RESEARCH

Submitted as a Requirement of the

Master of Management (Finance and Investment Management)

University of the Witwatersrand Business School

Johannesburg

Statistical Arbitrage on the FTSE/JSE TOP 40 Index

Mandlenkosi Svato Ngcobo: 0317279E

Supervisor: Dr. Blessing Mudavanhu

Abstract

The mid 2000's saw the materialization of research into the financial engineering field of high frequency trading. It is arguable that the most prominent model to emerge from the research has been pairs trading. This idea can be extended to allow for more than two assets in a modelling method now known as statistical arbitrage.

The research identifies a collection of assets with a deterministic component; it then follows a multiple linear regression to exploit persistent mispricings among these assets. Further, multiple linear regression metrics are used to identify the analytic form of the trading rule and to validate the performance of the model.

The first part of model constructs combinations of assets which contain a significant predictable component by co-integration, the second part builds a predictive models for the dynamics of the mispricing using statistical model.

The success of the model is demonstrated with reference to a statistical analysis of 5-minute closing prices on the Johannesburg Stock Exchange (JSE) TOP40 Index and the constituent shares of the JSE TOP40 Index.

Table of Contents

List of Figures

Acknowledgements

I extend my sincere thanks to my supervisor, Dr. Mudavanhu for the guidance and direction in the various phases of building this technical paper.
I would also like to exempt Professor Aligedede and Meisie Moya as exemplary, and thank you for running such a highly esteemed program.

For financially enabling this research I thank Barclays Global Africa Ltd. I love my family and thank you for believing in me, to see this work through to fruition.

Above all I praise God, my Lord and Saviour Jesus Christ, with Him all things are possible.

Chapter 1

This paper tests the existence of a modelled form of statistical arbitrage on the constituent shares of the FTSE/JSE Top 40. A divergence-convergence trading rule is applied to shares sampled at 5-minute frequencies, sampling intervals.

1.1 Origins

Statistical arbitrage was developed out of the archetype trading strategy, the pairs trade. Accounts differ in academic literature as to the origins of pair trading. It is however known that in the early 1980's NunzioTartaglia led a team of scientists and engineers to study time series of market prices to uncover arbitrage opportunities in financial markets, and in so doing created a program which traded shares in combination Gatev et al. (2006). Paul Wilmot offers the origins of pairs trading to Gerry Bamberger who by 1983 had initiated the Morgan Stanley pairs trading program, it is not clear from the literature whether the Tartaglia's program was distinct from Bamberger's program according to Pole (2007) .

The original concept proposed that a pair of tradable assets be held in combination, the illustration below is an example of the two such assets, in this paper the two tradable assets considered will be the Synthetic Asset and the TOP40 Index. The Synthetic Asset is a linear combination of tradable assets, filtered on the basis of being a constituent of the equity index.

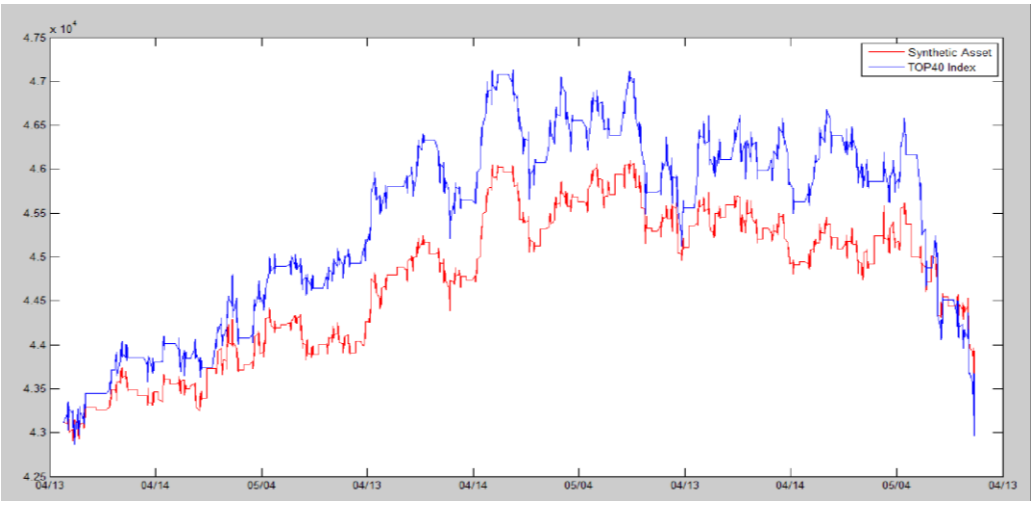

Figure 1: Price Time Series of Synthetic Asset vs. TOP40 Index over Research Horizon

1.1.2 Classical Pairs Trade

According to Avellaneda and Lee (2008) pairs trade, the "ancestor" of statistical arbitrage can be expressed according to the econometric model:

$$\ln\left(\frac{P_t}{P_{t_0}}\right) = \alpha(t - t_0) + \beta \ln\left(\frac{Q_t}{Q_{t_0}}\right) + \varepsilon_t \qquad (1.1)$$

where

t denotes the current time,

t_0 denotes the beginning of period time,

P_t and Q_t are current prices of tradable assets P and Q respectively,

ε_t is the white noise term,

$\alpha \; and \; \beta$ are constants

The parameters of such a model can be estimated using various techniques including classic linear regression Brooks (2008).In practice a pairs trade investment strategy is implemented with a contrarian view. With reference to the model above, if the residual ε_t is significantly small, one

would buy 1 Rand of share P and sell β Rand of Q. The portfolio produces a positive return as valuations converge as observed by Pole (2007).

To address the question related to statistical arbitrage, the "descendant" model will hold the same econometric form however the interpretation of the underlying instruments used in the pairs trade will change. P_t will be taken as the time series of the FTSE/JSE TOP40 Index and Q_t the time series of the synthetic portfolio. The synthetic index portfolio will consist of equity shares from the FTSE/JSE TOP40 Index, using the shares that are deemed statistically significant for the pricing of the index. Once this process is complete it is expected that Q_t will be correlated with P_t.

1.1.3 Definition of a Statistical Arbitrage

There is no universally accepted definition for statistical arbitrage, when used in this paper it will refer to one of a wide range of market neutral trading strategies which seek to exploit price deviations among tradable assets away from an economically significant relationship. This is consistent with the usage in Lo (2010). These strategies assume some degree of mean reversion, for a more detailed study as applied to the interest parity relation consult Cuthbertson and Nitzsche (2004).

1.2 Problem Statement

Existing computational modelling techniques allow practitioners to model regularities in asset price dynamics. Exploiting predictive information provided by such models is statistical arbitrage, which is considered as a subclass of arbitrage strategies in which systematic components in traded asset dynamics are exploited by market participants known as arbitrageurs. There is competition between arbitrageurs to identify and exploit pure arbitrage opportunities generated by employing market neutral strategies to be both self-limiting and restricted to relatively privileged market players who are geared to trade quickly, at low cost, and with sufficient financial leverage to make the process profitable. Statistical arbitrage opportunities however, are based on empirical regularities with no direct theoretical underpinning, and are more likely to be both more persistent and more prevalent in financial markets. Pure risk-free

arbitrage opportunities are rapidly eliminated by market activity, statistical arbitrage opportunities are of lesser importance to arbitrageurs and will thus be more persistent. More prevalent because in principle they may occur between any set of tradable assets rather than solely in cases where a suitable risk-free hedging strategy can be implemented. The debate continues on whether markets are so efficient that no predictable components can possibly exist. Regularities in asset price dynamics should exist however any easily identifiable effects will be arbitraged away in the very process of being exploited. By looking at the markets with the correct tools it may be possible to discover previously unknown "patterns" in market activity.

The main problem is to determine whether the TOP40 Index admits statistical arbitrages, assess the character of theses statistical anomalies in the framework of a multiple linear regression model.

1.3 Objectives

The objective of the study is to identify and exploit the opportunities identified by the model using statistical techniques. The methodology includes identifying statistical mispricings in a set of tradable assets, modelling the dynamics of the predictable components, and employing a predictive forecast through a suitable asset allocation strategy. Identifying statistical arbitrage opportunities in the financial markets is considered one of the most challenging tests of predictive modelling and is of high practical significance. The highly competitive nature of financial markets means that any predictable component in traded asset dynamics it is likely to be very small. The model performance relies only on its predictive ability, the market neutral positions eliminates the favourable performance of the individual holdings of the portfolio. The prospective gains of a significant model are large as seen in Figure 9, even when transaction costs are accounted for.

There are similar features among riskless arbitrage trading rules however they differ significantly in design. A riskless arbitrage modelling methodology will generally follow the three steps following steps:

A relationship describing the behaviour of the assets is determined. The main purpose of this part of our modelling process is to determine which permutation of assets contain an identifiable and thus possibly predicable component in the dynamics and whether said permutation is de-correlated with respect to major sources of economic risk.

Identification of specific arbitrage opportunities: This objective deals with of de-correlation of the target asset with respect to economic sources of risk. It is important as it aids in the diversification of risk sources, within the set of model for the fair price relation.

Implementation of an appropriate trading strategy: This objective implies that the best possible cumulative return from the trading strategies is directly dependant on the absolute size of the statistical mispricing. Performance can be increased by reducing data faults with financial engineering techniques in the predictive modelling and trading rule construction phase.

Determining the trading range for the anomaly and determining the distribution of the mispricing using statistical and financial engineering techniques taking into account considerations such as earnings, etc. are core to the analysis mean reversion component.

When the price of the statistical anomaly is less than the lower cut-off bound, the share is regarded as suitable for purchase, in anticipation of a rise in the statistical anomaly's value. When the price of the statistical anomaly is higher than the upper cut-off bound, the share is regarded as suitable for sale, in anticipation of a drop in the statistical anomaly's value. We thus on average expect the mispricing to revert to the long term average.

1.4 Expected Significance of the Study

This study expands on a growing body of knowledge pertaining to statistical arbitrage, with research in this field still coming to maturity Gamzo (2013). This paper further provides a premier in a fully parametric statistical arbitrage and trading rule, which at 5% significance level generate positive abnormal returns via the momentum.

The findings of this research project will demonstrate to both academia and financial engineering practitioners the capabilities of a linear model to identify and exploit statistical mispricings.

1.5 Research Question and Hypotheses

The main research question is: Does the FTSE/JSE TOP 40 admit statistical arbitrage opportunities?

The research hypothesis is:

H_{0i}**:** The FTSE/JSE TOP 40 does not admit in Implicit Statistical Arbitrage opportunities.
H_{1i}**:** The FTSE/JSE TOP 40 does Implicit Statistical Arbitrage opportunities.

1.6 Limitations of this Research

The data used to conduct the study is based on closing prices, and ignores the two key frictions realised in traded asset markets and that is the bid offer spread and the trading costs involved in entering a position. The effort made within the performance analysis to account for this was to make assume a negative drift to the cumulative returns, as demonstrated in Figure 9, assuming 20, 50 and 70 basis transaction costs, respectively.

Mid price data ignores slippage and margin costs which may vary across market makers. The study finding is profitability net of 'plausible but hypothesised' transaction costs. Historical intra-day transaction time series data is difficult to source from vendors as majority of commercial vendors store daily closing prices as a minimum frequency for their databases. The data that collected from the various vendors therefore cannot be back referenced against the vendors' database and is as reliable as the programs used to sample them. Sampling in discrete time will invariably mean that there will be tick data lost as sampling is not conducted in continuous time. Further, analysis is limited to synchronous statistical measures. We have limited the time path granularity due to our discrete time sampling technique.

This research continues as follows, in Chapter 2, a literature review is presented, Chapter 3 describes the data used and explains the methodology implemented. Chapter 4 documents the implicit statistical arbitrage model performance Chapter 5 concludes the research with findings and offers possible avenues of additional investigation.

Chapter 2: Literature Review

In this chapter we make a review of the existing literature on the key elements of the statistical arbitrage model to be researched. The econometric model is a multivariate regression, consequently the literature review will cover arbitrage pricing theory and Risk Arbitrage, mean reversion and co-integration; market neutrality and market efficiency as themes for understanding the models behaviour and performance.

2.1 Arbitrage Pricing Theory and Risk Arbitrage

A pure arbitrage opportunity (PAO) is a zero-cost trading strategy that offers the possibility of a gain with no possibility of a loss. As is well known, the existence of PAOs is incompatible with a competitive equilibrium in asset markets. The fundamental theorem of the financial theory establishes a link between the absence of PAOs and the existence of a positive pricing kernel which supports securities prices Bondarenko (2003), in the methodology portion of this research study the pricing kernel is referred to as the fair price relation. While the absence of PAOs is a necessary condition for any equilibrium model, this condition alone often yields pricing implications that are too weak to be practically useful. Cochrane and Saa-Requejo (2000) propose to rule out not only PAOs but also ``good deals" (GDs), or investment opportunities with high Sharpe ratios. Following Hansen and Jagannathan (1991), Cochrane and Saa-Requejo show that precluding GDs imposes an upper bound on the pricing kernel volatility and yields tighter pricing implications when markets are incomplete. Bernardo and Ledoit (2000) propose to rule out approximate arbitrage opportunities (AAOs), or investment opportunities which offer high gain-loss ratios, where gain (loss) is the expectation of the positive (negative) part of the excess payoff computed under a benchmark risk-neutral measure. They demonstrate that restricting the maximum gain-loss ratio implies, loosely stated, that an admissible pricing kernel cannot deviate too far from the benchmark pricing kernel. Similarly, in this research it is shown that that synthetic asset (benchmark pricing kernel) cannot deviate significantly from TOP40 Index (admissible pricing kernel) without admitting persistent exploitable opportunities. Bondarenko (2003) pg.876 - 879 shows analytically that if the fair price relation exists (or equivalently if the

aforementioned deviation is admitted) that the market of tradable assets admits statistical arbitrage opportunities.

Arbitrage pricing theory states that the returns on an the target asset (say TOP40 Index) can be modelled using linear combination of various economic elements or asset price returns, where models' sensitivity to changes in each predictor variable is represented by a predictor variable beta coefficient within fair price relation. The rate of return derived from the fair price relation is then to be used to price the target asset - the target asset returns should equate the average fair price relation returns for the period discounted at the rate indicated by the fair price relation. If significant divergences do prevail, the process of market participants exploiting this divergence should bring it back into line. The arbitrage pricing model was introduced by the economist Stephen Ross in 1976.

The premise of Arbitrage Pricing Theory states that if the returns of an asset can be matched by a combination of other assets then the returns of forming a portfolio which replicates the asset should be the same as the returns of the target asset. In an efficient market, no riskless arbitrage opportunities should occur. This means market participants should not be able to generate abnormal returns by trading in the same assets at different prices, by an amount that exceeds the transaction cost of entering such a trade. The constraint for no-arbitrage including transaction costs is depicted in the formula below:

$$\left|payoff\left(Y_t - SA(Y_t)\right)\right| < TC \tag{2.1}$$

Here Y_t is the target asset, $SA(Y_t)$ is a synthetic asset which is modelled to imitate the returns of Y_t where TC are the transaction costs, and can be considered as the net costs involved in constructing buying the synthetic asset and selling Y_t (or vice versa) as per Burgess (1996).

.

2.2 Mean Reversion and Convergence Trade strategies

In finance, mean reversion is the assumption that a share's returns will converge to an average return in the long run. Mean reversion is considered as a more scientific method of choosing asset trade levels in comparison to techniques such as technical analysis, where specific numeric trade levels are determined from historical data to identify trade signals, i.e. that is opposed to interpreting price movements using charts (technical analysis which is known as charting).

2.3 Co-Integration

The objective of the first part of our methodology is to support the pre-processing of financial time-series in a manner which allows the construction and identification of combinations of assets which contain a predictable component which can be exploited in a statistical arbitrage context. Such combinations are referred to throughout the thesis as statistical mispricings.

Possibly the most applied times series model structure in any field is the Autoregressive (AR). In these models future values are projected as weighted averages of recently exhibited values.AR models often appear in Auto Regressive Integrated Moving Average (ARIMA) models as cited by Box and Jenkins (1976). Here integrated refers to the difference operation applied to the times series before analysing the autoregressive structure. The Exponentially Weighted Moving Average (EWMA) provides a forecast function which can be shown to be optimal for integrated models.

Cointegration is a financial engineering tool based on determining long-term relationships between asset returns documented in the paper by Engle and Granger (1987). Two time series which are not stationary are co-integrated if a linear combination of the two exists which is stationary. The cointegration property will help us identify combinations of assets which are stationary. The research model will consider all the possible pairs within the FTSE/JSE Top 40 Index, not only co-integrated ones. By doing so the statistical arbitrage portfolio is allowed to be neutral to the broader market.

Johansen (1988) created another approach that can be used when we want to consider more than two financial assets simultaneously. A set of cointegrating vectors can be found in the system.

The most stable set of cointegrating vectors in the long-term is used to derive the spread between the assets is not the set of cointegrating vectors with the lowest variance, as was the case with the ordinary least squares case. As documented by Alexander (2001) on pg 361 the Engle and Granger (1987) methodology is favoured in financial modelling due to its stability and ease of use, this is an important factor to consider from a risk management point of view. In this research we deal with multiple share holdings and prefer the vectorised Engle and Granger (1987) methodology.

2.4 Market Neutrality

As result of the recent financial market crisis, it has been documented in many financial publications that the year 2007 was particularly difficult for hedge funds that employed quantitative techniques to generate cumulative return as per Khandani & Lo (2007), this includes statistical arbitrage hedge funds. Jacobs and Levy (1995) describe long-short share trading rule as being neutral to market risk. Strategies which are neutral to market risk maintain even exposure to these risks for all positions at all times. This trading strategy removes all exposure to directional risk from the market risk factors, That is, the realised return does not present correlation with the market index, which is the same as holding a portfolio with a beta of zero. The portfolio generates returns by isolating the alpha when adjusted by risk. Fung & Hsieh (1999) report that a strategy is market neutral if the return from such portfolio is independent of the relative return of the market. Hedge funds which are market neutral avoid systematic risk factors by taking bets on relative price movements.

There are many possible co-integration based market neutral trading rules, where a market participant enters into a new trade when the statistical mispricing is far away from a long-term average, and terminates the trade when it has returned is has returned the long-term average again. Burgess (2003), Lin et al. (2006) or the work of Galenko et al. (2007), call this filed high-frequency trading make use of daily closing prices among 4 world indexes, instead of real intraday continuous or intraday minute frequency data as used in this research.

2.5 Market Efficiency

When testing for statistical arbitrage we avoid the dilemma of a joint hypothesis which prevails the standard market efficiency tests because the definition of statistical arbitrage is free of any equilibrium model and its presence is incompatible with the Efficient Market Hypothesis.

In excess of a decade before the algorithmic trading era and focusing primarily on the implications of short term trading Froot, Scharfstein and Steinz (1992) conducted research into high frequency trading. They found that short investment horizons determine the nature of asset returns, resulting in a specific type of informational inefficiency. Consequently, short horizon speculators seem to manipulate a form of price behaviour that may result in a decision being made on the basis of information that is not related to the fundamental value of the asset. They then assert that short-term speculators rely on short term information excessively and, as a result, they reduce the quality of market price information and affect the Efficient Market Hypothesis adversely.

Various investigations have concluded that share prices appear to contradict the efficient markets hypothesis. Jegadeesh and Titman (1993) investigated a trading strategy that bought well-performing shares and sold poorly performing shares. They showed that this trading strategy generated average abnormal returns of 12% per year, where excess returns are defined relative to a standard capital asset pricing model. Lakonishok et al. (1994) arrive at a similar result by buying shares considered as value and selling shares considered as glamorous, identified with metrics like book-to-market values, dividends, price earnings ratios, cash flows, and sales growth. Chan et al. (1996) confirm that the abnormal returns can be generated for portfolios formed on the basis of past returns and earnings announcements

Prior research supports the possibility of many other persistent anomalies such as those provided by earnings announcements, issuance of shares, and dividends. Shleifer (2000) documents an apt review of this literature. Within context of this research we provide evidence to the persistence of anomalies pertaining to the fair price relationship and test whether deviations in this relation can be considered as persistent anomalies.

Chapter 3: Methodology

This chapter covers the data employed to conduct the research and make findings, and then details the approach followed to determine of fair-price relationships between assets and finally the method used to find the predictors with most explanatory

3.1 Data Description

Discrete time sampling allows us to apply the insights of Khandani and Lo (2007) in this study. Times series price data will be stored in spreadsheet format, as available from Bloomberg ® are the price vendors of choice for this research. Bloomberg® provides quote information for the TOP 40 Index (Y_t) and the TOP 40 Index constituent shares ($X_{i,t}$,) at the 5 minute frequency. Bloomberg® only stores high frequency quote information for the past 100 business days, which results in a research time horizon from 24 March 2014 to 03 October 2014. The data used will make use of Microsoft Visual Basic®, MATLAB® and E-Views® for the purpose of analysis and testing.

3.1.1 Data Transformation

If we consider lines 83 from the code in the appendix, notice that the price values are first transformed into log returns before the parameter estimation is conducted. This is the deviation from the Burgess (2006) methodology as the theory of M_t model is based on price. However Burgess later goes on to acknowledge that the models fitted on a returns basis produced superior performance to the ones fitted using price, due to the stationarity of the log return time series.

3.1.2 In-Sample Data

The sample date uses the first 2000 observation s and this period runs from 24 March 2014 to the 13 April 2014, this is approximately a 3 week period out of a total of 28 week research horizon.

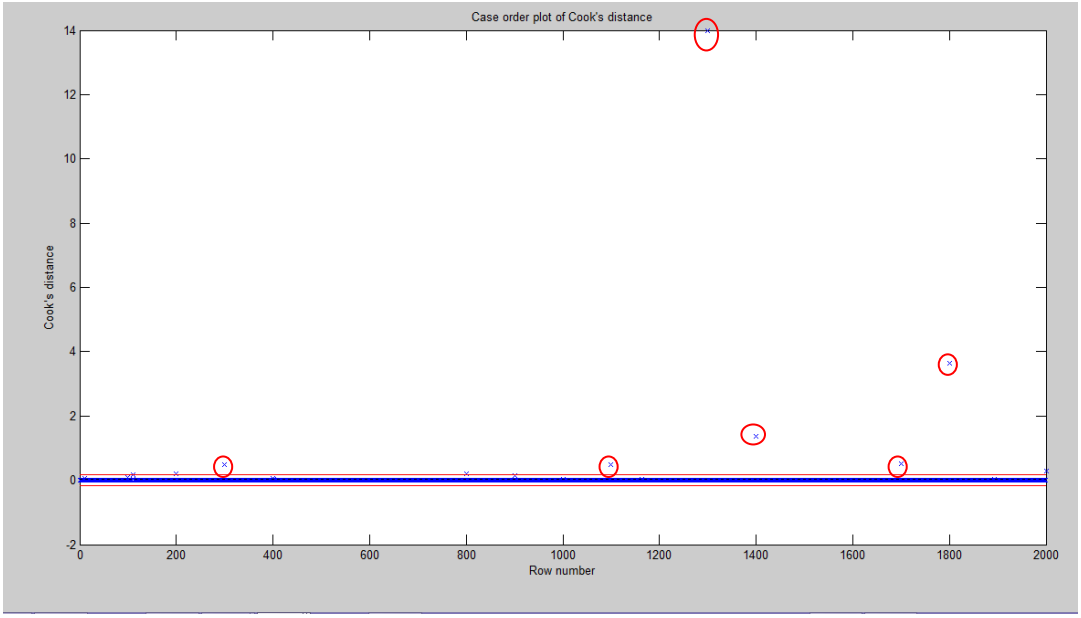

Figure 2: Cooks Distance for the In-Sample Observations

Although this is an ex-post exercise there are opportunities that occur either only so briefly or at the beginning/end of the trading day which under ordinary trading conditions would not allow for exploitation are not considered for the purpose of this research. We have filtered for these by excluding these opportunities from the data set ensuring a smoother look for the statistical mispricing which has a more statistically discernable structure. Please not this based on economically practical motivation.

From a statistical modelling perspective the Cooks Distance measure is an adequate means of identifying outliers in the constituent share values. It also shows the influence of each observation on the statistical mispricing M_t values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier as indicated in the graph above. The observations excluded according to Cooks distance can be found in the appendix, and amount to approximately 1 hour worth of trading data. This methodology is consistent with outlier treatment in Chatterjee and Hadi (1996).

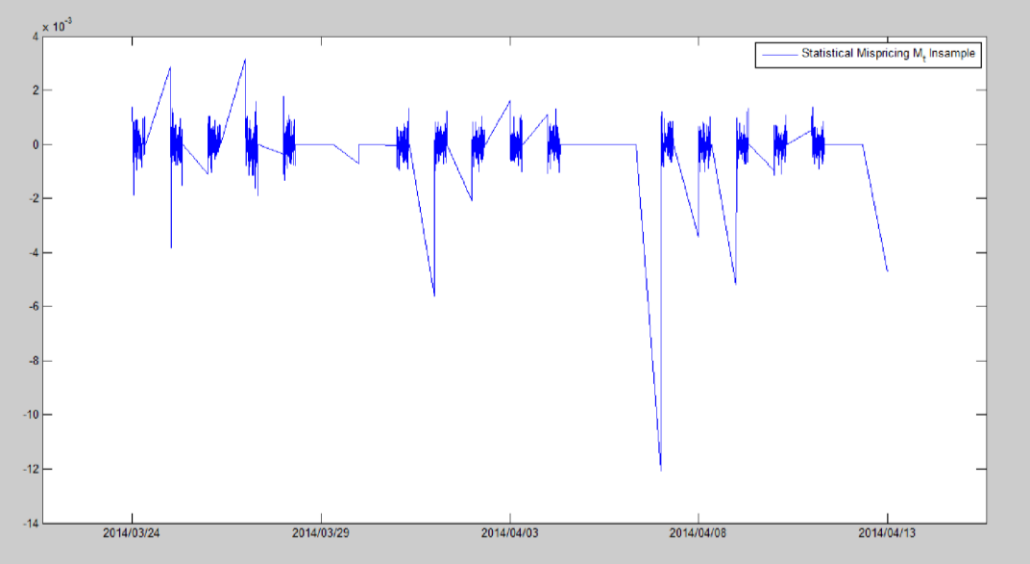

Figure 3: Statistical Mispricing Unadjusted

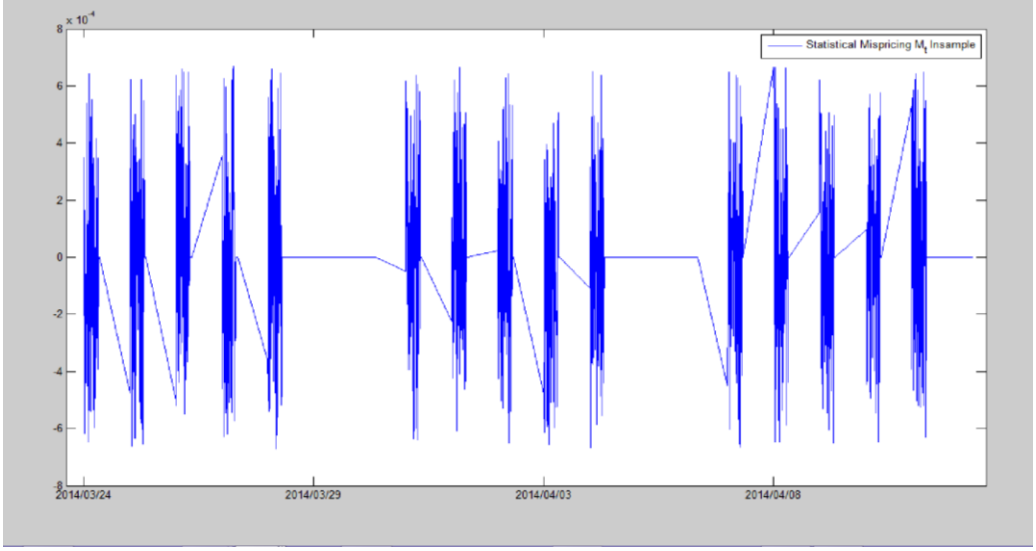

Figure 4: Statistical Mispricing adjusted for un-exploitable opportunities and outliers

Figure 4 demonstrates an improvement in the smoothness in the return time series achieved over Figure 3 which is a direct result of removing outliers from the in-sample data according to the Cooks Distance statistic. In figure 4 it easier to discern weekly and daily non trading periods, and the model provides coefficients with lower standard errors as modelling benefit.

3.2 Modelling

3.2.1 Constructing the Statistical Mispricing

The discussion from the introduction section lets us from a statistical point of view show, that the returns from the error time-series of the multiple linear regression model can be taken as a synthetic portfolio which displays mean-reversion, and some visible degree of predictable behaviour. The theory of the so called statistical pricing M_t is that, statistically predictable combinations of predictor returns can be exploited as the basis of an abnormal return generating trading strategy, regardless of the existence of an economically tractable theoretical fair-price relationship between the set of target asset and synthetic asset considered.

These trading strategies are considerably riskier than true arbitrage trading strategy. Given this, statistical misprings are likely to prevail more often than the latter and be more persistent in traded markets. The persistence is likely to be higher due the fact that true arbitrage anomalies disappear instantaneously in traded markets quickly. We expect them to prevail more in principle as they can manifest between combinations of predictor variables as opposed to in isolation.

We begin with finding the appropriate synthetic index portfolio. This will involve a regression of the constituents' returns in the index on the index itself. We begin with the standard mathematical formula for constructing a synthetic asset whose price is given by $SA(Y)_t$ as per Jordan and Miller (2008):

$$SA(Y)_t = \sum_{X_i \in X}^{N} w_i X_{i,t} \tag{3.1}$$

where

t denotes the current time,

t_0 denotes the beginning of period time,

$X_{i,t}$ is the current price of ith asset of the index,

ε_t is the white noise term,

w_i is the weight of the ith constituent asset of the index

Expressed in its econometric form:

$$SA(Y)_t = \sum_{X_i \in X} \beta_i X_{i,t} + \varepsilon_t \quad (3.2)$$

where

t denotes the current time,

$X_{i,t}$ is the current price of i[th] asset of the index,

ε_t is the white noise term,|

β_i is the co-efficient of the i[th] constituent asset of the index

The purpose of our procedure is to determine which set of predictor variables model the fair-price relation on which we will base statistical mispricing and trading rule. Technically speaking, given a set of constituent times series returns U_A and a given target asset, $Y \in U$, our aim is to model the synthetic asset $SA(Y)$ so that the value of the synthetic asset can be taken as a statistically speaking a "fair-price" for Y, shown mathematically in the equation below :

$$E[Y_t] = SA(Y_t) \quad (3.3)$$

Further, Equation (3.3) should hold strong enough that a deviation from this equation should be taken to be a statistical anomaly such that the behaviour of the statistical mispricing return time series is as follows:

$$M_t = Y_t - SA_t(Y) \quad (3.4)$$

this possesses a feature which can be exploited as the core logic of the trading strategy based on the fair price relation. The procedure for modelling m_t is rooted on the use of the cointegration to determine the fair price relation. The multiple linear cointegration regression documented by Granger (1983) is used to determine the optimal combination of predictor variables which possess the highest long-term correlation with Y

The coefficients of the linear combination are estimated by a regression of the historical returns of Y_t on the historical returns of a set of "constituent" assets $C \subset U_A - T$:

$$SA_t(Y) = \sum_{X_i \in X} \beta_i X_{it} \; such\; that \; \{\beta_i\} = arg_w \, min \, var\left(Y_t - \sum_{X_i \in X} \beta_i X_{t,i}\right) \quad (3.5)$$

With the cointegrating vector $\beta = [\beta_1 \; ... \; \beta_{42}]^T$ of constituent weights is given by:

$$\boldsymbol{\beta}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X} \boldsymbol{Y} \quad (3.6)$$

where

$\boldsymbol{Y}$ is the vector of target asset returns

$\boldsymbol{X}$ is the matrix of constituent asset returns

The synthetic asset model is defined as:

$$\boldsymbol{SA} = \left\{\boldsymbol{Y} \in U_A; \; \boldsymbol{X} \subset U_A - \{\boldsymbol{Y}\}; \boldsymbol{\beta} \in R^{[42]}\right\} \quad (3.7)$$

where

U_A is the asset universe of TOP40 Index constituent shares

$\boldsymbol{Y} \in U_A$ is the target asset i.e. the TOP40 Index

$\boldsymbol{\beta}$ is a vector of predictor variable weights

$\boldsymbol{R}^{[42]}$ is any vector of length of dimension 42 x 1 whose elements take on real values only

Given such a model, we can derive the time-series which represents the associated statistical mispricing, that is equation 3.4 can be re-written as:

$$\boldsymbol{M}_t = \boldsymbol{Y}_t - \sum_{X_i \in X} \boldsymbol{\beta}_i \boldsymbol{X}_{i,t} \quad (3.8)$$

M_t is accepted as an aggregate portfolio containing the explanatory and predictor variables $\{Y_t, X_1, X_2, .., X_{42}\}^T$ with weights $\{1, \beta_1, \beta_2, .., \beta_{42}\}^T$.The returns on this portfolio possess the residual value of TOP40 Index, relative to its fair price relation permutation of predictor variables $SA_t(T) = \sum_{X_i \in X} \beta_i X_{i,t}$ and it is considered to be a version of the explanatory variable Y_t without trend in stochastic terms.

For this framework the collection of predictor variables X is regarded to act as a proxy for the undetected risks so as to behave as a mutual time varying trend in market returns. In optimising the correlation between the explanatory variable and the fair price relation the construction methodology cannot explain the return dynamics specific to the explanatory variable, however should rather optimise the predictor variables' sensitivities to common risk sources. For Equation (3.8) the result of the optimisation methodology is to synthetically construct combinations of the explanatory variable Y_t and the fair price relation $SA_t(Y)$ which have comparable sensitivities to the underlying but not directly observable economic risks which drive the explanatory variable price time series behaviour.

3.2.2 Risk Aggregation

The inclusion of risk is treated in an aggregate manner not from an economic risk point of view where the tracking error between the explanatory variable Y_t and the fair price relation $SA(Y_t, w)$ is taken to be the measure of risk. The fair price relation parameters' are those, w^* which minimise the variance of the return deviation time-series for the in sample data, this is:

$$w^* = \arg_w \min var(Y_t - SA_t(Y, w)) \tag{3.9}$$

where

w* are optimised beta coefficients from the training sample data

Y_t are the returns from the index

$SA_t(Y, w)$ are the returns from the synthetic asset

Each of the sensitivities s_i is indirectly defined in the optimisation of residual return variance (when speaking about the fair price relation coefficients w^*) and will tend to diminish the aggregate sensitivity to market risk, while never forcing any limits on allowable values for any of the sensitivities s_i.

The importance of coefficients β_i is in the regression technique that can be determined from past returns using a cointegration regression as per Granger (1983). We employ this methodology for

the fair price relation and it is constructed by regressing the 5-minute past returns of TOP40 Index on the past returns of a set of predictor variablesX_t are the returns at time t of specific names provided in Appendix Table A and i = 1,2…42 according to Bloomberg® ticker, such that:

$$E[Y_i] = SA(Y)_t = \sum_{C_i \in C} \beta_i X_{i,t} \tag{3.10}$$

Theβ_i parameters were determined by applying the regression methodology below. The predictor variables are considered to be the ideal hedge in so much as the ordinary least squares methodology provides that the deviation between the return time-series is minimised in the mean squared error (MSE). With the fair price relationship of the form shown in Equation (3.8) the mispricing$M_t$is the deviation in the fair price relationship that is the residual of the explanatory variable and the synthetic counterpart $M_t = Y_t - \sum_{X_i \in X} \beta_i X_{i,t}$.

We will consider instances where the explanatory and predictor variables are few and the optimal predictor variable weights $\beta_i$are constant stochastically speaking, here the standard ordinary least squares technique is used to determine the fair price relation.

The TOP40 Index is a market capitalisation weighted index, meaning the weight of the i^{th} asset in the index is dependent on the price. Very large relative moves in a lower capitalised share are required to increase its weight in the index.

Optimally derived groupings of the predictor variables are independent of market risk and will capture the explanatory variable specific aspects of the return time-series behaviour. This combination of explanatory and predictor variables are responsive to statistical anomalies due to the fact that provide a chance to profit from discernable components in explanatory variable specific return time-series dynamics in a way that is free of changes in the price value of the market as a whole, or other market risks. Further, the explanatory variable specific element of the return time series behaviour is only implicitly observable by market participants, and only in aggregation with market wide events, so it is reasonable to suppose that normality in the

behaviour will persist from this perspective that have not yet been arbitraged away by arbitrageurs.

3.3 Multicollinearity

A variant of the stepwise regression procedure is used to generate the fair price relation pricing model. The TOP40 Index was taken as the explanatory variable, and the number of predictor variable n_c was chosen to be 5 for both practical and statistical modelling reasons given later in the study. In the process of choosing an acceptable amount of predictor variables for the model, the risk to the modeller is either over- or under-specifying the model both of which errors come with their respective risks. An over parameterised model will likely be economically, practically intractable and my contain multicollinearity, whereas an under parameterised model will suffer from low explanatory power as observed by the R^2 statistic of the model.

The mispricing model is given as:

$$M_t = Y_t - \sum_{i=0}^{5} X_{i,t}\hat{\beta}_{c(i,s),t} \tag{3.11}$$

$Y_{T,t}$ is the return time series for the TOP40 Index in the estimation model

$\hat{\beta}_{c(i,s),t}$ is the estimated weight of the ith constituent asset included in the estimation model

A backwards stepwise multiple linear regression of the TOP40 Index returns in the 19,293 by 1 vector Y_t on the 5 most influential predictive variables in the 19,293-by-6 matrix X_t, in a. Distinct predictive terms appear in different columns of X_t. $\hat{\beta}$ is a 6-by-1 vector of optimised parameters for all of the terms in X_t, with the first column an intercept column, that is a vector of ones. The stepwise fit calculates the coefficient estimate values in $\hat{\beta}$ as follows: the model will start by including all of the TOP40 Index constituent shares' returns and will sequentially delete constituent shares from the regression. The default criterion is to include variables if the p-value is less than 1%.

When a possible predictor variable is not included in the final model, then neither is its parameter estimate in $\hat{\beta}$. If a term is included, then the parameter is in $\hat{\hat{\beta}}_t$ and that term is a result of Equation (3.11), that is the method does not re-introduce the terms it previously excluded from the model while determining the parameter estimates, See appendix Full Regression Output for output results in column with header "Status". This modelling methodology is consistent with Draper (1998).

An alternative method for choosing predictor variables is by predictor importance. Predictor variable importance determines coefficients by predictor variable importance trees by summing changes in the MSE due to splits on every predictor and taking the ratio the sum by the number of tree nodes at that step. The importance vector has an element for each predictor variable in the data used in the methodology.

Although the methods differ in application the predictor variables chosen by the two different methods result in the same significant predictor variables.

3.4 Identification of specific arbitrage opportunity

When the mispricing M_t contains a significantly observable mean-reversion element, a series of statistical arbitrage rules are designed to exploit the effect without need for an explicit forecasting model. The core assumption of an implicit statistical trading rule is that future return will be such that they tend to an average to minimise the deviation between the explanatory variable and the fair price relation combination. Meaning that over a time trading rules based on selling overvalued sets of assets and buying the undervalued sets of assets should realise abnormal cumulative returns relative to transaction costs.

Statistical arbitrage trading rules work by treating M_t as a portfolio combining the of the explanatory and predictor variables $\{Y_t, X_1, X_2, \dots, X_5\}$ with weights $\{1, -\beta_1, -\beta_2, \dots, -\beta_5\}$. This is equivalent to purchasing Y_t and selling $SA(Y)_t = -\sum_{C_i \in C} \beta_i X_{i,t}$, β is the single action of

purchasing the M_t. Further, selling Y_t and purchasing the $SA(Y)_t$ is thought of asselling the mispricing M_t.

In this research the arbitrage trading rules are conducted by means of parameterisation, which defines the optimal holding in the mispricing portfolio as a function of the current level of the mispricing:

$$posn(M_t, k, h) = -\frac{\sum_{j=1}^{h} sign(M_{t-j})|M_{t-j}|^k}{h} \quad \text{if} \quad |M_{t-j}| > \text{Cut-Off Bound} \qquad (3.12)$$

h is the holding period of the trading rule
k is the sensitivity of the trading rule to the magnitude of the mispricing

The negative sign indicates that the position should be opposite to the sign of the current sign of the mispricing M_t. The variable h is used as the holding period which is effectively a smoothing factor for the output of the trading rule. It reduces the number of transactions which are generated. For the case of regressing returns as conducted in this research exercise it is advised that k remain k < 0 ensuring $posn(M_t, k, h) > 0$.

The sign indicates that M_t should be purchased when negative, and sold when positive. The variable k allows the position size proportionally to the size of the mispricing M_t.

For the implicit statistical arbitrage a two sided 1% cut-off band on the histogram of the training data is used as an indicator of significant divergence in the statistical mispricing M_t i.e. as a trade signal, where deviations outside the band signal the opening of a position in the portfolio with the position closed on the first sequential deviation following the trade signal, this is equivalent to the position formula above for k = 0 and h =1. In price terms which it is easier to think in, a 1% deviation is equivalent 450 point deviation in of the statistical mispricing P_{M_t}.

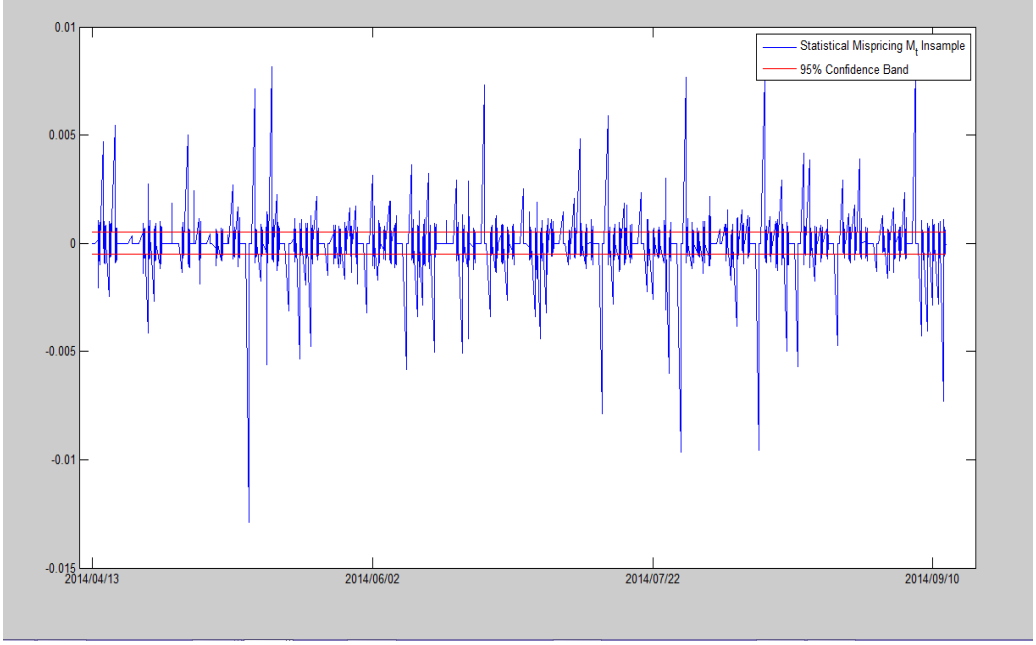

Figure 5: Time series of Mispricing M_t for entire Sample Research Period

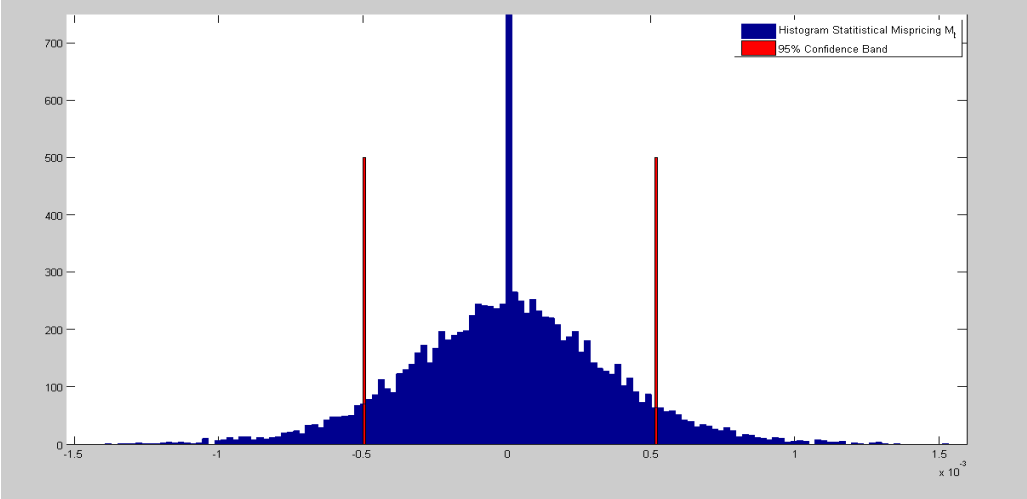

Figure 6: Histogram of Mispricing M_t for the In-Sample Research Period

Figures above demonstrate how the cut-off bands (in red) in the time series distribution (Figure 5) can be viewed in the empirical distribution (Figure 6). We observe that large deviations in the fair price relation have a lower weight than the more modest deviations, and consequently

extreme tail moves occur less frequently and do not persist through time. This observation is the basis on which the trading rule is constructed.

3.6 Summary

This section has covered the cointegrating vectors methodology used to determine the fair price relation for the statistical mispricing model M_t. The synthetic assets consists of optimal linear groupings of the predictor variable return time series and are created by a procedure that is based on the concept of a cointegrating vector regression, with possible additions in the form of stochastic parameters and high-dimensional models.
The motivation towards the pre-processing methodology is to find groupings predictor variables that are immune to market wide events and improve the possibly exploitable element of the explanatory variable return dynamics.

In the next section we review the results of the research, a range of tests are designed provided to categorise the presence of an exploitable element in the mispricings behaviour and uses tests for an autoregressive component, mean-reversion component and non-random walk type of behaviour.

Chapter 4: Results

4.1 Introduction

The results described in this section refer to the parameterised model described in Chapter 3The data consists of 5-minute frequency closing prices between 24 March 2014 and 03 October 2014. The 19,293 observations are divided into 2,000 in-sample observations used estimate the statistical mispricing model to price and 17,293 out-of-sample observations used to estimate the performance of the estimated model M_t. The in-sample regression uses logarithmic returns to estimate the co-efficient of the fair price relation and the out of sample study uses relative returns to asses model performance as the relative return are what the arbitrageur would realise had he held the mispricing portfolio.

4.2 Performance of the Fair Price Relation

As explained in methodology the fair price relation reduces to a multiple linear regression. The fair price relation is derived from the 2 000 in sample observations. The performance diagnostics are provided for the reduced model.

The model executes fitting the fair price relation well, by adhering to the predictor variable selection policy described in the section on multicollinearity. The p values for all the coefficients are significance level of$\alpha = 1\%$ for the t-Test on the individual coefficients.

```
Linear regression model:
    TOP40Index ~ 1 + BIL + NPN + OML + SAB + AGL

Estimated Coefficients:
                   Estimate       SE            tStat     pValue
    (Intercept)    9.9351e-06    5.8518e-06    1.6978       0.089704
    BIL                 0.173     0.0065838    26.276    8.4336e-131
    NPN              0.098439     0.0038369    25.656    1.4297e-125
    OML               0.08035     0.0061298    13.108     1.0479e-37
    SAB               0.14608     0.0062865    23.237    7.9991e-106
    AGL              0.088542      0.004858    18.226     9.7633e-69


Number of observations: 1986, Error degrees of freedom: 1980
Root Mean Squared Error: 0.000261
R-squared: 0.741,  Adjusted R-Squared 0.74
F-statistic vs. constant model: 1.13e+03, p-value = 0
```

Regression Output 1: Model Estimation and Regression Fit Diagnostics

The F-statistic Regression Output 1 is used to test the significance of the estimated regression or equivalently the joint significance of the predictor variables. F-statistic values in the ANOVA analysis from Regression Output 2 are used to assess the joint significance of the predictor variables in the model. Further F-Test affirms that the coefficients are jointly significant also at 99% confidence level. This statistic should be compared with an F(m, T − k), which in this case is an F(5, 1980). The test statistic values is1132.8, with p = 0. This suggests we reject the null hypothesis of joint insignificance. The test statistic clearly exceeds the critical values at both the 5% and 1% levels, and hence the null hypothesis is rejected. It would thus be concluded that the null hypothesis is not supported by the data. The fair price relation is statistically significant.

	SumSq	DF	MeanSq	F	pValue
Total	0.00051968	1985	2.618e-07		
Model	0.00038506	5	7.7012e-05	1132.8	0
Residual	0.00013461	1980	6.7986e-08		
. Lack of fit	0.00013461	1380	9.7546e-08	2.3499e+31	0
. Pure error	2.4906e-36	600	4.151e-39		

Regression Output 2: ANOVA (Analysis of Variance) Table

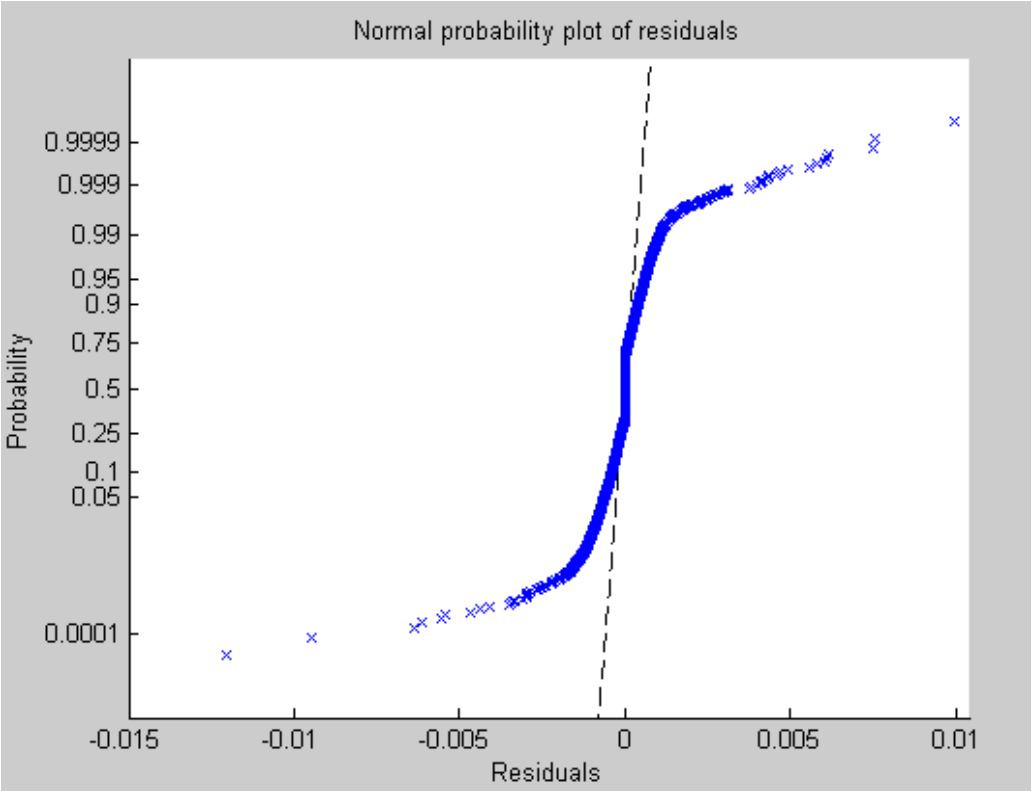

Figure 7: Quantile-Quantile plot of mispricing M_t Residuals vs. Standard Normal Distribution

For small magnitude deviations the mispricing M_t behaves similar to a Standard Normal Distribution this within the 0.25 to .9 probability bounds, however with this research we are not concerned with the explicit model performance but use the tails as an indicator of deviation. With regard to the implicit statistical arbitrage it S-shape informs us that significant absolute deviations occur more infrequently, with the "infrequence" positively correlated to the absolute magnitude of the residual.

Test which rely on the autocorrelation function (ACF) are usually implemented to find short term price behaviour meanwhile unit-root tests are designed to differentiate between long-term price behaviour. A test that relies on the autocorrelation function (ACF) are most effective at detecting short-term effects such as such as momentum or short-term correction effects. However the unit-root tests and variance ratio tests are effective at identifying long term effects, such as distinguishing between stationary and non stationary behaviour.

The graph below shows that the significant auto correlated lags are the 1st lag (5-minute) and the 12th lag (60th minute), with the short term significant lag negatively correlated (mean reversion) whereas the longer term significant lag positively correlated. It is this mean reversion property that is exploited in the contrarian trading strategy for the implicit statistical arbitrage.

The Durbin-Watson test assesses whether there is 1st order lag autocorrelation among the residuals or not. The test statistic DW_{M_t}= 2.1188 and p_{DW} = 0.6285 for the in-Sample observations when we perform a two-sided Durbin-Watson, the results suggest that the residuals are not auto correlated.

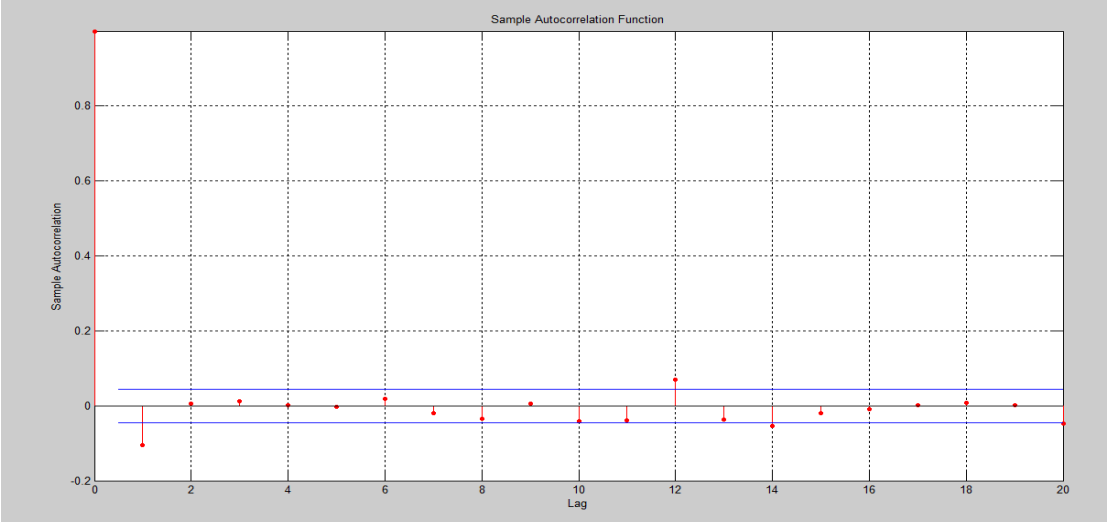

Figure 7: Autocorrelation Function for the mispricing M_t

The structure of the scatter plot (below) of the 1 period (5-minute) lagged residuals is evenly dispersed this informs us that the explicit residuals might expect to the explicit model to posses some forecasting ability.

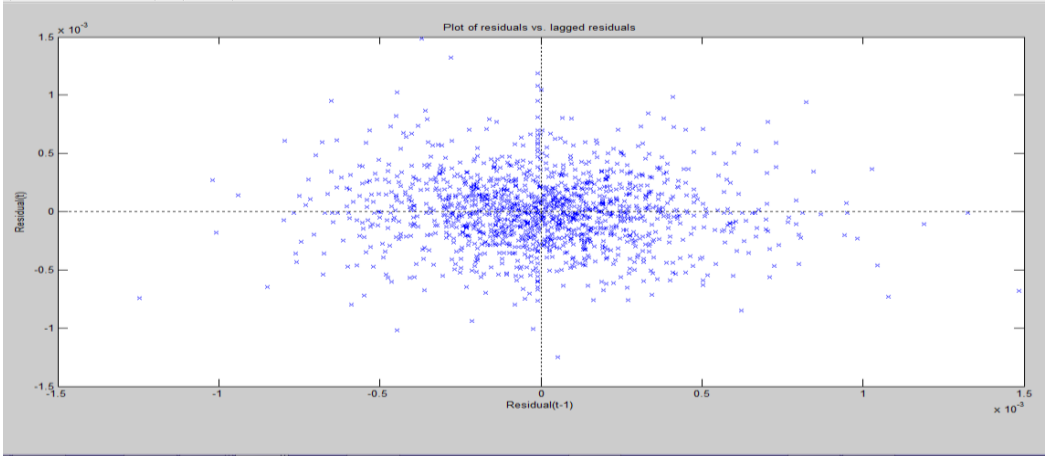

Figure 8: Scatter-plot of Residuals versus 1 Period (5-minute) lagged residuals

4.3 Out of Sample Performance

Tests and output designed examine the performance of the fair price relation over time. Refer back to Equation (2.1), where the risky statistical arbitrage is considered as abnormal returns beyond the transaction costs incurred. As shown in Figure 9, the out of sample performance we consider the performance of the trade strategy for varying transaction cost parameters (0.3%, 0.5%,0.7%), this demonstrates the sensitivity of the profitability of the implicit statistical arbitrage strategy to the magnitude of the transaction costs.

For the purpose of this research the transaction costs are assumed to be 0.30%. The cumulative returns net of transaction costs are 90.15% holding period return from 13 March 2014 to 03 October 2014; this is equivalent to a 283% effective annualised return. These returns exceed that of reference study Lakonishok et al. (1994), with the fair price relation being a less economically tractable relation in comparison to the CAPM.

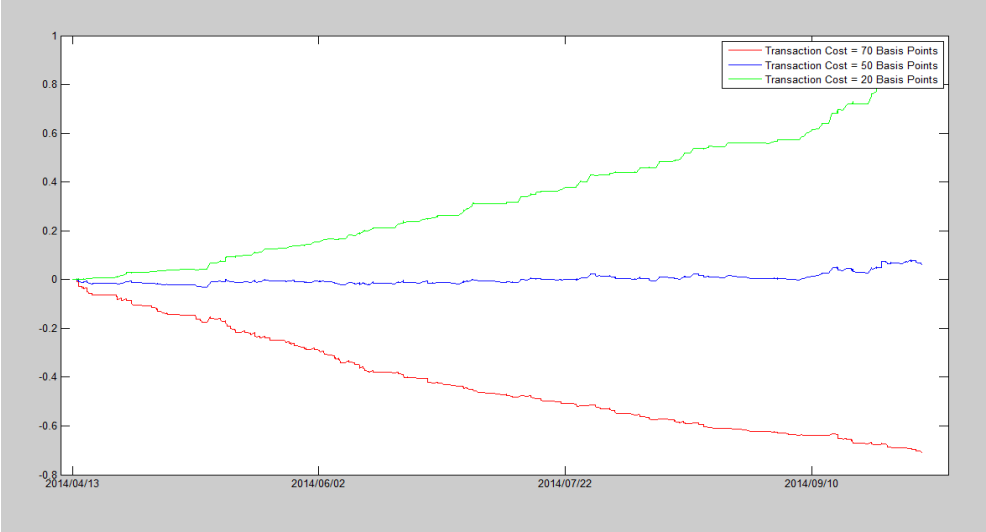

Figure 9: Cumulative Returns for varying Transaction Costs

Chapter 5: Conclusion

This research introduced and developed a model which uses computational modelling to exploit small but consistent regularities in traded asset price dynamics. Within the methodology different techniques are applied to parts of the modelling process and the performance analysis process as required. Extensions of the econometric methodology of cointegration were developed and are suitable for use in cases where the parameters of the underlying fair price relationship are stochastic, however for the purpose of this research static parameters were derived and applied to an out of sample set.

We consider statistical arbitrage as a generalisation of the pure riskless arbitrage strategies which are based on predefined relationships between financial assets, typically between derivative instruments such as options and futures contracts and the "underlying" assets upon which the derivatives are based. From the statistical arbitrage perspective, we apply our extended cointegration methodology to identify a statistical fair price relationship between a set of related asset prices. Deviations from a theoretical no-arbitrage relationship are considered as mispricings, which represent potential opportunities for riskless arbitrage, we considered deviations from the analogous statistical fair-price relationships as potential opportunities for statistical arbitrage. The research introduced the pair's trade as the original econometric model required to examine the fair price relation. The fair price relation and target asset were then used to derive the statistical mispricing. Based on the presence of an exploitable mean reversion component in the statistical mispricing and transaction cost assumption, we tested for the presence of abnormal returns as signalled by the statistical mispricing.

This approach can be considered the purest method of evaluating the added-value which is provided by a computational modelling approach to investment finance, since profits and losses of the resulting models are almost entirely independent of the underlying movements in the market as a whole and instead reflect only the informational advantage, if any, which is provided by the models themselves.

Furthermore, the significance of the resulting performance can be evaluated not only from a statistical perspective but also from a practical perspective in which the economic advantages of the models can be assessed after market frictions such as transaction costs have been taken into

account. In principle, the risk and return of the strategy can either be multiplied through leverage (up or down) and/or overlaid with a market-timing component on top of the existing static trading strategy. Thus the benefits of our approach are potentially of value to active fund managers in general, as well as arbitrageurs and hedge funds in particular.

In the remainder of this chapter we give the main results of the research and offer recommendations.

5.1 Research findings

The output provided from Chapter 4 supports existence of a statistically significant fair price relation and forms the basis for the implicit statistical arbitrage model. We take using a 9x longer out of sample data period than in sample data period as initial evidence and the ability to generate positive cumulative returns in excess of transaction costs as further evidence towards the parameters of the fair price relation as stationary through time.

The mean reversion element in the statistical mispricing M_t (the residuals of the fair price relation) has proved to be persistent while the magnitude M_t a reliable tool for generating of a contrarian trading rule in the event of large divergences. The potential to generate the momentum effect and correspondingly abnormal returns is extremely sensitive to the magnitude of the transaction costs incurred in implementing the implicit Statistical arbitrage strategy.

In line with Bernardo and Ledoit (2000) the research demonstrates after assuming a model form for the statistical arbitrage, 99^{th} percentile deviations in the synthetic asset versus the TOP40 Index pricing kernels, yield statistically significant cumulative returns over the research horizon provided market frictions are less than 54 basis points per transaction.

5.2 Further Research

The model identifies persistent anomalies assuming static parameter coefficients. The first extension offered is to allow for time adaptive model parameters through a rolling OLS

regression. The Equation (3.10) above can be applied in a classical linear regression framework over a rolling time window to identify the shares that will be used to construct the synthetic index portfolio. This research did not consider the Kalman filter extension as it allows parameters (β_i becomes $\beta_i(t)$) that is to be stochastic and it is preferred to the rolling ordinary least squares for time adaptive coefficient estimation as per Dunis and Shannon (2005). Details of the regression and modelling methodology can be found in Harvey (1981) and Hamilton (1994).The findings under the static coefficient model demonstrate the fair pair relation holds over research horizon, given only fractional estimation window, possibly suggesting stochastic coefficients may add to the performance of the basic model only marginally.

The third part of the methodology controls the risks posed by model selection and performance instability through actively encouraging diversification across a "portfolio of models". A novel population-based algorithm for joint optimization of a set of trading strategies is suggested and inspired both by genetic and evolutionary algorithms and by modern portfolio theory.

Appendix

Full Regression Output

Short Code	Full Name	Coefficient Estimate	Standard Error	Status	Importance Predictor Rank
BIL	BHP Billiton PLC	10.99%	0.20%	In	1
SAB	SABMiller PLC	9.52%	0.17%	In	4
CFR	Cie Financiere Richemont SA	9.22%	0.19%	Out	7
MTN	MTN Group Ltd	6.73%	0.17%	Out	8
NPN	Naspers Ltd	6.04%	0.11%	In	2
SOL	Sasol Ltd	5.61%	0.17%	Out	6
AGL	Anglo American PLC	5.52%	0.17%	In	5
OML	Old Mutual PLC	3.48%	0.21%	In	3
SBK	Standard Bank Group Ltd	2.43%	0.16%	Out	15
FSR	FirstRand Ltd	2.34%	0.18%	Out	18
BTI	British American Tobacco PLC	2.04%	0.22%	Out	28
REM	Remgro Ltd	1.85%	0.13%	Out	21
MNP	Mondi PLC	1.67%	0.16%	Out	30
APN	Aspen Pharmacare Holdings Ltd	1.50%	0.14%	Out	33
GRT	Growthpoint Properties Ltd	1.38%	0.17%	Out	12
ANG	AngloGold Ashanti Ltd	1.31%	0.09%	Out	22
SHF	Steinhoff International Holdings Ltd	1.30%	0.13%	Out	38
IMP	Impala Platinum Holdings Ltd	1.28%	0.11%	Out	15
SLM	Sanlam Ltd	1.28%	0.14%	Out	23
LHC	Life Healthcare Group Holdings Ltd	1.10%	0.11%	Out	25
SHP	Shoprite Holdings Ltd	1.08%	0.14%	Out	26

Short Code	Full Name	Coefficient Estimate	Standard Error	Status	Importance Predictor Rank
BGA	Barclays Africa Group Ltd	1.07%	0.10%	Out	16
REI	Reinet Investments SCA	1.07%	0.13%	Out	24
NED	Nedbank Group Ltd	1.06%	0.12%	Out	20
TBS	Tiger Brands Ltd	0.98%	0.12%	Out	35
INP	Investec PLC	0.94%	0.21%	Out	10
WHL	Woolworths Holdings Ltd/South Africa	0.94%	0.14%	Out	30
BVT	Bidvest Group Ltd	0.85%	0.16%	Out	29
KIO	Kumba Iron Ore Ltd	0.85%	0.11%	Out	39
ITU	Intu Properties PLC	0.83%	0.14%	Out	13
DSY	Discovery Ltd	0.79%	0.15%	Out	19
VOD	Vodacom Group Ltd	0.74%	0.14%	Out	41
RMH	RMB Holdings Ltd	0.71%	0.13%	Out	32
MDC	Mediclinic International Ltd	0.69%	0.11%	Out	34
IPL	Imperial Holdings Ltd	0.53%	0.12%	Out	14
INL	Investec Ltd	0.52%	0.19%	Out	36
AMS	Anglo American Platinum Ltd	0.38%	0.07%	Out	11
ASR	Assore Ltd	0.21%	0.05%	Out	27
CCO	Capital & Counties Properties PLC	0.11%	0.15%	Out	40
MPC	Mr Price Group Ltd	0.10%	0.12%	Out	42
MND	Mondi Ltd	0.09%	0.16%	Out	37
EXX	Exxaro Resources Ltd	-0.47%	0.12%	Out	31

Matlab Code for Research Study

```
%% Modelling and Results
disp('Running Regression')
BusHrsPriceMat  = priceMat(BusHrsIndFind,:);
BusHrsPriceTimeMat = timeMesh(BusHrsIndFind,:);
returnMat = log(BusHrsPriceMat(2:end,:)./BusHrsPriceMat(1:end-1,:));
returnTimeMat = BusHrsPriceTimeMat(2:end);
ALSI40 = returnMat(:,end);
Constituents = returnMat(:,1:end-1);

x_t = [ones(size(Constituents,1),1) Constituents];
ConstituentNo = size(Constituents,2);

tree = RegressionTree.fit(x_t,ALSI40);
imp = predictorImportance(tree);
a = sort(imp');
reducedRegressionColID = [find(imp'==a(43)) find(imp'==a(42)) find(imp'==a(41)) find(imp'==a(40)) find(imp'==a(39))];

trainingDataNo = 2000;
CooksOutliers = [10,110,200,300,400,800,900,1000,1100,1300,1400,1700,1800,2000];
OutofSampleDataNo = TimeIntervalNo - trainingDataNo;
InSampleRowIndex = 1:trainingDataNo;
InSampleRowIndex(CooksOutliers) =[];

[b,bint,residuals_linear,rint,stats] = regress(ALSI40(InSampleRowIndex),x_t(InSampleRowIndex,reducedRegressionColID),.01);
StaticM_t = ALSI40(trainingDataNo + 1:end) - x_t(trainingDataNo + 1:end,reducedRegressionColID)*b/sum(b);
StaticM_t_OutSampleTimeMat = returnTimeMat(trainingDataNo + 1:end);
StaticM_tInsample = ALSI40(InSampleRowIndex) - x_t(InSampleRowIndex,reducedRegressionColID)*b/sum(b);

stepmdl = LinearModel.stepwise(Constituents(InSampleRowIndex,reducedRegressionColID - 1),ALSI40(InSampleRowIndex));
ds = mat2dataset([x_t(InSampleRowIndex,reducedRegressionColID), ALSI40(InSampleRowIndex)],'VarNames',[PrintNames'; justNames(end)]);
mdl = LinearModel.fit(ds);
mdl = LinearModel.fit(ds);
%% Trading rule and Cumulative Returns
alphaStudy = 0.01;
TransactionCostsStudy = 0.01/100*30;
PositveDivergenceTradeSignalLevel = quantile(StaticM_t,1-alphaStudy);
NegativeDivergenceTradeSignalLevel = quantile(StaticM_t,alphaStudy);
PositveDivergenceTradeSignalLevelVec = PositveDivergenceTradeSignalLevel*ones(size(StaticM_t,1),1);
NegativeDivergenceTradeSignalLevelVec = NegativeDivergenceTradeSignalLevel*ones(size(StaticM_t,1),1);
posn1SignalPositiveIND = StaticM_t > PositveDivergenceTradeSignalLevel;
negan1SignalNegativeIND = StaticM_t < NegativeDivergenceTradeSignalLevel;


posn1SignPositiveDivergenceIND = [posn1SignalPositiveIND;0] - [0; posn1SignalPositiveIND];
posn1SignNegativeDivergenceIND = [negan1SignalNegativeIND;0] - [0; negan1SignalNegativeIND];
posn1SignPositiveDivergenceIND = abs(posn1SignPositiveDivergenceIND(1:end-1));
posn1SignNegativeDivergenceIND = abs(posn1SignNegativeDivergenceIND(1:end-1));


TransactionCostVec = zeros(OutofSampleDataNo,1);
for i = 2:OutofSampleDataNo
    if posn1SignPositiveDivergenceIND(i-1) ==  1 || posn1SignNegativeDivergenceIND(i-1) == 1
        TransactionCostVec(i) = TransactionCostsStudy;
    elseif posn1SignPositiveDivergenceIND(i) == -1 || posn1SignNegativeDivergenceIND(i-1) == -1
        TransactionCostVec(i) = TransactionCostsStudy;
    else
        i = i;
    end
end

for i = 2:OutofSampleDataNo
    if posn1SignPositiveDivergenceIND(i-1) == 1 && posn1SignPositiveDivergenceIND(i) == 0
    if posn1SignPositiveDivergenceIND(i-1) == 1 && posn1SignPositiveDivergenceIND(i) == 0
       posn1SignPositiveDivergenceIND(i) = 1;
    elseif posn1SignNegativeDivergenceIND(i-1) == 1 && posn1SignNegativeDivergenceIND(i) == 0
           posn1SignNegativeDivergenceIND(i) = 1;
    else
       i = i;
    end
end

i = 2;
for i = 2:OutofSampleDataNo
    if posn1SignPositiveDivergenceIND(i-1) == 0 && posn1SignPositiveDivergenceIND(i) ~= 0
       posn1SignPositiveDivergenceIND(i-1) = 0;
    elseif posn1SignNegativeDivergenceIND(i-1) == 0 && posn1SignNegativeDivergenceIND(i) ~= 0
           posn1SignNegativeDivergenceIND(i-1) = 0;
    else
       i = i;
    end
end
posn1SignPositiveDivergenceIND = abs(posn1SignPositiveDivergenceIND);
posn1SignNegativeDivergenceIND = -abs(posn1SignNegativeDivergenceIND);
posn1PosReturns = posn1SignPositiveDivergenceIND.*StaticM_t;
posn1NegReturns = posn1SignNegativeDivergenceIND.*StaticM_t;

i=0;
TotalCumReturns = [posn1PosReturns(1) + posn1NegReturns(1); zeros(size(OutofSampleDataNo-1,1))];
for i = 2:OutofSampleDataNo
    TotalCumReturns(i) = (TotalCumReturns(i-1)+1)*(1+posn1NegReturns(i)+posn1NegReturns(i))*(1-TransactionCostVec(i))-1;
end
```

Bibliography

1. Alexander C (2001). Market Models: A Guide to Financial Data Analysis, John Wiley & Sons Ltd.

2. Avellaneda M. and Lee J (2010).Statistical Arbitrage in the U.S Equity Markets. Quantitative Finance, 10, 761-782

3. Box P. and Jenkins M (1976). Time Series Analysis: Forecasting and Control (2nd Edition).

4. Brooks C (2008). Introductory Econometrics for Finance (Second Edition).

5. Burges A (1996). A Computational Methodology for Modelling the Dynamics of Statistical Arbitrage, University of London, London Business School.

6. Burgess A. N. (2003). Using Cointegration to Hedge and Trade International Equities. In Dunis, C., Laws, J. And Naïm, P. [eds.] Applied Quantitative Methods for Trading and Investment. John Wiley & Sons, Chichester, 41-69.

7. Burgess A. N. (2006). A computational Methodology For Modelling the Dynamics of Statistical Arbitrage. University of London, London Business School.

8. Chatterjee S and Hadi A.S. (1996). Influential Observations, High Leverage Points, and Outliers in Linear Regression." Statistical Science. Vol. 1, 1986, pp. 379–416.

9. Cuthbertson K and Nitzsche D (2004). Quantitative financial economics: Stocks, bonds and foreign exchange. Chichester, England: Wiley. Chicago (Author-Date, 15th ed.)

10. Dunis C. L. and Shannon G. (2005) Emerging Markets of South-East and Central Asia: Do They Still Offer a Diversification Benefit? Journal of Asset Management, 6, 3, 168-190.

11. Engle R and Granger J (1987). Co-integration and Error, Correction Representation, Estimation and Testing. Econometrica Vol 55, March 1987

12. Froot K, Scharfstein D and Stein J (1992). Herd on the Street: Informational inefficiencies in a Market with Short-Term Speculation, Journal of Finance XLVII (4): 1461-1484

13. Fung, W. and Hsieh, D. A. (1997). Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds, The Review of Financial Studies 10 (2): 275-302.

14. Galenko et al. (2007).Statistical arbitrage and high-frequency data with an application to Eurostoxx 50 equities

15. Gamzo (2013).Algorithmic Trading, Market Efficiency and the Momentum effect, University of the Witwatersrand.

16. Gatev, E Goetzmann W. N., &Rouwenhorst K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. Review of financial studies, 19(3), 797-827.

17. Granger, C. W. J. (1983), Cointegrated variables and error-correcting models, UCSD Discussion Paper.

18. Harvey A. C. (1981) Time Series Models, Philip Allan Publishers, Oxford. Hamilton

19. Jegadeesh N and Titman S (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. The Journal OF Finance , Vol. XLVIII, No. 1 March 1993

20. Johansen, S, (1988) Statistical Analysis of Cointegration Vectors. Journal of Economic Dynamics and Control

21. Jordan B. and Miller T (2008). Fundamentals of Investments - Valuation and Management (5th Edition), chapter 6 Toronto: McGraw-Hill Ryerson.

22. Jacobs, B.I., Levy, K.L (1995). More on long-short strategies. Financial Analysts Journal 51 (2), 88-90

23. Khandani E. and Lo A (2007). What Happened To The Quants In August 2007? Journal of Financial Markets, Elsevier, vol. 14(1), pages 1-46

24. Lakonishok J., Shleifer A Vishny R.W (1994). Contrarian Investment, Extrapolation, and Risk. Journalof Finance 49 (5), 1541–1578.

25. Lin Y.-X, Mccrae M. and Gulati, C. (2006). Loss Protection in Pairs Trading through Minimum Profit Bounds: A Cointegration Approach. Journal of Applied Mathematics and Decision Sciences, vol. 2006, 1-14

26. Lo A(2010). Hedge Funds: An Analytic Perspective (Revised and expanded ed.), Princeton University Press. p. 260.

27. Pole A (2007). Statistical Arbitrage: Algorithmic Trading Insights and Techniques. Wiley Publishers

28. Ross S (1976). The Arbitrage Theory of Capital Asset Pricing. Journal of Economic Theory 13 (3): 341–360.

29. Shleifer A (2000) Inefficient Markets An Introduction to Behavioral Finance,Clarendon Lectures in Economics, Oxford: Oxford University Press, 2000.