Historical Price and Return

I will exert historical prices (Gatev[52]) and returns (Chen[57]) correlation in my pairs trading strategy to find the optimal trading opportunity. Looking at historical prices or returns, securities that are deviating from their peers are distinguishable by computing the related gap, and testing if the security future price or return will converge to its peer, bring out a potential trading opportunity. In other words, trades securities that underperform or overperform, their peers have potential to have high return with reversal method explained earlier. Therefore, after fixing peers and co-moving groups with correlation-cointegration defined before, I will check performance of each security's price and return with regard to its belonging groups, to extract any possible trading opportunity.

In this chapter, I will consider the pairs-trading. I will mainly focus on model and notions provided by Flori [58] for pairs-trading, based on price and return paradigm as:

$$price\ gap \begin{cases} \Delta p_{i,t} = p_{i,t} - p_{CGi,t} \\ \Delta^{\beta} p_{i,t} = p_{i,t} - \beta_{p,i} p_{CGi,t} \end{cases} \qquad (4.3.1.1.)$$

$$return\ gap \begin{cases} \Delta r_{i,t} = r_{i,t} - r_{CGi,t} \\ \Delta^{\beta} r_{i,t} = r_{i,t} - \beta_{r,i} r_{CGi,t} \end{cases} \qquad (4.3.1.2.)$$

where $p_{i,t}$ and $p_{CGi,t}$ are respectively normalized adjusted price of security i and average normalized adjusted price of cointegration group i of the same security. While $r_{i,t}$ and $r_{CGi,t}$ are return for security i and average return of related cointegration group. In above mentioned gap formulas for each security i, $\beta_{p,i}$ and $\beta_{r,i}$ are respectively the regression over price and return with respect to the average price and returns of its group with a white noise as signal as in the formula:

$$\begin{cases} p_{i,t} = \beta_{p,i} p_{CGi,t} + \vartheta_{i,t}\,, & \vartheta_{i,t} \sim \aleph(0, \sigma_{\vartheta}^2) \\ r_{i,t} = \beta_{r,i} r_{CGi,t} + \tau_{i,t}\,, & \tau_{i,t} \sim \aleph(0, \sigma_{\tau}^2) \end{cases} \qquad (4.3.1.3.)$$

where using the ordinary least square, $\beta_{p,i} = ({p_{CGi,t}}^T \times p_{CGi,t})^{-1} \times p_{CGi,t} \times p_{i,t}$ and similarly for return $\beta_{r,i}$ I get its estimation.

To detect any possible divergence, I will consider checking the normalized price and return gaps mentioned before as $\tilde{\Delta}p_{i,t}$, $\tilde{\Delta}r_{i,t}$. Note that I removed the most of the noise parts in correlation construction process, and consequently, I expect that the white noises $\vartheta_{i,t}$ and $\tau_{i,t}$ are negligible and have little effects on divergence or convergence of two securities.

4.4. Artificial Neural Networks and LSTM

Artificial neural networks (ANN) inspired by biological neural system of human brains, are powerful tools that are recently being used pervasively in different researching areas of recognition, production, time-series prediction and etc.[60]. Artificial neural networks ease to extract complex pattern from huge datasets by computers without explicitly programming the training model. There are so many variant sort of neural networks called Recurrent Neural Network, Convolutional Neural Network, Residual Neural Network, Multi-layer Perceptron and Long-Short Term Memory, etc. where in this thesis I will take advantage of Long-Short Term Memory (LSTM). However, I use LSTM as another sorting criteria after detecting co-moving groups to evaluate best candidates to buy highest rated stocks and selling the lowest rated stock. More specifically I will evaluate also LSTM performance as a sorting criteria coupled with price gap and return gap sorting criteria.

4.4.1. Recurrent Neural Network Structure

Each neural network includes three layer of, input layer, one or more hidden layer for each time step and output layer, where each layer formed by neurons as illustrated in figure 401. To retrieve more complex patterns in data, one could increase number of layers and their neurons, but adding more layers and neurons makes it computationally expensive. Starting from input layer that receive information, it goes through hidden layer with different nodes with specific weight of importance, each layer will be activated with an activation function.

Considering that my training set constitute of n sample and each individual sample shown as X_i, concludes of m features ($X_i \in R^m$, $\forall i \in \{1,2,3,\dots,n\}$). Set of features will have one or more target features y_i that are the output of my model and purpose of network to be estimated. Starting with input layer by giving training set as entries to the network, they pass through each hidden layer l which depends on weight w_i^l and bias b_i^l that results in z_i^l which is the input for the next layer by following formula:

$$z_i^l = w_i^l . a_i^{l-1} + b_i^l \qquad (4.4.1.1.)$$

Where w_i^l is the weight assigned to connect nodes of previous layer to the current layer, $a_i^{l-1} = f(z_i^{l-1})$ is the activation of previous layer and finally b_i^l is the bias assigned to node i in current layer l. The result of output layer is $\hat{y}_i$ which is an approximation of the target feature y_i. However each node includes two parts as in Figure 402 (b), sum of outputs of previous layer z_i^l and a bias associated to it. This approach of neural network is called forward propagation.

Where w_i^l is the weight assigned to connect nodes of previous layer to the current layer, $a_i^{l-1} = f(z_i^{l-1})$ is the activation of previous layer and finally b_i^l is the bias assigned to node i in current layer l. The result of output layer is $\hat{y}_i$ which is an approximation of the target feature y_i. However each node includes two parts as in Figure 402 (b), sum of output of previous layer z_i^l and a bias associated to it. This approach of neural network is called forward propagation.

To measure the performance of my prediction I need a loss function to evaluate the goodness of my prediction $\hat{y}_i$ with the known target value y_i as $L(\hat{y}_i, y_i)$ for each sample i. The sum of loss over all samples, is called cost function:

$$J(w,b) = \frac{1}{m} L(\hat{y}_i, y_i) \qquad (4.4.1.2.)$$

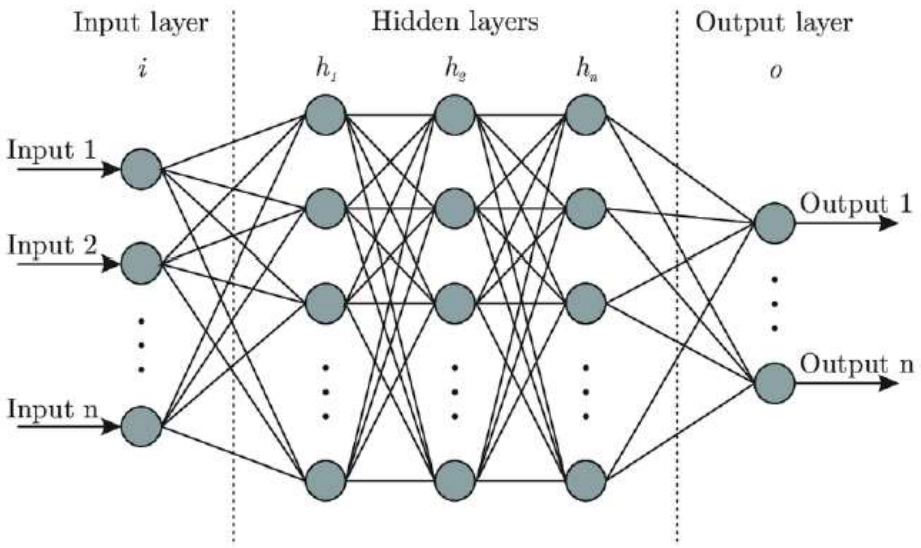


Figure 4.4.1. Neural network structure

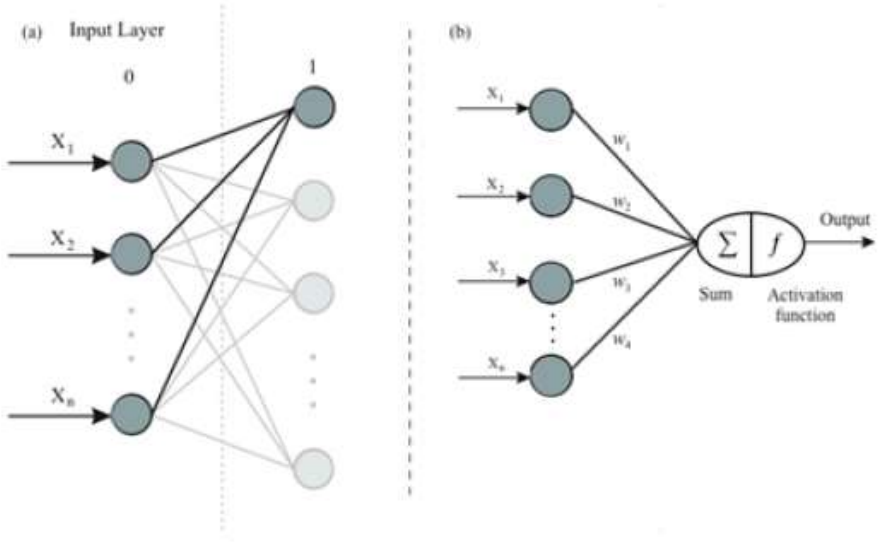


Figure 4.4.2. Neural network layers structure

To optimize the cost function given above, I need an algorithm to adjust parameters w and b related to weights and bias of each node, in a way that minimizes the cost function. Back propagation is a well know algorithm for feedforward neural networks to figure out how to change model internal parameters to minimize the cost function [61]. Back propagation uses Gradient Descent optimization method to find the direction vector in weight-bias plane(Figure 4.4.3.) through which the error will decrease.

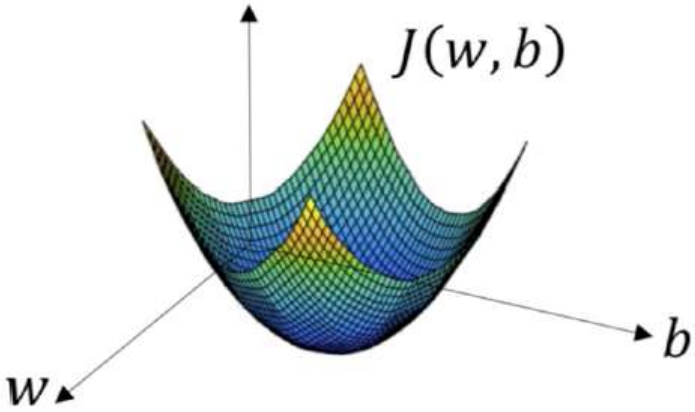


Figure 4.4.3.: Convex cost function[62]

For financial data set, as I am working with time-series, I need to extend my model to Recurrent Neural Network (RNN), where it facilitates to update context information computed from past inputs and use to investigate output. In this approach there is cyclic structure that allows algorithm to take advantage of past inputs for a dynamic size time window, where time window depends on the weights and information received within each node.

Consider time $t \in \{1,2,3, \dots, T_x\}$ and for each time t, time series input $X^{<t>} \in R^{n \times m}$ associated with target output $y^{<t>}$, the goal of recurrent neural network is to find the approximation $\hat{y}^{<t>}$, at each time t, $\hat{y}^{<t>}$ being studied by $X^{<t>}$ and activation information received from previous layer denoted by $a^{<t-1>}$ as illustrated in Figure 4.4.4. In this model I have w_{aa}, w_{ax} and w_{ya} as weights respectively connecting two consecutive network, input $X^{<t>}$ and network and finally network and estimated output $\hat{y}^{<t>}$.

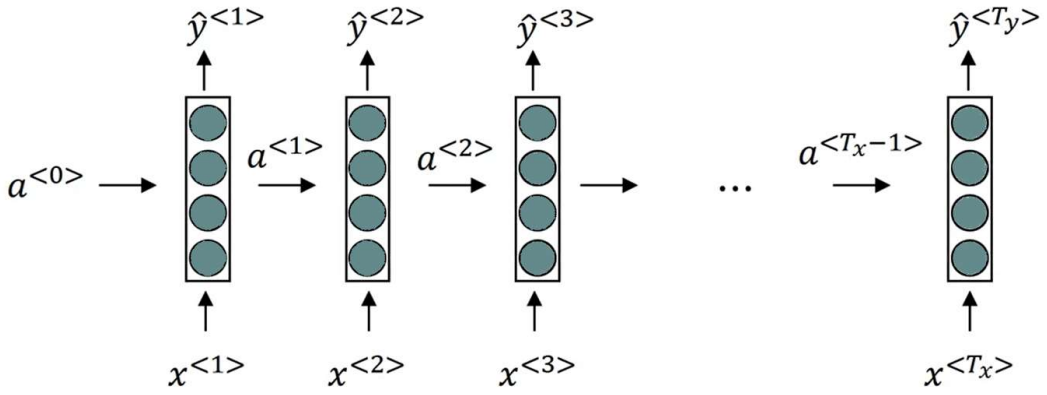


Figure 4.4.4.: Recurrent Neural Network Structure

General formula for activation $a^{<t>}$ and estimated output $\hat{y}^{<t>}$ is:

$$a^{<t>} = g_1(w_{aa}a^{<t-1>} + w_{ax}x^{<t>} + b_a) \qquad (4.4.1.3.)$$

$$\hat{y}^{<t>} = g_2(w_{ya}a^{<t>} + b_y) \qquad (4.4.1.4.)$$

where g_1 and g_2 are activation functions and b_a, b_y are biases. Similar definition for loss function and back propagation needs to be done as in ANN to find the minimized cost function.

However an efficient learning algorithm used in recurrent networks is to use gradient descent, to minimize the cost function and find the best possible output with respect to the weights of network. Nevertheless the basic RNN I have discussed, is not very effectual to capture long-term dependencies within network. In my case that I need to work with very deep network, using gradient descent, it can cause the vanishing gradient or exploding gradient as taking derivatives in gradient descent procedure, high number of layers could lead to the derivatives, to grow exponentially or decrease exponentially. It does not mean that it is impossible to train a recurrent network, while it states that when the temporal span of the dependencies are high the gradient descent is an inefficient method.

4.4.2. Gated Recurrent Units

To resolve vanishing/exploding gradient problem faced with recurrent neural network, I will extend my model to Gated Recurrent Units(GRU) and more specifically Long Short Term Memory(LSTM). Gated recurrent network is a modification of RNN to capture long range connections and resolve vanishing gradient problems. The principal idea is to introduce a more sophisticated activation function consists of affine transformation coupled with a simple nonlinear element-wise units named gate units[63].

In GRU method I am going to use a new unit called memory cell (c) that provide a bit of memory to members. After defining memory cell, at each step I am going to replace and overwrite the memory cell with a new memory candidate $\hat{c}^{<t>}$ using previous memory $c^{<t-1>}$ and $X^{<t>}$ using a $tanh$ activation function as:

$$\hat{c}^{<t>} = \tanh(W_{cc}\Gamma_r \otimes c^{<t-1>} + W_{cx}X^{<t>} + b_c) \qquad (4.4.2.1)$$

$$\Gamma_r = \sigma(W_{rc}c^{<t-1>} + W_{rx}X^{<t>} + b_r) \qquad (4.4.2.2.)$$

where Γ_r is the relevance unit that demonstrates the relevance of $c^{<t-1>}$ to $\hat{c}^{<t>}$ and $\otimes$ is element wise multiplication operation.

Another essential element to be added to GRU is the update gate with a value between zero and one using a sigmoid function. This updating gate unit, will let us to decide whether update memory cell $c^{<t-1>}$ with $\hat{c}^{<t>}$ or not and how to modulate the flows of information at earlier stages. Updating gate that decides how much the unit updates its activation or content is defined as below:

$$\Gamma_u = \sigma(W_{uc}c^{<t-1>} + W_{ux}X^{<t>} + b_u) \qquad (4.4.2.3.)$$

And finally the actual value of $c^{<t>}$ will be updated as:

$$c^{<t>} = \Gamma_u \otimes \hat{c}^{<t>} + (1 - \Gamma_u) \otimes c^{<t-1>} \qquad (4.4.2.4.)$$

With considering $c^{<t>}$ to have the same role as activation $a^{<t>}$ in RNN, one could proceed to find the estimation $\hat{y}^{<t>}$ and following backward propagation to optimize the cost function (Figure 4.4.2.2.).

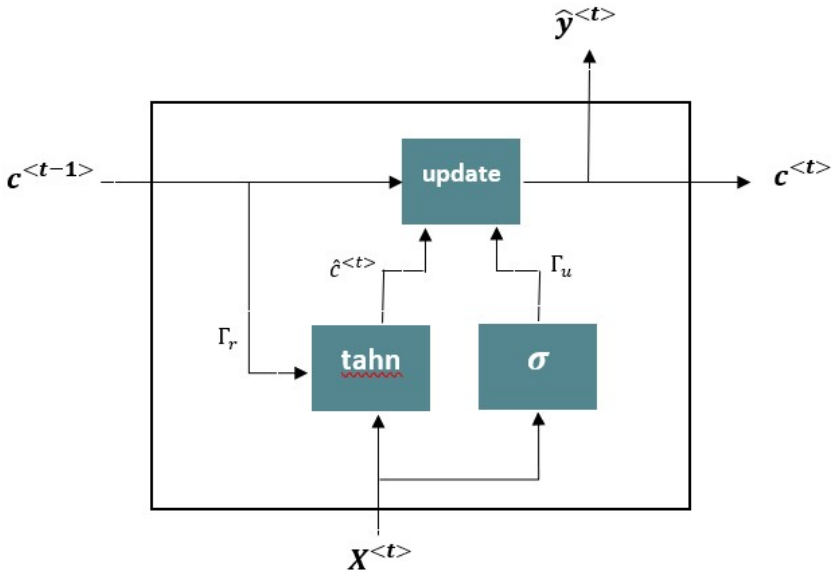
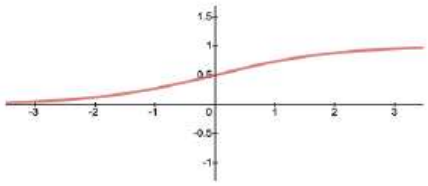
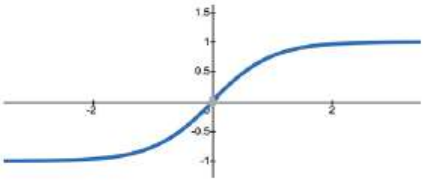


Figure 4.4.2.1.: Gated Recurrent Unit Structure

where in aforementioned formulas, $\sigma(z) = \frac{1}{1+e^{-z}}$ and $\tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$ are the sigmoid and tanh activation functions as shown in Figure 405.5.



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

Figure 4.4.2.2.: sigmoid and tanh function representation

4.4.3. Long Short Term Memory

To further extend my model, I will introduce the Long Short Term Memory (LSTM) unit which was initially proposed by Hochreiter and Schmidhuber [64]. The main idea is to add two Output Gate and Forget Gate, where the output gate adjusts the amount of memory content exposure and forget gate, helps to update memory cell $c^{<t>}$ by partially forgetting part of useless information within previous memory states:

$$\hat{c}^{<t>} = \tanh(W_{ca}a^{<t-1>} + W_{cx}X^{<t>} + b_c) \quad (4.4.3.1.)$$

$$\Gamma_u = \sigma(W_{ua}a^{<t-1>} + W_{ux}X^{<t>} + b_u) \quad (4.4.3.2.)$$

$$\Gamma_f = \sigma(W_{fa}a^{<t-1>} + W_{fx}X^{<t>} + b_f) \quad (4.4.3.3.)$$

$$\Gamma_o = \sigma(W_{oa}a^{<t-1>} + W_{ox}X^{<t>} + b_o) \quad (4.4.3.4.)$$

$$c^{<t>} = \Gamma_u \otimes \hat{c}^{<t>} + \Gamma_f \otimes c^{<t-1>} \quad (4.4.3.5.)$$

$$c^{<t>} = \Gamma_o \otimes a^{<t>} \quad (4.4.3.6.)$$

Note that $c^{<t>}$ is no more equal to activation $a^{<t>}$ unlike GRU and is controlled by output unit Γ_o. In LSTM model, memory cell $c^{<t>}$ being updated by update gate effect over memory candidate $\hat{c}^{<t>}$ and effect of forget gate to previous state memory cell(Figure 4.3.1.1).

4.4.1. Loss Function and Gradient Descent

As mentioned earlier to check the performance of my long short term memory model, I need to define a loss function as a evaluation to my predicted target values with known targets. For this purpose, I use cross-entropy loss function $L(\hat{y}_i, y_i)$ as a measurement of accuracy of my estimation $\hat{y}_i$ from true value y_i. Suppose that I want to calculate the cross-entropy loss

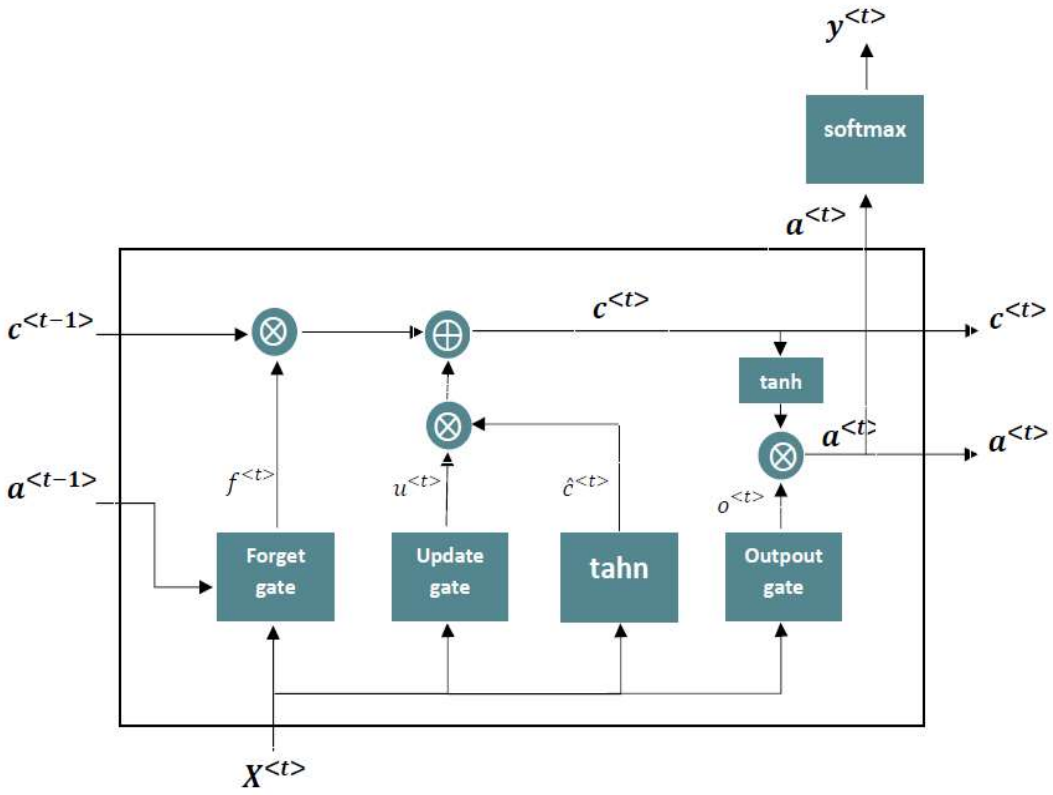


Figure 4.3.1.1. : Long Short Term Memory Structure

function for a single neuron (Figure 4.4.1.1.)., for a single neuron I have inputs $x_1, x_2, x_3, \ldots$ and corresponding weights $w_1, w_2, w_3, \ldots$ and bias b then the cross-entropy cost is:

$$L(y_i, \hat{y}_i) = \frac{-1}{n}\sum_{x}[y_i \ln \hat{y}_i + (1 - y_i)\ln(1 - \hat{y}_i)] \qquad (4.4.1.1.)$$

by summing over all inputs and n is number of inputs.

To detect optimal weights and bias, I use back propagation approach and a version of Gradient Descent method called Root Mean Squared Propagation (RMSProp) to achieve best possible minimization for cost function. RMSProp is designed to accelerate the optimization process in two ways: by improving capability of optimization method based on Gradient Descent with Momentum approach, and also by decreasing the number of cost function evaluation to reach the optimum.

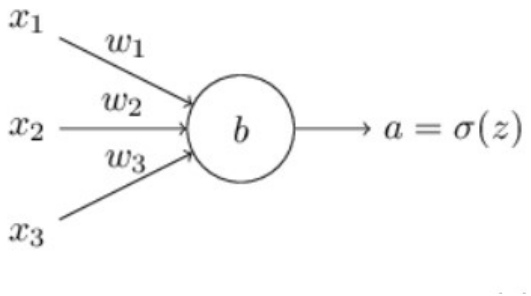


Figure 4.4.1.1.: Single Neurone Structure

Considering cost function $J(w, b)$ in simple gradient descent method to find the optimized cost function, at each step new weights w and bias b are updating as below:

$$w_{t+1} = w_t - \alpha \frac{\partial J(w,b)}{\partial w} \qquad (4.4.1.1.)$$

$$b_{t+1} = b_t - \alpha \frac{\partial J(w,b)}{\partial b} \qquad (4.4.1.3.)$$

where α is constant learning rate. If one of $\frac{\partial J(w,b)}{\partial b}$ or $\frac{\partial J(w,b)}{\partial w}$ has relatively small value, it will be updated very slow and make a lot of oscillations before reaching optima. The idea of RMSProp is to speed up the learning process and bringing back the gradient to the steepest direction, by dividing the gradient by an exponentially weighted average of its recent magnitude and adding two new variables S_{dw} and S_{db} to tackle the weighted average dw^2 and db^2:

$$S_{dw} = \beta S_{dw} + (1 - \beta) dw^2 \qquad (4.4.1.4.)$$

$$S_{db} = \beta S_{db} + (1 - \beta) db^2 \qquad (4.4.1.5.)$$

$$w_{t+1} = w_t - \alpha \frac{dw}{\sqrt{S_{dw}}} \qquad (4.4.1.6.)$$

$$b_{t+1} = b_t - \alpha \frac{db}{\sqrt{S_{db}}} \qquad (4.4.1.7.)$$

4.4.2. Hyperparameters

In almost every machine learning method there is a set of parameters that are constant and are defined a priori by user. Search for optimal hyperparameters commonly performed experimentally or by hand with testing for an specific gird point, while there are some tuning techniques for some model that are self- adjusting model parameters[65]. As examples of Hyperparameters, number of hidden layers, number of neurons, learning rate α and decay coefficient λ are relevant in my LSTM model[66].

I used 3 hidden LSTM layer and one Dense output layer within Keras python library. In each LSTM hidden layer, 50 nodes are defined with 20% dropout, where at each layer there is 20% chance to drop each node, Dropout is known as a simple and powerful regularization method in Deep learning field. With dropout, within learning process of network when it is trying to tune neuron weights, it avoids to build a fragile method by a less sensitive reaction of network to weights of specific neurons and allows other neurons to step in.

However, α is the tunning parameter that I use in RMSProp to adjust the gradient step and improve convergence speed. With large values of α it could fail to converge to optima, while choosing small α will converge slowly. While decay factor is set a priori to avoid overfitting my LSTM network, by adding a regularization factor of neuron weights to the cost function. Where it avoids assigning large weights. In my model I consider learning rate of 0.001 and decay factor of 0.9.

5. Integrated Data-driven Pairs-Trading approach

In this chapter, I am going to introduce the data being exerted within this thesis and to test and further investigate my theory. For current study I wrote a Command Line Interface (CLI) application , using python to retrieve stock data through Yahoo Finance API. Further analysis is done from correlation calculation, co-moving groups construction, LSTM method and etc., all of which implemented within the same application. I will start introducing financial data used, and go through all the steps in details to assay my pairs-trading theory. This chapter is organized as below:

1. S&P 500 as case study
2. Building proper co-moving groups using correlation-cointegration approach
3. Pairs-Trading approach with either return or price gaps sorting
4. LSTM Network architecture, training and performance check as a sorting approach
5. Portfolio construction

I will demonstrate, step by step, the implementation of my approach and gather all needed information and structures required to be followed by experimental results. Further empirical analysis will be accomplished in next sections.

5.1. Stock Information

I will investigate my theory over Standard and Poor's (S&P 500) market index, tracking the performance of the largest companies listed on United States stock exchange market. I will propound and test my theory over S&P data between 2010 and 2021 which includes also the most recent economic recession related to COVID-19, which has been detected by expertise and researchers as biggest economic crisis after Great Depression. Where I collect stock profile information from Yahoo Finance API within my CLI application. Nevertheless, I collect yearly S&P500 securities, based on their availabilities in Yahoo Finance API and their changes over

period 2010 and 2021. Figure 5.1.1 represents an example of 10 stock adjusted close prices within 2020.

Within my correlation-cointegration strategy, introduced in previous chapters, I am considering 3 years' time window with rolling one year. By 3 years' time window and one year rolling within 2010 and 2021 period, I will get 8 years of out-of-sample examples to be analyzed within my methodology.

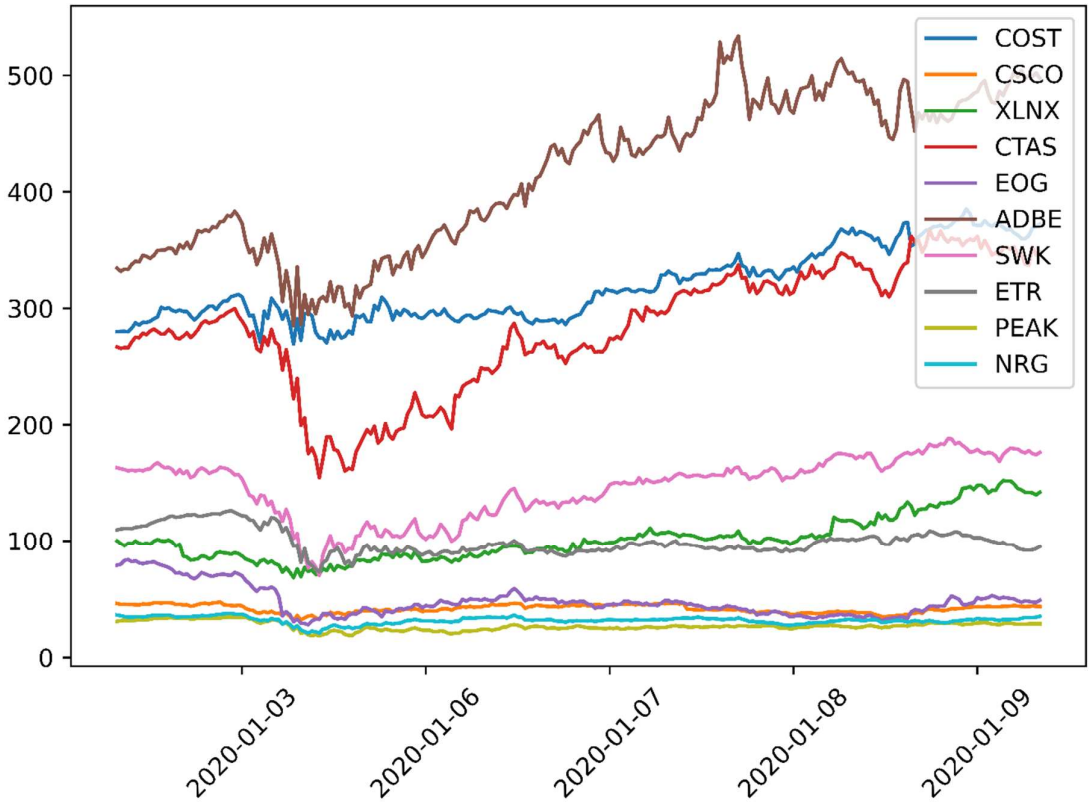


Figure 5.1.1: Stock price e.g. 2020

5.2. Building Co-moving Groups

After exploiting the required stock data, the next step is to find proper co-moving groups. In my technique, I first need to retrieve the Pearson's correlation for the related time-window (Figure 5.2.1: Correlation heat map for 10 stocks within 2020). Correlation as a pre-selection method to capture potential co-moving members, coupled with cointegration, to detect comoving groups.

In cointegration , on January in each year, I look back at the last 3 years historical adjusted-close-prices of S&P stocks in [t-3, t-1] named $S_{i,t}$. Primarily, by calculating the correlation, I pre-select the most correlated securities as potential candidates. For each pairs of $(S_{j,t}, S_{i,t})$ I checked the results of both Engle-Granger and Johansen tests with significance level of 1%. For each stock i, I built the cointegration group $CG_{i,t}$ by testing stock i with respect to all other available stocks j.

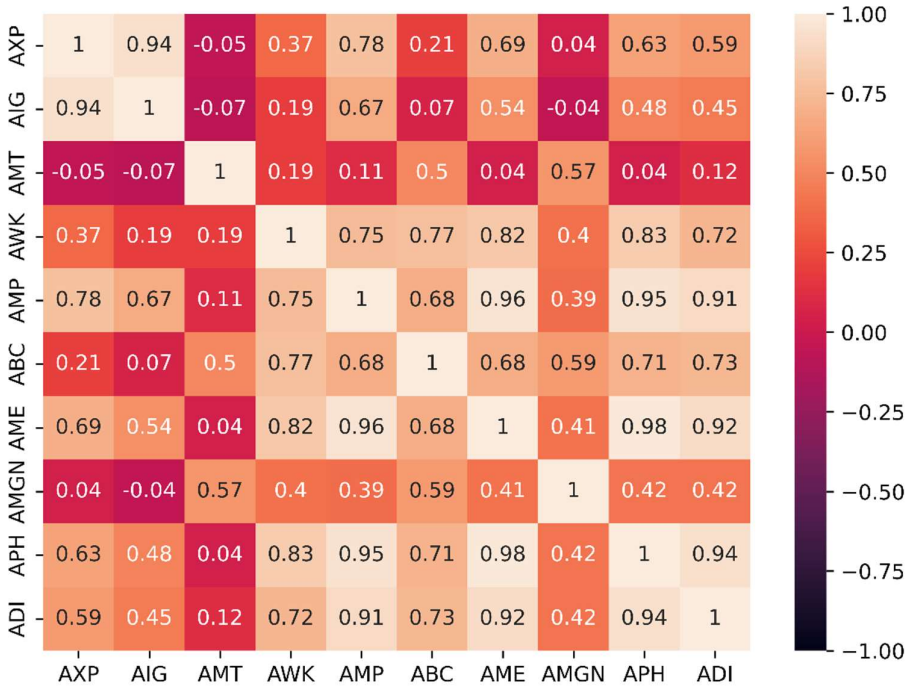


Figure 5.2.1: Correlation heat map for 10 stocks within 2020.

Remind that the cointegration relation is not an asymmetric relation, meaning that if the stock j belongs to $CG_{i,t}$, it does not necessarily mean that the stock i should belong to $CG_{j,t}$.

2010-2013	2011-2014	2012-2015	2013-2016	2014-2017	2015-2018	2016-2019	2017-2020
127	108	96	80	68	58	44	35

5.2.1. Yearly comoving group numbers with more than two securities.

As shown in the Table 5.2.1., it seems that the co-moving groups with more than two securities are ranging from 127 in period 2010-2013 to 35 in period 2017-2020 based on two step correlation-cointegration approaches. Where I set 85% as correlation threshold for detecting candidates with Pearson's correlation, and then deploying both Engle-Granger and Johansen cointegration with 1% threshold. It shows that co-moving groups with more than two securities are following a strictly decreasing pattern between period 2010 till 2020.

5.3. Pairs Trading Approach and Metric Selection

To pursue my pairs-trading strategy, I considered "*price gap*" ($\Delta p_{i,t}$ *and* $\Delta^{\beta} p_{i,t}$) and "*return gap*" ($\Delta r_{i,t}$ *and* $\Delta^{\beta} r_{i,t}$) as measures to investigate the deviation of each stock with respect to its belonging peer as in below formulas (considering the normalized prices at each training time-window to get the same scaling for all securities).

$$price\ gap \begin{cases} \Delta p_{i,t} = p_{i,t} - p_{CGi,t} \\ \Delta^{\beta} p_{i,t} = p_{i,t} - \beta_{p,i} p_{CGi,t} \end{cases} \qquad (5.3.1.1.)$$

$$return\ gap \begin{cases} \Delta r_{i,t} = r_{i,t} - r_{CGi,t} \\ \Delta^{\beta} r_{i,t} = r_{i,t} - \beta_{r,i} r_{CGi,t} \end{cases} \qquad (5.3.1.2.)$$

However, after detecting properly the co-moving stocks within each period, then, I calculated "*price gap*" and "*return gap*" elements with β (regressing formulas 5.3.1.3.) which are calculated using 3-yeard' training data and fed into network as test data.

$$\begin{cases} p_{i,t} = \beta_{p,i} p_{CGi,t} + \vartheta_{i,t}\,, & \vartheta_{i,t} \sim \aleph(0, \sigma_{\vartheta}^2) \\ r_{i,t} = \beta_{r,i} r_{CGi,t} + \tau_{i,t}\,, & \tau_{i,t} \sim \aleph(0, \sigma_{\tau}^2) \end{cases} \qquad (5.3.1.3.)$$

I have used ordering based on "*price gap*" and "*return gap*", with h=1 as investment time horizon and by opening buy position for top decile and sell position for bottom decile, without considering transaction costs. From my results as shown in table 5.3.1., comparing $\Delta p_{i,t}$, $\Delta r_{i,t}$, $\Delta^{\beta} p_{i,t}$ and $\Delta^{\beta} r_{i,t}$, it is observable that $\Delta^{\beta} p_{i,t}$ and $\Delta^{\beta} r_{i,t}$ contributing almost the same results and returns in comparison to $\Delta p_{i,t}$ and $\Delta r_{i,t}$, which was expectable as calculated βs are slightly different from 1 for each cointegration group and each training period.

From Table 5.3.1. I can observe that portfolios built based on return gap (Δr), result in strictly higher returns in comparison to price return, except for year 2018. Where, using the return gap results on average 5% higher return in period 2013-2020 and is a better criteria in comparison to price gap, but other financial factors like *"Maximum Draw Down", "Expected Shortfall", "Sharpe ratio"* and "*Volatility*" need to be investigated for further decisions.

	2013	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
$\Delta \mathbf{p}$	4.46	4.19	-0.77	6.92	9.50	10.13	4.37	0.53
$\Delta \mathbf{r}$	14.98	7.03	14.21	8.89	17.98	7.37	6.49	3.67
$\Delta^{\boldsymbol{\beta}} \boldsymbol{p}$	4.33	3.90	-0.82	6.86	9.23	10.24	4.29	0.54
$\Delta^{\boldsymbol{\beta}} \boldsymbol{r}$	13.82	7.16	14.16	8.50	17.55	7.01	5.98	3.87

Table 5.3.1. portfolio returns based on $\Delta p_{i,t}$, $\Delta r_{i,t}$, $\Delta^{\beta} p_{i,t}$ and $\Delta^{\beta} r_{i,t}$ sorting criteria. (First numbers are related to top decile return and numbers in parenthesis are related to bottom decile portfolios)

I can also conclude that considering the returns using price gap or return gap, it seems that within crisis years of 2013 and 2017, the returns achieved by top-bottom portfolios are not affected, while in COVID-19 year (2020), the returns are low in both criterions. Rather, as in Figure 5.3.1., at years 2013 and 2017, the returns based on the return gap, have higher values. As $\Delta^{\beta} p_{i,t}$ and $\Delta^{\beta} r_{i,t}$ returns are very close respectively to those of $\Delta p_{i,t}$, $\Delta r_{i,t}$, for further analysis I will only consider of $\Delta p_{i,t}$, $\Delta r_{i,t}$ to be compared together and with LSTM criterion in next section.

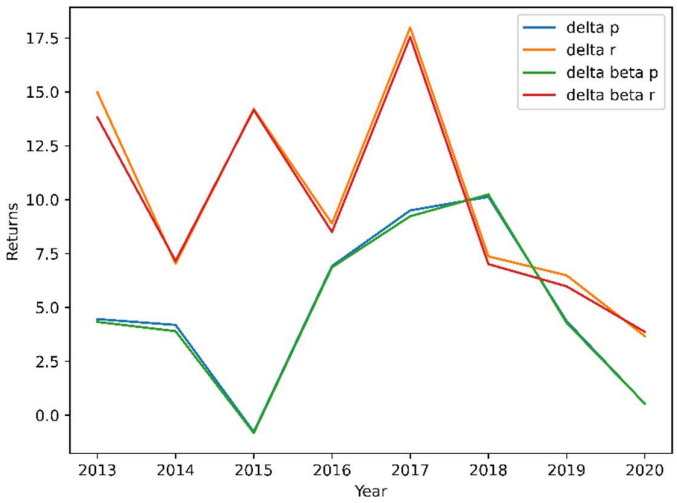


Figure 5.3.1: Comparison of returns based on different criterions

5.4. LSTM Network Architecture

To feed my LSTM network, I need some features to be selected. For each security i, I will consider multivariate time series of Δr_i gap of prices with respect to its peers, V_i as past trading volumes and p_i as past prices. Moreover, the multivariate time series for stock i has 1-year lookback, $\tau = 240$ days, which means for each time-window of 3-years, I collect about 750-240 =510 sequence of features as the first year used to form the first sequence and it is moving daily. The input features at each time θ for 3-year training time window are of the form:

$$\begin{pmatrix} \Delta r_{i,\theta} & \Delta r_{i,(\theta+1)} & \cdots & \Delta r_{i,(\theta-\tau+1)} \\ V_{i,\theta} & V_{i,(\theta+1)} & \cdots & V_{i,(\theta-\tau+1)} \\ p_{i,\theta} & p_{i,(\theta+1)} & \cdots & p_{i,(\theta-\tau+1)} \end{pmatrix}$$

For building LSTM network, I consider 50 nodes for single hidden layer. Furthermore, I used 20% dropout technique to normalize weights assigned to each neuron within network. With dropout, within learning process of network when it is trying to tune neuron weights, it avoid to build a fragile method by a less sensitive reaction of network to weights of specific neurons and allow other neurons to step in.

However, I use cross-entropy as loss function and RMSprop as stochastic descent learning algorithm for minimizing the loss function. To find the related LSTM layers weights, I set the learning rate as 0.001 in RMSprop and 0.9 decay factor. Another approach is to take 10 percent of the training set as validation set, to check the possibility of early stop, in the sense that the training process will be stopped after 10 epoch without any significant amelioration in loss function, to avoid long time running. This approach is similar to pruning roots in Random Forest method, where at each node if no further improvement achieved, it will be automatically pruned.

	2013	2014	2015	2016	2017	2018	2019	2020
LSTM	19.72	14.41	8.28	8.80	7.33	27.71	0.98	3.35

Table 5.4.1. Portfolio returns based on LSTM sorting criteria

From Table 5.4.1. I can observe that during period 2013-2020, using LSTM sorting criteria in top-bottom portfolio construction, it achieves a better performance in comparison to price gap and return gap criterions. Considering to feed LSTM network with volume, price and delta return, it outperforms the other two price gap and return gap investigated earlier. Where in this period, LSTM criteria, achieves 2% higher return with respect to return gap and 7% higher in

comparison to price gap within h=1 investment time horizon (Figure 5.4.1. Comparison of LSTM, return gap and price gap returns).

Figure 5.4.1: Comparison of returns based on price/return gaps and LSTM criterions

As in case of price gap and return gap, I observe also in LSTM within distress periods of financial crisis in years 2013 and 2017, the performance and returns of top-bottom portfolios is not impressed, while in COVID-19 economic recession, my results show poor performance, with lowest returns in 2020. However, with LSTM criteria, the returns experience higher amplitude where it attains highest value of 27.71 and lowest value of 0.98 in two consecutive years of 2018 and 2019.

Moreover, I made further analysis by calculating Standard Deviation of returns (σ), the Sharpe ratio (SR), the Expected Shortfall (ES) and finally the Maximum Draw Down (MDD) for all the investigated sorting criterions within top-bottom strategy (Table 5.4.2.).

		2013	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
Δp	return	**4.46**	**4.19**	**-0.77**	**6.92**	**9.50**	**10.13**	**4.37**	**0.53**
	σ	12.21	13.67	1.1	16.48	9.37	11.7	1.8	2.33
	SR	19.62	6.97	-5.63	22.8	4.24	20.25	138.5	2.17
	ES	8.94	6.28	-2.72	12.41	0.27	11.66	80.93	1.36
	MDD	3.79	4.31	-0.57	4.85	11.44	5.42	1.49	2.32
Δr	*return*	**14.98**	**7.03**	**14.21**	**8.89**	**17.98**	**7.37**	**6.49**	**3.67**
	σ	7.23	7.63	9.93	22.26	14.68	8.45	10.53	6.32
	SR	13.18	76	111.3	28.84	17.86	172.3	32.93	11.45
	ES	123.7	18.51	50.55	12.38	10.75	89.67	10.42	8.26
	MDD	7.59	6.87	13.33	9.64	22.92	4.57	11.85	8.33
LSTM	*return*	**19.72**	**11.84**	**8.28**	**8.80**	**7.33**	**27.71**	**0.98**	**3.35**
	σ	11.84	14.37	10.56	12.86	15.97	18.43	7.2	11.3
	SR	108.3	133.3	86.91	22.38	84.88	173.1	73.77	31.28
	ES	121.7	34.98	38.37	10.02	49.43	102.5	31.44	9.51
	MDD	8.71	6.24	10.70	8.81	24.35	18.71	6.25	4.17

Table 5.4.2. Risk level analysis of different criterions based on Standard Deviation of returns (σ), the Sharpe ratio (SR), the Expected Shortfall (ES) and the Maximum Draw Down (MDD)

Considering the volatility of returns, the values are more steady for return and price gaps, with 9-10% on average within this period, while for LSTM, except for 2019, all volatilities are higher than 10.5%. This pattern shows that the LSTM method based on price, volume and delta return, tends to select more volatile stocks within top-bottom strategy (Figure 5.4.2. Comparison of LSTM, return gap and price gap volatilities).

Figure 5.4.2: Comparison of volatility based on price/return gaps and LSTM criterions

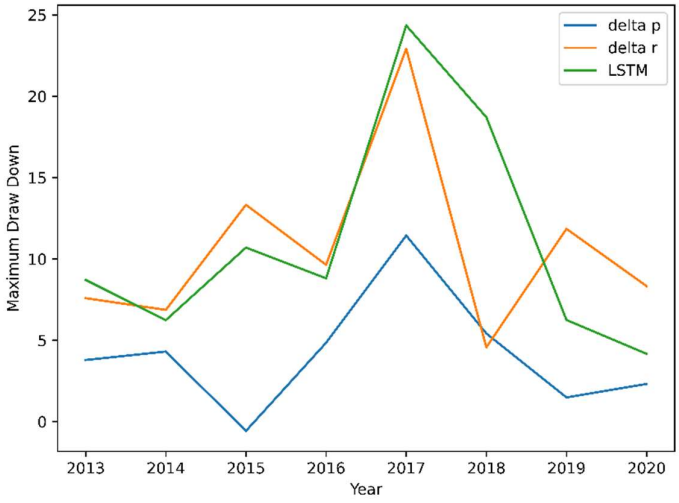


Figure 5.4.3: Comparison of MDD based on price/return gaps and LSTM criterions

Furthermore, analysis of Maximum Draw Down (MDD), as maximum loss observed in each year from a peak of a portfolio to trough, before a new peak attained, shows that regardless of the criterion, within financial crisis years, it reaches the highest value, especially within 2017. In addition, excluding year 2018, on average, LSTM based strategy (9.8%) has lower MDD in comparison to return gap based strategy (11.6 %), while both, show higher maximum draw

downs with respect to price gap based strategy (4 %) (Figure 5.4.3. Comparison of LSTM, return gap and price gap maximum draw downs.).

I also used Sharpe ratio as a measure of risk-adjusted returns, that evaluates my portfolios excess return in terms of risk. A portfolio with greater Sharpe ratio has a better risk-adjusted performance, while negative Sharpe ratio shows that the portfolio's return is expected to be negative. Considering Sharpe ratio, my results demonstrate that LSTM criteria risk adjusted returns, overperform the other two. Instead, Expected Shortfall (ES) or Conditional Value at Risk (CVaR), as measure of risk assessment that quantifies the tale risk investment beyond 99% threshold, is calculated by averaging the extreme losses. From my observations and results, LSTM based method, experiences slightly higher ES in comparison to return gap based method and both, extremely higher ES values with respect to price gap.

6. Conclusion

We know that there have been introduced and examined many pairs-trading approaches that can appropriately capture and predict short future securities prices and behaviors. By the nature of pairs-trading, its needed to detect groups of co-moving stocks and consequently divergence behaviors with respect to their belonging peers. Thanks to recent statistical techniques, live access to huge amount of financial data from different resources and also to modern neural network methods, with aim of fast computers, it facilitates to foresee financial market behaviors in near future. Correlation and cointegration are known as efficient measures that can capture the co-movements behavior of stocks. However, as in financial data signal to noise ratio is low, it needs a reasonable methodology and framework to discover supreme candidates to be placed in pairs-trading approach. Also, undefinable characteristics of financial data which makes it difficult to capture main effective features, stimulates the usage of data-driven modeling. Which comes with the capability to exploit in a good extent, hidden structure of financial data.

Taking advantage of LSTM, as a powerful supervised Machine Learning technique, the non-linear dependencies are captivated. By using LSTM based sorting criteria over co-moving groups of stocks within S&P 500 data, it improve results in comparison to "price gap" or "return gap" techniques for portfolio construction. My results in this thesis revealed that constructing a portfolio based on LSTM sorting criteria, results on average to higher return with respect to price and return gap by 7% and 2%. Although using LSTM based top-bottom strategy, considers to pick more volatile stocks, but not only the maximum draw down is lower in comparison to return gap method, but also better performance in risk assessment measures.

Another achievement in my thesis is that, considering correlation to pre-select possible candidates for group construction, meliorates the strategy. The approach I used in this thesis, is able to not only improve the co-moving group detection with using correlation-cointegration method, but also with LSTM it covers the shortcomings arises by uncertainty and noisy financial data that makes it difficult to capture hidden structure laid behind.

Concluding that the LSTM has a better performance in pairs-trading approaches, beside correlation pre-selection that improved co-moving groups. Another interesting result of thesis is the effect of COVID-19 recession as a distress period. The methods used in this thesis, show

that within distress period of financial crisis in years 2013 and 2017, the returns and results are not impressed, while in COVID-19 year of 2020, they have a poor performance.

However, further researches needed to assess the performance of this approach on stock data combined with “Alternative Data”. Alternative data and NLP as a tool to make investment decision and portfolio construction, is a source of signal generation. Through alternative data, strong signals with high financial impacts, if being properly analyzed, improves signal to ratio and consequently with employing sequence to sequence LSTM , it is expected to augment future price forecasts and pairs selection. Worth to mention that one could also consider to amend signal to ratio by utilizing denoise and detone technique by fitting Kernel Density Estimation (KDE) to stock data, and the Marcenko-Pastur probability density function that also can remove Market Component. Market component may cover useful signal part, and high market component can make pairs-trading approaches to struggle while finding comoving securities.

By increasing the accessibility of financial data and developing modern statistical techniques as Artificial Intelligence, there exist a high potential to attain higher accuracy in short term stock predictions. Considering that 90% of the data produced worldwide was created and became accessible within past few years. Also there is a potential to engage methodologies that are able to recover hidden information and features within high dimensional data to be subject of financial analysis.

7. Bibliography

[1] H. Markowitz, "Portfolio Selection", The Journal of Finance, 1952.

[2] R. C. Merton, "On Estimating the Expected Return on the Market: an Exploratory Investigation", National Bureau of Economic Research, 1980.

[3] J. D. Jobson, B. Korkie, "Estimation for Markowitz Efficient Portfolios", Journal of the American Statistical Association, 1980.

[4] R. O. Michaud, R. O. Michaud, " Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation ", Oxford University Press, 2008.

[5] V. DeMiguel, L. Garlappi, R. Uppal, " Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? ", Oxfor University Press, 2007.

[6] H. Konno, H. Yamazaki, "Mean-Absolute Deviation Portfolio Optimization Model and Its Application to Tokyo Stock Market",Management Science, 1991.

[7] W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk", The Journal of Finance, 1964.

[8] M. Rossi , "The Capital Asset Pricing Model: a Critical Literature Review.", Global Business and Economics Review, 2016.

[9] R. T. Rockafellar, S. Uryasev, "Optimization of Conditional Value-at-Risk", Research Report, Center for Applied Optimization, University of Florida, 2003.

[10] M. R. Young,"A Minimax Portfolio Selection Rule with Linear Programming Solution ", Management Science , 1998

[11] E. F. Fama, "The behavior of stock-market prices". The Journal of Business, (1965) .

[12] N. Jegadeesh, S. Titman, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency", The Journal of Finance, 1993.

[13] S. Cheng, A. Hameed , A. Subrahmanyam, S. Titman, "Short-term reversals: The effects of past returns and institutional exits". Journal of Financial and Quantitative Analysis, 2017.

[14] D. Avramov, T. Chordia , A. Goyal, "Liquidity and autocorrelations in indi- vidual stock returns". The Journal of Finance, (2006).

[15] R.F. Stambaugh , "Predicting Returns in the Stock and Bond Markets." Journal of Financial Economics17, 1986

[16] A.W. Lo, "Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test.", The Review of Financial Studies(1), 1988.

[17] D. W. Blackburn, N. Cakici, "Overreaction and the cross-section of returns: International evidence", Journal of Empirical Finance, 2017.

[18] M. Cremers, A. Pareek, "Short-Term Trading and Stock Return Anomalies: Momentum, reversal, and Share Issuance", Review of Finance, 2014

[19] D. Stattman, “Book Values and Stock Returns.”, The Chicago MBA: A Journal of Selected Papers. 1980.

[20] J. Lintner, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets", The Review of Economics and Statistics., 1965.

[21] J. Tobin, "Liquidity preference as behavior towards risk.", Review of Economic Studies, 1958.

[22] G.W. Douglas, "Risk in the equity markets: an empirical appraisal of market efficiency", Yale Economic Essays, 1969.

[23] O. Ledoit, and M. Wolf "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." Journal of Multivariate Analysis, 2004.

[24] Miller, Scholes, "Rates of return in relation to risk: a reexamination of some recent findings", Studies in the Theory of Capital Markets, 1972.

[25] F. Black, M. Jensen, M. Scholes, "The capital asset pricing model: some empirical tests", Studies in the Theory of Capital Markets, 1972.

[26] Basu, Sanjay. "Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis." Journal of Finance. 1977.

[27] Banz, W. Rolf 1981. "The Relationship Between Return and Market Value of Common Stocks.", Journal of Financial Economics. 1981.

[28] Bhandari, Laxmi Chand. "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence." Journal of Finance. 1988.

[29] D. Stattman, "Book Values and Stock Returns." The Chicago MBA: A Journal of Selected Papers. 1980.

[30] R. Barr, K. Reid, R. Lanstein. "Persuasive Evidence of Market Inefficiency. ", Journal of Portfolio Management. 1985.

[31] F. E. Fama, K. R. French. "Common Risk Factors in the Returns on Stocks and Bonds." Journal of Financial Economics. 1993.

[32] M. M. Carhart, "On persistence in mutual fund performance." The Journal of Finance, 1997

[33] E. F. Fama, K. R. French, "A five-factor asset pricing model." Journal of Financial Economics, (2015)..

[34] R. N. Mantegna, "Hierarchical Structure in Financial Markets ", The European Physical Journal B., 1999.

[35] M. M. Lopez de Prado, "Machine Learning for Asset Managers", Elements in Quantitative Finance, Cambridge University Press. 2020.

[36] M. S. Perlin, "Evaluation of Pairs-trading Strategy at the Brazilian Financial Market", Journal of Derivatives and Hedge Funds, 2008.

[37] J. Wang , "Pairs Trading with Robust Correlation." University of British Columbia, May 2009.

[38] J. Hauke, T. Kossowski , "Comparison of Values of Pearson's and Spearsman's Correlation Coefficient on the same Sets of Data", 2011.

[39] D. Wen , C. Ma, G. Wang, S. Wang, "Investigating the Features of Pairs Trading Strategy: A Network Perspective on the Chinese Stock Market.", 2018

[40] J. P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, A. Kanto,"Asset Trees and Asset Graphs in Financial Market", Physica Scripta, 2003.

[41] J. P. Ramos-Requena, J. E. Trinidad-Sevogia, M. A. Sanchez-Granero, "Some Notes on the Formation of a Pair in Pairs Trading.", Mathematics, 2020.

[42] D. Snow, "Machine Learning in Asset Management.", FirmAI, 2019.

[43] G. Y. Ban, N. El Karoui, A. E.B. Lim, "Machine Learning and Portfolio Optimization", Managemenet Science, 2016.

[44] W. F. Sharpe, "The Sharpe Ratio", Journal of Portfolio Management, 1994.

[45] A. Hedayati Moghaddam, M. Hedayati Moghaddam, M. Esfandyari, "Stock Market Index Prediction Using Artificial Neural Network", Journal of Econometrics, Financial and Administrative Science 21, 2016.

[46] V. N. Vapnik, "The Nature of Statistical Learning Theory.", Springer NY, 2013.

[47] O. Hegazy, O. S. Soliman, M. Abdul Salam, "A Machine Learning Model for Stock Market Prediction.", International Journal of Computer Science and Telecommunications, 2013.

[48] G. Zhigiang, W. Huaiqing, L. Quan, "Financial Time Series Forecasting Using LLP and SVM Optimized by PSO", Springer. 2008.

[49] Z. Zhang, S, Zohren, S. Roberts, "Deep Learning for Portfolio Optimization.", University of Oxford, 2020.

[50] S. Hochreiter, J. Schmidhuber. "Long short-term memory" Neural computation, 1997

[51] C. Krauss, "Statistical arbitrage pairs trading strategies: Review and outlook." Journal of Economic Surveys, 2017.

[52] E. Gatev, W. N. Goetzmann, K.G. Rouwenhorst, "Pairs trading: Performance of a relative-value arbitrage rule Review of Financial Studies, 2006.

[53] S. Mudchanatongsuk, J. A. Primbs, W. Wong, "Optimal Pairs Trading: A Stochastic Control approach," Proceeding of the American Control Conference, 2008.

[54] R. F. Engle, C. W. Granger, "Co-integration and error correction: Represen- tation, estimation, and testing.", Econometrica: Journal of the Econometric Society, 1987.

[55] S. Johansen, "Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models." Econometrica: Journal of the Econometric Society , 1991.

[56] F. Bilgili, "Stationarity and cointegration tests: Comparison of Engle - Granger and Johansen methodologies", Erciyes University, Faculty of Economics and Administrative Sciences, 1998.

[57] H. Chen, S. Chen, Z. Chen, F. Li,. "Empirical investigation of an equity pairs trading strategy." Management Science, 2017.

[58] A. Flori, D. Regoli, "Revealing pairs trading opportunities with long short-term memory networks", European Journal of Operational Research, 2021.

[59] D. Blitz, J. Huij, S. Lansdorp, M. Verbeek, "Short-term residual reversal." Journal of Financial Markets, 2013.

[60] C. Bishop, P. Bishop, G. Hinton, O. U. Press, "Neural Networks for Pattern Recognition." Advanced Texts in Econometrics, Clarendon Press,1995.

[61] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, 2015.

[62] Andrew Ng. , "Neural Networks and Deep Learning.", (accessed July 2, 2020).

[63] J. Chung, C. Gulcehre, K. H.Cho, Y. Benglo, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.", In NIPS 2014 Workshop on Deep Learning, December 2014

[64] J. Schmidhuber , S. Hochreiter, "long short term memory." Neural Computation , 1997.

[65] M. Claesen and B. De Moor, “Hyperparameter search in machine learning,” 2015.

[66] A. Krogh and J. Hertz, “A simple weight decay can improve generalization.” Advances in Neural Information Processing Systems 4, 1991.

[67] D. Delpini, S. Battiston, G. Caldarelli, M. Riccaboni, "Portfolio Diversification, Differentiation and the Robustness of Holdings Networks", Applied Network Science, 2020.

[68] J. D. Curto, P. Serrasqueiro, "The impact of COVID-19 on S&P500 sector indices and FATANG stocks volatility: expanded APARCH model", Finance Research Letters, 2021.