KU LEUVEN

FACULTY OF SCIENCE

Statistical Arbitrage in the European Equity Markets

Renat Kozhakhmetov

Supervisor: Prof. Wim Schoutens

Thesis presented in
fulfillment of the requirements
for the degree of
Master of Science in Statistics

Academic year 2018-2019

AFDELING
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be

© Copyright by KU Leuven
Without written permission of the promotors and the authors it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to KU Leuven, Faculteit Wetenschappen, Geel Huis, Kasteelpark Arenberg 11 bus 2100, 3001 Leuven (Heverlee), Telephone +32 16 32 14 01.

A written permission of the promotor is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

Contents

Preface

This paper has been written to introduce and explain the details of one of the first trading techniques based on statistical models with a specific focus on the European equity exchanges.

I would like to thank KU Leuven and LStat centre for giving me an opportunity to learn. I would like to express my high gratitude to Prof. Wim Schoutens for his advices, guidanaces over my studies and believing in me.

Summary

This masters thesis will study statistical arbitrage, one of the important quantitative concepts used by capital management firms. The principles of this concept serve as a basis for other arbitrage type trading methods.

Preceded by a brief introduction, a conceptual explanation of Pairs Trading strategy will be carried out in Chapter 2, with description of its main features. Chapter 3 studies Time Series models, one of the common modelling methods applied for arbitrage trading with backtesting conducted in the end of this chapter. Chapter 4 summarizes various types of statistical arbitrage used nowadays. This is followed by Chapter 5, where the factor model based statistical arbitrage will be discretely studied with a focus on defactored returns. Statistical arbitrage using defactored returns will be applied in practice in the end of the chapter by utilizing examples of securities from the main European exchanges.

Glossary

Arbitrage Pricing Theory or APT is an asset pricing model based on the idea that an asset's returns can be predicted using the linear relationship between factors representing systematic risk and expected return for an asset.

Capital Asset Pricing Model or CAPM is a theory which describes the relationship between systematic risk and expected return for assets, particularly stocks.

Efficient Market Hypothesis or EMH is a theory which implies the impossibility to make economic profits by trading on the basis of information.

Hedge Fund is an investment fund which pools capital from accredited private investors or institutional investors and invests in a variety of assets by employing risk management techniques.

Short Selling is the sale of borrowed stocks and other financial instruments with expectation that they can be bought back later at a lower price.

1. Introduction

Since its invention three decades ago at ”Morgan Stanley” banks trading operations desk, statistical arbitrage evolved from simple pairs trading strategy into different complex mathematical algorithms automated by computers. Despite these changes, the main principle of historical based reversion to the mean remains the same. Movements of two highly correlated securities should be monitored and appropriate decisions are made based on these movements [21].

Reversion to the mean or mean reversion is a financial economics concept, suggesting that prices of securities (usually returns) are distributed around the historical averages, the thought is that any price that deviates far from these averages will again rebound to its presumed conditions. The main advantage of the strategy is that it can be applied in any types of market trends: upward, downward, sideways and can be used both for buying and selling. This strategy is often referred to as ’market neutral’ or a strategy with a zero beta according to the Capital Asset Pricing Model (CAPM) [17]. They are strategies that are neutral to market returns, this means, that the return from the strategy is uncorrelated with the return from markets. Often associated with Hedge Funds, market neutral strategies largely rely on Short Selling, which is mostly banned in major European security exchanges.

One good example is that short selling is completely prohibited in Belgium since August 2008, which can be connected to the bankruptcy of ”Fortis” bank, which used to be the 20th largest financial institution in the world by revenue before the financial crisis of 2008. The other three European countries followed this example, when in 2011, France, Spain and Italy announced that they have accepted the same regulatory policy. However, in the major exchanges of Europe, for example in London and Frankfurt such operations are still allowed.

In the early days of the statistical arbitrage, majority of returns from employing the strategy were highly profitable by bringing concerns of various financial intermediaries on why shares of some large companies were being sold despite promising fundamentals? As time passed, the pattern of the strategy became obvious and many small capital financial organizations followed the movements of hedge-funds and large banks, employing the strategy by making inefficient markets - efficient, the well known observation from the Efficient Market Hypothesis theory introduced by Eugene Fama [8] and later formalized by Burton Malkiel [10]. The theory could explain why most trading strategies stop profiting after some periods and why prices on financial market’s distribution can be compared to a random-walk.

In 1900, a French mathematician, Louis Bachelier introduced for the first time a mathematical model of Brownian motion and its use for valuing stock options. This was the first paper to use advanced mathematics for finance; rendering Bachelier a pioneer in the study of stochastic processes. The work became an inspiration for many random-walk theorists, but was criticized by many financial economists as well [2].

In 2003, Malkiel published an article, where he has adjusted his claims that market movements can be predictable up until some certain degree [11]. Among famous critics of the theory to mention, is one of the most successful investors of all time and the third wealthiest person in the World (according to Forbes, June 2019) - Warren E. Buffet. Moreover, how consistent profits of 'Morgan Stanley' bank's team or other firms using quantitative methods could be explained in the first place? The theory about random walks of prices on securities exchanges is still under research and strong controversial.

Let us follow to the next chapters, where we will breakdown the mechanisms of the statistical arbitrage and will observe the driving forces behind the technique and study its various opportunities and disadvantages.

2. Concepts of Statistical Arbitrage

2.1 Pairs Trading

Pairs trading is a market neutral strategy in its most primitive structure. The market neutral portfolios are constructed using two correlated securities, consisting of a long position in one security and a short position in the other, in a predetermined ratio. At any given time, the portfolio is associated with a quantity called the spread, which is the difference between two stocks. This quantity is computed using the quoted prices of securities usually forms a time series. The spread is in some ways related to the residual return component of the return. Pairs trading involves putting on positions when the spread differs substantially from its mean value, with the expectation that the spread will revert back. The positions can also be put on the opposite direction, with the expectation of divergence [21]. For example, let us observe two stocks from The Euro Stoxx 50 Index, which is a stock index of Eurozone.

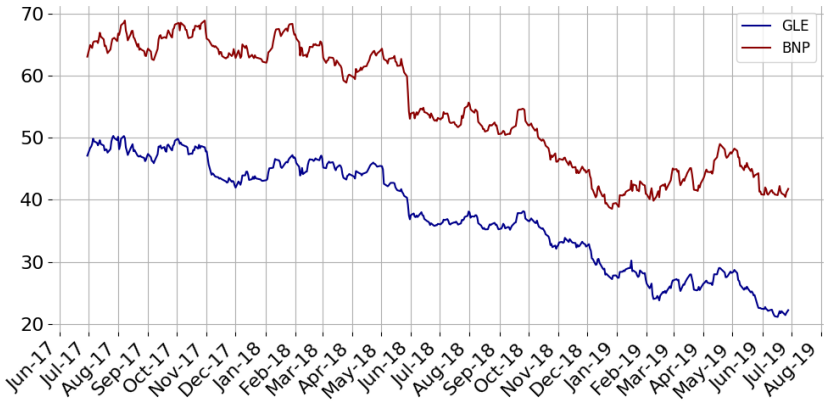


Figure 2.1: Daily Closing Prices of GLE and BNP.

As we can observe on Figure 2.1, from July 2018 until June 2019, the adjusted close prices for the shares of ”Societe Generale S.A.” (GLE) and ”BNP Paribas S.A.” (BNP), both from European financial sector. Prices moved in a similar pattern until some convergence and divergences of range between stocks. Both companies showed a downward trend, which can be linked to the European Central bank’s decision to target the position of holding low interest rates in order to stimulate the Eurozone’s slowly declining economy and major global concerns about Brexit.

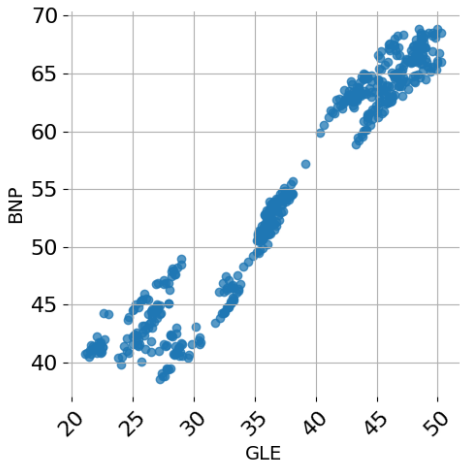


Figure 2.2: Scatter-plot of GLE and BNP.

On Figure 2.2 above, scatter plot is given for these adjusted closing prices for both axis. We can observe strong positive correlation between the two stocks with the coefficient equal to 0,86. However, correlation is not a strong enough requirement for the strategy, a more stronger relation is required - such as cointegration, which will be later discussed in detail.

2.1.1 Market Neutral Strategies

Market neutral strategies are the strategies exploited by investors, which can gain profits despite directions of the markets by constructing such portfolios to ultimately decrease directional risks (caused by market). There are two main market-neutral strategies: statistical arbitrage and fundamental arbitrage. Fundamental market-neutral investors use analysis of financial statements to make predictions, whilst the former method is based on quantitative methods.

For example, let us consider two portfolios A and B, with positive betas β_A and β_B and with returns r_A and r_B

$$r_A = \beta_A r_m + \theta_A$$
$$r_B = \beta_B r_m + \theta_B \ ,$$

where r_p is the return on the asset, r_m is the return on the market portfolio, and the β serves as a leverage number of the asset return over the market return, therefore βr_m is the market or systematic component of the return and θ_p is a non-systematic component [21].

We now construct a portfolio AB, by taking a short position on r units of portfolio A and a long position on one unit of portfolio B. The return on this portfolio is given as $r_{AB} = -r.r_A + r_B$. Substituting for the values of r_A and r_B, we have:

$$r_{AB} = -r.(\beta_A r_m + \theta_A) + (\beta_B r_m + \theta_B)$$

after opening the brackets and switches we have the following formula:

$$r_{AB} = (-r\beta_A + \beta_B).r_m + (-r.\theta_A + \theta_B) \ . \tag{2.1}$$

Thus, the combined portfolio has an effective beta of $-r\beta_A + \beta_B$. This value becomes zero, when $r = \frac{\beta_B}{\beta_A}$. Therefore, by a reasonable choice of the value of r in the long-short portfolio we have created a market neutral portfolio. In such aspect, we can protect our portfolio from directional risks of the market [21].

2.1.2 Noise Models

The first generalized rule of statistical arbitrage was primitive and very profitable. The original strategy worked as follows: find a correlated pair of stocks, calculate the spread, calculate the mean spread and standard deviation for some optional time period. Further, buy the lower priced stock and sell the higher priced stock when the spread is equal (or higher) to the mean value plus one standard deviation and vice-versa; buy the higher priced stock and sell the less priced stock when the spread is equal (or lower) to to the mean value minus one standard deviation.

Let us exercise the rule on the shares of our financial companies. Below on Figure 2.3, the plot for the spread between GLE and BNP banks for the second quarter of 2019 is given.

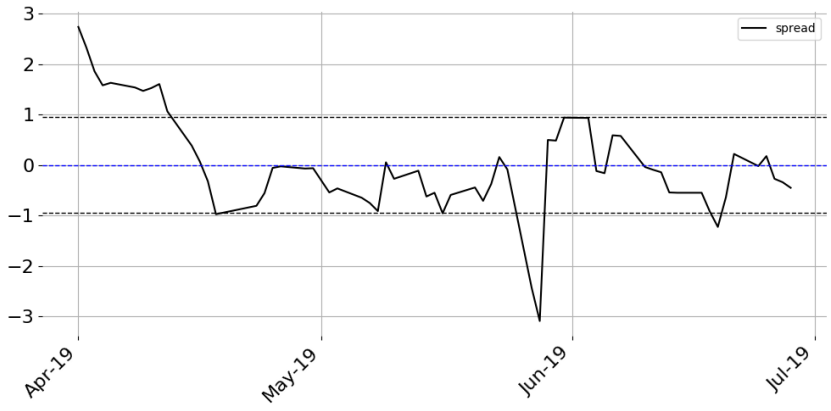


Figure 2.3: Spread for Daily Closing Prices of GLE and BNP.

As we can see from from the plot, we had several trading signals, where the spread value fluctuated outside (or equal to) our deviation regions, where we could either sell the spread at upper dashed line (+$\sigma_S = +1$) and buy spread below at lower dashed line (-$\sigma_S = -1$) with both types of trade should be closed at the blue line ($\mu_S = 0$). Moreover, we could increase our signals by trading on the divergence of the spread or waiting to reach the opposite standard deviation level by doubling our profit [13].

However, one obstacle is, by increasing the number of signals, we could also increase our risks, which means incurring losses, which may exceed profits and having a final negative portfolio return. Another obstacle is that the spread margins fluctuate;

such that the mean value shifts from one value to another value. Therefore solutions such as the calibration of time periods (90 days in our case) are required with corresponding time period for calculating the mean and standard deviation of the present period [13]. Let us explain the algorithm of pairs trading statistical arbitrage in a more mathematical form.

2.1.3 Theoretical Framework

Let $p_{it} = ln(P_{it})$ be equivalent to logarithmic adjusted close prices and we assume that p_{it} follows a random-walk model, where $p_{it} = p_{i,t-1} + r_{it}$, with r_{it} equal to the return as a sequence of uncorrelated innovations. Time series of log-returns are anticipated to be mean-reverted, but we cannot trade on the mean-reversion of returns [5].

It was explained in the noise models section that the general idea is to buy an under-priced stock and to sell an over-priced stock, so we base our decisions on the oscillations in the spread value. Therefore, for our spread value (s_t), we can construct such a linear combination $s_t = p_{1t} - \beta p_{2t}$, which we assume is stationary and has a mean-reverting feature because both stocks are driven by a common component (e.g. financial sector).

Let us consider a portfolio with buying 1 share of company A and selling β shares of company B, which makes it equal to the difference of the spread in the time period. The return of our portfolio will have the following view:

$$r_{p,t+i} = (p_{1,t+i} - p_{1,t}) - \beta(p_{2,t+i} - p_{2,t+i}) \tag{2.2}$$

$$= (p_{1,t+i} - \beta p_{2,t+i}) - (p_{1t} - p_{2t}) \tag{2.3}$$

$$= s_{t+i} - s_t \ . \tag{2.4}$$

by using above expression we linked non-stationary time series to stationary series of portfolio's logarithmic returns.

Remaining aspect is an interpretation of the coefficient β, which is called cointegration coefficient, the concept, comparable to correlation and it implies that the ratio between two time series fluctuates around the mean. Presence of cointegration allows us to apply forecasting models, its formal approach of testing will be discussed in a later section.

The return on the portfolio is the increment to the spread value in the period i, thus our strategy will be focused solely on the equilibrium value (mean) of the spread equal to $E(s_t) = \mu_s$ and we will enter into the trade when the spread has substantial deviations (σ_s) from its mean value. The amount of this deviation should cover all the costs of trading and should generate realistic profits [19].

Let us allow η be equal to the trading costs and let $\triangle$ be our target deviation of s_t from its expected value. Then, the trading rules are following:

- When at time t: $s_t = p_{1t} - \beta p_{2t} = \mu_w - \triangle$, we buy a 1 share of company A and sell β shares of company B.

- When at time $t+i$: $s_{t+i} = p_{1,t+i} - \beta p_{2,t+i} = \mu_w + \triangle$, we sell back a 1 share of company A and buy back β shares of company B,

under conditions that $2\triangle > \eta$ and $i > 0$.

Therefore, the main task of statistical arbitrage pairs trading can be summarized into: finding cointegrated stocks, estimating cointegration coefficient β and single trade log-return $\triangle$, to propose valid forecasting models, to trade. Further, bid-ask spreads, broker's commissions can be calculated and stop-loss orders can be employed, which are orders for a stock-broker to sell a security when it reaches certain price levels in order to prevent large capital drawdowns [19].

Now, because our main interest will be the point forecast we can turn to explaining the operation of time series models and apply this model in a practical example.

3. Time Series Models

A stochastic process is a sequence of stochastic variables: $Y_1, Y_2, Y_3 ... Y_i$, where we observe the process from $i = 1$ until $i = T$, yielding a sequence of numbers: $y_1, y_2, y_3, ... y_T$, called time series [6]. Stochastic process is a theoretical or population counterpart and time series themselves can be defined as its sample [20].

Time series often have the same type of random behavior from one time period to the next modeled by stationary stochastic processes and the concept of (weak) stationarity is a foundation of statistical inference in time series analysis. A stationary stochastic process is a probability model for a time series with time-invariant behavior. We usually assume a weaker version of stationarity, which means that its mean and variance are time-invariant and the correlation between two observations depends only on the lag (time interval between them) [20]:

$$E[Y_i] = \mu, \ \forall i \tag{3.1}$$

$$Var(Y_i) = \sigma^2, \ \forall i \tag{3.2}$$

$$Cov(Y_i, Y_j) = \gamma(|i - j|), \ \forall i, j \ . \tag{3.3}$$

Weak stationarity provides basic framework for prediction and stationary time series are found to be mean-reverting time series. Financial time series usually non-stationary, otherwise it would be possible trade assets at lower price, wait for reversion to the mean price, then profit at higher prices [20]. However, returns of a portfolio can distribute around a mean of zero, hence a combination of two or more non-stationary price series can be applied to create a stationary portfolio (cointegration) [5].

The function $\gamma(h) = Cov(Y_t, Y_{t-h})$ is the autocovariance function of the process. The function $\rho(h) = Corr(Y_t, Y_{t-h})$ is the autocorrelation function of the process, which gives insight in the dependency structure of the process and the basic tool to study time series. If Y_t is weakly stationary, then the autocorrelation function or serial correlation $p(h)$ at lag h equals:

$$Corr(Y_t, Y_{t-h}) = \rho(h) = \frac{Cov(Y_t, Y_{t-h})}{\sqrt{Var(Y_t)Var(Y_t - h}} = \frac{Cov(Y_t, Y_{t-h})}{Var(Y_t)} = \frac{\gamma(h)}{\sigma^2} \ . \tag{3.4}$$

The plot of the calculated correlation against time intervals forms an estimation of the autocorrelation function, called the correlogram. It gives insight about the dependency structure of the process [6], where the x-axis specifies the h number of lags and y-axis specifies the correlations of series with itself $\hat{\rho}(h)$, at lag determined by x. We often see two lines, corresponding to the critical values of the test statistic $\sqrt{T}\hat{p}(h)$ for testing $H_0 : \rho(h) = 0$ for a specific value of h [6].

For creating useful time-series models, we need a stationary process without autocorrelation (a white noise); this is called an innovation process. Below are the condition of a time series Y_i to be considered as a (weak) white noise process with mean μ and

variance σ^2:

$$E[Y_i] = \mu, \ \forall i,$$
$$Var(Y_i) = \sigma^2, \ \forall i,$$
$$Cov(Y_i, Y_j) = 0, \ \forall i \neq j.$$

Future values of a white noise process are independent of the past and the present. Past values of a white noise process have no information, for this reason any future value of the process is the mean value μ [20].

Let us proceed further, by assuming that $Y_1, ..., Y_N$ are observations from a stationary process, we can estimate their parameters by applying formulas below:

$$\hat{\mu} = \bar{Y} = \frac{1}{N} \sum_{t=1}^{N} Y_t$$
$$\hat{\sigma}^2 = S_Y^2 = \frac{1}{N-1} \sum_{t=1}^{N} (Y_t - \bar{Y})^2$$
$$\hat{\gamma}(h) = \frac{1}{N-h} \sum_{t=1}^{N-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y}), \quad 0 \leq h \leq N-1$$
$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\frac{1}{N-h} \sum_{t=1}^{N-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})}{\frac{1}{N-1} \sum_{t-1}^{N} (Y_t - \bar{Y})^2} \ .$$

Some assumptions motivated by parsimony: we do not need a different expectation or variance for every Y_t; since the covariance between Y_t and Y_{t+h} does not depend on t, all $n - h$ pairs of data points that are separated by lag h time units can be used to estimate $\gamma(h)$. Generally, $\frac{1}{N-h}$ is replaced by $\frac{1}{N}$, since h is mostly very small compared to N [20].

Stationary process requires estimation of an infinite number of parameters (e.g. $\rho(1), \rho(2), ..$), thus we need models, a class of stationary processes entirely characterized by a (small) finite number of parameters and constructed using innovations [20].

One common approach in economics and finance is using linear time series models, which can provide accurate approximations in real applications [6]. The time series is considered as linear if it is written in the form as:

$$Y_t = \mu + \sum_{i=1}^{\infty} \psi_i \epsilon_{t-i} \ , \tag{3.5}$$

where μ is mean of Y_t and ϵ_t is independent and identically distributed white noise with mean 0 and variance σ_ϵ^2.

If Y_t is stationary, we can obtain its mean and variance:

$$E(Y_t) = \mu \tag{3.6}$$

$$Var(Y_t) = \sigma_\epsilon^2 \sum_{i=0}^{\infty} \psi_i^2 \tag{3.7}$$

$$Cov(Y_t, Y_{t-l}) = \gamma(l) = \sigma_\epsilon^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+l} \tag{3.8}$$

$$\rho(l) = \frac{\gamma(l)}{\gamma(0)} = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+l}}{\sum_{i=0}^{\infty} \psi_i^2} \quad \text{for } l \geq 0 \ . \tag{3.9}$$

For stationary time series, $\rho(l)$ converges to 0 as l increases [20].

In time series models, differencing operation permits to convert non-stationary series into stationary, where new time series consist of the changes of the original series [20]. The differencing operator is defined by $\triangle = 1 - L$, where L is the lag operator so that $\triangle Y_t = Y_t - LY_t = Y_t - Y_{t-1}$. The k-th order differencing operator is defined by:

$$\triangle^k Y_t = (1 - L)^k Y_t = \sum_{l=0}^{k} = \binom{k}{l} (-1)^l Y_{t-l} \ . \tag{3.10}$$

Commonly, a unit root test is used in order to choose whether we should apply differencing operation or not. Tests such as: Dickey-Fuller (DF) and augmented Dickey-Fuller (ADF) help to detect the permanent effect of innovations and the presence of deterministic or stochastic trends [20].

3.1 Classical Models

3.1.1 Autoregressive Process

When Y_t has a statistically significant lag 1 autocorrelation, the lagged value Y_{t-1} can be useful in forecasting Y_t [20]. Simple model that makes use of such predictive power can be an autoregressive process of the first order, a regression model that utilizes the dependent relationship between a current observation and observation over a previous period.

The process $\{Y_t\}$ is considered as a first order autoregressive process AR(1) if:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t, \quad \forall t \ . \tag{3.11}$$

Assuming that series are weakly stationary, we have:

$$E(Y_t) = \mu, \quad \forall t$$

$$Var(Y_t) = \gamma(0) = \frac{\sigma_\epsilon^2}{1 - \phi_1^2}, \quad \forall t \Rightarrow \gamma(0) > \sigma_\epsilon^2, \text{ unless } \phi_1 = 0$$

$$Cov(Y_t, Y_{t-h}) = \gamma(h) = \frac{\sigma_\epsilon^2 \phi_1^{|h|}}{1 - \phi_1^2}, \ \forall t, h$$

$$Corr(Y_t, Y_{t-h}) = \rho(h) = \phi_1^{|h|}, \ \forall t, h \ .$$

Since $E(Y_t) = \phi_0 + \phi_1 E(Y_{t-1})$, we have $\mu = \phi_0 + \phi_1 \mu$, therefore $E(Y_t) = \mu = \frac{\phi_0}{1-\phi_1}$, with necessary and sufficient conditions for AR(1) to be weakly stationary is $|\phi_1| < 1$. Otherwise, if $\phi = 1$, then it takes a random walk form: $Y_t = Y_{t-1} + \epsilon_t$.

The general form of the process, which can be considered as a multiple linear regression model with lagged values of the time series as predictor variables, can be written in the following way:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + ... + \phi_p Y_{p-1} + \epsilon_t \ , \tag{3.12}$$

where $\phi_0 = (1-(\phi_1+...+\phi_p))\mu$. In order to determine the order of a process, partial autocorrelation function is used, which is the correlation between Y_t and Y_{t-h}, conditional given $Y_{t-1}, ..., Y_{t-h+1}$ [6].

Parameter estimation of autoregressive process can be conducted by using conditional least-squares method and maximum likelihood estimation, both usually calculate identical results. The least squares estimation of ϕ_1 and μ minimizes:

$$\sum_{t=2}^{N}((Y_t - \mu)) - (\phi_1(Y_{t-1} - \mu)))^2 \ .$$

If errors are Gaussian white noise, then conditional least squares outputs the same estimates as maximum likelihood method. The residual series estimate $\epsilon_2, ..., \epsilon_N$ and are defined by:

$$\hat{\epsilon_t} = (Y_t - \hat{\mu}) - \hat{\phi_1}(Y_{t-1} - \hat{\mu}), \ t \geq 2 \ .$$

For an adequate model, residuals should perform as a white noise process. Further to mention, by using repeated substitutions of an above expression, we can find $Y_t = \mu + \sum_{h=0}^{\infty} \phi_1^h \epsilon_{t-h}, \ \forall t$, which is an infinite moving average representation of a stationary Y_t.

Let us proceed to another type classical time series model [20].

3.1.2 Moving Average Process

A process $\{Y_t\}$ is a moving average process if Y_t can be expressed as a weighted average of the past values of an innovation process ϵ_t or extensions of white noise series [19]. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. Moving-average models are always weakly stationary. For the first order moving average process, MA(1), the current value of time series is equal to sum of the current innovation realization and the product of a coefficient with the past innovation process one time-step back [20].

The process $\{Y_t\}$ is a first order moving average process, if:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} \ . \tag{3.13}$$

Assuming that series are weakly stationary, we can obtain:

$$
\begin{gathered}
E(Y_t) = \mu \\
Var(Y_t) = (1 + \theta_1^2)\sigma_\epsilon^2 \\
\gamma(1) = \theta_1 \sigma_\epsilon^2, \quad \gamma(h) = 0, \quad \forall |h| > 1 \\
\rho(0) = 1, \quad \rho(1) = \frac{\theta_1}{1 + \theta_1^2}, \quad \rho(h) = 0, \quad \forall |h| > 1 \ .
\end{gathered}
$$

Autocorrelation function of MA(1) cuts off at the first lag. The general form of the process:

$$
Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \ , \tag{3.14}
$$

where μ and $\theta_1, ..., \theta_q$ are unknown parameters and the autocorrelations of an MA(q) process are equal to zero for lags larger than q [20]. If the correlogram plot for autocorrelation function shows a strong decline and becomes non-significant after lag q, then there is evidence that the series was generated by an MA(q) process [6].

For moving average models, maximum likelihood is frequently used for estimations, however two approaches are used for evaluation of the likelihood function.

The first approach assumes that initial innovation realizations (i.e. ϵ_t for $t \leq 0$) are zero. Therefore, innovations needed for estimation of likelihood function calculation are obtained recursively from the model, beginning with $\epsilon_1 = Y_1 - \mu$ and $\epsilon_2 = Y_2 - \mu + \theta_1 \epsilon_1$, calculated estimates are conditional-likelihood estimates.

The second way of estimation, which is known as exact-maximum likelihood takes initial innovations a_t, $t \leq 0$, as an additional parameters of the model and estimate them jointly with other parameters. Results of the latter method are more preferred in practice, particularly when the MA model is close to be non-invertible [19].

3.1.3 Autoregressive Moving Average Process

Finally, autoregressive moving average process combines both previous models. The process $\{Y_t\}$ is considered as an autoregressive moving average model, ARMA(1,1) if:

$$
Y_t = \mu + \phi_1 (Y_{t-1} - \mu) + \epsilon_t + \theta_1 \epsilon_{t-1} \ , \tag{3.15}
$$

with $\phi_1 \neq \theta_1$, if not, cancellations in the equation and process reduces to white noise series. As we can see it has a presence of both MA(1) and AR(1) component parts [19]. Given that series are weakly stationary, by taking expectations and assuming

$\mu = 0$ we have:

$$E(Y_t) = \mu = \frac{\phi_0}{1-\phi_1}$$
$$Var(Y_t) = \phi^2 Var(Y_{t-1}) + \sigma_\epsilon^2 + \theta^2\sigma_\epsilon^2 + 2\phi\theta Cov(Y_{t-1}, \epsilon_{t-1})$$
$$\gamma(0) = \frac{(1+\phi^2+2\phi\theta)\sigma_\epsilon^2}{1-\phi^2}$$
$$\gamma(1) = \phi\gamma(0) + \theta\sigma_\epsilon^2 = \frac{(\phi+\theta)(1+\phi\theta)}{1-\phi^2}\sigma_\epsilon^2$$
$$\gamma(h) = \phi\gamma(h-1),\ \forall h \geq 2$$
$$\rho(1) = \frac{(\phi+\theta)(1+\phi\theta)}{1+\theta^2+2\phi\theta}, \quad \rho(h) = \phi\rho(h-1),\ \forall h \geq 2\ .$$

The general form of the process has the following view:

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + ... + \phi_p(Y_{t-p} - \mu) + \epsilon_t + \theta_1\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q}\ , \tag{3.16}$$

where Y_t depends on lagged values of itself and on lagged values of an innovation process. When there are common factors between AR and MA polynomials, order (p, q) is reduced [20]. The autocorrelation and partial autocorrelation function plots are not descriptive for determining the order of a model, nonetheless if neither of them give visual information, then ARMA model with the least number of parameters have to be chosen. Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are useful for determining the order of the process, where models with the smallest values are favoured [20].

Differencing operator can be applied, where one extends the autoregressive polynomial to have 1 as a characteristic root, hence model will become autoregressive integrated moving average process (ARIMA). We can apply unit root tests to decide how we should model the time series (e.g. Dickey-Fuller test) [20].

Dickey-Fuller test checks for the null-hypothesis that a process has a unit root (i.e. $\phi = 1$), an alternative hypothesis is that a process is stationary. The test is based on AR(1) process test and works as rewriting the process as:

$$\triangle Y_t = (\phi - 1)Y_{t-1} + \epsilon_t = \pi Y_{t-1} + \epsilon_t\ , \tag{3.17}$$

where $H_0 : \pi = 0$ vs. $H_1 : \pi < 0$, therefore we regress $\triangle Y_t$ on Y_{t-1} and testing H_0. Correspondingly, for an augmented Dickey-Fuller test we add a drift with linear time trend because series is trending from an explicit function of time and add lagged values of $\triangle Y_t$:

$$\triangle Y_t = \beta_0 + \beta_1 t + \pi Y_{t-1} + \sum_{j=1}^{m} \gamma_j \triangle Y_{t-j} + \epsilon_t\ , \tag{3.18}$$

with $H_0 : \pi = 0$ vs. $H_1 : \pi < 0$ ($m = trunc(\sqrt[3]{N-1})$) [6].

Autoregressive moving average models can be estimated using maximum likelihood method or using two-stage least squares regression as well.

3.1.4 Model Selection

Different model selection criteria based on forecast errors are available. For contrasting between models we can use: in-sample and out-of-sample selection rules. Two model comparison approaches differ in objectives, the former gives insights on dynamic structures of the series, the latter is focused more on forecasting performance [20].

In-sample approach includes methods, which are based on forecast errors and forecast itself is made from the model estimated using all the available data. Standard methods for this approach include: mean squared error (MSE), Akaike's information, Bayesian information criteria. MSE is an average of squares of the errors, it advantages more complex models and we choose a model with the smallest value [6]. AIC and BIC in contrast, penalize complex models, where the latter method adds even more penalization for complexity [20]. Below are calculations for each measure are given:

$$MSE = \frac{\sum_{t=1}^{N} e_t^2}{N}$$
$$AIC = log(\hat{\sigma}^2) + \frac{2p}{N}$$
$$BIC = log(\hat{\sigma}^2) + \frac{plog(N)}{N} \ .$$

Out-of-sample approach is based on splitting the series into estimation and forecasting sub-samples: For the first sub-sample, series $y_1, y_2, ...y_t$ will be used to estimate the model and this model is used to make h-step-ahead predictions. For the second sub-sample, $y_{S+1}, y_{S+2}, ..., y_{S+h}$ will be used to validate the model and the h-step-ahead forecast error is computed as:

$$e_{S+h}^h = y_{S+h} - \hat{y}_{S+h}^{(S)} \ .$$

Splitting the series repeatedly at all time points between S and $N - h$ yields a sequence of h-step (artificial) out-of-sample forecasts and these forecast errors are used to calculate prediction accuracy measures such as: root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). Below are calculations for each measure:

$$RMSE_h = \sqrt{\frac{\sum_{t=S+h}^{N} (e_t^h)^2}{N - S - h + 1}}$$
$$MAE_h = \frac{\sum_{t=S+h}^{N} |e_t^h|}{N - S - h + 1}$$
$$MAPE_h = \frac{1}{N - S - h + 1} \sum_{t=S+h}^{N} \frac{|e_t^h|}{y_t} \ .$$

The out-of-sample approach requires to re-estimate the model numerous times and the value for S and $N - S - h + 1$ have to be large enough [6].

For every model, obtained residuals should examined on whether the series are coming from white noise stochastic process. Commonly autocorrelation function's plot is checked, however more standard statistical test, such as Q-statistic (Ljung-Box test) is recommended to apply [6]. Our main interest is in testing simultaneously that several autocorrelations e_t are 0.

$$H_0 : \rho(1) = \rho(2) = ... = \rho(m) = 0$$
$$H_1 : \rho(i) \neq 0 \text{ for } i \in \{1, ..., m\}.$$

Test statistic Q:

$$Q(m) = N(N+2) \sum_{h=1}^{m} \frac{\hat{\rho}(h)^2}{N-h} \, ,$$

where N is a sample size, $\hat{\rho}$ is a sample autocorrelation function as lag h and m is a number of lags being tested. The null-hypothesis should be rejected if $Q(m) > \chi^2_{\alpha,m}$ at the significance level of α [6].

3.1.5 Forecasting

The classical problem for forecasting can be breakdown in the subsequent way: we have an access to historical time series data with values up to the ongoing present time [20]. We are needed to predict the value of the following time step value as nearly as achievable. Using past, present values and specifying the model, we need to forecast $\hat{Y}_{N+k}$, with N as an origin of prediction and k as a horizon of prediction. The value for $\hat{Y}$ is selected such that the mean-squared error (MSE) is minimized:

$$E\big[(Y_{N+k} - \hat{Y}_{N+k}|\mathcal{F}_N)\big] \leq \min_f E\big[(Y_{N+k} - f)^2|\mathcal{F}_N)\big] \, , \tag{3.19}$$

where f is a function of the current information set $\mathcal{F}_N$ and the most optimal predictor equals to conditional expectation:

$$E\big[(Y - f(X))^2\big] \geq E\big[(Y - E(Y|X))^2\big], \; \forall f(X) \, ,$$

what makes $\hat{Y}_{N+k} = E[Y_{N+k}|\mathcal{F}_N]$ as a best predictor in terms of minimizing the mean squared error [20].

For a general AR(p) process, 1-step-ahead forecast estimates have the following view:

$$\hat{Y}_{N+1} = \hat{\mu} + \hat{\phi}_1(Y_N - \hat{\mu}) + \hat{\phi}_2(Y_{N-1} - \hat{\mu}) + ... + \hat{\phi}_p(Y_{N-p+1} - \hat{\mu}) \, , \tag{3.20}$$

with k-step forecast error calculated as:

$$e_{N+k} = Y_{N+k} - \hat{Y}_{N+k} \tag{3.21}$$
$$e_{N+k} \approx \phi^{k-1}\epsilon_{N+1} + \phi^{k-2}\epsilon_{N+2} + ... + \phi\epsilon_{N+k-1} + \epsilon_{N+k} \, , \tag{3.22}$$

with its variance equal to:

$$Var(e_{N+k}) = \Big(\phi^{2(k-1)} + \phi^{2(k-2)} + ... + \phi^2 + 1\Big)\sigma_\epsilon^2 = \frac{1-\phi^{2k}}{1-\phi^2}\sigma_\epsilon^2 \ , \tag{3.23}$$

which converges to $\gamma(0) = \frac{\sigma_\epsilon^2}{1-\phi^2}$ as horizon $k \to \infty$. Confidence intervals of the process are given by $\hat{Y}_{N+k} \pm z_{\frac{\alpha}{2}}\sqrt{Var(e_{N+k})}$, where using of $z_{\frac{\alpha}{2}}$ assumes that $\{\epsilon_t\}$ is Gaussian white noise. Uncertainty in predictions increases with an increase of a forecast horizon [20].

For a general MA(q) process, 1-step-ahead prediction is calculated as:

$$\hat{Y}_{N+1} = \hat{\mu} + \hat{\theta}_1\hat{\epsilon}_N + ... + \hat{\theta}_q\hat{\epsilon}_{N+1-q} \ , \tag{3.24}$$

with multi-step-ahead prediction errors:

$$e_{N+k} \approx \epsilon_{N+k} + \theta_1\epsilon_{N+k-1} + ... + \theta_{k-1}\epsilon_{N+1}, \ \forall k \leq q \tag{3.25}$$

$$\approx \epsilon_{N+k} + \theta_1\epsilon_{N+k-1} + ... + \theta_q\epsilon_{N+k-q}, \ \forall k > q \ , \tag{3.26}$$

with its variance equal to $\big(1 + \theta_1^2 + ... + \theta_q^2\big)\sigma_\epsilon^2$, which goes to the variance of the series.

For a general autoregressive moving average processes p, q:

$$Y_{N+k} = \hat{\mu} + \hat{\phi}^k(\hat{Y}_N - \hat{\mu}) + \hat{\phi}^{k-1}\hat{\theta}\hat{\epsilon}_N, \ k \geq 2 \ , \tag{3.27}$$

with residuals estimated as:

$$e_{N+k} \approx \sum_{j=1}^{k-1} \phi^{k-j-1}(\phi + \theta)\epsilon_{N+j} + \epsilon_{N+k} \ . \tag{3.28}$$

Variance of (e_{N+k}) of an ARMA process converges to a finite value as horizon of prediction k increases [20]. As we covered main time series models, now we can discuss cointegration concept.

3.1.6 Cointegration

There are different quantitative techniques for selecting pairs for trading. Main methods are: stochastic approach, distance approach and cointegration.

Distance Approach is based on selecting pairs with the lowest Euclidean distance. Stochastic approach uses the Ornstein-Uhlenbeck process to model the mean reverting behaviour of a spread, where the speed of reversion can be estimated [1].

Let us start with an explanation of cointergration concept suggested by Engle and Granger, which we briefly mentioned in the theoretical framework and we will still later discuss its alternative approach [7].

The main idea of the approach is based on the concept of a linear relationship between two variables represented as time series, where we can obtain stationary time series as a combination of both [21].

We can explain more formally, by assuming two time series X_t and Y_t integrated of order one and regressing one variable on another:

$$Y_t = c + \beta X_t + \epsilon_t$$

We can claim that X_t and Y_t are cointegrated if there exist a constant c and coefficient β, which makes ϵ_t in the above model achieve stationary properties. The long-run equilibrium relationship between two variables is $Y_t = c + \beta X_t$, is calculated using ordinary least squares method and ϵ_t is the deviation from this equilibrium, where both variables adjust themselves to restore and the term is called as error correction [21].

Error correction and cointegration are actually analogous representation often called Granger representation theorem, which we can illustrate as following.

Let ϵ_{x_t} be the white noise process of time series $\{x_t\}$ and let ϵ_{y_t} be the white noise process corresponding to the time series $\{y_t\}$. Error correction representation can be written as:

$$\begin{aligned} y_t - y_{t-1} &= \alpha_y(y_{t-1} - \gamma x_{t-1}) + \epsilon_{y_t} \\ x_t - x_{t-1} &= \alpha_x(y_{t-1} - \gamma x_{t-1}) + \epsilon_{x_t} \end{aligned} \tag{3.29}$$

The left part of the equation is the increment to the time series one a single step, the right part is the sum of the error correction part and the white noise part. Expression can be breakdown as following: $\alpha_y(y_{t-1} - \gamma x_{t-1})$ is the error correction part, the term $y_{t-1} - \gamma x_{t-1}$ is the deviation from the long-run equilibrium and γ is the cointegration coefficient with a_y is the error correction rate or a speed to achieve an equilibrium [7].

Therefore, as original series go through the time, deviations from the long-run equilibrium caused by white noise series corrected in the future periods [21]. We can advance to full practical backtesting of our pairs of stocks using all the information we have already covered.

3.2 Backtesting

3.2.1 Data

Let us combine our theoretical framework for statistical arbitrage pairs trading with the above discussed methods for time series analysis. We will use companies from the financial sector that are included in the 'Euronext 100' index.

Our portfolio considers shares of companies such as: "Societe Generale S.A.", "BNP Paribas S.A.", "Banco Santander, S.A." (SAN), "ING Groep N.V." (INGA), "Credit Agricole S.A." (ACA), "Aegon N.V." (AGN), "Ageas S.A" (AGS), "AXA S.A" (AXA), "Gecina" (GFC), "Icade" (ICAD), "KBC Groep N.V." (KBC), "Natixis S.A." (KN) and "SCOR SE" (SCR).

Our time period is expanding from July 2017 until July 2019 (approximately half of the length in primary market trends) and data was obtained from the financial and economic data server "Quandl".

We will work with adjusted close prices as their calculations consider aspects such as dividends, stock splits and rights offerings to correctly determine the price per share.

3.2.2 Analysis

We start our analysis with a focus on finding cointegrated pairs of stocks and below is our heatmap with pairs (Figure 3.1). This is based on the Engle-Granger two-step cointegration test [7], where the null hypothesis states no cointegration.

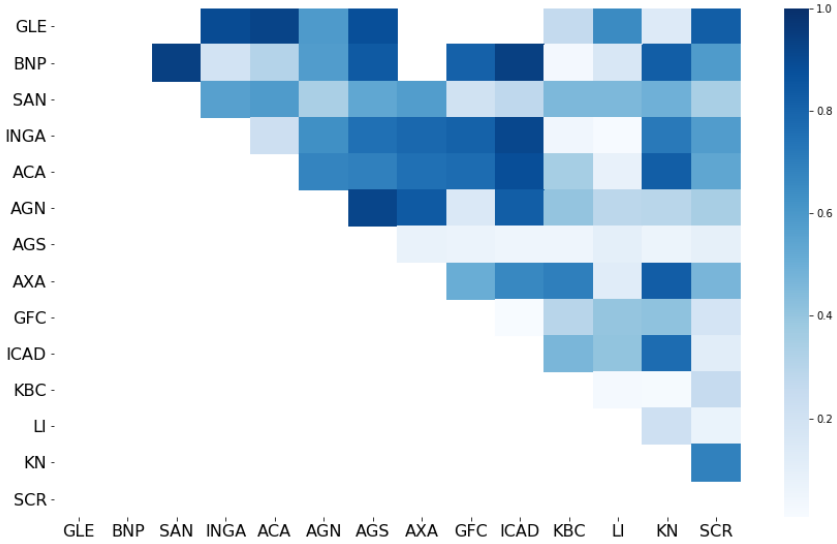


Figure 3.1: Cointegrated Pairs Heatmap.

Intensity of colors on heatmap indicates an increase of p-values (white colors indicate p-values over 0.95), where can notice that suggested cointegrated pairs are: INGA/LI, GFC/ICAD and KBC/KN. For the task, we chose a pair with the highest combined cumulative return, which was ”Gecina”/ ”Icade” pair, both French public real estate investment companies. Engle-Granger two-step cointegration test showed their p-value was equal to 0.007, thus the null hypothesis of no cointegration was rejected.

On the next page, on Figure 3.2, the plot for daily adjusted log closing prices is given, where we can notice a very similar trend pattern for each of the companies.

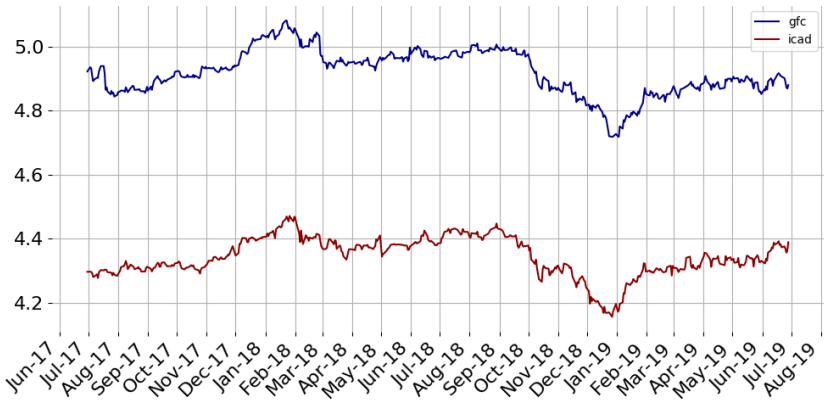


Figure 3.2: Daily log Closing Prices of GFC and ICAD.

The combined cumulative gross return was equal to 4.3%, which is not large, however it is positive compared to the overall negative returns of the financial sector (assuming we are holding equal weights of stocks for both companies). The plot for the gross cumulative returns is given below, where we can see closely identical patterns.

Figure 3.3: Log-returns of GFC and ICAD.

Let us proceed further using logarithmic prices of our stocks and fit a simple linear regression model $p_{1t} = \beta_0 + \beta_1 p_{2t} + s_t$, where s_t stands for residuals series. The stock of "Gecina" company will be used as an explanatory variable and our least squares method estimated model has the following fit:

$$icad_t = 0.4507 + 0.7916\ gfc_t + \hat{s}_t, \quad \hat{\sigma}_s = 0.0237\ ,$$

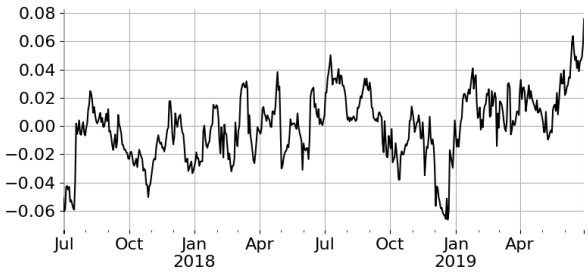
p-value of $0.00 < 5\%$ indicating a significant relationship between the logarithmic prices of both variables. On the next page (Figure 3.4) is a quick summary output for our model, where can see that our adjusted coefficient of determination equals to 0.842, which means 84% of the variance in the response is predictable using an explanatory variable.

Dep. Variable:	icad	R-squared:	0.842
Model:	OLS	Adj. R-squared:	0.841
Method:	Least Squares	F-statistic:	2682.
Date:	Thu, 25 Jul 2019	Prob (F-statistic):	3.67e-204
Time:	12:24:56	Log-Likelihood:	1179.8
No. Observations:	507	AIC:	-2356.
Df Residuals:	505	BIC:	-2347.
Df Model:	1		
Covariance Type:	nonrobust		

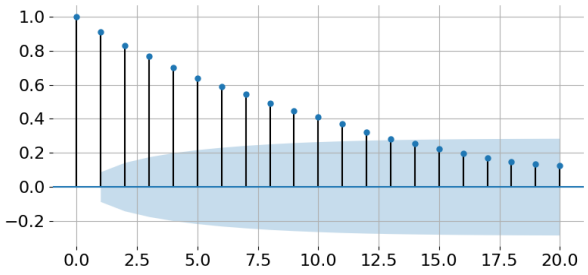
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4507	0.075	5.993	0.000	0.303	0.598
gfc	0.7916	0.015	51.788	0.000	0.762	0.822

Figure 3.4: Ordinary Least Regression Results.

In accordance with our trading strategy, we have $s_t = icad_t - 0.7916\ gfc_t$, with $\hat{\mu}_s = 0.4507$ and $\triangle = 0.0237$, we enter into the trade by selling the spread if $s_t > \mu_s + \triangle$ and oppositely we buy the spread if $s_t < \mu_s - \triangle$. Buying the spread means we sell one stock of ICAD and buy 0.79 fraction of a share of GFC. We cannot buy fractions, thus the number of shares should multiplied by 100 and log-return of a single pair is equal to $2\triangle = 0.0474$.



(a) Time Series of Spread.



(b) Autocorrelation Plot.

Figure 3.5

In order to check for stationarity of spread values, we can refer to time series of residuals plot and autocorrelation function plot. According to the first plot, it seems that residuals have a pattern of stationarity (Figure 3.5, a) and we can see exponential decay on the second plot for autocorrelations (Figure 3.5, b), suggesting stationarity. As stationarity was achieved, we can say that the two stocks are cointegrated, furthermore augmented Dickey-Fuller confirmed that data has no unit root and is a stationary process.

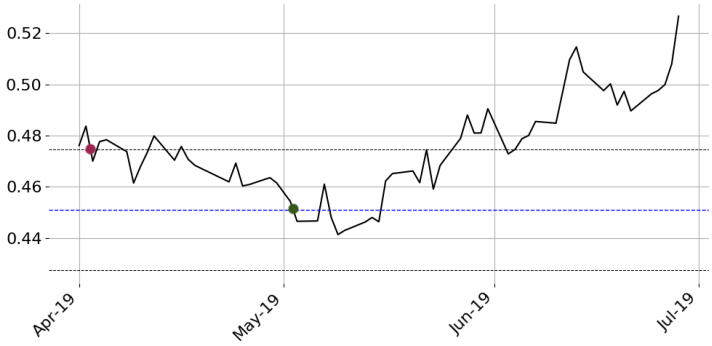


Figure 3.6: Spread for Daily Closing Prices of gfc and $icad$.

On Figure 3.6, the estimated spread series for the period from April to July (2019) is given. We had a signal to sell the spread at the beginning of April and buy it back by closing the trade at the beginning of May with a profit of 2.3%. As compared to 4.3% of 2 years by using direct investment, single trade only for the second quarter is marginal.

Let us proceed to applying our forecasting methods and start with the observation time series of obtained residuals. We already noticed the mean-reverting phenomenon by observing autocorrelation correlogram and applying the ADF test, thus no differencing is required to obtain stationarity.

Now our goal is to find a valid model, which we can use for predicting future values of our time series of spread. We can start by observing the partial autocorrelation plot below (Figure 3.7). This suggests a fit for the AR(1) model, as we see that the first lag is outside significance bound, which means that higher-order autocorrelations are adequately explained by the lag 1 autocorrelation:

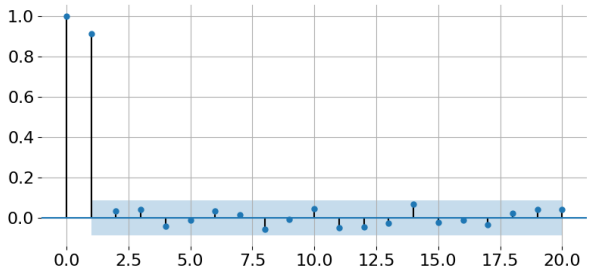


Figure 3.7: Partial Autocorrelation Plot.

We fitted the following first order autoregressive model:

$$\hat{s_t} = -0.0038 + 0.9139 s_{t-1}, \ \forall t$$

The choice for first order autoregressive model was also confirmed by the automatic step-wise model selection function based on Akaike's information criterion equal to -2908.24 and Bayesian information criterion equal to -2895.79, where for both criteria AR(1) model was favored. Below is the summary for our fitted model:

Dep. Variable:	Last	No. Observations:	446
Model:	ARMA(1, 0)	Log Likelihood	1457.047
Method:	css-mle	S.D. of innovations	0.009
Date:	Mon, 29 Jul 2019	AIC	-2908.093
Time:	18:17:27	BIC	-2895.792
Sample:	06-30-2017	HQIC	-2903.243
	- 04-01-2019		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0038	0.005	-0.760	0.448	-0.013	0.006
ar.L1.Last	0.9139	0.020	46.770	0.000	0.876	0.952

Figure 3.8: AR(1) Model Results.

To validate the models we have to test if the residuals are white noise by assessing the residuals' correlogram and using the Ljung-Box Q-statistic. Both testing procedures confirmed that residuals can be considered as a white noise process.

In order to evaluate the model forecasting performance we used out-of-sample criteria, splitting the data-set in an estimation and a validation subsets with proportion of 90-10, where we used the last two months as a test sample. With the multi-step ahead forecast errors, we computed the mean squared error, root mean squared error, mean absolute error for the model and their calculated results are given below:

MSE	RMSE	MAE
0.00097	0.031586	0.02445

By observing the plot for our forecast on Figure 3.9, we can see that the prediction interval widens over time because the further away we forecast, the more uncertainty exists. This can be seen in the inaccuracy observed in this model during late June.

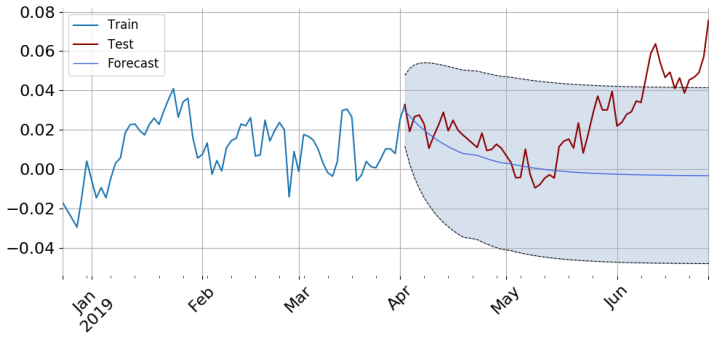


Figure 3.9: AR(1) Model Forecast.

Let us proceed to the next chapter, where we will discuss the different types of statistical arbitrage.

4. Types of Statistical Arbitrage

Statistical arbitrage trading can be summarized into the following basic steps:

- identification of potential stocks (e.g. to find strongly covariate pairs for pairs trading)
- systematising of trading rules
- selecting suitable portfolio of stocks based on risk/return.

As mentioned above, pairs trading constitutes the foundation of statistical arbitrage and the basis from which other methods evolved from, which are listed below. Methods such as: risk arbitrage, cross market arbitrage, exchange traded fund arbitrage and factor model arbitrage are among the methods used in the industry. This thesis will now discuss these methods in greater depth.

4.1 Risk Arbitrage

Risk arbitrage is an investment strategy related to corporate events such as mergers and acquisitions which change the capital structure of the company [21].

Let us first to define the sides involved in such corporate operations. The first side is: an acquirer or a bidder, which is a main company that takes over a smaller one (the ”target” company). The bidder company can cover the deal by acquiring the target company with a cash only position, or exchange its own stocks at some fixed ratio or use a combination of these methods. In all cases, the main task is the evaluation and acquisition price set by a bidder company, which is often different to the market price traded at exchanges.

The algorithm of the risk arbitrage strategy is focused on the spread as similar to pairs trading:

spread = exchange value of a target share - market price of target share

As the price of the target stock should emerge with the price of a bidder company, this means that the spread should approach zero by the end of the acquisition. However, the spread often remains positive due to the risk that counter-parties may not reach an agreement and merger can be cancelled. The spread cannot be negative as well as arbitrageurs could sell the target share at market price and buy it back at exchange value [21].

The main strategy therefore, is to sell the stocks of a bidder company and buy the shares of a target company and profit from the spread when the deal is finally completed. If the deal is cancelled however, where the shares of the target company will decline and the shares of target company will increase, an investor may incur large

losses. The strategy is based more on fundamental valuation of the company than a quantitative approach, therefore insider trading might take place as well [3].

4.2 Cross Market Arbitrage

Cross market arbitrage is based on the assumption that assets should have identical fundamental value on different securities or mercantile exchanges across the world. Thus, the trading strategy is based on the existing spread due to demand and supply between the same asset but listed and traded on different markets.

For example, if company A is traded at €10 per share in Frankfurt, it might be traded at €9.5 for the same company in London, due to difference in currency conversion, difference in supply and demand or different expectations of market participants. Therefore, it is possible to sell the stock in Frankfurt and buy it back in London to make an immediate profit of €0.5. As the spread is usually non-significant, the transaction involves large volume of trades in order to cover the currency conversion, broker commission and other fees involved in a single trade. Therefore, large risk exists that during operation the spread might disappear during the transaction, which often is the case. As a result, the investor might incur losses and expose himself/herself to being in an open trade position. High frequency trading might be a solution for such cases.

4.3 Exchange Traded Fund Arbitrage

Exchange traded fund (or ETF) is an investment fund which tracks certain securities such as indexes, baskets of stocks from a sector, bonds or synthetic financial products. ETFs may also track prices for commodities in exchange for its own securities (stocks) issued by the fund.

Arbitrage opportunities might appear due to the mispricing of an ETF stock itself or securities it tracks or between another exchange traded funds tracking the same assets. Commonly there is a fixed ratio between the ETF's stock traded and the security or commodity it tracks, therefore arbitrage opportunities might exist due to a change in this equilibrium between these two components. For example, the stock for ETF might be traded at a value much larger than the combined value of stocks it includes. In order to benefit from such an arbitrage opportunity an investor could short the stocks of the ETF [14].

The strategy is similar to cross market arbitrage involving large volumes of capital to finance and technological speed to react to created arbitrage opportunities, thus institutional investors are often involved in this type of arbitrage.

4.4 Factor Model Arbitrage

Finally, factor model arbitrage is a type of statistical arbitrage based on the selection of individual stock price series analyzed as a collection of stocks with an attempt to explain risk and return characteristics. A similar concept was alluded to when market neutral strategies were discussed; particularly in the discussion around neutralising market impacts on portfolios (coefficient β).

However, stock movements in a portfolio can be explained by more than one factor model, starting from market indices, industries or more underlying fundamental factor models, which supposed to be determined.

The basic idea is that baskets of stocks can be broken down into an element that is determined by one or more underlying factors in the market and an element characteristic to the stock, or called specific returns [13]:

$$\text{stock return} = \text{market factor return} + \text{specific return}$$

For this model, the factors are estimated from the historical stock return data and a stock's return can be connected with several of these factors through exposure/ sensitivity values. For several explanatory factors, the return of a stock will be a cumulative of the return contributions of the factors scaled in proportion with sensitivity/ factor exposure.

The next section will consider the concepts underpinning factor model arbitrage in detail.

5. Concepts of Factor Model Arbitrage

Depending on the type of the factors used, factor models can be split into three main categories: macroeconomic factor models, fundamental factor models and statistical factor models.

The macroeconomic factor models are set up using historical stock returns and observable macroeconomic variables. Few examples of proprietary type macroeconomic factor models are Burmeister, Ibbotson, Roll and Ross models. The factors in these listed models often consisted indicators such as: short-term bond yield changes, long-term bond yield changes, dollar value versus other currencies, investor confidence, and changes in long-run economic growth [21].

The fundamental factor model exploits company and industry specifics and market data as raw descriptors to explain the returns. Good commercially available models are the BARRA and Wilshire Atlas models. Indicators used for these models are typically industry factors to which companies are relevant and operate. Fundamental factors based on financial indicators such as: price/earnings ratio, the price/book ratio, particularly, indicators attributing to the capital structure of the company (e.g. debt/equity ratios) are also included in this type of fundamental factor.

The thesis will focus on statistical factor models, where factors are called as *eigen* portfolios, which are portfolios with an attribute that their returns are uncorrelated with each other, returns on any portfolio can be expressed as a linear combination of the returns on these *eigen* portfolios [21]. The main difficulty with statistical factor analysis is an interpretation of factors from a qualitative side, which is present in both macroeconomic and fundamental factor models.

The theoretical concept on which all factor models are based on can be explained by Arbitrage Pricing Theory [21]. Therefore, let us proceed to the detailed explanation of this theory.

5.1 Arbitrage Pricing Theory

Arbitrage Pricing Theory (APT) is a theory originally proposed by Stephen A. Ross [15].

The theory explains that asset can be completely defined by its factor exposure/ sensitivity profile, where every factor contributes to the overall asset return, which is an aggregate of these contributions [21]. The concept similar to CAPM and often considered as extension.

Let us now describe some terminology behind APT. We begin with factor exposures as: $(\beta_1, \beta_2, \beta_3, ...\beta_k)$ and $(r_1, r_2, r_3, ..., r_k)$ are the return contributions of each factor, thus

the overall asset return is:

$$r = \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + ... + \beta_k r_k + r_e \tag{5.1}$$

where r_e is the specific (or idiosyncratic) return or specific return on the stock, which cannot be explained by the factors and it is expected to be zero.

One of the important assumptions of arbitrage pricing theory is that this specific return cannot correlate with factor returns and specific returns of others stocks [21]. We know that risk in a stock can be measured as a variance of return, where smaller variance is associated with smaller risk and larger variance is associated with greater risk and greater returns as well (e.g. variance of a risk-free asset is 0).

This mentioned approach of measuring the risk by using variance introduced by Harry Markowitz is widely accepted method of risk evaluation in nowadays [12]. We can explain how risk/ variance of returns are calculated in APT based on this theory. Thus, let us give an example by using two factor model below:

$$r = \beta_1 r_1 + \beta_2 r_2 + r_e \tag{5.2}$$

as we know that risk has an analogous concept as variance, let us expand the squared return:

$$\begin{gathered} r^2 = (\beta_1 r_1 + \beta_2 r_2 + r_e)^2 \\ r^2 = \beta_1^2 r_1^2 + \beta_2^2 r_2^2 + 2\beta_1\beta_2 r_1 r_2 + 2\beta_1 r_1 r_e + 2\beta_2 r_2 r_e + r_e^2 \end{gathered} \tag{5.3}$$

further, the following equation can be written as:

$$Var(r) = \beta_1^2 Var(r_1) + \beta_2^2 Var(r_2) + 2\beta_1\beta_2 Cov(r_1, r_2) + Var(r_e)\ , \tag{5.4}$$

where r_e is uncorrelated with both r_1 and r_2, therefore we can omit their products and further rewrite the Equation 5.4 in a matrix form below:

$$Var(r) = [\beta_1\beta_2] \begin{bmatrix} Var(r_1) & Cov(r_1, r_2) \\ Cov(r_1, r_2) & Var(r_2) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} Var(r_e) \tag{5.5}$$

Expression in square brackets includes variance of the factor returns on its diagonal and their covariance on its off-diagonal elements and is commonly referred as covariance matrix of factor returns and it is important in the definition of risk. Vector $[\beta_1, \beta_2]$ is the factor exposure/ sensitivity vector and $Var(r_e)$ is the stock specific variance.

Knowledge of the covariance matrix with the factor exposure/ sensitivities and specific variances is sufficient to define arbitrage pricing theory completely, thus we summarize Equation 5.5 with an expression below [21]:

$$\sigma_{ret}^2 = \sigma_f^2 + \sigma_{specific}^2\ . \tag{5.6}$$

Let us proceed to next section where we will explain cointegration concept based on common trends.

5.2 Cointegration and APT

We have discussed the concept of cointegration by Engle and Granger, however let us refer to an alternative approach in the definition of cointegration proposed by Stock and Watson known as common trends model and which can be linked with APT [18].

The main concept is based on expression, where time series is represented as a sum of stationary and non-stationary component. We can consider that if two time series are cointegrated then these series negate non-stationary components, leaving only stationary [21]. Consider two time series:

$$\begin{aligned} y_t &= n_{y_t} + \epsilon_{y_t} \\ z_t &= n_{z_t} + \epsilon_{z_t}, \end{aligned} \tag{5.7}$$

where ϵ_{y_t} and ϵ_{z_t} are stationary components of series (specific components), and n_{y_t} and n_{z_t} are non-stationary components (common trends).

Let us rewrite Equation 5.7 by using cointegrated linear combination of stationary time series $y_t - \gamma z_t$, where after expansion and rearrangements we can get the following expression:

$$y_t - \gamma z_t = (n_{y_t} - \gamma n_{z_t}) + (\epsilon_{y_t} - \gamma \epsilon_{z_t}) \tag{5.8}$$

By referring to above equation, if the left part of the equation have to be a stationary process, thus non-stationary part have to be zero $n_{y_t} = \gamma n_{z_t}$. In other words the trend component of the first series have to be a scalar multiple of trend component of the second series, which means their trend have to be identical up to a scalar in order to achieve cointegration, where γ is their cointegration coefficient [21].

We have mentioned in the section for autoregressive processes about random walk process, therefore let us apply on our common trends part in the form below:

$$\begin{aligned} n_{y_{t+1}} - n_{y_t} &= r_{y_{t+1}} \\ n_{z_{t+1}} - n_{z_t} &= r_{z_{t+1}}, \end{aligned}$$

where according to 5.8, we have $n_{y_t} = \gamma n_{z_t}$.

Further, by requiring $n_{y_{t+1}} = \gamma n_{z_{t+1}}$, we can obtain $r_{y_t} = \gamma r_{z_t}$ and finally, cointegration coefficient can be obtained by regressing one innovation process on another, which can give us:

$$\gamma = \frac{Cov(r_y, r_z)}{r_z} \ .$$

In order to connect both theories of arbitrage pricing and cointegration concept we can use the following example by using random walk process of logarithm of stock prices. Analogously to Equations 5.7 and 5.8, we can write them as a sum of random walk and stationary series:

$$log(price_t) = n_t + \epsilon_t \ , \tag{5.9}$$

where ϵ_t is the random walk, and n_t is the stationary component. In order to obtain logarithmic returns differencing operation is required, which yields below expression:

$$log(price_t) - log(price_{t-1}) = n_t - n_{t-1} + (\epsilon_t - \epsilon_{t-1})$$
$$r_t = r_t^c + r_t^s \ ,$$

where r_t^c is the return due to the stationary component, r_t^s is the return due to the non-stationary trend component. As we can observe that return from trend component is similar to the innovation from the trend, in other words their common trends are identical up to a scalar, therefore they have a direct connection to common factor returns of APT [21]. Now, we can proceed to explanation and application of factor models.

5.3 Factor Model

We have discussed in the beginning of the chapter about the concept a factor model, their main types and their importance in analysis of portfolio returns, therefore let us mathematically formulated the theory. Let us assume that there are k assets and T time periods and let r_{it} be the return of asset i in the time period t.

A general form for the factor model is:

$$r_{it} = \alpha_i + \beta_{i1} f_{1t} + \cdots + \beta_{im} f_{mt} + \epsilon_{it}, \quad t = 1, ..., T; \quad i = 1, ..., k, \tag{5.10}$$

where α_i is a constant representing the intercept, $f_{jt}|j = 1, ..., m$ are m common factors, β_{ij} is the factor loading for asset i on the jth factor and the specific factor of asset i is represented by ϵ_{it}.

For returns of assets, the factor $f_t = (f_{1t}, ..., f_{mt})$ is assumed to be an m-dimensional stationary process such that

$$E(f_t) = \mu_t$$
$$Cov(f_t) = \Sigma_f, \text{ an } m \times m \text{ matrix}$$

and the asset specific factor ϵ_{it} is a white noise series and uncorrelated with the common factors f_{jt} and other specific factors. Specifically, we assume that

$$E(\epsilon_{it}) = 0 \quad \forall\, i, t,$$
$$Cov(f_{it}, \epsilon_{is}) = 0 \quad \forall\, j, i, t, \text{ and } s,$$
$$Cov(f_{it}, \epsilon_{is}) = \begin{cases} \sigma_i^2, & \text{if } i = j \text{ and } t = s, \\ 0, & \text{otherwise} \end{cases}$$

Therefore, there is no correlation between common and specific factors and furthermore, the specific factors are also uncorrelated among each other. However, common factors in contrast do not have to be uncorrelated with each other in some factor models [19].

Let us proceed by rewriting Equation 5.10 in a matrix form as following:

$$r_{it} = \alpha_i + \beta_i f_t + \epsilon_{it},$$

where $\beta_i = (\beta_{i1}, ..., \beta_{im})$ is a row vector of loadings and the joint model for the k assets at time t is:

$$r_t = \alpha + \beta f_t + \epsilon_t, \quad t = 1, ..., T \tag{5.11}$$

where $r = (r_{1t}, ..., r_{kt})'$, $\alpha = (\alpha_1, ..., \alpha_k)'$, $\beta = [\beta_{ij}]$ is a $k \times m$ factor loading matrix and $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{kt})'$ is the vector of errors with Cov(ϵ_t) = D = diag$\sigma_1^2, ..., \sigma_k^2$, a $k \times k$ diagonal matrix. The covariance matrix of the return r_t is then

$$Cov(r_t) = \beta \Sigma_f \beta' + D.$$

If the factors f_{jt} are observed, Equation in 5.11 above is considered as a cross-sectional regression form [19].

Furthermore, by considering the factor model in 5.10 in the time series form, we can obtain:

$$R_i = \alpha_i 1_T + F\beta_i = E_i,$$

for the i-th asset $i = 1, ..., k$, where $R_i = (r_{i1}, ..., r_{iT})'$, 1_T is a T-dimensional vector of ones, F is $T \times m$ matrix whose t-th row is f_t', and $E_i = (\epsilon_{i1}, ..., \epsilon_{iT})')$. The covariance matrix of E_i is $Cov(E_i) = \sigma_i^2 I$, a $T \times T$ diagonal matrix.

Eventually, expression 5.12 can be written as:

$$r_t = \xi g_t = \epsilon_t,$$

where $g_t = (1, f_t')'$ and $\xi = [\alpha, \beta]$, which is a $k \times (m+1)$ matrix. Taking the transpose of the prior equation and combining all data together, we have:

$$R = G\xi' + E, \tag{5.12}$$

where R is a $T \times k$ matrix of returns whose t-th row is r_t' or ,equivalently, whose i-th columns is R_i, G is $T \times (m+1)$ matrix whose t-th row is g_t', and E is a $T \times k$ matrix of specific factors whose t-th row is ϵ_t'. If the common factors f_t are observed, then above equation is a special form of the multivariate linear regression model, where covariance matrix of ϵ_t do not have to be diagonal. [19].

Above, we have explained a generalization of factor model for asset returns, our factor model trading strategy further will focus on defactorization of these returns and we have to apply our statistical methods in order to obtain these factor returns. Let us therefore proceed to the next chapter, where we can discuss statistical factor models, which focuses on covariance structure of the series.

5.3.1 Statistical Factor Analysis

Factor analysis is a technique to measure and develop scales, was originally a method to establish a connection between students performances in various subjects and their general intelligence. It has many similarities with another data reduction statistical technique - principal component analysis, however both methods are certainly distinct. [16].

The objective of statistical factors analysis is to use computed covariance (correlation) matrix [16]:

- To identify common factors explaining the most variation in covariance matrix
- To identify via factor rotations plausible factor solutions
- To estimate pattern and structure loadings, communalities and variances of indicators
- To conduct interpretation of factors
- To estimate their factor scores (if necessary).

Statistical factor analysis is divided into exploratory type, where factors are unobserved and confirmatory type, where factors are observed (e.g. macroeconomic, fundamental).

We will aim to find unobserved factors for our portfolio of stocks and apply defactored model statistical arbitrage, however let us conduct mathematical discussion of the method by further discussing theoretical framework of the strategy [13].

Let us consider the return $r_t = (r_{1t}, ..., r_{kt})'$ of k assets at time period t and assume that the return series r_t is weakly stationary with mean μ and covariance matrix r_t. The statistical factor model supposes that r_t is linearly dependent on a few unobservable random variables $f_t = (f_{1t}, ..., f_{mt})$ and k additional noises $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_k t)$. Here $m < k$, f_{it} are the common factors, and ϵ_{it} are the errors.

Mathematically, the statistical factor model cal also take the form of Equation 5.10 with an exception that the intercept α is replaced by the mean return μ. Therefore, a statistical factor model can take the following form:

$$r_t = \mu + \beta f_t + \epsilon_t, \tag{5.13}$$

where $\beta = [\beta_{ij}]_{k \times m}$ is the matrix form of factor loadings, β_{ij} is the loading of the ith variable on the jth factor, and ϵ_{it} is the specific error of r_{it}.

The main characteristic of the statistical factor model is that the m factors f_{it} and the factor loadings β_{ij} are not unobserved, which prevents Equation 5.13 from being a multivariate linear regression model [19].

The model in Equation 5.13 is an orthogonal factor model if the following assumptions are satisfied:

$$E(f_t) = 0 \text{ and } Cov(f_t) = I_m, \text{ is the } m \times m \text{ identity matrix.}$$
$$E(\epsilon_t) = 0 \text{ and } Cov(\epsilon_t) = D = \text{diag}\{\sigma_1^2, ..., \sigma_k^2\} \ (i.e., \text{ D is a } k \times k \text{ diagonal matrix}).$$
$$f_t \text{ and } \epsilon_t \text{ are independent, thus } Cov(f_t, \epsilon_t) = E(f_t, \epsilon_t') = 0_{m \times k}.$$

We can deduce using previous assumptions, that:

$$\begin{aligned} \Sigma_r = Cov(r_t) &= E[(r_t - \mu)(r_t - \mu)'] \\ &= E[(\beta f_t + \epsilon_t)(\beta f_t + \epsilon_t)'] = \beta\beta' + D \end{aligned}$$

and

$$Cov(r_t, f_t) = E[(r_t - \mu)f_t'] = \beta E(f_t f_t') + E(\epsilon_t f_t') = \beta. \tag{5.14}$$

Using above equations, we can deduce for the Equation 5.15,

$$\begin{aligned} Var(r_{it}) &= \beta_{i1}^2 + ... + \beta_{im}^2 + \sigma_i^2, \\ Cov(r_{it}, r_{jt}) &= \beta_{i1}\beta_{j1} + ... + \beta_{im}\beta_{jm}, \\ Cov(r_{it}, f_{jt}) &= \beta_{ij}. \end{aligned} \tag{5.15}$$

The quantity $\beta_{i1}^2 + \cdots + \beta_{im}^2$, which is the portion portion of the variance of r_{it} contributed by the m common factors, is called the communality. The remaining portion σ_{i2} of the variance of r_{it} is termed as specific variance [19].

It is not uncommon that covariance matrix Σ has an orthogonal factor representation as it consists of a random variable r_t, which does not have orthogonal factor representation and moreover it is non-unique. In other words, for any $m \times m$ orthogonal matrix P satisfying $PP' = P'P = I$, let $\beta^* = \beta P$ and $f_t^* = P' f_t$. Then

$$r_t - \mu = \beta f_t + \epsilon_t = \beta PP' f_t + \epsilon_t + \epsilon_t = \beta^* f_t^* + \epsilon_t.$$

In addition, $E(f_t^*) = 0$ and $Cov(f_t^*) = P'Cov(f_t)P = P'P = I$. Therefore, β^* and f_t^* form another orthogonal factor model r_t. Non-uniqueness gives an opportunity to choose subjective number of factors, however it also allows us to conduct rotations of factors in order to achieve interpretability as we know that P is an orthogonal matrix, then the transformation $f_t^* = P' f_t$ is a rotation in the m-dimensional space [19].

As we work with covariance matrix Σ_r or correlation matrix ρ_r, the matrix has to checked for its "factorability" or in other words, the possibility of obtaining factors and two methods exist to perform such test: Bartlett's test of sphericity and Kaiser-Meyer-Olkin Measure of Sampling Adequacy.

Bartlett's test of sphericity tests the null hypothesis that correlation matrix ρ_r is an identity matrix, which indicates that variables are not related and therefore incompatible for factorization.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) is a statistic, which evaluates the proportion of variance in variables might be caused by unobserved factors, therefore higher values are preferred and values below 0.5 are considered as unacceptable. Below, are the guidelines are suggested by Kaiser and Rice [9]:

KMO Measure	Recommendation
≥ 0.90	Marvellous
$0.80+$	Meritorious
$0.70+$	Middling
$0.60+$	Mediocre
$0.50+$	Miserable

5.3.2 Factor Estimation

The orthogonal factor model in Equation 5.13 can be calculated via: principal components method and maximum-likelihood method.

The first method does not depend on the assumption of normality and it is the main method for exploratory factor analysis as we do not posses any information about factors and their number.

The second method in contrast is more applied for confirmatory factor analysis and requires normally distributed data. In this thesis we will omit the second method as we are interested only in obtaining common latent factors.

Let $(\hat{\lambda}_1, \hat{e}_1), ..., (\hat{\lambda}_k, \hat{e}_k)$ be pairs of the eigenvalues and eigenvectors of the sample covariance $\hat{\Sigma}_r$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \hat{\lambda}_k$. Let $m < k$ be the number of common factors. Thus, the factor loadings matrix is given by

$$\hat{\beta} \equiv [\hat{\beta}_{ij}] = \Big[\sqrt{\hat{\lambda}_1}, \hat{e}_1 \mid \sqrt{\hat{\lambda}_2}, \hat{e}_2 \mid \cdots \mid \sqrt{\hat{\lambda}_m}, \hat{e}_m\Big]. \tag{5.16}$$

The calculated specific variances are the diagonal elements of the matrix $\hat{\Sigma}_r - \hat{\beta}\hat{\beta}'$ or, $\hat{D} = \text{diag}\{\hat{\sigma}_1^2, ...\hat{\sigma}_k^2\}$, where $\hat{\sigma}_i^2 = \hat{\sigma}_{ii,r} - \sum_{j=1}^{m} \hat{\beta}_{ij}^2$, where $\hat{\sigma_{ii,r}}$ is the (i,i)-th element of $\hat{\Sigma}_r$. The communalities are calculated as follow:

$$\hat{c}_i^2 = \hat{\beta}_{i1}^2 + \cdots + \hat{\beta}_{im}^2.$$

Whilst the error matrix is approximated below (defactored returns):

$$\hat{\Sigma}_r - (\hat{\beta}\hat{\beta}' + \hat{D}).$$

In the perfect case the matrix should be close to 0. It is also possible possible to show the the sum of squared elements of $\hat{\Sigma}_r - (\hat{\beta}\hat{\beta}' + \hat{D})$ is less than or equal to $\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_k^2$. Thence, the approximation error is bounded by the sum of squares of the overlooked eigenvalues. According to solution in Equation 5.16, the calculated factor loadings based on the principal component method remain unchanged with an increase of the number of common factors m [19].

5.3.3 Factor Selection

Our next step is to determine the number of factors need to be retained. We can refere to heuristic rules such as eigenvalue-greater-than-one rule and Catell's scree plot [16].

Eigenvalue-greater-than-one rule is based on the rule of dropping all components with eigenvalues below one. The idea for this rule is that for standardized data, the amount of variance explained by each factor should at least equal to the variance of at least one variable [16].

Catell's scree plot is the plot of the factors on the x-axis against corresponding eigenvalues on the y-axis. As one moves to the right, toward later factors, the eigenvalues drop, thus when the drop makes an "elbow" shape or less steep decline, scree test suggests to remove all further further factors [4].

Additionally to above rules we can choose the number of factor depending on number of factors explained, this value could be as low as 50%.

5.3.4 Factor Rotation

The goal of factor rotation is to find groups of large and insignificant coefficients in any column of the rotated matrix of factor loadings. For a rotation of factors, for any $m \times m$ orthogonal matrix P,

$$r_t = \mu + \beta f_t + \epsilon_t = \beta^* f_t^* + \epsilon_t, \tag{5.17}$$

where $\beta^* = \beta P$ and $f_t^* = P' f_t$. In addition,

$$\beta\beta' + D = \beta PP'\beta' + D = \beta^*(\beta^*)' + D. \tag{5.18}$$

above result suggests that the communalities and the specific variances do not change from orthogonal transformation, which are identical to rotating the common factors in the $m-$dimensional space and infinite rotations are possible in order to achieve good interpretations [19].

Different rotation methods exist such as: varimax, quartimax, equimax, oblimin, promax, each serving its own purpose. Let us skip an explanation of rotation methods as we are not interested in interpretation of our factors, but rather estimation of our defactored returns, for which we need to estimate their factor scores.

5.3.5 Factor Scores

In order to estimate factor scores, multiple regression have to be used [16]. Below is one example of estimation, where factor score for stock return i, for a given factor j can be defined as:

$$\hat{f}_{ij} = \hat{\beta}_1 r_{i1} + \hat{\beta}_2 r_{i2} + \cdots + \hat{\beta}_p r_{ip} \ ,$$

where $\hat{f}_{ij}$ is the estimated factor score for factor j for stock returns i, $\hat{\beta}_p$ is the estimated factor score coefficient for company p and r_{ip} is the p-th observed return for company

i. This equation can be represented in matrix form as:

$$\hat{F} = R\hat{B} \ ,$$

where $\hat{F}$ is an $n \times m$ matrix of m factor scores for the n companies. R is an $n \times p$ matrix of observed returns and B is a $p \times m$ matrix of estimated factor score coefficients, which can be estimated using maximum-likelihood or ordinary least squares methods.

5.3.6 Defactored Model

Specific returns from the fitted regression model (stocks regressed on estimated factors) are referred to as defactored returns (we could also use the term as idiosyncratic returns from our arbitrage pricing theory). For example, for a company i this may be expressed algebraically as:

$$dr_{s_i} = r_i - r_{f_1} + r_{f_2} + \cdots + r_{f_m} \ ,$$

which is the difference between observed stock returns and weighted factor scores. Factor loadings/ exposures have to be updated periodically to maintain any reasonable expectation of forecast (and thence trading performance). Since the statistical factors directly reflect structure in historical prices of companies, it is obvious to have periodical changes in this structure [13].

Vector of defactored returns, computed for each day in the sample, produces raw time series from which the prediction model can be built. For example, for an autoregressive model, an entry in the regression for day t of company j is:

$$\sum_{a=1}^{k} dr_{j,t-a} = \beta_1 dr_{j,t-k} + \cdots + \beta_q dr_{j,t-k-q+1} + \epsilon_{j,t}$$

Above equation shows that k-day cumulative defactored return to day t is regressed on the q daily defactored returns immediately preceding the cumulation period [21]. Let us follow to the next section, where we can summarize all the above discussed knowledge into quantitative trading strategy [1].

5.3.7 Theoretical Framework

We skip the drift term α, we will focus on sums of our defactored or idiosyncratic returns dr_t, an increment of stationary stochastic process over the related period, which we expect to be mean-reverting. We know from arbitrage pricing theory and statistical factor analysis assumptions that $E(dr_t) = 0$. Based on statistical tests we choose the stock following the model and apply our forecasting techniques from Chapter 3. Let s_t be equal to the sum of cumulative defactored returns $\sum_{t=1}^{N} dfr_t$ for which we have to calculate empirical standard deviation $\triangle = \sqrt{\frac{\sum_{t=1}^{N}(s_t - \mu_t)^2}{N-1}}$ required for construction of trading rules. Therefore, trading rules are generated as following:

- When at time t: $s_t > +\triangle$, we sell the respective stock.
- When at time t: $s_t < -\triangle$, we buy the respective stock.

Trading positions are closed when returns are reverted back to the levels of standard deviation $\triangle$ [1]. Let us proceed to the next section where we will apply on practice our factor model arbitrage.

5.4 Backtesting

5.4.1 Data

We will use the daily adjusted closing prices of stocks traded at London Stock Exchange starting from the period of 30th of June 2015 until the 30th of June 2019. This four-year period is justified by the changes in cyclical trends of stock market trends. The following companies are included in our analysis: "3i Group plc", "Admiral Group plc", "Anglo American plc", "Antofagasta plc", "Ashtead Group", "Associated British Foods plc", "AstraZeneca plc", "Auto Trader Group plc", "AVEVA Group plc", "Aviva plc", "BAE Systems plc", "Barclays plc", "Barratt Developments plc", "The Berkeley Group Holdings plc", "BHP", "BP plc", "British American Tobacco plc", "The British Land Company plc", "BT Group plc", "Bunzl plc", "Carnival Corporation & plc", "Centrica plc", "Coca-Cola HBC A.G.", "Compass Group plc", "CRH plc", "Croda International plc", "DCC plc", "Diageo plc", "Direct Line Insurance Group plc", "EVRAZ plc", "Experian plc", "Ferguson plc", "Fresnillo plc", "GlaxoSmithKline plc", "Glencore plc", "Halma plc", "Hargreaves Lansdown plc", "Hiscox Ltd", "HSBC Holdings plc", "Imperial Brands plc", "Informa plc", "InterContinental Hotels Group plc", "International Airlines Group S.A.", "Intertek Group plc", "ITV plc", "JD Sports Fashion plc", "Johnson Matthey", "Just Eat plc", "Kingfisher plc", "Legal & General Group plc", "Lloyds Banking Group plc", "London Stock Exchange Group plc", "Marks & Spencer plc", "Melrose Industries plc", "Micro Focus plc", "Mondi Group", "Wm Morrison Supermarkets plc", "National Grid plc" (NG), "Next plc", "NMC Health", "Ocado", "Pearson plc", "Persimmon plc", "Prudential plc", "Reckitt Benckiser Group plc", "RELX plc", "Rentokil Initial", "Rio Tinto", "Rightmove plc", "Rolls-Royce Holdings plc", "Royal Bank of Scotland Group plc", "Royal Dutch Shell PLC", "RSA Insurance Group plc", "The Sage Group plc", "Sainsbury's", "Schroders plc", "Scottish Mortgage Investment Trust", "SEGRO plc", "Severn Trent plc", "Smith & Nephew plc", "DS Smith plc", "Smiths Group plc", "Spirax-Sarco Engineering plc", "SSE plc", "Standard Chartered plc", "Standard Life Aberdeen plc", "St. James's Place plc", "Taylor Wimpey plc", "Tesco plc", "TUI AG", "Unilever", "United Utilities Group plc", "Vodafone plc", "Whitbread plc" and "WPP plc".

5.4.2 Analysis

To start our analysis, let us observe our sample correlation matrix plot of daily log-returns on Figure 5.1:

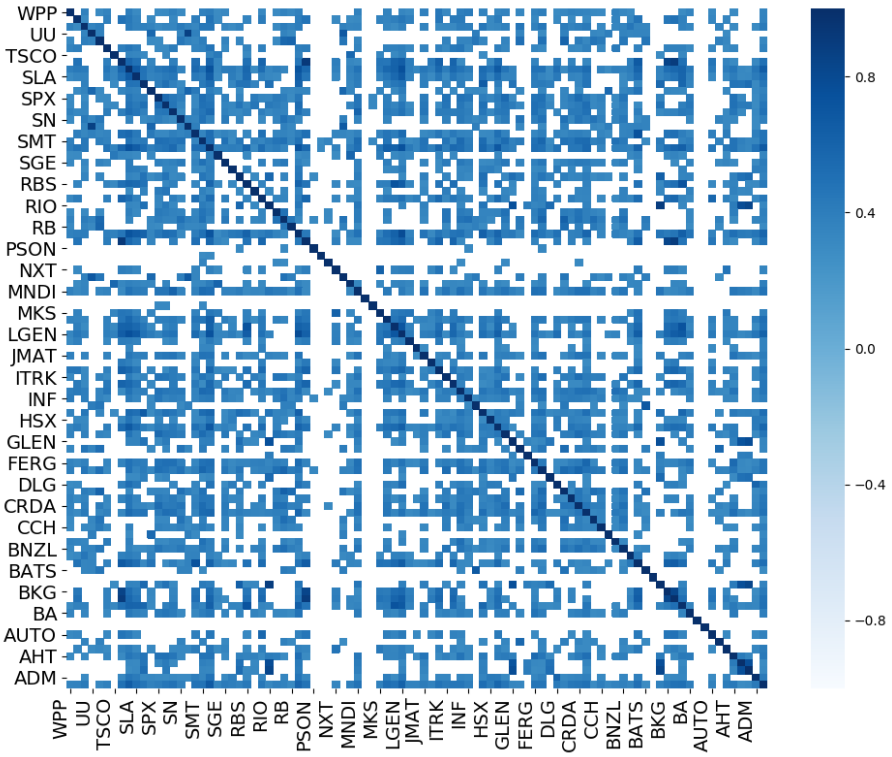


Figure 5.1: Correlation plot.

The plot masks correlations below 0.3 and we can see that most elements of our matrix are correlated from moderate to strong and we can see most variables are correlated, which means that we can apply our factorization method. However, presence of correlation is not enough, therefore we have to refer more to statistical tests. We applied Bartletts sphericity test and found it statistically significant with a p-value of less than 0.05. The next test applied was the Kaiser-Meyer-Olkin measure of adequacy test which showed that 0.96 proportion of the variance among variables that might be common variance and caused by latent factors.

Further, we performed factorization without rotation as we are more interested in obtaining factor returns rather than their interpretation. Several factorization procedures were applied by using the number of factors equal to the number of companies (95) as an initial procedure, then using 40 factors explaining 80% of the variance, 14 factors suggested by the scree plot and 10 factors approximately equal to the number sectors. The decision was to use 10 factors.

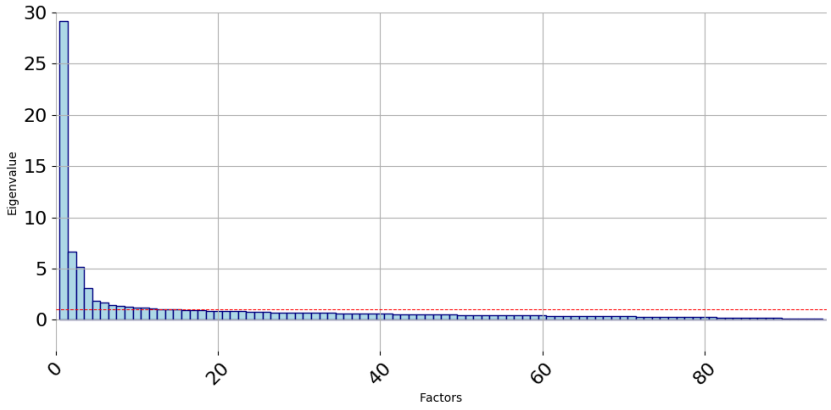


Figure 5.2: Scree plot.

The plot for the number of factors against their eigen values (Figure 5.2) shows that 14 factors have eigen values larger than 1, however 14 factors explained only 60% of the variance and minimum 40 factors explained variance more than 80%.

In order to obtain these specific returns we must regress factors scores on each stock returns by using ordinary least squares. As an example, let us illustrate on "3i Group plc". elow are the results for "3i Group plc" (III) company regressed by 10 obtained factors.

Dep. Variable:	III	R-squared:	0.592
Model:	OLS	Adj. R-squared:	0.588
Method:	Least Squares	F-statistic:	149.6
Date:	Thu, 08 Aug 2019	Prob (F-statistic):	6.87e-193
Time:	12:05:47	Log-Likelihood:	-1012.8
No. Observations:	1043	AIC:	2048.
Df Residuals:	1032	BIC:	2102.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.25e-17	0.020	-2.14e-15	1.000	-0.039	0.039
f1	0.7477	0.020	37.590	0.000	0.709	0.787
f2	-0.0860	0.020	-4.324	0.000	-0.125	-0.047
f3	-0.0100	0.020	-0.505	0.614	-0.049	0.029
f4	-0.0074	0.020	-0.372	0.710	-0.046	0.032
f5	-0.0564	0.020	-2.834	0.005	-0.095	-0.017
f6	-0.0641	0.020	-3.221	0.001	-0.103	-0.025
f7	0.0229	0.020	1.154	0.249	-0.016	0.062
f8	-0.0179	0.020	-0.898	0.369	-0.057	0.021
f9	0.1189	0.020	5.975	0.000	0.080	0.158
f10	-0.0532	0.020	-2.674	0.008	-0.092	-0.014

Figure 5.3: Ordinary Least Regression Results.

Factor loadings serve the same purpose as regression coefficients, therefore fitted values of returns for each stock can be obtained by further calculating (specific) residuals and their cumulative sum. Our fitted model for the company "3i Group plc" has the following view:

$$iii_t = 0.747fr_{1t} - 0.086fr_{2t} - 0.01fr_{3t} - 0.0074fr_{4t} - 0.0564fr_{5t} - 0.0641fr_{6t} + ...$$
$$0.0229fr_{7t} - 0.0179fr_{8t} + 0.1189fr_{9t} - 0.0532fr_{10t} + \hat{dfr}_t$$

Next, on Figure 5.4, the plot for cumulative defactored returns for the second quarter of 2019 is given, where 10 factor returns subtracted from III's returns. The obtained residual returns were summed and standard deviation was calculated for creating trading signals. The left triangle shows the signal for opening the trade by buying the stock when defactored return values deviates larger than one standard deviation for an arbitrarily large value, which is 1.5 in our case. Afterwards, we anticipate mean-reversion and close the trade by selling back the stock when the value for specific return is inside the standard deviation region (right triangle). When deciding whether to sell the stock, the opposite rule applies.

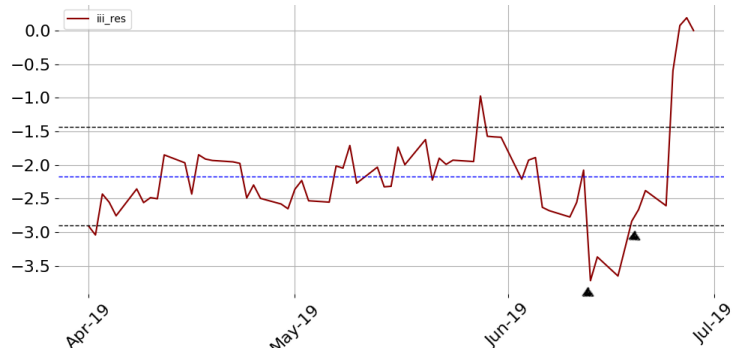


Figure 5.4: Cumulative defactored returns for *iii*

Below, we plotted the idiosyncratic returns for all the stocks in our data-set and we can see the value of their returns fluctuate around 0. The series widen over time by finally returning to equilibrium value.

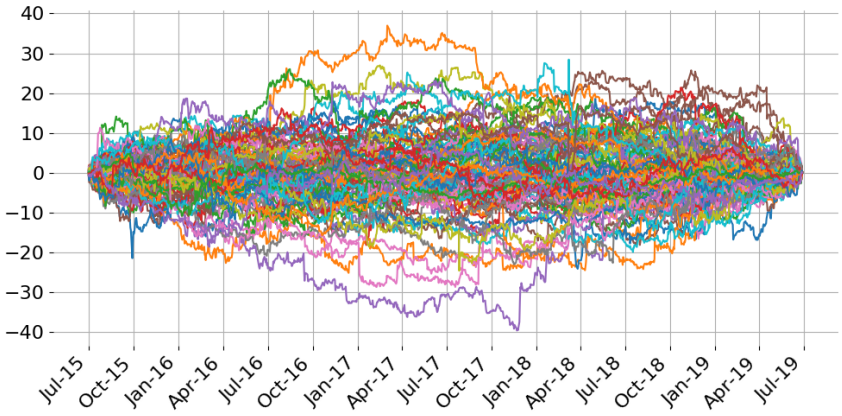
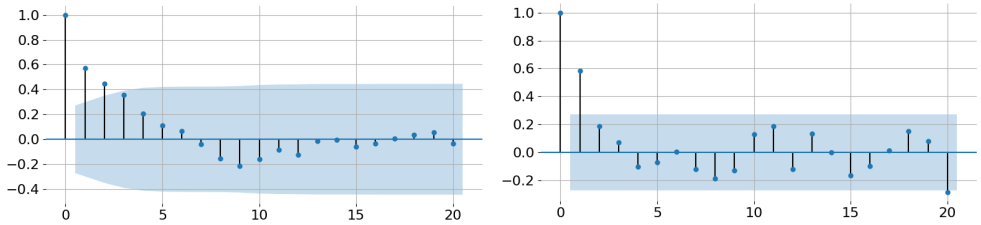


Figure 5.5: Cumulative defactored returns.

Furthermore, we can choose the companies based on assumptions and gaining positive cumulative returns for the given period (second quarter 2019) in order to apply our time series models. We will use the last two weeks of June as an out-of-sample period forecasting period. Let us continue our example with 3i Group plc company, where the same procedure can be applied to every company in our data-set. We will try to select time series model for predicting our average cumulative return for the last two

weeks of second quarter, thus let us explore autocorrelation and partial autocorrelation plots (Figure 5.6).



(a) Autocorrelation Plot. (b) Partial Autocorrelation Plot.

Figure 5.6

On both plots above we can see that autocorrelation plot (a) shows lag 2 autocorrelation and lag 1 for partial autocorrelation plot (b). Augmented Dickey-Fuller confirmed that data has no unit root and is a stationary process. Furthermore, the Ljung-Box test was applied in order to check if the series is a white noise process, where the null hypothesis (that the first 20 autocorrelations are jointly zero) was rejected.

Now, because we have a stationary time series with statistical properties such as constant mean and constant variance over time, we can proceed to choose a model based on Akaike's and Bayesian information criteria. The following models were considered: AR(1), MA(1), AR(2), MA(2), ARMA(1,1) and ARMA(2,2).

	AR(1)	MA(1)	AR(2)	ARMA(1,1)	ARMA(2,1)
AIC	36	45	34	35	36
BIC	42	51	42	43	46

Table 5.1: Model Selection.

According to Table 5.1, the second order autoregressive model was chosen as it gave the lowest value for AIC and a similar value for BIC along with the first order autoregressive model. Further, to validate the model we checked if the residuals follow white noise process, where the null hypothesis of Ljung-Box test was not rejected. Additionally, to our model choice process we applied automatic step-wise model selection function and the model choice was confirmed, thus below is our fitted model:

$$\sum_{a=1}^{k} \hat{dr}_t = -2.2696 + 0.4683 dr_{t-1} + 0.2655 dr_{t-2}$$

Below, is the plot for our predicted values using our selected model for the last two weeks of June 2019 using the data from April 1st until June 15th as a training sample.

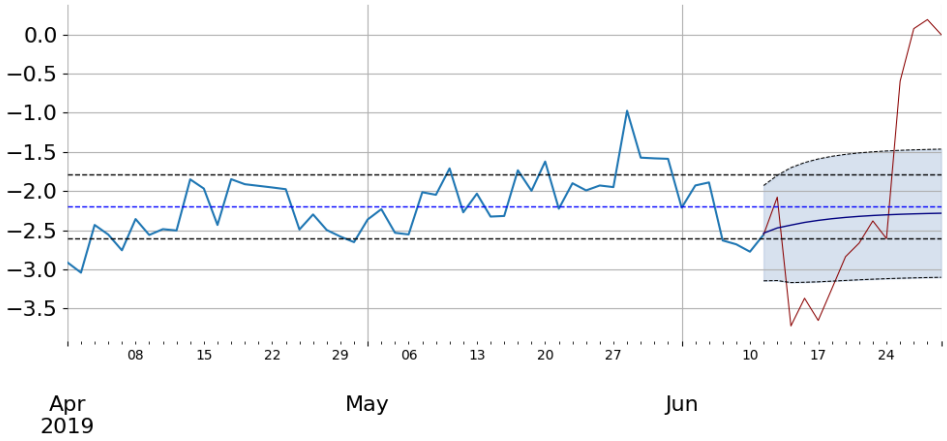


Figure 5.7: AR(2) Model Forecast.

We can see from the plot that the dark blue line shows convergence to the same level as the mean value for cumulative defactored returns for the second quarter of 2019 and the lighter blue regions shows 95% confidence interval.

6. Conclusion

The first section of this thesis discussed the concepts of statistical arbitrage and time series models. In particular, the main features of the pairs trading strategy - the foundation of most arbitrage methods - were outlined in depth.

We have explained the concept of market neutral strategies and noise models by gradually moving to discuss the theoretical framework, where the pairs trading strategy was explained in detail starting from the cointegration coefficient and the calculation of a single profit from every trade [21].

Further, we have used major European stocks from the financial sector by finding cointegrated pairs and applying our time series technique to conduct forecasting. We chose cointegrated pairs from a given basket and chose pairs with a highest cumulative logarithmic return, which was not significant due to the temporary downtrend movement in the European financial sector. We have shown that a single profit for the second quarter of 2019 for a pair of ”Gecina” and ”Icade” companies was already equal to 2.3% compared to 4.3% of a direct investment. To list the total amount of return, the pair has made over 40% (41.4%) for the period between January 2017 and July 2019, which is remarkable compared to decreasing prices of the sector, which can be connected with the theory of market neutral strategies.

In the second part of the thesis we have discussed various types of statistical arbitrage with further explanation of the concept of factor model arbitrage, in particular we have explained the arbitrage pricing theory, the link between cointegration and the arbitrage pricing theory, which is closely interconnected with pairs trading [21] as well. Further, we have explained the statistical method of factorization with a further focus on defactorized returns and applying the trading strategy on defactorized stocks from London's Stock Exchange from the period of July 2015 to July 2019.

We have worked with specific returns of stocks, where we have calculated their cumulative and their deviation in order to set entrance and exit points for trading. The trading strategy - based on the assumption of the arbitrage pricing theory [1] - was shown to result in a total cumulative logarithmic return of 37.4%. However, direct investment into stocks was shown to produce a 92.8% return over a four-year period. We should mention that one might prefer direct investment into assets, however the return gained from defactored arbitrage strategy is market neutral, therefore it is not completely correct to compare with market returns as the latter depends on macroeconomic conditions in a certain country (or of the EU). Furthermore, the time series models were applied to predict the mean value for cumulative logarithmic returns.

The trading strategies exploited in our paper have many flaws and many improvements can be added in order to improve their profitability and robustness. Robustness of strategies could be more important due to the need for consistency and predictability of returns, however riskier strategies may promise higher profits. Robust methods may include robust estimation of covariance matrix. Alternatively, non-parametric estima-

tion methods might be applied, such as neural networks in order to obtain defactored returns.

Profits can be controlled by changing the number of factors, where specific returns might be more affected by market gains, by using opposite reversion signals or by choosing for trading more volatile stocks. However, one might be more affected by the problem of non-constant variances. Non-constant variances is a common problem in financial time series, which means that changes in volatility over time cause problems for classical time series models. Therefore, models such as generalized autoregressive conditional heteroskedasticity may be able to adapt to changes in volatility.

We have not mentioned optimal periods for updating estimated values in our pairs trading or defactored model arbitrage strategies, where periods may vary from seconds to quarter. In the real trading environment, trading strategies have to be tested for simulated stock prices, with further calculation of maximum draw-down and statistical significance of returns [5]. Additional costs such as broker commission or 'slippage' costs have to included for calculating net profits for single trades. Stop loss orders should also be added in order to mitigate against event risks and large capital draw-downs.

7. References:

[1] Marco Avellaneda and Jeong-Hyun Lee. "Statistical Arbitrage in the U.S. Equities Market". In: *Quantitative Finance* 10.7 (2008), pp. 761–782.

[2] Louis Bachelier. "Theorie de la Speculation". In: *Annales Scientifiques de l'Ecole Normale Superieure* 3.17 (1900), pp. 21–86.

[3] Ivan F. Boesky. *Merger Mania: Arbitrage : Wall Street's Best Kept Money-Making Secret*. Holt Rinehart & Winston, 1985. ISBN: 978-0-030-02602-7.

[4] R. B. Cattell. "The Scree Plot Test for the Number of Factors". In: *Multivariate Behavioral Research* 1.2 (1966), pp. 140–161.

[5] Ernie Chan. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business*. Wiley Trading Series. Johnn Wiley & Sons, Inc., 2008. ISBN: 978-0-470-28488-9.

[6] Christophe Croux. *Time Series Analysis, Course, D0M63B, KU Leuven*. Sept. 2018.

[7] Robert F. Engle and C. W. J. Granger. "Co-Integration and Error Correction: Representation, Estimation, and Testing". In: *Econometrica* 55.5 (1987), pp. 251–276.

[8] Eugene Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *Journal of Finance* 25.2 (1970), pp. 383–417.

[9] John Kaiser Henry F.; Rice. "Little Jiffy, Mark IV". In: *Educational and Psychological Measurement* 34.1 (1974), pp. 111–117.

[10] Burton G. Malkiel. "The Efficient Market Hypothesis". In: *Science* 240.4911 (1989), pp. 1424–1425.

[11] Burton G. Malkiel. "The Efficient Market Hypothesis and Its Critics". In: *Journal of Economic Perspectives* 17.1 (2003), pp. 59–82.

[12] Harry Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[13] Andrew Pole. *Statistical Arbitrage: Algorithmic Trading Insights and Techniques*. Wiley Finance Series. Johnn Wiley & Sons, Inc., 2007. ISBN: 978-0-470-13844-1.

[14] Laurence Rosenberg. *ETF Strategies and Tactics: Hedge Your Portfolio in a Changing Market*. McGraw-Hill Finance & Investing, 2008. ISBN: 978-0-071-49734-3.

[15] Stephen Ross. "The Arbitrage Theory of Capital Asset Pricing". In: *Journal of Economic Theory* 13.3 (1976), pp. 341–360.

[16] Subhash Sharma. *Applied Multivariate Techniques*. Johnn Wiley & Sons, Inc., 1996. ISBN: 978-0-471-31064-8.

[17] William F. Sharpe. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk". In: *Journal of Finance* 19.3 (1964), pp. 425–442.

[18] Watson M. Stock J. "Testing for Common Trends". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1097–1107.

[19] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics. Johnn Wiley & Sons, Inc., 2010. ISBN: 978-0-470-41435-4.

[20] Tim Verdonck. *Statistical Tools for Quantitative Risk Management, Course, G0Q24A, KU Leuven*. Sept. 2018.

[21] Ganapathy Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*. Wiley Finance Series. Johnn Wiley & Sons, Inc., 2004. ISBN: 978-0-471-46067-1.