Project on

# STATISTICAL ARBITRAGE

Submitted by

PRATIK. P. PORE

Internship at

ECKOVATION CAREERS

Course Name

MACHINE LEARNING

# INDEX


| Sr. No. | Content | Page No. |
|---|---|---|
| 1 | Abstract | 3 |
| 2 | Introduction | 4 |
| 3 | Reviews of Definition | 6 |
| 4 | Literature Review Of Strategies | 9 |
| 5 | What is SA? | 12 |
| 6 | Theil-Sen Regression | 14 |
| 7 | Coding the Strategy | 15 |
| 8 | Conclusion | 19 |
| 9 | References | 19 |

# ABSTRACT

Statistical Arbitrage (SA) is a common financial term. However, there is no common definition in the literature while investors use the expression SA for a variety of different strategies. So, what is SA? In order to answer this question, we investigate SA strategies across equity, fixed income and commodity. The analysis of strategies key features indicates that no existing definition fully describes them. To bridge this gap, we identify a general definition and propose a classification system that encompasses the current forms of SA strategies while facilitating the inclusion of new types as they emerge.

We investigate statistical arbitrage strategies when there is ambiguity about the underlying time-discrete financial model. Pricing measures are assumed to be martingale measures calibrated to prices of liquidly traded options, whereas the set of admissible physical measures is not necessarily implied from market data. Our investigations rely on the mathematical characterization of statistical arbitrage. In contrast to pure arbitrage strategies, statistical arbitrage strategies are not entirely risk-free, but the notion allows to identify strategies which are profitable on average, given the outcome of a specific σ-algebra. Besides a characterization of robust statistical arbitrage, we also provide a super-/sub-replication theorem for the construction of statistical arbitrage strategies based on path-dependent options. In particular, we show that the range of statistical arbitrage-free prices is, in general, much tighter than the range of arbitrage-free prices.

# INTRODUCTION

The technique of statistical arbitrage is the systematic exploitation of perceived mispricings of similar assets. A trading strategy built around statistical arbitrage involves three fundamental pillars:

(1) a measure of similarity of assets,

(2) a measure of pricing mismatch, and

(3) a confidence metric for each mismatch.

Traditional statistical arbitrage techniques, like “Pairs Trading”, employ these three pillars, holding long-short positions in a pair of strongly “similar” assets. The covariance (or correlation) between two assets is a widely used metric of their “similarity”. However, recent studies have used additional attributes such as co-integration tests to select asset pairs that are better suited for a statistical arbitrage trading strategy.

The concept of arbitrage is fundamental in financial literature and has been used in classical analysis of market efficiency, whereby arbitrage opportunities are quickly exploited by investors. However, pure arbitrage opportunities are unlikely to exist in a real trading environment. An arbitrageur typically engages in a trade that involves some risks. In the specific case where these risks are statistically assessed, then it is appropriate to use the term statistical arbitrage (SA). SA has been broadly investigated in literature, however, scholars either focus on definitions or on developing and testing investment strategies, while we are not aware of any attempt to reconcile these two areas of research. On the one hand, several studies introduce definitions extending the concept of arbitrage through statistics but with little emphasis on strategies. On the other hand, research on statistically determined arbitrage strategies focuses on models and investment opportunities with little or no discussion on definitions and theoretical framework. This leads us to our research question. Model-independent and robust finance aims, in particular, at calculating arbitrage-free price bounds for exotic derivatives, such that these bounds do not depend on any specific model assumptions. This approach, however, lacks applicability, since the price bounds turn out to be rather broad. In this paper, we instead consider a weaker notion of arbitrage, the so-called G-arbitrage, referring to strategies which are profitable on average, given information contained in the σ-algebra G. Further, we show how G-arbitrage strategies can be explicitly constructed based on path-dependent options.

What is SA? This paper addresses this question with an in-depth investigation of SA. We begin by reviewing existing definitions of arbitrage, which are reduced to a common framework to analyze and compare them. We survey statistically determined arbitrage strategies analyzing both the academic and financial industry research. In total, we review 165 articles on the subject, published between 1995 and 2016. Particular attention is paid to hedge funds techniques, market

neutral investment strategies and algorithmic trading. The strategies are discussed in a standardized way analyzing equity, fixed income and, for the first time, commodity. We find that these strategies show significant similarities and common features that define them. The comparison of theoretical definitions and strategies' key features indicates that no available definition appropriately describes SA strategies. To bridge this gap, we propose a general definition, which more closely reflects investors strategies. In addition, we suggest that, instead of searching for a definitive theoretical definition of SA, scholars should instead agree on a classification system that encompasses the current forms of SA while facilitating the inclusion of new types as they emerge. We propose a simple system for classifying strategies that takes into account the strategies risk and return profile. We illustrate the advantages of this approach by demonstrating how it can guide theoretical development and empirical testing. We also provide examples of potential future research directions.

We build on the above outlined theory but replace the definition of arbitrage by the weaker notion of G-arbitrage. The definition of G-arbitrage for self-financing trading strategies which was defined in assumes a fixed physical measure P as well as a specific risk-neutral measure Q used for pricing. We consider a model-independent and robust version of G-arbitrage in the following sense. First, and in line with the model-independent setting, all martingale measures calibrated to option prices are considered as potential pricing measures. Second, instead of assuming a fixed real-world measure, we allow for an arbitrary set P of admissible physical measures. In this way, we extend the notion of G-arbitrage to a setting which is model independent w.r.t. Q and robust w.r.t. P. We refer to corresponding strategies consisting of self-financing positions in the underlying security and static positions in liquid options as Probust G-arbitrage strategies. A frequently considered special case of G-arbitrage is statistical arbitrage. Corresponding strategies are especially important when it comes to applications as they induce an average gain regardless of the outcome of the underlying security at terminal time. In practice, the term statistical arbitrage is often associated with pairs trading strategies which build on cointegrated pairs of assets. These strategies, however, do not fall within the rigorous definition of statistical arbitrage, which is also in line with our framework.

The paper is organized as follows.
In **Section 2**, we review existing definitions of SA producing a comprehensive mapping.
In **Section 3**, we report a survey of statistically determined arbitrage strategies.
In **Section 4**, we identify the key features which are common to the various strategies. We combine the findings of the previous sections and propose a general definition and classification system.
**Section 5** concludes the paper.

# Review of Definitions

It is commonly accepted that Statistical Arbitrage (SA) started with Nunzio Tartaglia who, in the mid-1980s, assembled a team of quantitative analysts at Morgan Stanley to uncover statistical mispricing in equity markets. However, SA came to the fore as a result of Long-Term Capital Management (LTCM), a hedge fund founded in 1994, where Nobel Prize winners Sholes and Merton both worked. The company developed complex SA strategies for fixed income which were initially extremely successful. However, in 1998, as a result of the financial crises in East Asia and Russia, LTCM's arbitrage strategies started producing large losses which endangered global markets and forced the Federal Reserve Bank of New York to organize a bailout in order to avoid a wider financial collapse. Nevertheless, SA continued to grow in popularity with applications progressively expanding to all asset classes. SA has become one of the main investment strategies in investment banks and mostly for hedge funds. In particular, the term SA is used to denote hedge funds that aim to exploit pricing anomalies in equity markets. Technological developments in computational modelling have also facilitated the use of SA in high frequency trading and with the so-called machine learning methods, such as neural networks and genetic algorithms. In more recent years, SA has seen renewed interest in emerging areas such as bitcoin, big data and factor investing.

The literature on the limits of arbitrage is quite broad and provides some insights on why SA opportunities exist. Mou reports how arbitrageurs have to face three different types of risks: fundamental risk, noise trader risk and synchronization risk. Duffie describes the risks arising from inattentive investors. Finally, behavioral effects can generate additional risk and asset bubbles. On the one hand, these risks create SA opportunities. On the other hand, the same risks can undermine arbitrageurs' efforts and cause delays in correcting market anomalies.In this section, we review all definitions of arbitrage available in literature which may be suitable to define SA. Our analysis encompasses both alternative definitions of arbitrage as well as definitions of statistical arbitrage. Before reviewing the various definitions, we briefly recall the four types of definitions that are commonly used:

1) lexical
2) conceptual
3) abstract and
4) operational.

Lexical definitions use simple terms for a wide audience. Conceptual definitions describe a concept in a way that is compatible with a measurable occurrence. Abstract definitions are used when the meaning cannot be measured empirically. Finally, operational definitions provide a clear and concise meaning of a concept in a way that can be measured. Operational definitions

clearly specify the object and criteria of measurement which makes them particularly suitable for scientific investigation. We find that existing definitions can be categorized as lexical, conceptual or operational while there are no abstract definitions.

### 2.1 Lexical Definition of SA

Some lexical definitions tend to be vague and lack formalism because traders, for good commercial reasons, tend to be obscure about their investment methods. Pole for example writes that SA uses mathematical models to generate returns from systematic movements in securities prices. According to Avellaneda and Lee, the term statistical arbitrage encompasses a variety of strategies characterized by systematic trading signals, market neutral trades and statistical methods. Montana defines SA as an investment strategy that exploits patterns detected in financial data streams. Burgess defines statistical arbitrage as a framework for identifying, modelling and exploiting small but consistent regularities in asset price dynamics. Other definitions are centered on the concept of mispricing. Thomaidis and Kondakis define SA as an attempt to profit from pricing discrepancies that appear in a group of assets. Do, Faff and Hamza claim that SA is an equity trading strategy that employs time series methods to identify relative mispricings between stocks. Burgess also describes statistical arbitrage as a generalization of a traditional arbitrage where mispricing is statistically determined through replicating strategies. In using derivatives, Zapart describes statistical arbitrage as an investment opportunity when perfect hedging is not possible. A general definition of SA strategy should describe what SA is and its objectives. We find instead that some definitions focus on specific implementations and techniques. In particular, in a broad range of papers, SA is associated with pairs trading and cointegration.

### 2.2 Conceptual Definitions of SA

Another set of definitions can be classified as conceptual as they can be associated with specific measures. In reviewing Hedge Funds (HFs) strategies, Connor and Lasarte use the probability of a loss in defining SA as a zero-cost portfolio where the probability of a negative payoff is very small but not exactly zero. Stefanini uses the expected value in noting that SA seeks to capture imbalances in expected value of financial instruments, while trying to be market neutral. For Saks and Maringer, SA accepts negative payoffs as long as the expected positive payoffs are high enough and the probability of losses is small enough. Focardi, Fabozzi and Mitov focus on uncorrelated returns reporting that SA strategies aim to produce positive, low-volatility returns that are uncorrelated with market returns.

## 2.3 Operational Definition of SA

We next discuss the various extensions of arbitrage available in the literature that are used mainly in asset pricing. All definitions can be classified as operational and are mathematically formulated. Here, we provide a description of the various arbitrages while we refer to the relative papers for a more rigorous formulation. We first introduce the classical definition of arbitrage, defined as a zero-cost trading strategy with positive expected payoff and no possibility of a loss. The absence of arbitrage is a necessary condition for equilibrium models, however this condition alone is often too weak to be practically useful for certain applications such as option pricing. A first attempt to provide a new definition of arbitrage is made by Ledoit who defines δ-Arbitrage (δA) using the Sharpe ratio. Ledoit defines δA as an investment strategy having a Sharpe ratio above a constant and strictly positive level δ. They define a strategy as a Good Deal (GD) if its market price lies outside the range of plausible prices as determined by the various discount factors. Bernardo and Ledoit introduce the Approximate Arbitrage (AA) as they note that the Sharpe ratio is not a good measure of the attractiveness of an investment opportunity. If returns are not normally distributed strategies can have arbitrarily low Sharpe ratios, hence the introduction of a gain-loss ratio. AA is defined as an investment strategy whose maximum gain-loss ratio is above a predefined constant value greater than one. Instead of using the Sharpe ratio or the gain-loss ratio, Carr, Geman and Madan base their definition of Acceptable Opportunity (AO) on two distinct sets of probability measures (valuation and stress measures). AO is defined as an investment strategy having a non-negative expected value under each valuation measure and losses capped under a set of stress measures. In other words, AO is an investment opportunity acceptable to a wide variety of reasonable individuals as it has expected non-negative payoff with losses capped under probability measures reflecting stressed conditions (stress measures). Bertsimas, Kogam and Introduce $\varepsilon\varepsilon$-Arbitrage (εA) referring to replication strategies for derivatives. An εA occurs whenever the price of a derivative significantly differs from the least costly optimal replication strategy. In the literature, there are two definitions of Statistical Arbitrage (SA) which differ significantly from each other. Bondarenko's SA is a trading strategy which can have negative payoffs, as long as the average payoff is non-negative for a given augmented information set. Key in the definition is the introduction of the augmented information set, which, in addition to the market information at time t, also includes the knowledge of the final price. Hogan's SA is a long horizon trading opportunity that, at the limit, generates a risk-less profit. According to this definition SA satisfies four conditions:

1) it is a zero-cost, self-financing strategy, that in the limit has
2) positive expected discounted payoff
3) a probability of a loss converging to zero, and
4) a time averaged variance converging to zero if the probability of a loss does not become zero in finite time.

# Literature Review of Strategies

The existing literature on SA includes a small number of reviews of arbitrage strategies which cover only single asset classes. In fixed income, Duarte,Longstaff and Yu conduct an analysis of the risk and return characteristics of the most widely-used fixed income arbitrage strategies. In equity, Do, Faff and Hamza analyze different approaches to pairs trading: distance approach, cointegration approach, stochastic spread approach and stochastic residual spread approach. Again, focusing on equities, Pole elaborates on pairs trading as well as statistical models for time series analysis. There are no reviews for commodities, where studies primarily focus on modelling spreads and term structures for single commodities. In our review, for the first time, we look at SA across all asset classes to identify common features and defining elements. We review the existing literature on statistically determined arbitrage strategies and, particularly, on those labelled as SA.

| SA strategy | Equities | Bonds | Commodities | Volatility | FX | Mix | Total |
|---|---|---|---|---|---|---|---|
| Pairs trading | 103 | | 6 | | 1 | 2 | 112 |
| Capital structure arbitrage | | 30 | | | | | 30 |
| Volatility arbitrage | | | | 9 | | | 9 |
| Term structure arbitrage | 1 | 4 | 3 | | | | 8 |
| Swap spread arbitrage | | 3 | | | | | 3 |
| Mortgage arbitrage | | 3 | | | | | 3 |
| **Total** | **104** | **40** | **9** | **9** | **1** | **2** | **165** |

Ornstein-Uhlenbeck is a model used to describe the multivariate dynamics of financial variables.

We identify 165 articles in literature discussing SA strategies spanning from 1995 to 2016. The surveyed studies focus on equities], followed by bonds while other asset classes appear only in a small number of articles: commodities, volatility and FX. Just two articles discuss pairs trading across asset classes (mix): investment grade credit default swaps versus equity and gold miners versus gold.

We categorize the various strategies based on the classification proposed by Duarte, Longstaff and Yu who identify five different types of SA strategies in fixed income:

1) swap arbitrage strategies,
2) term structure arbitrage (or yield curve arbitrage),
3) mortgage arbitrage,
4) volatility arbitrage and
5) capital structure arbitrage.

We add equity pairs trading to the classification for fixed income of Duarte, Longstaff and Yu. The term SA is used very frequently in particular in relation to pairs trading which includes pairs trading between indices, ETFs and spread trading between commodities. Various articles focus on cointegration, the Ornstein-Uhlenbeck stochastic process and, more recently, high frequency trading. Pairs trading is predominantly an equity strategy. Capital structure arbitrage is the second most documented strategy which includes primarily convertible arbitrage strategies. Term structure strategies are documented only in eight studies of which four analyze bonds. Swap spread arbitrage and mortgage arbitrage are discussed in three studies each.

### 3.1 Review of Strategies

We next describe the six identified trading strategies. Pairs trading is a SA strategy which is particularly popular in equity. In its simplest formulation, pairs trading aims to identify pairs of stocks whose prices have historically moved together. When the spread between the two components of the pair significantly widens, the strategy sells the best performing security to buy the laggard. If the spread reverts to the mean the trade will be profitable regardless of market trends. This strategy relies on the assumption of a (long-term) equilibrium in the investigated spreads which can be detected through a variety of statistical methods. Long and short positions can be combined in a ratio which makes the trade market-neutral (with a neutral beta position versus the market) or dollar-neutral. The use of pairs trading is not limited to stocks. There are applications to other areas such as spreads between different commodities, commodity future contracts and freight markets. Pairs trading can also be used to model the spread between different portfolios. Term structure arbitrage is a common SA strategy which typically involves taking market-neutral long-short positions at different points of a term structure as suggested by a relative value analysis. Positions are held until the trade converges and the mispricing disappears. Term structure arbitrage is particularly common in fixed income (also called yield curve arbitrage) and commodities. In spite of being one of the most common SA strategies, the literature on implementations of yield curve arbitrage is quite limited and mostly focuses on interest rates models. Term structure arbitrage in commodities uses models (similar to the one used in rates) to identify relative value opportunities across the curve. An implementation of term structure arbitrage in commodities is described by Mou who identifies investment

opportunities arising from the futures rolling of the main commodity indices. In credit, SA opportunities in the term structure of CDS are studied by Jarrow, Li and Ye. Volatility arbitrage is a popular and widely used strategy. Its implementations are structured to be pure bets on volatility and should not be influenced by the actual direction of the underlying. Similarly to other types of arbitrage, volatility arbitrage refers to a wide range of different strategies which can be classified into:

1) gamma trading,
2) volatility surface arbitrage,
3) cross asset volatility trading and
4) dispersion trading.

Gamma trading plays the implied volatility versus the historical volatility on the same asset. If the realized volatility exceeds the volatility implied in the option price, arbitrageurs can profit by buying an option and hedging the delta in the underlying market. The positive income is proportional to where S is the price of the underlying, and Γ is the gamma of the option. Volatility surface arbitrage is a relative value strategy trading the implied volatilities on the same underlying in different points of the volatility surface. Arbitrageurs identify anomalies in implied volatilities across different strike prices and maturities and profit from buying (selling) options whose implied volatility is excessively low. Cross-asset volatility trading plays the implied volatility of an asset versus the implied volatility of another asset through traditional long-short trades. Finally, dispersion trading (also known as decorrelation trading) trades the volatility of a basket of securities (generally and index) against the volatilities of the components of the same basket. The volatility of an index is a function of the volatilities of the constituents and the correlations between them. Greater correlations translate into less diversification and higher index volatility. Decorrelation is traded by selling index variance swaps and buying single stock variance swaps. Swap spread arbitrage is another popular fixed income strategy which bets on the difference between a fixed and a floating yield. It is structured in two parts. Entering this part of the trade the arbitrageur earns the treasury rate TR and pays the repo rate tr . The overall cash flow of the trade is the fixed interest rate component (also known as swap spread) and L r t t − is the floating rate part which needs to be rolled periodically (typically every three months). The strategy generates a positive income as long as the floating yield exceeds the fixed one. Swap spread arbitrage is immune from interest rate risk if both the repo rate and LIBOR (which generally have the same maturity and rolling dates) react similarly to a move in rates. Mortgage arbitrage consists of buying mortgage-backed securities (MBSs) while hedging their interest rate exposure primarily through derivatives. The strategy provides a positive carry as the yield on MBSs is typically higher than that of comparable treasury bonds. As the spread earned is generally small, arbitrageurs use leverage to enhance returns. Mortgage arbitrage strategies can be classified based on the different types of MBS used. A popular implementation of the strategy is with pass-through MBSs which pass all of the interest and principal cash flows of a pool o

mortgages to the pass-through investors. Capital structure arbitrage involves taking long and short positions in the various instruments of a company's capital structure. This includes a variety of strategies between equity, debt and credit instruments of a given company. Some of the most popular strategies are credit arbitrage and convertible arbitrage. Credit arbitrage (also known as capital structure arbitrage) usually refers to strategies that aim to exploit mispricing between a company's credit default swap (CDS) and its equity. Arbitrageurs use the information on the equity price and the capital structure of an obligor to compute its theoretical CDS spread. The theoretical CDS is then compared with the level quoted in the market. If the market spread is higher (lower) than the theoretical spread, then the strategy goes short (long) on the CDS contract while simultaneously hedging the equity with a short (long) position. Convertible Arbitrage is one of the most popular capital structure strategies and involves buying a portfolio of convertible bonds while selling short the underlying stocks. Intuitively, if the stock increases in price, the bonds will appreciate and if the stock falls the short position will profit. In some versions, the interest rate risk is hedged with treasury futures or interest rate swaps. In addition to credit arbitrage and convertible arbitrage, other capital structure arbitrage strategies focus on the spread between bonds and equities of the same company. In particular Schaefer and Strebulaev show that structural models provide accurate predictions of the sensitivity of corporate bond returns to changes in the value of equity (hedge ratios). Other strategies instead focus on the spread between CDS and corporate bonds or different types of credit default.

### 3.2 What is SA?

In this section, we define SA strategies. We identify those features which are common to the surveyed arbitrage strategies. We compare them with the available definitions and provide a new definition in conjunction with a classification scheme. The new definition incorporates all strategies' key elements and the classification scheme encompasses the important dimensions of SA while being flexible and easy to use.

If we can make consistent small profits on a small time interval, we can theoretically generate large profits from the accrued small returns. We obtain strong signals, however we find that we cannot easily profit them. The predominant issue is that the expected returns of each trade in the high frequency domain is on the same order as the spreads of the returns (more on this in section 6). In addition, there is a large trade-off between quick order executions and spread-associated costs. Thus, without the guarantee that we trade at the desired price given by the signal, we will lose money on the trade. Furthermore, there is a minimum tick size, which is too large for our strategy to make a profit on. Even if we could overcome these issues, we would need to generate substantial profits from each trade due to transaction costs.

| Strategy | Descriptions |
|---|---|
| Pairs trading | Plays mean reversion in the spreads of two securities |
| Term structure arbitrage | Takes long-short positions across the term structure |
| Volatility arbitrage | Plays the spread of implied vs. realized volatility of the same security or implied vs. implied volatility of the same or different securities |
| Swap spread arbitrage | Profits from the spread between a fix and a floating leg by entering a short (long) Treasury position and simultaneously buying (selling) an IRS |
| Mortgage arbitrage | Buys MBS hedging the interest rates exposure |
| Capital structure arbitrage | Takes long-short positions on different instruments of a company (credit arbitrage and convertible arbitrage) |

Volatility arbitrage identifies relative value opportunities between volatilities. Swap spread plays a fixed spread versus a floating spread. Mortgage arbitrage models the spread of MBS over treasury. Capital structure arbitrage profits from the spread between various instruments of the same company. Spreads trading involves taking long-short positions in order to profit from spreads or simply to bet on a security while being market-neutral. Not all strategies guarantee gains but rather offer positive expected excess returns with an acceptably small potential loss. Arbitrageurs require a positive expected excess return over the risk free to compensate for risk. The potential loss must be acceptably small in order to qualify the strategy as arbitrage rather than simple investment. Although not all the academic literature reports it, trades always have taken profit and stop loss features. The take profit identifies when a trade no longer offers positive expected excess returns. A take profit is triggered in case there is reversion to the mean (pairs trading, term structure arbitrage, volatility arbitrage and capital structure arbitrage) or when the positive carry disappears (swap spread arbitrage and mortgage arbitrage). The stop loss quantifies when a loss is no longer acceptably small and results from investors' risk tolerance.

From the previous analysis, it is possible to conclude that three key factors define statistically determined arbitrage opportunities:
1) relative value,
2) positive expected excess returns and
3) acceptably small potential loss.

Take profit and stop loss are features which enable to operationalize SA strategies.

# THEIL-SEN REGRESSION

The Theil-Sen estimator is an exceptionally simple and robust linear regression estimator, affording estimates of slope and intercept that are virtually identical to their ordinary least squares counterparts in the absence of outliers, but which do not change appreciably in the presence of outliers. In fact, with univariate data, it improves on ordinary least squares in almost every way imaginable, and it is therefore a striking fact that this remarkable estimator is not universally known and used. It can be used to derive robust estimates of beta and the correlation coefficient that are virtually identical to their classical counterparts when asset returns are normally distributed, and which are significantly more robust when asset returns are highly skewed or contaminated with outliers.

In robust statistics, **robust regression** is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as ordinary least squares, have favourable properties if their underlying assumptions are true, but can give misleading results if those assumptions are not true; thus ordinary least squares is said to be not robust to violations of its assumptions. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process.

In particular, least squares estimates for regression models are highly sensitive to outliers. While there is no precise definition of an outlier, outliers are observations which do not follow the pattern of the other observations. This is not normally a problem if the outlier is simply an extreme observation drawn from the tail of a normal distribution, but if the outlier results from non-normal measurement error or some other violation of standard ordinary least squares assumptions, then it compromises the validity of the regression results if a non-robust regression technique is used.

As defined by Theil (1950), the Theil–Sen estimator of a set of two-dimensional points $(x_i,y_i)$ is the median $m$ of the slopes $(y_j - y_i)/(x_j - x_i)$ determined by all pairs of sample points. Sen (1968) extended this definition to handle the case in which two data points have the same $x$ coordinate. In Sen's definition, one takes the median of the slopes defined only from pairs of points having distinct $x$ coordinates.

Once the slope $m$ has been determined, one may determine a line from the sample points by setting the $y$-intercept $b$ to be the median of the values $y_i - mx_i$. The fit line is then the line $y = mx + b$ with coefficients $m$ and $b$ in slope–intercept form.[11] As Sen observed, this choice of slope makes the Kendall tau rank correlation coefficient become approximately zero, when it is used to compare the values $x_i$ with their associated residuals $y_i - mx_i - b$. Intuitively, this suggests that how far the fit line passes above or below a data point is not correlated with whether that point is on the left or right side of the data set. The choice of $b$ does not affect the Kendall coefficient, but causes the median residual to become approximately zero; that is, the fit line passes above and below equal numbers of points.A confidence interval for the slope estimate may be determined as the interval containing the middle 95% of the slopes of lines determined by pairs of points[12] and may be estimated quickly by sampling pairs of points and determining the 95% interval of the sampled slopes. According to simulations, approximately 600 sample pairs are sufficient to determine an accurate confidence interval.[10]

# CODING THE STRATEGY

## 1. Importing the dataset

In this model, we are going to use daily NSE data for the stock of trading on NSE for the time period from January 2016 to December 2016. We import our training.CSV file named 'NSE 2016.csv' saved in the personal drive in your computer. This is done using the pandas library, and the data is stored in a dataframe named dataset. We then drop the missing values in the dataset using the dropna() function. We choose only the NSE data from this dataset, which would also contain the Date, Adjusted Close and Volume data.

```
df1=pd.read_csv("C:\\Users\\Shubham\\nse_2016.csv")
df1.head()
```

| | SYMBOL | SERIES | OPEN | HIGH | LOW | CLOSE | LAST | PREVCLOSE | TOTTRDQTY | TOTTRDVAL | TIMESTAMP | TOTALTRADES | ISI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 20MICRONS | EQ | 33.90 | 35.00 | 32.25 | 33.0 | 33.30 | 33.75 | 48174 | 1593805.60 | 2016-04-21 | 253 | INE144J0102 |
| 1 | 3IINFOTECH | EQ | 4.40 | 4.45 | 4.30 | 4.3 | 4.30 | 4.35 | 684070 | 2991352.25 | 2016-04-21 | 360 | INE748C0102 |
| 2 | 3MINDIA | EQ | 13939.70 | 14200.00 | 13324.00 | 13424.1 | 13400.05 | 14080.50 | 1464 | 20233135.70 | 2016-04-21 | 722 | INE470A0101 |
| 3 | 8KMILES | EQ | 2018.95 | 2029.10 | 1956.20 | 1973.4 | 1970.25 | 2000.85 | 28967 | 57987253.05 | 2016-04-21 | 3264 | INE650K0101 |
| 4 | A2ZINFRA | EQ | 26.40 | 26.80 | 25.80 | 25.9 | 25.80 | 26.30 | 91366 | 2388936.45 | 2016-04-21 | 576 | INE619I0101 |

## 2. Taking User Input:

```
In [*]: user_input = (input("Enter Stock name: "))
        stocks = df2.get_group(user_input)

        Enter Stock name:
```

We take the input from the user.

# 3. Data Pre-Processing

```
In [24]: train.head()
```

Out[24]:

| | OPEN | HIGH | LOW | CLOSE | TOTTRDQTY | Date | PREVCLOSE | TOTTRDVAL | TOTALTRADES | HL_PCT |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 335.60 | 336.50 | 327.60 | 329.95 | 4629838 | 2016-04-21 | 335.50 | 1.528778e+09 | 65059 | 2.716728 |
| 1 | 318.00 | 321.70 | 315.15 | 319.35 | 6293941 | 2016-05-13 | 318.20 | 2.006523e+09 | 51607 | 2.078375 |
| 2 | 312.95 | 319.40 | 306.00 | 317.70 | 10389150 | 2016-05-06 | 317.65 | 3.277715e+09 | 120299 | 4.379085 |
| 3 | 321.15 | 327.05 | 319.00 | 320.00 | 12005375 | 2016-03-29 | 322.45 | 3.867055e+09 | 70702 | 2.523511 |
| 4 | 254.55 | 256.75 | 252.20 | 253.70 | 8328306 | 2016-08-17 | 255.10 | 2.117731e+09 | 90919 | 1.804124 |

```
In [25]: train.shape
```

Out[25]: (247, 10)

```
In [26]: test.head()
```

Out[26]:

| | OPEN | HIGH | LOW | CLOSE | TOTTRDQTY | Date | PREVCLOSE | TOTTRDVAL | TOTALTRADES | HL_PCT |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 311.6 | 311.6 | 307.10 | 308.35 | 7919072 | 2017-06-28 | 312.35 | 2.451289e+09 | 73681 | 1.465321 |
| 1 | 259.1 | 259.2 | 256.00 | 257.00 | 17508901 | 2017-11-28 | 259.25 | 4.499432e+09 | 100436 | 1.250000 |
| 2 | 263.0 | 264.6 | 260.90 | 262.20 | 8240790 | 2017-02-28 | 264.60 | 2.166087e+09 | 95489 | 1.418168 |
| 3 | 285.8 | 286.5 | 277.35 | 278.00 | 12374841 | 2017-04-28 | 285.80 | 3.452998e+09 | 120143 | 3.299081 |
| 4 | 282.9 | 284.2 | 277.80 | 281.90 | 15897260 | 2017-03-20 | 281.25 | 4.452069e+09 | 115321 | 2.303816 |

```
In [27]: test.shape
```

Out[27]: (249, 10)

In this we split the data according to the year 2016 & 2017. Then we split this data into training & testing set. The stock data from 2016 goes into the training set & the stock data from 2017 goes in the testing set. To train the data we use 'DictVectorizer'. The class DictVectorizer can be used to convert feature arrays represented as lists of standard Python dict objects to the NumPy/SciPy representation used by scikit-learn estimators.While not particularly fast to process, Python's dict has the advantages of being convenient to use, being sparse (absent features need not be stored) and storing feature names in addition to values.DictVectorizer implements what is called one-of-K or "one-hot" coding for categorical (aka nominal, discrete) features.

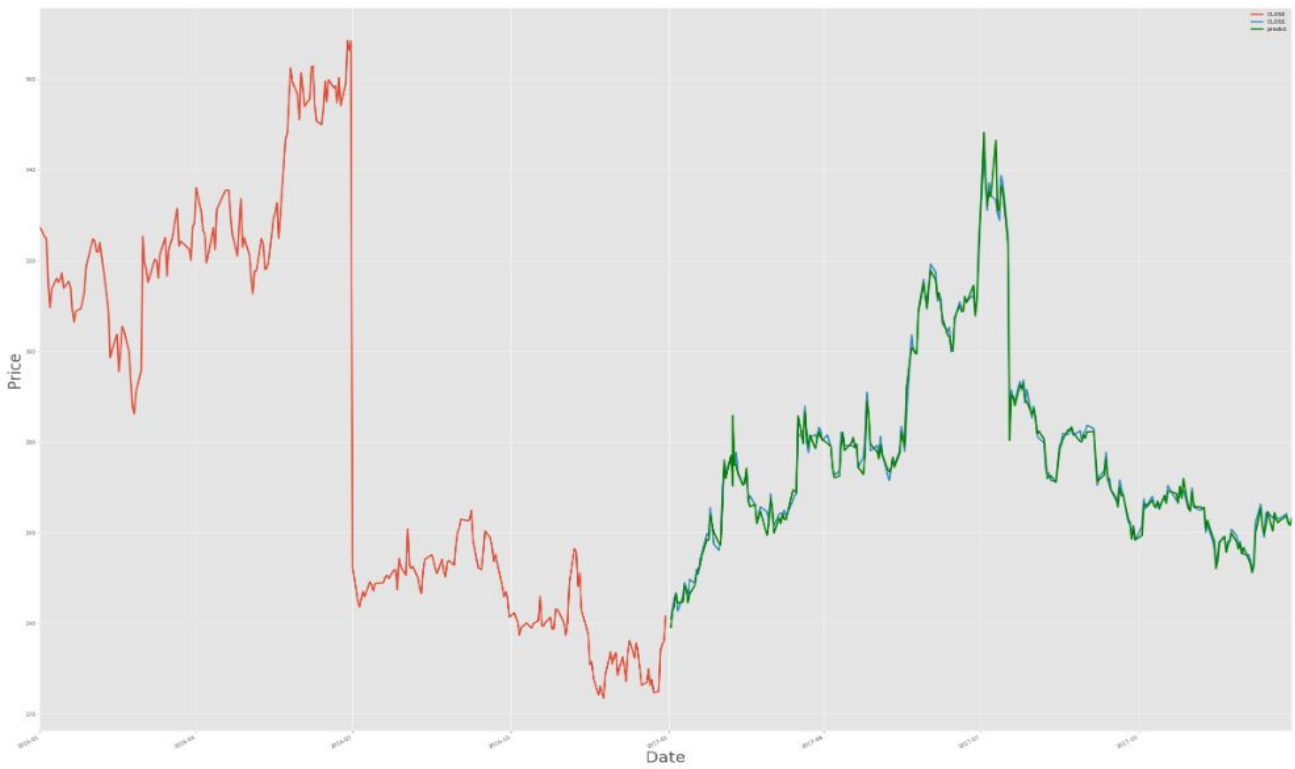
# 4. Training & Fitting the Model :

To train the data we use 'DictVectorizer'. The class DictVectorizer can be used to convert feature arrays represented as lists of standard Python dict objects to the NumPy/SciPy representation used by scikit-learn estimators.While not particularly fast to process, Python's dict has the advantages of being convenient to use, being sparse (absent features need not be stored) and storing feature names in addition to values.DictVectorizer implements what is called one-of-K or "one-hot" coding for categorical (aka nominal, discrete) features. The trained model is then fitted to the Theil-Sen Regression. Theil-Sen estimator is an exceptionally simple and robust linear regression estimator, affording estimates of slope and intercept that are virtually identical to their ordinary least squares counterparts in the absence of outliers, but which do not change appreciably in the presence of outliers.After fitting the test set to the trained model we get the score of the predicted model.

```
In [28]: #Classifier
         clf = TheilSenRegressor()
         clf.fit(X, Y)
         predict = clf.predict(x)
         print("Accuracy of this Statistical Arbitrage model is: ",clf.score(x,y))

         Accuracy of this Statistical Arbitrage model is:  0.9928573776929362
```

## 5. Plotting the Result :



# Conclusion

In this paper, we investigate the concept of statistical arbitrage (SA). As there is no agreement in literature on a common definition, we review both the theoretical and empirical works on SA since its introduction. In particular, we look at all those definitions, which may be suitable to identify this class of strategies. We produce a review of all strategies which may be associated with the concept of statistically determined arbitrage opportunities. We identify those common features which define the concept embedded in investors thinking. As no definition is suitable to describe this type of strategies, we introduce a general definition and propose a classification system that encompasses the current forms of SA strategies while facilitating the inclusion of new types as they emerge. Our study makes several contributions to the existing literature. We bridge the gap existing between the literature on arbitrage definitions and SA strategies. We perform an innovative investigation of SA both in academic and financial industry research analyzing, for the first time, SA across all asset classes (equity, fixed income and commodity). We find a general definition, which includes all SA strategies and propose a classification system measuring the strategies' risk and return profile. This facilitates the inclusion of new strategies and measures as they emerge. Our analysis allows investors to have a common framework to evaluate investment opportunities and brings clarity in SA investing, guiding theoretical development and empirical testing. We also provide examples of potential future research directions.

# References

[1] Fama, E. (1969) Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25, 383-417. https://doi.org/10.2307/2325486

[2] Ross, S. (1976) The Arbitrage Theory of Capital Asset Pricing. Journal of Economic Theory, 13, 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

[3] Shleifer, A. and Vishny, R. (1997) The Limits of Arbitrage. The Journal of Finance, 52, 35-55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[4] Philips, T. K., "Robust Estimates of Betas and Correlations Using Theil-Sen Regression", Encyclopedia of Financial Models, Vol. 2, (2012).

[5] Stefanini, F. (2006) Investment Strategies of Hedge Funds. John Wiley & Sons.

[6] Pole, A. (2007) Statistical Arbitrage. John Wiley & Sons, Hoboken, NJ.

[7] Duarte, J., Longstaff, F.A. and Yu, F. (2006) Risk and Return in Fixed Income Arbitrage: Nickerls in Front of a Steamroller? The Review of Financial Studies, 20, 769-811. https://doi.org/10.1093/rfs/hhl026

[8] Brogaard, J., Hendershott, T. and Riordan, R. (2014) High-Frequency Trading and Price Discovery. The Review of Financial Studies, 27, 2267-2306. https://doi.org/10.1093/rfs/hhu032

[9] Chaboud, A.P., Chiquoine, B., Hjalmarsson, E. and Vega, C. (2014) Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market. The Journal of Finance, 69, 2045-2084. https://doi.org/10.1111/jofi.12186

[10] Payne, B. and Tresl, J. (2015) Hedge Fund Replication with a Genetic Algorithm: Breeding a Usable Mousetrap. Quantitative Finance, 15, 1705-1726. https://doi.org/10.1080/14697688.2014.979222