LUT University

School of Business and Management

Master's Degree in Strategic Finance and Business Analytics

Master's Thesis

Pairs trading revisited - the case of OMX Helsinki

Sami Kohvakka, 2020

Supervisor: Eero Pätäri D.Sc. (Econ.)

2nd examiner: Sheraz Ahmed D.Sc. (Econ.)

ABSTRACT

Author:	Sami Kohvakka
Title:	Pairs trading revisited - the case of OMX Helsinki
Year:	2020
Faculty:	School of Business and Management
Major:	Strategic Finance and Business Analytics
Master's Thesis:	Lappeenranta-Lahti University of Technology LUT 99 pages, 23 figures, 30 tables, and 5 appendices.
Examiners:	Professor Eero Pätäri & Associate Professor Sheraz Ahmed
Keywords:	pairs trading, stock markets, copula, cointegration

This thesis examines pairs trading opportunities in OMX Helsinki Stock Exchange. Pairs trading is a self-financing trading strategy, where trader enters a long position and offsetting short position simultaneously in two correlated or otherwise related assets. It draws from the relative value between the assets and ought, in theory, provide positive returns independent of market returns. In practice, this strategy is executed by selling the overvalued asset and purchasing the undervalued asset.

The performance of pairs trading rules is compared between distance-based, cointegration-based and copula-based trading signal generation. Data consists of companies listed at OMX Helsinki main list between 2004 and May 2020. To overcome survivor bias, few companies that went bankrupt during the time period were added to the pool of possible pairs. Pairs were limited to allow only companies within the same main industry classification group.

In general, pairs that are made of different share classes of one company are quite suitable for pairs trading. Both distance-based and cointegration-based screening criteria favored such pairs over pairs formed of two separate companies. Copula method seemed to be the weakest both in terms of the number of trading opportunities and the average profit per trade. In general, only the five most cointegrated or closely related pairs at the fitting period are suitable for trading. Distance method seems to create more consistent returns than cointegration method.

TIIVISTELMÄ

Tekijä:	Sami Kohvakka
Otsikko:	Parikaupankäynti Helsingin pörssissä
Vuosi:	2020
Tiedekunta:	School of Business and Management
Pääaine:	Strateginen rahoitus ja liiketoiminta-analytiikka
Pro gradu -tutkielma:	Lappeenrannan-Lahden teknillinen yliopisto LUT 99 sivua, 23 kuvaa, 30 taulukkoa ja 5 liitettä.
Tarkastajat:	Professori Eero Pätäri ja apulaisprofessori Sheraz Ahmed
Hakusanat:	parikaupankäynti, osakkeet, copula, yhteisintegraatio

Tutkielmassa selvitetään parikaupankäynnin mahdollisuuksia Helsingin pörssissä. Parikaupankäynti on itse itsensä rahoittava kaupankäyntistrategia, jossa samanaikaisesti avataan toisensa kumoava lyhyt ja pitkä positio korreloituneissa tai muuten toisistaan riippuvissa kohteissa. Strategia perustuu kaupankäynnin kohteena olevien arvopapereiden tai hyödykkeiden suhteelliseen arvoon, ja tarjoaa siten periaatteessa markkinoista riippumatonta tuottoa. Strategiassa siis myydään yliarvostettua ja ostetaan aliarvostettua hyödykettä samanaikaisesti.

Parikaupankäynnin kannattavuutta tutkitaan etäisyyspohjaisen, yhteisintegraatiopohjaisen ja copulapohjaisen kaupankäyntisignaalien luonnin kautta. Data sisältää päälistalle listatut yritykset vuoden 2004 alusta vuoden 2020 toukokuuhun. Selviytyjävinouman vuoksi dataan lisättiin pörssistä konkurssin vuoksi poistuneita yrityksiä. Parien muodostusta rajoitettiin siten, että molempien osakkeiden on oltava samalta pääsektorilta.

Käytännössä saman yrityksen eri osakesarjoista muodostuvat parit osoittautuivat hyviksi parikaupankäynnin kohteiksi. Nämä parit valikoituivat muita useammin sekä etäisyyteen että yhteisintegraatioon perustuvassa valintatavassa. Copula-menetelmä osoittautui huonoimmaksi sekä kaupankäyntimahdollisuuksien että keskimääräisen voiton perusteella mitattuna. Käytännössä kaupankäyntiin soveltuu vain viisi lähiten toisiaan seuraavaa paria. Etäisyysmenetelmä vaikuttaa tarjoavan hieman vakaammat tuotot kuin yhteisintegraatioon perustuva menetelmä.

PREFACE

A lot has happened since I began writing this thesis in February 2019. Starting a new job, moving to a different city and buying a house while studying for not one but two master's degrees can easily overwhelm the most of us. In March 2020, the global COVID-19 epidemic reached Finland and turned our lives upside down. Abruptly social isolation at home was the new normal and everyone who could work from home began to work from home. Acknowledging my privileged position of being young and healthy, I was able to cope with the new situation quite well. Despite all the restrictions, I kinda liked that it eliminated commutes and left me more time to write this thesis.

I am expressing my gratitude towards everyone who's been riding this journey with me. My parents, for being there for me and my supervisor Eero Pätäri who provided me with much needed guidance on this thesis and showed extraordinary patience with me when writing my thesis took significantly longer than expected.

Without professor Pätäri's patience and understanding I probably would not have graduated. Finally, I want to thank my fiancée, who kept my work life, studies and free-time in balance. It made it possible to concentrate and keep myself motivated during times when finding motivation seemed difficult or free-time activities too tempting.

Lahti, September 24th, 2020

Sami Kohvakka

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS AND SYMBOLS

ADF	Augmented Dickey-Fuller
AIC	Akaike information criterion
BIC	Bayesian information criterion
DW	Durbin and Watson
SSD	Sum of squared differences
$\mathbb{R}$	Real numbers
$\mathbb{R}^+$	Positive real numbers

1 Introduction

The Finnish Stock market has been struggling to keep up with other developed economies in recent decades, and 12 years since the previous crash in 2007, the market indices barely rose above the level prior to the financial crises of 2007 before the COVID-19 pandemic struck. The overall economy had been slowing down in Euro area and China even before the pandemic. The government debt of Finland has doubled since the financial crisis of 2007, and we might be on the edge of yet another recession, this time likely far worse than the previous one. To boost the economy after COVID-19, the government plans to increase the budget by 33%, or 18,8 billion euros, funded entirely by debt. (HE 88/2020, p. 37). This serves as a strong incentive for investors to seek market-neutral trading strategies that are profitable independent of the current market conditions.

Regulating the markets limits volatility in the name of stability and pushes risks further to the tails, making the economy more fragile. When it crumbles, it crumbles big time, as we saw globally in 2007 and in Egypt in 2011. (Taleb and Blyth 2011). To protect themselves from economic downturn, investor can choose from two strategies. The first one is tail risk hedge, discussed in Litterman (2011). Tail hedge can be thought of as an insurance - it has low constant expected negative return, the price of the insurance, but in rare cases when tail risk realizes it gives a substantial one-time positive return. The other is identifying trading strategies that are market neutral.

Market-neutrality refers to trading strategies that draw from the relative performance of the assets instead of the absolute performance, as in conventional trading strategies. The total return of the portfolio is a function of the return differential between long and short assets. For a perfectly balanced market-neutral portfolio, gains in one asset are offset by losses in another asset, and therefore, total portfolio returns equal to zero. For a managed market-neutral portfolio, gains on the long asset are expected to outperform losses in the shorted asset in rising markets, and the short to outperform the long in falling markets, thus creating a consistently positive return regardless of the overall market direction. (Ehrman 2006, pp. 3–5, 27–33).

In Sharpe's capital asset pricing model (CAPM) terms, market-neutrality refers to portfolios that have zero beta. The CAPM decomposes portfolio returns to two components – one indicating the overall market returns, the systematic component, and other indicating independent returns of the assets, described by the residual and refereed to as nonsystematic component. The CAPM equation is usually written as $r_p = \beta r_m + \theta_p$, where r_p is portfolio excess returns, r_m is excess market return component and θ_p is the residual. Gradient β describes the leverage

of the portfolio over the market. A one percent increase in market returns increases returns of the systematic component by beta times one percent. Key assumptions of CAPM state that those two components are uncorrelated, and that mean value of residual is zero. This implies that the residual series must be mean-reverting and oscillate around zero. (Vidyamurthy 2004, pp. 3–7).

A key concept in creating market independent trading strategies is *statistical arbitrage*. Per Krauss, Do, and Huck (2017), statistical arbitrage refers to quantitative trading strategies often used by hedge funds and characterized by the following features: trading signals are systematic, as opposed to driven by fundamentals, constructed portfolio is market-neutral, and the mechanism for generating excess returns is statistical. Systematic refers to algorithm-based signals drawn from the data ignoring the fundamental characteristics.

Given the significant improvements in communications during the past few decades, investors are struggling to keep up with the speed at which new information is absorbed by the markets. An article in The Economist (2019) states that nowadays only 10% of institutional trading in America is done by traditional equity fund managers. Most floor traders have faced job extinction and been replaced by computers.

The world is very different for independent investors, often operating with relatively small cash amounts. While stock market operators and large institutional investment companies have built a high-speed network to enable fast and direct market access suitable for high-speed trading, their offering to individuals is very slim. Some companies, such as Lynx do provide computers direct access to the markets by exposing the market prices in machine-readable formats through application programming interfaces (APIs), but it can cost thousands of dollars per year.

Even though not officially marketed as API services, technologically inclined traders can level the playing field by converting any data source used to populate public web pages to machine-readable data streams through reverse engineering. These sources include pages drawing graphs about historic stock prices as well as stock brokers' official web pages and their mobile applications. Some brokers allow computers to place bids independently, others can be exploited to allow algorithmic trading through robot frameworks that parse HTML pages and emulate user actions by sending keystrokes and mouse clicks to a headless browser instance. To run trading algorithms, any computer connected to the internet would suffice. This brings algorithmic trading within the reach of most traders willing to commit enough effort to climb the learning-curve in basic programming and statistics.

Pairs trading is one of the possible trading methods aiming to find short-term statistical

arbitrages and exploit them for profit chasing. It is by construction market-neutral, as it draws from the relative performance of assets instead of market trends or fundamental factors. It is an investing strategy based on an assumption that there are long-run equilibrium levels in the valuation of two somehow related securities. Essentially, these stocks move in same direction by approximately same magnitude over time. In short term, there are random deviations from this equilibrium level. When such deviations occur, arbitragers are trying to capitalize on them by going long on the undervalued security and going short on the overvalued security. If the equilibrium level truly exists, relative values of those two securities will converge back to the equilibrium level. Two of the most cited publications related to pairs trading are Gatev, Goetzmann, and Rouwenhorst (1999; 2006).

Previously, pairs trading has been studied in OMX Helsinki by Kupiainen (2008), Harju (2016), and Rinne and Suominen (2017). Kupiainen focused only on distance method, but found it lucrative. Harju expanded previous research by including cointegration method and copula method, but he made some unorthodox choices regarding the fitting and trading period length and did not provide much insight to pair selection. Rinne and Suominen found an average transaction return of 2.4% for an arbitrary pair in OMX Helsinki using the distance method but did not proceed to examine the returns of other possible pairs.

Contrary to Harju (2016), this thesis aims to replicate the setting in Gatev et al. (1999; 2006) by using similar window lengths and aggregating the results to form multiple trading periods. It explores the uncharted territories of Kupiainen (2008) and Harju (2016) by applying *distance-based* trading rules to pairs selected by *distance* criterion and *cointegration-based* trading rules to pairs formulated by *cointegration* criterion.

1.1 Objectives and Restrictions

This research focuses on examining market-neutral trading strategies and testing if such strategies are feasible for small and large investors after transaction costs. This thesis tries to find statistical arbitrages in the Finnish stock markets and aims to construct a profitable beta-neutral portfolio using pairs trading strategies. The main research question in this thesis is:

Is it possible to construct market neutral, consistently profitable portfolios using pairs trading in Finnish stock markets?

Supporting research questions are:

What are the main methods of pairs trading?

Is some method superior to others in OMX Helsinki stock exchange?

1.2 Structure of the Thesis

This thesis begins with introduction, is followed by literature review discussing the most common methods of pairs trading, continues with empirical part applying those methods to OMX Helsinki and ends with a brief summary of findings. Literature review focuses around three main methods of pairs trading. These are distance method, cointegration method and copula method. All of these have been studied extensively in American stock markets. This section also examines briefly other emerging methods of pairs trading, such as stochastic control theory and machine learning.

The empirical section discusses about implementing those three main methods in the OMX Helsinki stock exchange and presents a summary of results when those methods are applied to the same market. Results are discussed in terms of what kinds of pairs different selection criteria favors, how many trading opportunities they create and what is the average return per opened trade. The empirical section discusses how the results obtained in this thesis compare with results presented by Harju (2016) and Rinne and Suominen (2017) as well as what could be some future research directions.

At the end of this thesis there are some supporting material, listing the trading periods and companies used, for which periods the data was available for each of those companies and what chart patterns typical pairs look like.

2 Literature review

This chapter examines the previous literature on pairs trading. It formulates an overall understanding on what types of trading strategies exist and how trading signals can be generated in this domain. The main focus of this chapter revolves around three of the most established signal-creation methods - the distance method, cointegration method and copula method.

2.1 Theoretical background

Focardi, Fabozzi, and Mitov (2016) argue that attractive investments attract investors and thus their prices increase. Progressively, this yields to less attractive, overpriced investments. As the investors realize their assets are overpriced, they will try to sell them, pushing the prices lower. This in turn increases the attractiveness of these investments. Natural price fluctuations like these are the source of mean reversion and statistical arbitrage in stock markets. By modelling these fluctuations investors should be able to make consistent profit.

Statistical arbitrage refers to consistently profitable trading rules that generate risk free profits. (Hogan et al. 2004). It often involves opening related and offsetting positions that can be closed for profit at a later time. Arbitragers drive the markets to be more efficient by exposing significant mispricings. For example, index futures arbitragers open positions when absolute deviation from fair value exceeds the transaction costs of arbitrage. If the contract can be liquidated early, the value of an option to do so is added to the absolute value of the deviation. (Neal 1996).

According to Huck and Afawubo (2015) pairs trading strategies can be grouped to three categories:

- The minimum distance approaches
- Multi-criteria decision methods
- The modelling of mean reversion

Of these three groups, the minimum distance approach was presented in Gatev, Goetzmann, and Rouwenhorst (1999), which is widely considered as the seminal paper about pairs trading. While technically also modelling mean reversion, it is therefore considered as a separate

group often serving as a benchmark for other methods. Multi-criteria decision methods are the most novel group of these, with little experimental support and no established signal creation methods. (Huck 2015).

Pairs trading is based on finding a pair of stocks whose prices have moved in harmony throughout history. When prices diverge, trader takes a short position on winner and goes long on the loser. When prices converge, the positions are closed. (Gatev, Goetzmann, and Rouwenhorst 2006). The direction of movement is irrelevant, as the trader speculates only on the spread of the asset prices. The underlying assumption is that there is an equilibrium level around which the spread fluctuates, which is why these strategies are sometimes referred to as *relative value* based trading strategies. (Triantafyllopoulos and Montana 2011).

According to Krauss (2017), several authors have since built on Gatev's paper, and enriched the concept of pairs trading by introducing more complex approaches. These approaches are listed in Table 1.

Table 1. Pairs trading approaches presented in literature

Approach	**Description**	**Examples**
Distance	Pairs are identified by using distance metrics. This is perhaps the simplest approach.	Gatev, Goetzmann, and Rouwenhorst (2006)
Cointegration	Cointegration tests are applied to identify pairs and generate signals.	Chiu and Wong (2015), Yu and Lu (2017)
Copula	Trading signals are generated by relative value drawn from estimating the joint probability distribution of returns.	Liew and Wu (2013), Xie et al. (2016)
Time series	Focuses on generating trading signals by time series analysis. Often ignores formation period.	Kim and Heo (2017)
Stochastic control	Uses stochastic control theory in determining value and policy functions for this portfolio problem. Ignores formation period.	C. W. Chen et al. (2017) Göncü and Akyildirim (2016)
Other	Experimental frameworks with less supporting literature. These approaches include machine learning and principal component analysis.	Huck (2010)

Pairs trading is not limited to the stock markets, and several attempts have been made to incorporate these practices on other asset classes as well. For example, Göncü and Akyildirim (2016) applied pairs trading rules on commodity futures markets. As another example, Montana and Parrella (2009) constructed an artificial asset representing the estimated fair

market valuation of a real asset and paired it against a tradable ETF. Lintilhac and Tourin (2017) applied cointegration based strategies to bitcoin markets.

Blázquez, Cruz, and Román (2018) found out that the pair of stocks with the highest correlation is also the one with the least distance between them, indicating that the correlation and the distance methods systematically choose the same pair of stocks in the same order.

An alternative pairs trading strategy was examined by Bolgün, Kurun, and Güven (2012), who engaged in long position on a synthetic Turkish ETF and short in Turkish Derivatives Exchange index futures contract.

2.2 Profitability of pairs trading

In a comprehensive analysis of pairs trading profitability, Jacobs and Weber (2015) studied 34 international stock markets and found abnormal returns to persist across those markets. Their analysis spanned from January 2000 to December 2013, and they used similar distance-based method of constructing pairs than Gatev et al. (2006), who had previously found pairs trading profitable with an average of 11% p.a return in the US markets.

Pairs trading is a self-financing, *dollar neutral* strategy. Funds obtained from short selling are used to create a long position on another asset. When the positions are closed, income from closing the long position is used to close the short position. Provided that the trades are profitable, this creates leverage as investor can create much more larger portfolios than conventional long-long portfolios. The size of a long-short portfolio is limited only by the margin requirements. (Ehrman 2006, pp. 63–65).

Rad, Low, and Faff (2016) studied the profitability of pairs trading in US markets. During their sample from 1962 to 2014, all three of the common pairs trading methods showed mean monthly excess returns from 91 to 43 basis points. However, the frequency of pairs trading opportunities showed significant decline for distance and cointegration methods starting from 2009. Similar observations were presented previously by Do and Faff (2010), who compared the profitability of Gatev's trading rule in US markets over three different time periods - 1962 to 1988, 1989 to 2002 and 2003 to 2009. Mean excess returns declined by 57 percent between the first two periods, and shrank to 0.24 percent in the last period. Tianyong, Ming, and Liang (2013) found pairs trading profitable in Shanghai stock market during their sample period from 2003 to 2008.

In general, profitability ranking of pairs trading methods varies in literature. Lei and J. Xu (2015) found co-integration based strategies more profitable than distance based strategies in Chinese stock market using dual-listed Chinese companies as tradable pairs. Smith and X. Xu (2017) determined that cointegration method was profitable in US markets only back in the 1980s. Under their parametrization, distance method outperformed cointegration method from 1980 to 2014. Intuitively, a less diversified portfolio yields higher returns than a larger portfolio, mainly because the average quality of pairs deteriorates as more and more pairs are accepted to the portfolio, but the lower number of pairs also bears higher risk.

Huck (2015) found the cointegration method superior to the distance based method between July 2003 and June 2013 when trading the components of the S&P 500 and the Nikkei 225. Both methods performed well during the 2008 financial crisis, and volatility timing using VIX index did not improve the performance of the cointegration method. Clegg and Krauss (2018) note that cointegration is not a permanent phenomenon between two series, which might explain why the efficiency of the cointegration method varies a lot in literature.

Mikkelsen (2018) compared the profitability of distance and cointegration methods on 18 seafood companies traded on the Oslo Stock Exchange. Neither of the strategies generated significant excess returns between January 2005 and December 2014.

Stübinger and Bredthauer (2017) examined pairs trading profitability in high frequency context, and discovered that despite declining profitability, profitable pairs trading strategies existed among the S&P500 constituents between 1998 and 2015. The best-performing pairs achieved an annualized Sharpe ratio of 8.14 and returns of 50.50 % p.a. after transaction costs. The relative performance of pairs trading was exceptionally good during market turmoils, such as dot-com crisis and the global financial crisis.

Rinne and Suominen (2017) argue that pairs trading returns can be justified by the liquidity pairs traders bring to the markets. When studying the stock prices of two of the largest domestic pulp manufacturers, Stora Enso and UPM, they found pairs trading returns exceptionally high on days of high trading volume. On those days, nearly 45 percent of traders engaged in pairs trading.

When trading substitutes in commodity markets, or derivatives using the same underlying asset, or different share classes of the same company, profitability of pairs trading is often attributed to enforcing the Law of One Price. Economic substitutes ought to be priced equally when market frictions are eliminated. To pairs trading this translates as having an equilibrium price that describes the equal utility of any asset in a given pool of similar assets when costs attributed to using such assets are factored in. (Hain, Hess, and Uhrig-Homburg 2018).

2.3 Risks of pairs trading

Since Gatev et al. (1999), authors have selected the assets from the same sectors to remain consistent with the original paper. Ehrman (2006, p. 65) elaborates on this by defining sector-neutrality as a condition of market-neutrality. If the assets were from different sectors, investor would be exposed to sector-specific risk and a sector-wide market swing could have a major effect on performance.

When entering into long and short positions, a pairs trader is in belief of having identified a temporary mispricing in relative prices of those assets. This belief is backed up by statistical probability of mean reversion. The greatest risk of such speculator is nonconvergence, which happens if trading period ends before convergence occurs, or price generating stochastic process changes in such way that convergence is no longer statistically likely. An extreme example of this is the bankruptcy of the statistically undervalued company. Exposure to nonconvergence can be limited by employing a stop-loss strategy (Shen and Wang 2001).

While market-neutral strategies generally reduce systematic risk, it exposes investors to different kind of risks. Most dominants of these are model risk and execution risk. Model risk can be thought of as a sort of *black box* - the investor relies on trading signals generated by a machine following a protocol he or she might not understand, essentially making blind decisions. Therefore, investor has little to no way of knowing whether the model is faulty or not before they start losing money. Execution risk refers to liquidity concerns, commission restraints, short sale rules and margin ability issues. Trading often means paying commissions often, which reduces the gain potential compared to *buy-and-forget* strategy. Liquidity problems may prevent the trader exiting the trade and realizing the gains altogether. (Ehrman 2006, pp. 39–41).

2.4 Distance approaches

The minimum distance approach is perhaps the best known approach to pairs trading. It is also one of the earliest and the simplest way of selecting pairs and deciding entry points. While modelling of mean reversion relies on statistical stationarity and cointegration tests, minimum distance approach selects pairs by minimizing the sum of squared differences and entry points by comparing the current difference in correlated prices to historical standard deviation of these prices.

The distance method is discussed in detail by Gatev et al. (2006) who open long and short positions when prices diverge more than by two historical standard deviations, and close the position when prices cross. Authors assume that the pairs are cointegrated of order one and divisions from the equilibrium are mean reverting.

Distance method is based on finding the top pairs that minimize the sum of squared differences (SSD). Essentially, finding two series that follow each other as closely as possible. Definition for SSD is given in Equation (1). (Huck 2013; Gatev et al. 2006).

$$SSD_{i,j} = \sum_{t=1}^{T} (P_t^i - P_t^j)^2 \tag{1}$$

In Equation (1), P_t^i and P_t^j are normalized prices for stocks i and j on day t, T being the number of trading days in the formation period. Trading signals are generated when difference in normalized prices reaches a predefined threshold, usually a multiplication of historical standard deviations. Trading happens either on the day of trigger signal or the next possible trading day.

Figure 1 illustrates distance method in practice. Solid red and black lines represent normalized share prices of two pseudorandomly generated assets. At the beginning of the trading window defined by the black rectangle, short position is opened on relatively overvalued red asset and long position is opened on relatively undervalued black asset. Trading window is defined by the spread between the assets. Positions are opened when difference in normalized prices exceeds two historic standard deviations, and is closed when prices converge. Prices converge on day 126, and that is when positions are closed. Short position on red asset is closed for a loss because price of the shorted asset is higher at the end of the trading window than it was at the beginning of it. Long position on black asset is closed for profit, because the price of the black asset is higher at the end of the trading period than it was at the beginning of it. When profits and losses are combined, we see that profit was made as the increase on red assets value (loss) is smaller than the increase on black assets value.

Ignoring trading costs, profit is always made when short position is opened on a relatively overvalued asset and long position is opened on a relatively undervalued asset given that prices eventually converge no matter how small the relative mispricing is. In practice, profits are determined not only by this convergence but also the costs associated with opening and closing the positions and holding an open overnight short position (rent for borrowing the asset). To ensure a better chance of profit, position should be opened only for sufficiently

Figure 1. Distance method

large mispricings.

The length of formation period varies in literature, but Gatev's original length of one year is frequently used, for example in Rad et al. (2016), but Huck (2013) found Gatev's original parametrization to produce less excess returns compared to 6 months or 18 months formation periods. Some could argue that there are more traders in the markets using Gatev's method with standard parametrization, which has competed the excess returns away. In a more recent paper, Huck (2015) notes that the length of trading period is generally set to six months, as in the Gatev's original paper. Although an arbitrary choice, it should allow trades enough time to occur yet keeping the selection relatively fresh.

Stübinger and Bredthauer (2017) apply Gatev's method to high frequency data. In intraday trading, 2.5 times the standard deviation seems to be a better threshold for generating trading signals than Gatev's original two standard deviations. Similar observations were presented by Huck (2013), who found three and four standard deviations to produce more excess returns than two standard deviations.

There is no consensus on the impact of fitting and trading period lengths as well as the optimal opening thresholds to pairs trading returns. In pairs trading literature these are often set to equal those defined in Gatev's original paper (Yan-Xia Lin, McCrae, and Gulati 2006). According to Huck (2013), returns of the distance method are highly sensitive to the length of the formation period. Appropriately selected length of the formation period yields positive excess returns even after compensated for data snooping bias. D. Chen et al. (2017) studied the impact of these parameters in Chinese commodity futures markets and concluded that it is best to set opening threshold to anything between 1.5 and 2.5 standard deviations and the length of training period to anything between 250 and 340 trading days, or 1 to 1.4 years.

2.5 Cointegration based approaches

Cointegration, discussed in detail by Engle and Granger (1987), refers to a situation where a linear combination of nonstationary time series is stationary. That is, series $(X_1, X_2, \ldots, X_n)$ are all integrated of order d, and the linear combination $\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ is integrated of order $d - 1$. Major emphasis is put on the special case where $d = 1$, meaning that the original series are integrated of order one. If such cointegration exists, there is a long run equilibrium between the series and deviations from equilibrium are stationary with finite variance.

In time series context, integration means simple difference between two consecutive values of the series. For series Z, first difference $w_t = Z_t - Z_{t-1}$, second difference $q_t = w_t - w_{t-1}$ and so on. Order of integration refers the ordinal number of the difference. (Roy 1977). Stationarity in time series refers to a process, which is free of trends, shifts and periodicity. It yields series that fluctuate around constant mean with finite, time-invariant variance. Therefore, random shocks will fade away quickly and the series will return to the long-term balance as time passes. (Watsham and Parramore 1997).

Cointegration trading begins by identifying cointegrated assets. Methods vary, but the Engle-Granger Augmented Dickey-Fuller test is common. The optimal hedge ratio discussed later in this chapter can be directly extracted from the first part of EG-ADF test. Regression model

describes the fair value of one asset relative to the other. It establishes an equilibrium level around which true market value fluctuates. (D. Chen et al. 2017; Tourin and Yan 2013).

Testing for cointegration begins by examining the order of integration in individual time series. If they are all integrated of the same order, there might be a cointegrating factor that makes the linear combination of these series integrated of order less than the individual series. In practical terms, cointegration method is based on formulas that imply that all deviations from the theoretical equilibrium level between the prices of two assets will in general revert back to this equilibrium level as the time passes. (Engle and Granger 1987).

Testing for stationarity is often based on augmented Dickey-Fuller test (ADF) proposed in Dickey and Fuller (1979). The null hypothesis of ADF is that the unit root is present in a time series sample, meaning that the sample is nonstationary and integrated of order one. The alternative hypothesis varies by case, and can either be stationarity, trend-stationarity or explosive, the fist two of these being more common than the last one.

Plotting the correlation coefficients of autocorrelation function (ACF) yields an autocorrelation plot, known as a correlogram. In correlogram, bars decrease quickly for a stationary series. (Kirchgässner, Wolters, and Hassler 2013).

Engle and Granger (1987) propose a simple, two-step method for testing the cointegration. First part of this test consists of running an ordinary least squares (OLS) regression of form

$$Y_t = \beta_0 + \beta_1 X_t + z_t \tag{2}$$

to estimate coefficient β_1 and enable computing the residual series of $z_t = Y_t - \beta_1 X_t$. In second part, the stationarity of these residuals is assessed by ADF. This is known as the Engle-Granger Augmented Dickey-Fuller (EG-ADF) test for cointegration.

For assumed cointegrated regression $y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_p x_{p,t} + u_t$, the Durbin-Watson (DW) test statistics for first order autocorrelation should not significantly differ from zero under the null hypothesis of no cointegration, indicating that $x_{1,t}$ is random walk, $\beta_1 = \cdots = \beta_p = 0$, and $\hat{u}_t$ becomes a random walk process with theoretical first order autocorrelation equal to unity. The process of calculating DW statistic is discussed in detail by Durbin and Watson (1950) and Durbin and Watson (1951). According to Leybourne and McCabe (1994), cointegrating regression Durbin-Watson (CRDW) and Augmented Dickey-Fuller tests (CRADF) both favor the null hypothesis of no cointegration. Thus, they encourage authors to supplement the results from those tests with their alternative approach, which defines cointegration as null hypothesis with an alternative hypothesis of no cointegration.

Vidyamurthy (2004, pp. 75–84) explores cointegration strategies with practical examples. A cointegrated time series can be decomposed to a stationary component and a nonstationary component. The cointegrating vector nullifies the nonstationary components, leaving only the stationary components. For cointegrated time series

$$y_t = n_{y_t} + \epsilon_{y_t} \tag{3}$$
$$z_t = n_{z_t} + \epsilon_{z_t}$$

where n_{y_t} and n_{z_t} are nonstationary random walk components, and ϵ_{y_t} and ϵ_{z_t} are the stationary components, the linear combination $y_t - \gamma z_t$ can be expanded and rearranged as

$$y_t - \gamma z_t = (n_{y_t} - \gamma n_{z_t}) + (\epsilon_{y_t} - \gamma \epsilon_{z_t}) \tag{4}$$

where nonstationary components must be zero for the series to be cointegrated. This entails that $n_{y_t} = \gamma n_{z_t}$, i.e. the trend component of one series must be a scalar multiple of the trend component in the other series.

Cointegration model can be applied directly to log-returns, provided that those are nonstationary. Assuming logarithm of stock returns as random walk is common in literature. Error-correcting representation of stocks A and B is written in Vidyamurthy (2004, p. 80) as

$$\log(p_t^A) - \log(p_{t-1}^A) = \alpha_A \log(p_{t-1}^A) - \gamma \log(p_{t-1}^B) + \epsilon_A \tag{5}$$
$$\log(p_t^B) - \log(p_{t-1}^B) = \alpha_B \log(p_{t-1}^A) - \gamma \log(p_{t-1}^B) + \epsilon_B$$

where $\log(p_{t-1}^A) - \gamma \log(p_{t-1}^B)$ is the long-run equilibrium of the cointegrated series. The model is defined by a cointegration coefficient γ and error correction constants α_A and α_B. The long-run equilibrium is the scaled difference of the logarithm of price. The return of the portfolio described in Equation (5) is determined by the change in spread between the assets, as indicated in Equation (6).

$$[\log(p_{t+i}^A) - \gamma log(p_{t+i}^B)] - [\log(p_t^A) - \gamma \log(p_t^B)] = spread_{t+1} - spread_t \tag{6}$$

Rearranging the terms in Equation (2) and using log-price notation from previous equations, the equilibrium value μ emerges as the intercept of first-stage regression in Engle-Granger cointegration test

$$\log(p_t^A) - \gamma \log(p_t^B) = \mu + \epsilon_t \tag{7}$$

The intercept value can be thought of as a premium paid for holding stock A over an equivalent position of stock B. Such premium could be explained by higher liquidity, higher voting power or the possibility of being a takeover target. (Vidyamurthy 2004, pp. 106–107).

With cointegration, it is possible to generalize the concept of pairs trading to construct portfolios of more than two securities, often referred to as basket trading. Given that $\boldsymbol{x_t} = x_{1t}, \ldots, x_{pt})$ is a multivariate time series of nonstationary cumulative returns of individual assets, in cointegrated portfolio of these p securities, each security is weighted by the corresponding coefficient in the cointegrating vector $\boldsymbol{b}$, the resulting basket $z_t = \boldsymbol{b'x_t}$ is a stationary time series equal to the total value of the basket at time t, provided that $\boldsymbol{x_t}$ follows geometric Brownian motion. In other words, any deviation of a security's price from a linear combination of the prices of other securities is temporary and reverting. If the deviation is significant enough, it can be exploited to generate trading signals. However, the feasibility of basket trading is limited by the possibility of a non-zero beta, exposing the investors to non-diversifiable systematic risk. (Yu and Lu 2017).

Cointegration can be applied to commodity futures markets as well. For example, Hain et al. (2018) examined cointegration based trading strategies on economic substitutes using European energy futures. In theory, it does not matter in which form energy is initially stored as long as it can be converted to a consumable form with reasonable costs. Energy has utility value equal to the amount of work produced when consuming the energy. By the Law of one Price, produced utility can be used to determine equilibrium level in raw energy prices when costs associated with transforming the stored energy to work are factored in. Temporary deviations from this equilibrium level can be traded for profit. For example, if oil is too cheap relative to coal, profit can be made by going long on oil futures and short on coal futures.

Clegg and Krauss (2018) applied partial cointegration (PCI) model to S&P 500 constituents. PCI is a weakened form of cointegration, allowing the residual series to have both mean-reverting and random walk components. Law, Li, and Yu (2018) propose an alternative, single-stage fuzzy approach to cointegration-based pairs trading as opposed to conventional two-stage binary approach.

Cointegration of prices conflicts with random walk hypothesis, because cointegration assumes that asset specific, or *idiosyncratic* price shocks are of transient nature but random walk hypothesis states that all price shocks are permanent. In cointegration setting, prices of assets should be driven by common factors like the overall demand for produced goods. There are some evidence on small and short-lived transient shocks, which presents a perfect setting for pairs trading due to quick convergence. (Farago and Hjalmarsson 2019).

2.6 Copula method

The copula method is based on utilizing bivariate copulas to generate trading signals when highly correlated securities diverge. Essentially, a copula is a multivariate probability distribution of a continuous and strictly decreasing convex generator function ϕ from I to $[0, \infty)$ such that $\phi(1) = 0$. It describes the relationship between two variables with uniform probability distributions. (Stander, Marais, and Botha 2013).

According to Sklar's theorem, any multivariate distribution function F can be written in terms of its marginals using a copula representation in Equation (8). (Ané and Kharoubi 2003).

$$F(x_1, x_2, \ldots, x_n) = C[F_1(x_1), F_2(x_2), \ldots, F_n(x_n)] \tag{8}$$

where F_i is an arbitrary marginal distribution function defined as

$$F_i(x_i) = P(X_i \leq x_i) for 1 \leq i \leq n \tag{9}$$

and

$$C(u_1, u_2, \ldots, u_n) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_n \leq u_n). \tag{10}$$

Assuming that the marginal distributions F_i are continuous, then F has an unique copula C, which is defined by the cumulative distribution functions $f_i(x_i)$ of marginals F_i when the copula C and marginals F_i are differentiable. The joint density $f(x_i, x_2, \ldots, x_n)$ is of form

$$f(x_i, x_2, \ldots, x_n) = f_i(x_1) \times f_2(x_2) \times \cdots \times f_n(x_n) \times c[F_i(x_1), F_2(x_2), \ldots, F_n(x_n)] \tag{11}$$

and

$$c(u_1, u_2, \ldots, u_n) = \frac{\partial^n (u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \ldots \partial u_n}, \tag{12}$$

which states that the joint density can be evaluated as a product of the marginal densities and the copula density. It is apparent, that the copula density $C(u_i, u_2, \ldots, u_n)$ contains information about the dependence structure of the X_is while the f_is describe the marginal behaviors. (Ané and Kharoubi 2003)

Of all the copulas, bivariate Archimedean copulas are the most relevant in finance. (Stander et al. 2013). According to Nelsen (2006, p. 110), the Archimedean copula function C for generator function ϕ is given as $C(u, v) = \phi^{[-1]}(\phi(u) + \phi(v))$, where $\phi^{[-1]}$ is the pseudo-inverse of ϕ, defined as

$$\phi^{[-1]}(t) = \begin{cases} \phi^{(-1)}(t) & , 0 \leq t \leq \phi(0) \\ 0 & , \phi(0) \leq t \leq \infty. \end{cases}$$

Some commonly used generator functions for Archimedean copulas include

Gumbel:	$\phi(t) = (-\ln t)^\alpha$	$, \alpha \in [-1, \infty)$
N14:	$\phi(t) = (t^{-1/\alpha} - 1)^\alpha$	$, \alpha \in [1, \infty)$
Clayton:	$\phi(t) = 1/\alpha(t^{-\alpha} - 1)$	$, \alpha \in [-1, \infty)/\{0\}$
Joe:	$\phi(t) = -\ln(1 - (1 - t)^\alpha)$	$, \alpha \in [-1, \infty).$

The true form of the generator function is usually unknown, and must be estimated. Estimation procedures are discussed in Genest and Rivest (1993). Their proposed solution is based on decomposing Kendal's tau, trying copula functions from different families and relying on χ^2 goodness-of-fit statistics. An alternative, graphical estimation procedure based on Deheuvels' empirical copula is proposed in Kharoubi-Rakotomalala and Maurer (2013).

Other relevant copulas include elliptical copulas, most notably the Gaussian copula and the Student t-copula. The Gaussian copula relates closely to the Pearson correlation, and as such represents the dependence structure of two normal marginal distributions. (C.-W. Huang, Hsu, and Chiou 2015).

Nelsen (2006) defines Gaussian copula as

$$C(x,y) = \int_{-\infty}^{\Phi^{-1}(x)} ds \int_{-\infty}^{\Phi^{-1}(y)} dt \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right\} = \Phi_\rho(\Phi^{-1}(x), \Phi^{-1}(y)) \tag{13}$$

where Φ is the univariate standard normal distribution function and Φ_ρ denotes the joint distribution function of the bivariate standard normal distribution with correlation coefficient $-1 \leq \rho \leq 1$.

According to Huang et al. (2015), the Student t-copula captures the tail dependence much better than the Gaussian copula when $\rho \neq 1$. It is defined as a differential equation using multivariate t-distribution in Equation (14).

$$C^t_{v,\rho}(x,y) = \int_{-\infty}^{t_v^{-1}(x)} \int_{-\infty}^{t_v^{-1}(y)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left\{1 + \frac{s^2 - 2\rho st + t^2}{v(1-\rho^2)}\right\}^{-\frac{v+2}{2}} dsdt \tag{14}$$

where $t_v : \mathbb{R} \to \mathbb{R}^+$ is the Student t-distribution function, t_v^{-1} is the inverse of t_v and $t_{\rho,v}$ is the bivariate t-distribution with parameters $\rho \in [-1, 1]$ and $v \in \mathbb{R}^+$. For untransformed series, parameter ρ is the linear correlation coefficient between the two series. For lognormal returns which are obtained by applying a nonlinear transformation, linear correlation between the series is not preserved and ρ becomes less than the linear correlation coefficient. (Krauss and Stübinger 2017). In such situation, rank correlation coefficient such as Kendall's tau is more useful in describing the dependence between the series. (Kendall 1938).

Joe (1996) and Joe and Hu (1996) define three lesser known families of bivariate copulas. Nikoloulopoulos, Joe, and Li (2012) call them BB1, BB4 and BB7. These families are similar to t-copulas but they introduce asymmetries to tail dependence.

In copula based trading, conditional probability functions are used to determine over- and under-valuation. The conditional probability functions $P(U \leq u|V = v)$ and $P(V \leq v|U = u)$ are defined as partial derivatives of the copula with respect to u and v in Equations (15) and (16). Stocks are identified as relatively undervalued when the conditional probability is less than 0.5, and overvalued when it is greater than 0.5. (Aas et al. 2009; Liew and Wu 2013). Positions should be taken when one of the values is close to 1, as the magnitude above 0.5 can be interpreted as how likely the stock is overvalued relative to the other. In general, positions are opened when (u, v) falls outside both confidence bands derived by

$P(U \leq u|V = v) = 0.05$ and $P(V \leq v|U = u) = 0.95$ or vice versa. The position should be closed when the conditional probability decreases back to 0.5. (Ferreira 2008; Krauss and Stübinger 2017).

$$P(U \leq u|V = v) = \frac{\partial C(u, v)}{\partial v} \tag{15}$$

$$P(V \leq v|U = u) = \frac{\partial C(u, v)}{\partial u} \tag{16}$$

The partial derivative of t-copula in Equation (14) with respect to y is given in Equation (17).

$$\frac{\partial}{\partial y} C_{\rho,v}(x, y) = t_{v+1}\left(\frac{t_v^{-1}(x) - \rho t_v^{-1}(y)}{\sqrt{1-\rho^2}} \sqrt{\frac{v+1}{v + t_v^{-1}(y)^2}}\right) \tag{17}$$

The rational behind copulas in finance is that securities' empirical returns are not Gaussian, unlike classical financial theories assume. In a non-Gaussian universe, where skewness and/or kurtosis of returns exceeds the limitations set by the normal distribution, copulas are the simplest way of modelling multivariate probability distributions. As an additional benefit, the dependence structure conveyed by a copula function is preserved under non-linear strictly increasing transformations, such as logarithmic transformation of return series. (Kharoubi-Rakotomalala and Maurer 2013).

Ané and Kharoubi (2003) notes that tail dependence plays an important role in modelling stock returns, and it is often overlooked by other methods. The issue with tails is that most methods assume thin tails and therefore tend to underestimate the impact of extreme values (Haug and Taleb 2011). Xie et al. (2016) demonstrate that although quite similar to Gaussian distribution, Student's t distribution as a marginal and joint distribution better captures the tail dependence of returns due to commonly having fatter tails than the Gaussian distribution.

Figure 2 displays the contour plots of the most common copula types under standard normal marginals. It illustrates the elliptical nature of Gaussian and Student's t-copula, as well as the asymmetrical nature of Clayton, Gumbel, Joe and BB-copulas. Figure 3 displays the density plots of the same copulas and gives perhaps a little better illustration of independence copula and the difference between Frank and independence copula. Asymmetric copulas can be rotated to obtain a better fit in some situations. BB-copulas introduced by Joe and Hu (1996) are modifications of Joe-copula and appear thus seemingly similar.

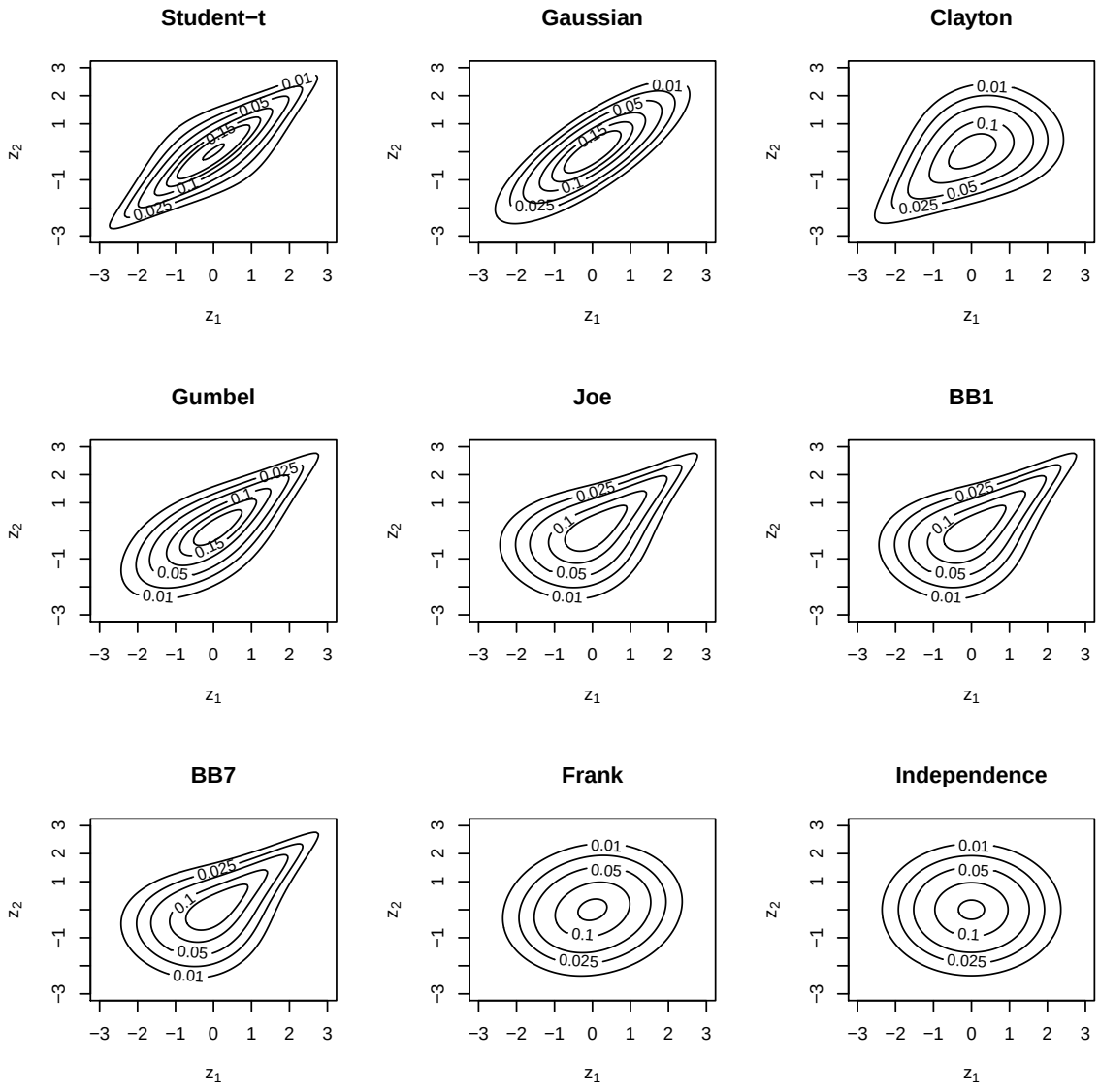
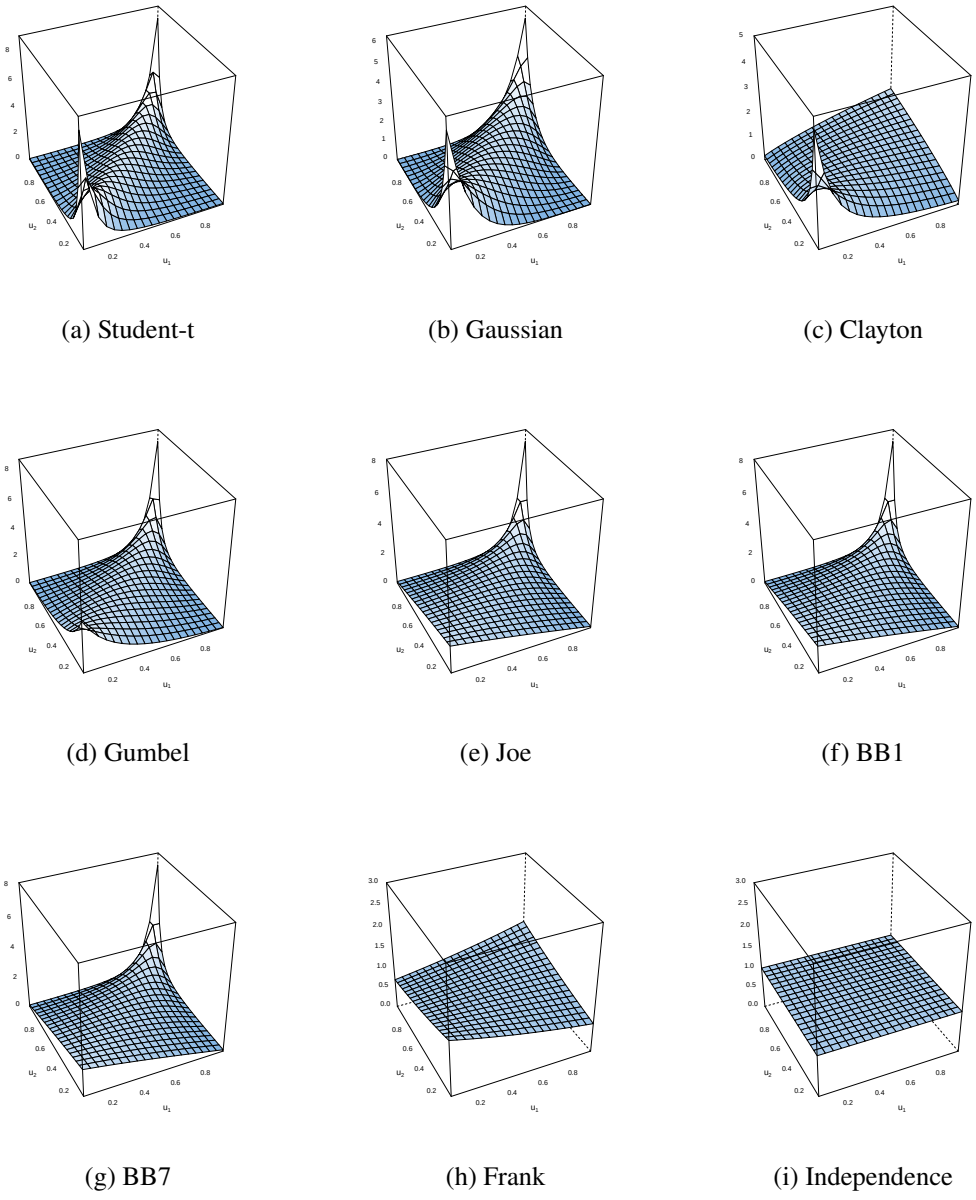


Figure 2. Contour plots of different copula types



(a) Student-t (b) Gaussian (c) Clayton

(d) Gumbel (e) Joe (f) BB1

(g) BB7 (h) Frank (i) Independence

Figure 3. Density plots of different copula types

Krauss and Stübinger (2017) tested the goodness-of-fit of five Archimedean copulas, two elliptical copulas and four extrema value copulas on DAX 30 constituents (435 pairs) and found out that in 71,26% of cases t-copula ranks first and Gaussian copula is clearly the second best choice winning 9,20% of cases making the elliptical family superior to other copula families. In non-elliptical copulas, there is no clear winner that is superior to other non-elliptical copulas most of the time. A non-elliptical copula might be the perfect choice for an individual pair, but one can safely assume t-copula to be the best choice in most of pairs and perform rather well in remaining pairs.

Table 2. Selected copulas in Krauss and Stübinger (2017). The column *Average* denotes the average rank a copula achieves, ranging from 1 to 11. The column *Winner* denotes the empirical probability for each copula to achieve the first rank.

Copula	**Average**	**Winner**
Archimedean copulas		
Ali-Mikhail-Haq	6.55	4.60%
Clayton	6.40	0.69%
Frank	4.26	4.37%
Gumbel	7.50	0.92%
Joe	10.51	0.00%
Elliptical copulas		
Gaussian	2.63	9.20%
Student's t	**1.23**	**71.26%**
Extreme value copulas		
Galambos	6.66	2.35%
Hüsler-Reiss	8.85	0.00%
Tawn	4.06	5.52%
t-EV	7.35	1.15%

Copula strategies can be divided to two groups. The more common, *returns-based* version, such as Liew and Wu (2013), loses the time structure as entry and exit signals are generated based on the last return without assessing how each pair trades subsequent to such signals. The other, *level-based* method, as in Rad et al. (2016), tries to generate some sort of return indices based on accumulated mispricings. (Krauss and Stübinger 2017).

2.7 Other approaches

Multi-criteria decision methods consist of ensembles of different selection criteria and neural network-based selection methods. For example, Huck (2010) used Electre III method to rank S&P 100 stocks by expected returns and form pairs by going long on the highest ranked

shares and shorting the lowest ranked shares. This method does not require estimation of equilibrium levels and is, by construction, dollar neutral.

Triantafyllopoulos and Montana (2011) extended state-space framework for modelling spread processes to introduce time-dependency in the model parameters. Their model was mainly motivated by exploiting temporary market inefficiencies through high-frequency trading.

Montana and Parrella (2009) used data stream analysis techniques to generate an artificial asset, which would be paired against a real, tradable asset. They paired the tradable asset against an artificial proxy composed of prices of other assets, market indices etc. that possess some explanatory power in relation to the real asset. By regarding the artificial asset as the *fair* price of the real asset, one could exploit the short-term divergences of the asset price from the computational, fair value of the asset. In commodity futures based approach, Göncü and Akyildirim (2016) assumed an *Ornstein–Uhlenbeck Lévy* process for the spread and gained relatively good results by trading crude oil and gasoline futures.

Experimental approaches with various success rates include Bayesian Neural Networks (Ruxanda and Opincariu 2018), ARMA based linear state space models with the Kalman filter (de Moura, Pizzinga, and Zubelli 2016) and quasi-variational inequalities (Song and Zhang 2013). All of these have been proven to be profitable within a single time frame at a specific marketplace, but the literature is rather limited and no generalizations on their profitability can be made.

3 Methodology and Data

This chapter describes data and introduces different statistical methods used in this thesis. It outlines and justifies the limitations set for pair selection and discusses how these methods were implemented in selected statistical software.

3.1 Data

Historical Finnish stock prices were fetched from Nasdaq Nordic database. The data consists of time series of all currently traded Finnish companies' stock prices from 2004 to mid 2020. In September 2019, there were 143 shares listed on the main list of OMX Helsinki. The total number of possible pairs at that time point can easily be calculated as 2-combination of 143 assets.

$$\binom{143}{2} = \frac{143!}{2!(143-2)!} = 10\,153$$

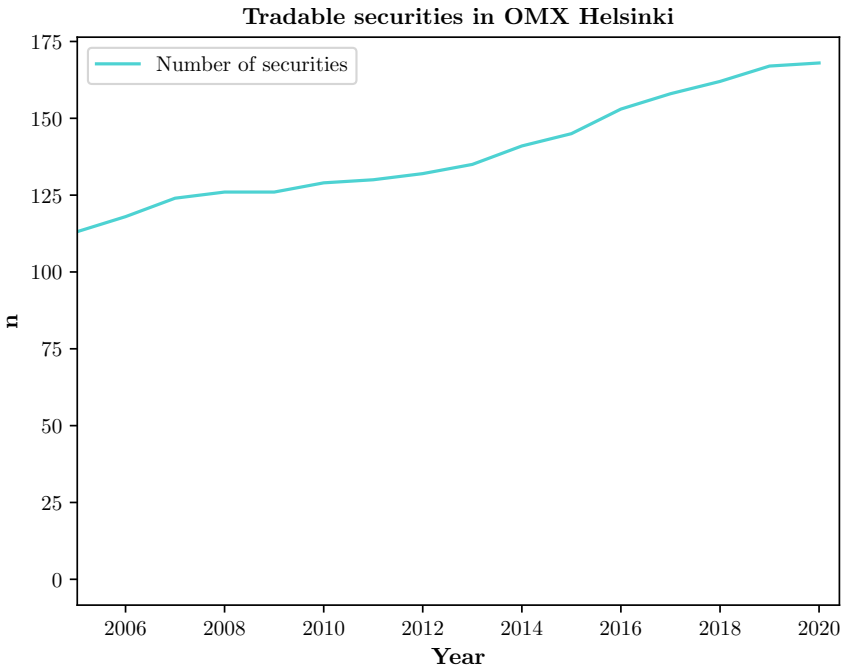


Figure 4. Number of tradable securities by year

Of all currently listed stocks, 78 were listed before year 2000, 30 were listed between 2000 and 2010 and remaining 35 were listed in 2010 or later. Thus, the true value of available pairs varies over time throughout the sample period. For each period, there is an ample pool of possible pairs from which to select the best 20 pairs.

This thesis aggregates results from partially overlapping trading windows. Each window consists of one year fitting periods followed by a 6-month trading period. Using a 3-month interval, there are 66 of these windows. Rolling windows are illustrated in Figure 5.

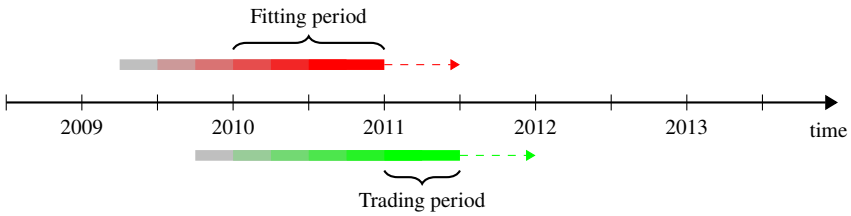


Figure 5. Overlapping training periods

Figure 6 displays overall market performance from 2004 to 2020. Market returns for each period are shown in Figure 7, which reveals that most trading periods provided medium to low returns, and some periods significant losses. Of all 66 periods, 68% provided profits. Unannualized mean return per period was 2,44%. Largest loss was −64,88% and biggest gain was 36,46 %. Several institutions provide OMXH 25 based index funds, so this index will be used as a market benchmark for *buy and hold* strategy.

To overcome survivor bias, a list of companies removed from the main list was extracted from a blog post by Osakekeisari (2018). This list is presented in Appendix A2.1. It mostly contains companies that were acquired by some other company or merged with another company. It also contains some companies that went bankrupt during the period. Past data was still available at Nasdaq Nordic Database for some of these companies. Those are listed in Table 3.

Of all companies traded during the observation period, five were identified to have been declared bankrupt. These are listed in Table 4. For marketing communications agency Evia and paperboard manufacturer Stromsdal data was no longer available.

Daily closing prices adjusted for dividends and splits were used in the analysis. All prices are nominated in Euros. Stock data is combined with Industry Classification Benchmark (ICB) table to divide instruments to different bins based on their industry. This classification was extracted from Nasdaq's list of companies listed on Nasdaq Helsinki. Full list of used

Figure 6. OMX Helsinki 25

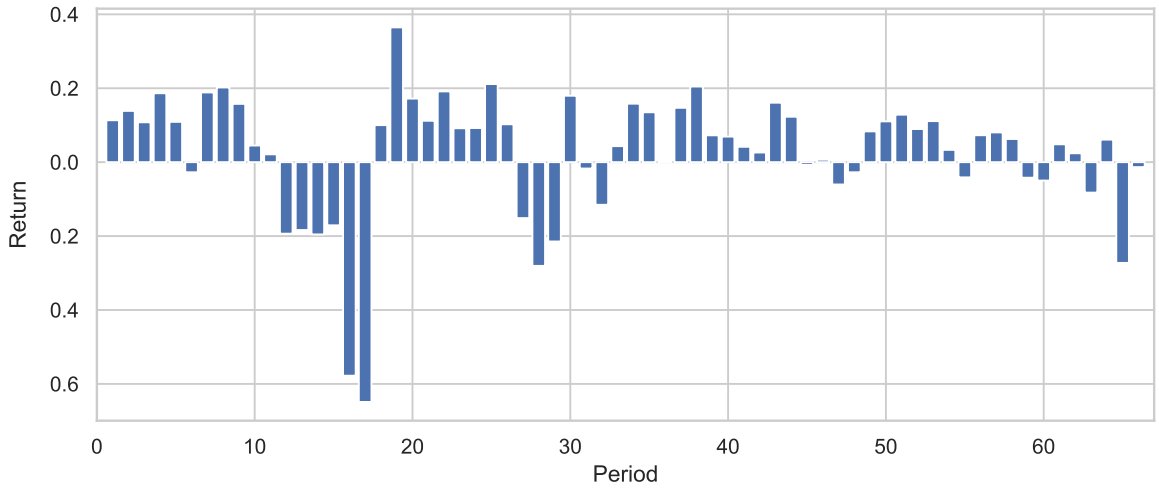


Figure 7. OMX Helsinki 25 returns per period

Table 3. List of removed companies for which data was available

Company	Date	Symbol	Sector	Reason
Ahtium Oyj	2018-03-15	AHTIUM	Industrials	bankruptcy
Affecto Oyj	2018-02-21	AFFECTO	Technology	acquisition
Lemminkäinen Oyj	2018-01-31	LEM1S	Industrials	merger
PKC Group	2017-07-09	PKC1V	Industrials	acquisition
Comptel	2017-06-29	CTL1V	Technology	acquisition
Norvestia	2017-06-09	NORVE	Financials	merger
Okmetic	2016-08-11	OKM1V	Industrials	acquisition
Biotie Therapies	2016-09-30	BTH1V	Health Care	acquisition
Turvatiimi	2015-04-09	TUT1V	Consumer Services	acquisition
Vacon	2015-05-18	VAC1V	Industrials	acquisition
Oral Hammaslääkärit	2014-12-19	ORA1V	Health Care	acquisition
Tiimari	2013-10-10	TII1V	Consumer Goods	bankruptcy
Nordic Aluminium	2012-12-15	NOA1V	Industrials	acquisition
Aldata Solution	2012-08-08	ALD1V	Technology	acquisition
Elcoteq SE	2011-11-17	ELQAV	Industrials	bankruptcy
Salcomp	2011-09-23	SAL1V	Technology	acquisition
Pohjola	2006-06-14	POH1S	Financials	acquisition

Table 4. List of bankrupt companies

Bankrupt date	Company	Data available
2018-03-15	Ahtium Oyj	True
2013-10-10	Tiimari	True
2011-11-17	Elcoteq SE	True
2009-02-07	Evia	False
2008-11-12	Stromsdal	False

companies, their ticker symbols and main business sectors per ICB classification is found in Appendix A1.1.

Cointegration is often restricted to allow only pairs composed of stocks belonging to the same GICS sector, to improve computational feasibility. Clegg and Krauss (2018) estimate that even after this sector restriction, it would take approximately 15 days to process all possible pairs in S&P 500 using parallel processing on an Intel i7-4790K with 8 threads and clock speed of 4 GHz. However, required computational resources decrease sharply when the universe of possible shares shrinks, as the number of possible combinations is a combinatorially increasing function of the batch size.

Although not necessary for computational reasons, similar restriction is placed here, as employed in Gatev's original paper. This limitation was motivated by the assumption that firms operating under the same sector share industry risk as well as market risk and it was also applied by Figuerola-Ferretti, Paraskevopoulos, and Tang (2018) on their research about cointegration in STOXX Europe 600 constituents. After imposing this limitation, the number of possible pairs in the OMX Helsinki decreases from 10 153 to 1 600 (Table 5). This allows to examine different time frames and aggregate results from multiple periods to obtain a more robust estimate of model performance.

Industrials is the largest sector, with 43 different securities. Besides Utilities and Oil & Gas, all sectors are large enough for intrasector trading. Neste and Fortum will be excluded from the study for being the only companies in those to sectors.

Table 5. Distribution of companies by sector

Sector	**Count**	**Combinations**
Industrials	43	903
Financials	19	171
Technology	18	153
Consumer Goods	17	136
Consumer Services	16	120
Basic Materials	13	78
Health Care	9	36
Telecommunications	3	3
Utilities	1	0
Oil & Gas	1	0
Total	140	1600

Naïve extrapolation will be used to account for missing values - for those days that did not see a trade, the price is assumed to be unchanged. Similar assumption is made in Mikkelsen (2018).

3.2 Methodology

For each training period, only stocks that had data for the entire period were considered. This eliminates shares that entered or exited the main list during the fitting period. However, stocks that exited during the trading period were included in the sample, as this represents future information which cannot be known by the trader when selecting the pairs.

Although Mikkelsen (2018) included all pairs that were found cointegrated at a 10 % significance level, only the best 20 pairs were considered in this thesis for each trading period. This restriction is placed to enforce computational feasibility over multiple trading periods and it should provide enough insight to the profitability of less than optimal pairs.

3.3 Normalization of prices

Because nominal prices vary a lot between different assets, prices must be normalized for methods relaying on distance between the asset prices. As in Rad et al. (2016) and Do and Faff (2010), the normalized price is defined as the cumulative return index.

For small x, such as daily stock returns, Kirchgässner et al. (2013, p. 7) states that

$$\frac{y_t - y_{t-1}}{y_{t-1}} \approx \ln(1 + x) \approx x \tag{18}$$

This holds when changes are at most ±20%. For larger deviations, properties of log-returns hold but interpretation is not as straightforward. For example, $\ln(1 + (-0.7)) \approx -1.2$. However, log-returns aggregate easily over time as the total logarithmic return is always the sum of individual log returns. (Equation 19). For computational purposes this is far better than the multiplicative approach required by simple returns. (Cryer and Chan 2008, p. 99).

$$\sum_{i=1}^{n} R_{\log,i} = R_{\log,1} + R_{\log,2} + R_{\log,3} + \cdots + R_{\log,n} \tag{19}$$

Combining above information, it can be stated that

$$\ln(y_t) - \ln(y_{t-1}) \approx \frac{y_t - y_{t-1}}{y_{t-1}} \sim N(\mu_{daily}, \sigma^2_{daily}) \tag{20}$$

and that logarithmic, or continuously compounded returns can be calculated as a natural logarithm of terminal price divided by the initial price (see Equation 21), log normalization was selected for the basis of normalization.

$$R_{\log} = \ln(\frac{P_t}{P_{t-n}}) \tag{21}$$

3.4 Computation of returns

Returns are computed using the method presented in Gatev et al. (2006) and Clegg and Krauss (2018). In this method, one Euro is allocated at the beginning of the trading period for each pair that opens during the trading period. This is known as the fully invested capital method.

For each pair there are two sources of cash flows. The first source is randomly distributed throughout the trading period, occurring every time the pair closes. The second source is cash flows at the end of the trading period when still open positions are closed because of termination of trading period. The excess return on a pair during a trading interval is the sum of the payoffs during the interval. This is a conservative approach because it ignores the fact that some returns are obtained during the trading period instead of at the end of it. (Gatev et al. 1999).

The formula for calculating the excess returns is defined in Xie et al. (2016) based on Gatev's paper as

$$r_{P,t} = \frac{\Sigma_{i \epsilon P} w_{i,t} r_{i,t}}{\Sigma_{i \epsilon P} w_{i,t}} \tag{22}$$

$$w_{i,t} = w_{i,t-1}(1 + r_{i,t-1}) = (1 + r_{i,1}) \ldots (1 + r_{i,t-1})$$

where r denotes returns and w weights, and daily returns are compounded to obtain monthly returns.

Win ratios are calculated for each method as suggested in D. Chen et al. (2017). Win ratio

describes the percentage of profitable trades. Lower win ratio requires higher profit per trade to match the profitability of better win ratio.

Because multiple overlapping periods are used, returns of the individual periods are averaged to obtain overall performance of each trading method. Because log-returns aggregate over time but not over different assets, log-returns must be converted to simple returns before aggregating them. Annualized simple returns also give a better understanding of profitability to most people when compared to continuously compounded alternative.

3.5 Data snooping bias

Extensive searching of potential trading rules within a data set is prone to spurious relationships. Seemingly profitable strategies often emerge when data is mined long enough, yet only few, if any of them are valid in out-of-the sample predictions. (Kuang, Schröder, and Wang 2014). For example, Yang, Cabrera, and Wang (2010) noted that adjusting for data-snooping bias using White's Reality Check substantially weakens otherwise strong predictability of ETF returns.

One of the best known tests was formulated by White (2000). He presents two approaches, Monte Carlo Reality Check and Bootstrap Reality Check, because analytical approaches are not feasible to obtain a consistent estimate of a p-value for $H_0 : E(f^*) \leq 0$ due to the unknown extreme value of a vector of correlated normals for the general case. Bootstrap method is discussed in Sullivan, Timmermann, and White (1999) based on White's then to-be-published working paper on Reality Check.

Data snooping bias can be tested with StepM test introduced in Romano and Wolf (2005). Stepwise Multiple testing framework extends White's (2000) method, and by design rejects all hypotheses that Bootstrap Reality Check rejects.

Hansen (2005) highlights some flaws in White's approach, and presents an alternative test known as superior predictive ability test (SPA). Hsu, Hsu, and Kuan (2010) argue that Hansen's test is more powerful than White's reality check.

3.6 Measures of profitability

Profitability of the trading methods was evaluated using a similar procedure than Gatev used in his original paper, as it seems to serve as a some sort of baseline in pairs trading literature.

Other methods considered for evaluating the profitability of trading rules include Jensen's alpha, historic Sharpe ratio and Omega ratio. Omega ratio was discarded due to the requirement of taking a stance on required rate of return, but it is discussed here because of similarities in assumptions to copula-based trading rules.

Omega ratio, presented by Keating and Shadwick (2002) was inspired by the empirical observation that joint distribution of returns form individual securities is not normally distributed. Thus, higher moments than simple mean and variance are required for complete description of returns. Note that this observation is similar to the one copula method is based on. Omega ratio acknowledges the information that the Sharpe ratio discards. While Sharpe ratio and its refinements minimize the potential for loss (and also the gains), Omega ratio adjusts to the individual preferences of required return. The ratio in Equation (23) is defined as a probability weighted ratio of gains to losses relative to a desired rate of return r.

$$\Omega(r) = \frac{\int_r^b (1 - F(x))dx}{\int_a^r F(x)dx} \tag{23}$$

where $F(x)$ is empirical probability density function of returns and a and b are the lower and upper boundaries of the returns. At mean μ, Omega ratio makes no difference between portfolios as $\Omega(\mu) = 1$. The Sharpe ratio would select the security with the lowest σ, and the lowest gain potential.

Of the legacy measures, Jensen's alpha, defined in Jensen (1968), is calculated from theoretical return R_i of the portfolio using the estimated β_{iM} of the portfolio and risk free rate R_f and market return R_M as defined in Equation (24).

$$\alpha_J = R_i - (R_f + \beta_{iM} \cdot (R_M - R_f)) \tag{24}$$

Jensen's alpha has been criticized mainly for its sensitiveness to the choice of benchmark model. It yields different results when arbitrage pricing theory is used instead of the capital asset pricing model. (Murthi, Choi, and Desai 1997).

According to Grau-Carles, Doncel, and Sainz (2019), Jensen's alpha and Sharpe ratio have very low Spearman and Kendall correlation, indicating that they favor different types of investments. The Sharpe ratio, Gaussian Value-at-Risk analysis and partial moments based measures such as Sortino Ratio and Omega Ratio yield very similar results.

The historic, or *ex post*, Sharpe ratio, originally published in Sharpe (1966) and revised in Sharpe (1994) is defined as the square root of average differential return $\bar{D}$ of portfolio returns R_{Ft} and benchmark returns R_{Bt} divided by the historic standard deviation σ_D, as presented in Equation (25).

$$S_h = \sqrt{\frac{\bar{D}}{\sigma_D}} \tag{25}$$

where

$$\sigma_D = \sqrt{\frac{\sum_{t=1}^{T}(D_t - \bar{D})^2}{T-1}} \tag{26}$$

$$\bar{D} = \frac{1}{T}\sum_{t=1}^{T} D_t \tag{27}$$

$$D_t = R_{Ft} - R_{Bt} \tag{28}$$

The Sharpe ratio relies on assumption of an elliptical return distribution. It is thus biased in the case of non-normally distributed returns. (Grau-Carles et al. 2019).

3.7 Modelling squared differences

The distance method is perhaps the simplest of all three major methods presented in this thesis, and it is relatively straightforward to implement in Python using *Pandas* and *NumPy*. Pandas is an open source, *panel data* focused, software library offering indexable on-memory data structures similar to data frames in R. With datetime indexing, slicing observation windows and calculating log returns is a breeze. NumPy is another open source library focusing on numerical computing and linear algebra. It provides efficient functions for logarithmic transformations in vectors. Both of these libraries have mappings to lower level C functions and can utilize multiple processor cores to increase processing speed.

The data was sliced to partially overlapping fitting periods of 12 months, each separated by

12 weeks. These periods were then looped over to obtain the ranking of pairs for each period. Ranking is based on minimizing the sum of squared differences in return index. For each training period, best 10 pairs were selected, visualized and their profitability was evaluated during the following trading period.

Trades were performed when the spread between the normalized asset prices exceeded two historic standard deviations to either direction from the mean value of the spread. Positions were closed when the spread minus the mean value of the spread changed in sign.

Pairs formed of different share classes of the same companies are often good choices for distance based trading. *Figure 8* illustrates the structure of return series in pairs selected by minimizing the sum of squared differences. In general these series follow each other very strictly, but temporary deviations occur at random intervals.

Trading strategy is explained here with SSAB Class A and Class B shares. *Figure 9* displays the spread between SSAB Class A and SSAB Class B shares, as well as the mean value and the general trigger levels of two times the spread's standard deviation. The spread fluctuates around its mean. The mean value is slightly positive, indicating that, in general, Class A shares have been a bit more expensive. Although not tested for statistical significance, this would be a reasonable assumption as each Class A share entitles holder to have one vote in general meeting, but each Class B share gives holder only $1/10^{th}$ of a vote (SSAB 2020).

Table 6. Summary statistics of SSABAH and SSABBH returns

	SSABAH	**SSABBH**	**spread**
count	262	262	262
mean	1.099	1.084	0.015
std	0.084	0.081	0.014
min	0.914	0.924	-0.029
25%	1.037	1.012	0.009
50%	1.111	1.092	0.016
75%	1.163	1.144	0.024
max	1.285	1.272	0.051

Figure 8. Illustration of typical distance pairs formed of different share classes of one company.

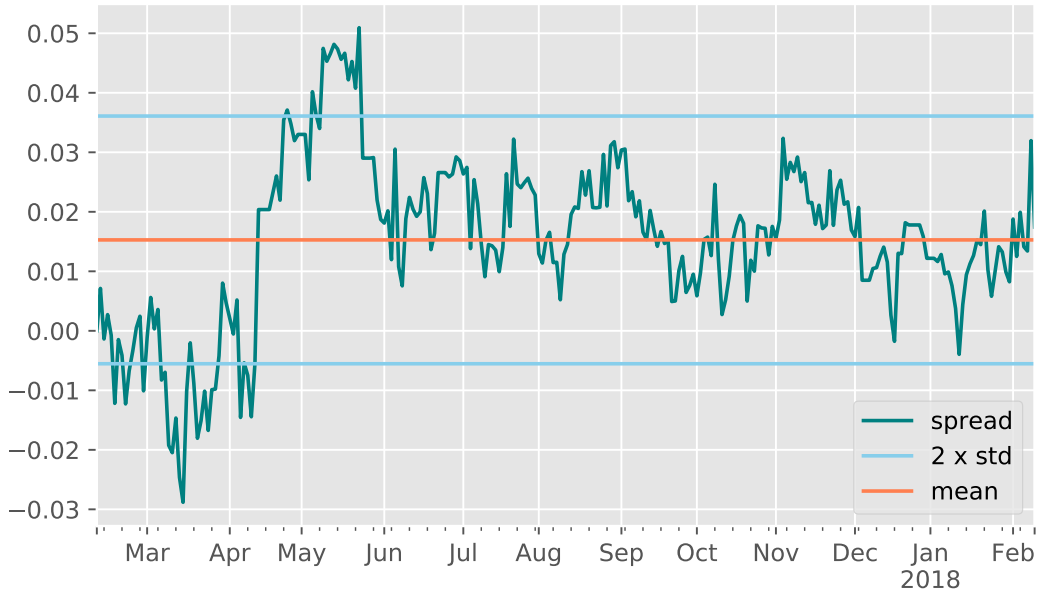


Figure 9. Spread of SSABAH and SSABBH with mean and opening thresholds at two standard deviations.

3.8 Modelling cointegration

The cointegration strategy was implemented as presented in D. Chen et al. (2017). This is consistent with Figuerola-Ferretti et al. (2018) and roughly follows the method defined in Vidyamurthy (2004, pp. 82–84).

1. Identify pairs that could be cointegrated. This can be based on the stock fundamentals or statistical approach from historical data. Vidyamurthy suggests using fundamental information for pair selection. This thesis uses fundamental selection criterion as described in Gatev et al. (1999).

2. Verify the proposed hypotheses of cointegration by applying statistical tests to historical data. Determine the cointegrating coefficient and examine the spread series to ensure it is stationary and mean-reverting.

3. Examine cointegrated pairs to determine suitable level for delta and start trading.

A cointegrated pair A, B satisfies the condition $A_t - \beta B_t = u_t$ where u_t is stationary. This can be rearranged to say that the fair price of A at time t is $\beta B_t + u_t$. The following example if purely fictional, uses nominal prices for the ease of understanding and is intended for illustrational purposes only.

Suppose that we have estimated from historic data that

Cointegration Ratio β	= 1.5
Premium u_t on asset A	= 0 €
Trading threshold	±0.8 €

And we know that

Price of A at time t	= 20.30 €
Price of B at time t	= 14.10 €

The computed fair price for A at time t is 1.5×14.10 € + 0€ = 21.15 €. Because this is higher than the actual market price of A at time t, it can be said that A is *undervalued* relative to B, or B is *overvalued* relative to A. The magnitude of this deviation $20.30 - 21.15 = -0.85$ is compared to the trading threshold, which is set at two times the standard deviation of past deviations. Since $|-0.85| > 0.80$, trader buys A and shorts B in cointegrating ratio of 1 to 1.5.

Now suppose that prices develop so that at time $t + 1$

Price of A at time $t + 1$	= 20.79 €	(+2.4%)
Price of B at time $t + 1$	= 13.86 €	(-1.7%)

The positions are then closed, because the prices have converged to the equilibrium level where 1.5×13.86 € = 20.79 €.

The total profit from the trade is the profit on A plus β times the profit on B:

$$= (20.79 \text{ €} - 20.30 \text{ €}) + 1.5 \times (14.10 \text{ €} - 13.86 \text{ €})$$
$$\approx 0.85 \text{ €}$$

And the return, ignoring all costs and margin requirements, is

$$= \ln(\tfrac{20.79 \text{ €}}{20.30 \text{ €}}) + 1.5 \times \ln(\tfrac{14.10 \text{ €}}{13.86 \text{ €}})$$
$$\approx 0.0239 + 1.5 \times 0.0172$$
$$\approx 4.1\%$$

For cointegration tests, Engle-Granger Augmented Dickey-Fuller test was implemented in Python using *statsmodels* package. To obtain the optimal hedge ratios and residual series, simple OLS regression was fitted first. Because *statsmodels* does not include constant by default, a user must remember to add a vector of ones to the matrix of exogenous variables. If one does not remember to do so, the residuals of the OLS regression will be biased and therefore ADF test might yield incorrect results. OLS regression was followed by ADF test on residual series, using AIC for determining the optimum number of lags. Tests were run for each of the qualifying pairs present during the fitting period.

After running the cointegration tests, several cointegrated pairs emerged with a negative cointegrating coefficient. Such pairs are diverging and violate dollar-neutrality, as those would imply long-long or short-short portfolios. The test were run again this time considering only pairs with positive cointegrating relationship.

The best 10 pairs for each period were selected by approximating MacKinnon's p-values for t-values from ADF regressions. This procedure is discussed in details in MacKinnon (1994). MacKinnon aims to solve the problem of nonstandard asymptomatic distributions in ADF tests for which only few critical values have been tabulated, mainly by Fuller in his book *Introduction to Statistical Time Series* (1976). MacKinnon's p-values are drawn from his approximations of the asymptotic distribution functions for these tests.

After ranking each pair, trades were performed on the best 10 pairs of each time frame. In this phase, the cointegrating parameter of the model was used to calculate the fair value of the first asset based on the current value of the second asset by assuming a long-run equilibrium to exist between the assets. The current value of the first asset was then compared to the computed fair value for the asset to obtain the relative value of the first asset. Positions were opened when the current value differed enough from the computed value and closed when the relative value changed from overpriced to underpriced or vice versa. For opening threshold, two standard deviations were used based on recommendations in D. Chen et al. (2017). Trades were closed at convergence.

Cointegration strategy is illustrated here with the most cointegrated pair from an arbitrarily chosen trading period, in this case period 56 where the fitting range was from from August 2016 to August 2017 and the trading period from August 2017 to February 2018. The most cointegrated pair for that range based on MacKinnon's statistic was Oriola Oyj A - Oriola Oyj B. Log-prices for that pair are presented in Figure 10.

The price of both share classes shows significant decline during the fitting period, and this decline steepens during the trading period. Cointegrating linear relationship between the two

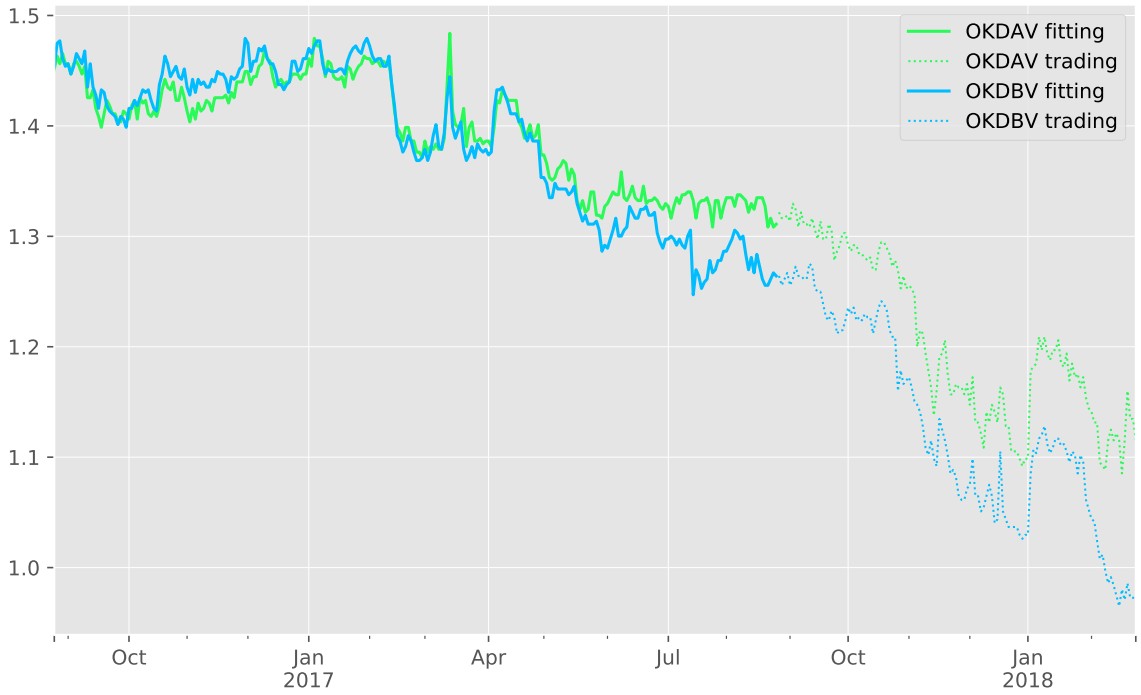


Figure 10. The most cointegrated pair for trading period 56

share classes is displayed in Figure 11. Class A shares trade at a small premium, which is most likely explained by higher voting power at general meetings (Oriola 2020).

Figure 12 illustrates cointegration based trading strategy on that pair. Positions are opened at vertical solid green lines, and closed at the red lines following these green lines. Dotted line represents computed equilibrium price of Class A shares. The direction of opening trade is illustrated with a green triangle pointing to desired direction of price movement i.e. down for short position and up for long position. Red $\times$ marks the price at which positions are closed at convergence. Profit is made when either both prices move in desired direction or one of the prices moves to desired direction more than the other does in undesired direction.

Detailed description of each trade in Figure 12 is presented in Table 7. Asset prices are represented as natural logarithms, and can be converted to real prices by raising e to the power of log-price if desired. For most though, log-representation in Table 7 is likely more understandable as $(e^{1.31} - e^{1.28})/e^{1.31} = 1.31 - 1.28 \approx 3\%$.

On November 6$^{\text{th}}$ spread changed sign and caused the open positions to be closed due to the convergence and the reverse position to be simultaneously opened due to significant change in opposite direction. Because of declining trend during the trading period, practically all of the profits were made through selling short the overvalued asset.

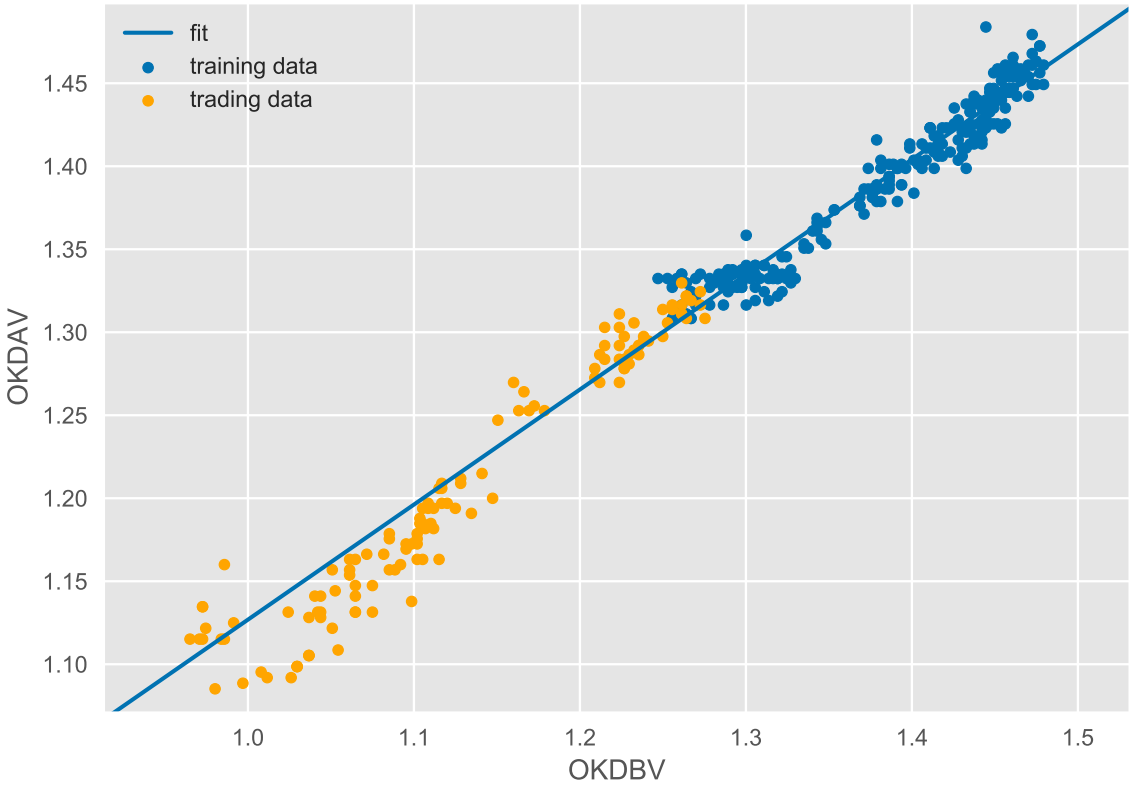


Figure 11. Fit of pair in Figure 10

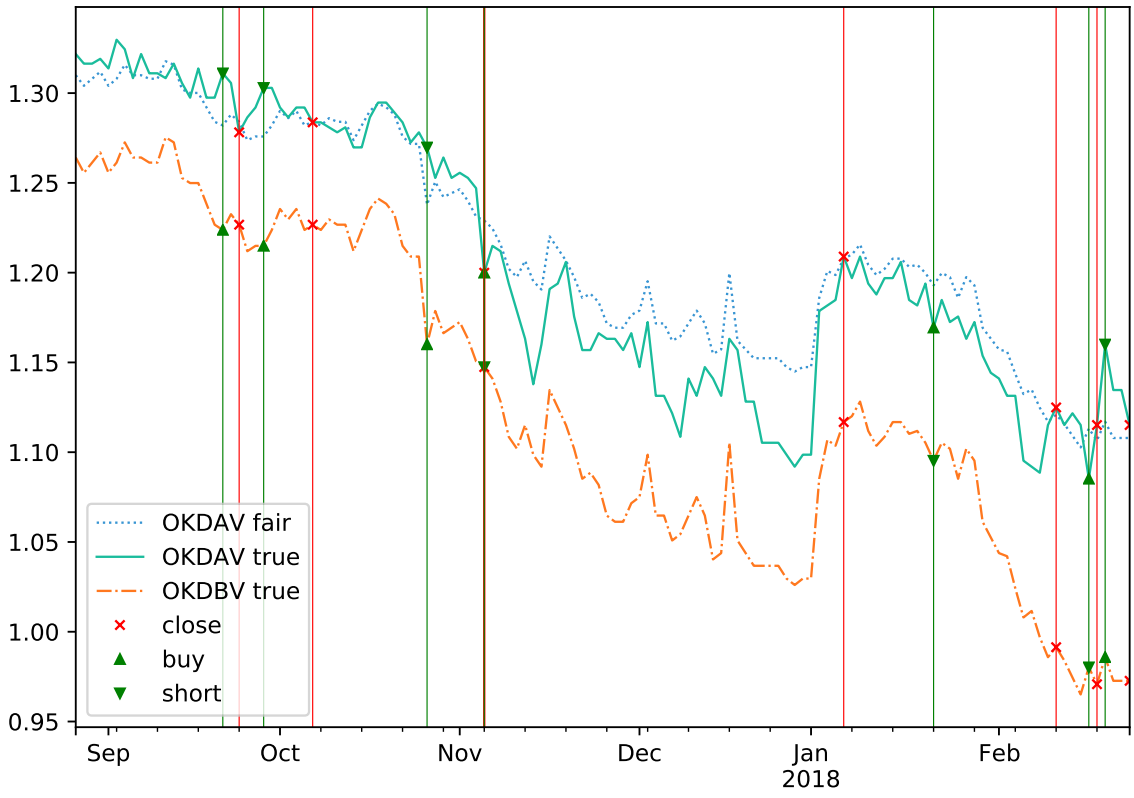


Figure 12. Trades on trading period 56 for pair in Figure 10

Table 7. Cointegration trades with Oriola shares between August 2017 and February 2018, *cointegrating coefficient = 0.69*

DATE	OKDAV	OKDBV	OKDAV_fair	valuation	spread	action	long profit	short profit	total profit
2017-09-21	1.31	1.22	1.28	over	-0.03	open			
2017-09-25	1.28	1.23	1.28	under	0.01	close	0.00	0.03	0.03
2017-09-28	1.30	1.21	1.28	over	-0.03	open			
2017-10-06	1.28	1.23	1.28	under	0.00	close	0.01	0.02	0.03
2017-10-26	1.27	1.16	1.24	over	-0.03	open			
2017-11-06	1.20	1.15	1.23	under	0.03	close + open	-0.01	0.07	0.06
2018-01-05	1.21	1.12	1.21	over	-0.00	close	0.01	0.02	0.03
2018-01-22	1.17	1.10	1.19	under	0.02	open			
2018-02-12	1.12	0.99	1.12	over	-0.00	close	-0.04	0.07	0.03
2018-02-16	1.09	0.98	1.11	under	0.03	open			
2018-02-19	1.12	0.97	1.11	over	-0.01	close	0.03	0.01	0.04
2018-02-20	1.16	0.99	1.12	over	-0.04	open			
2018-02-23	1.12	0.97	1.11	over	-0.01	close	-0.01	0.04	0.04
TOTAL							-0.01	0.26	0.26

3.9 Modelling copulas

Copulas are best illustrated with some examples. By choosing arbitrarily Orion Corporation Class A and Class B shares for this purpose, let us first plot the adjusted closing prices for arbitrary time interval from October 2016 to November 2019 (Figure 13).

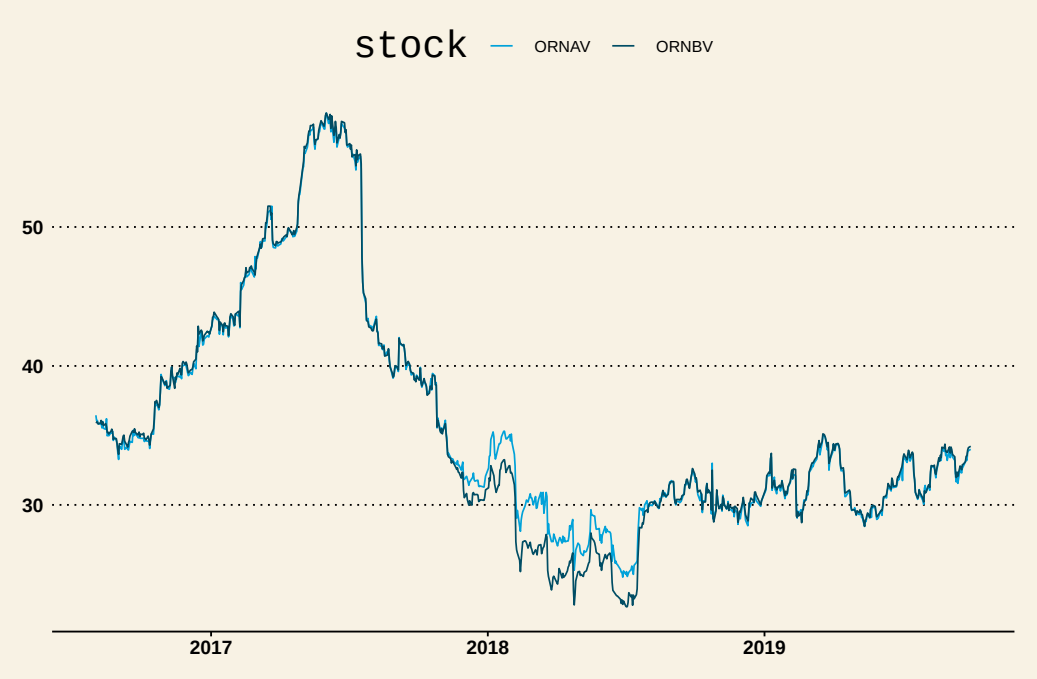


Figure 13. Adjusted closing prices

The prices are highly correlated and have Pearson coefficient of 0.876 and Spearman coefficient of 0.827. Log-returns plotted in Figure 14 are seemingly stationary and random. When plotted against each other, a dependency structure appears as indicated by correlation coefficients. (Figure 15).

In order to fit copulas, returns must first be scaled to interval [0,1]. After that, selection of copula family is rather straightforward. The resulting object indicates that based on Akaike information criteria, Student's t-copula would be the best fit to Orion shares. The density function of the fitted Student's t-copula is plotted in Figure 16. This resulting 3-dimensional plot can be converted to a contour plot, which clearly demonstrates the elliptical nature of Student's t-copula. (Figure 17).

Now this copula can be used to sample more observations with similar statistical properties than the original log-return series. For a better visualization, 10 000 points were sampled in Figure 18 based on the copula.

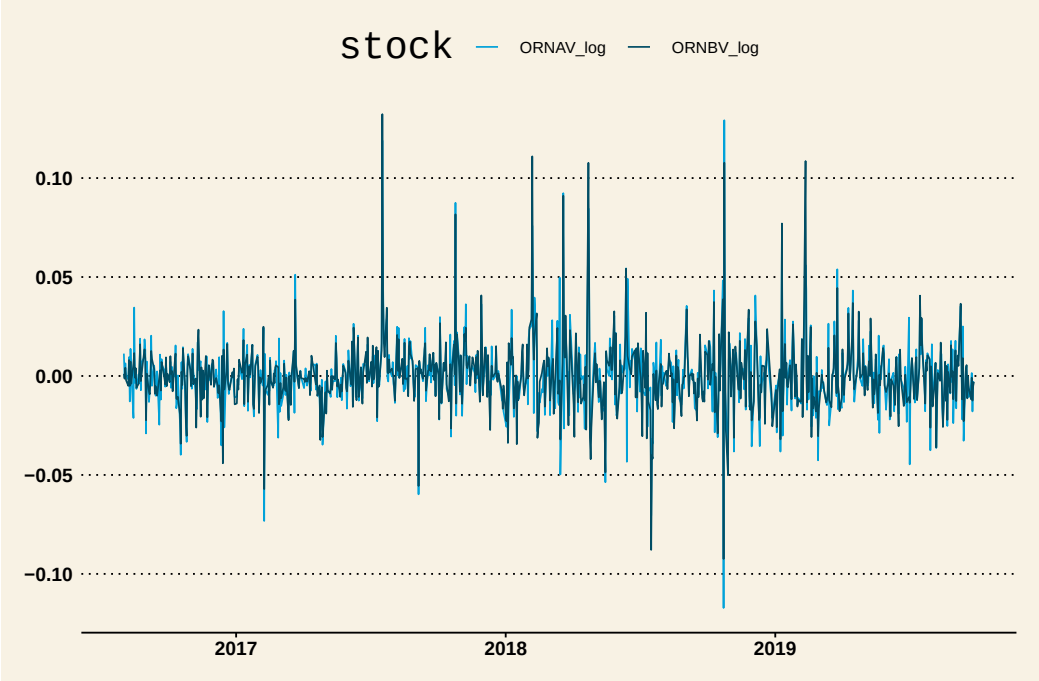


Figure 14. Daily log returns of Orion A and B

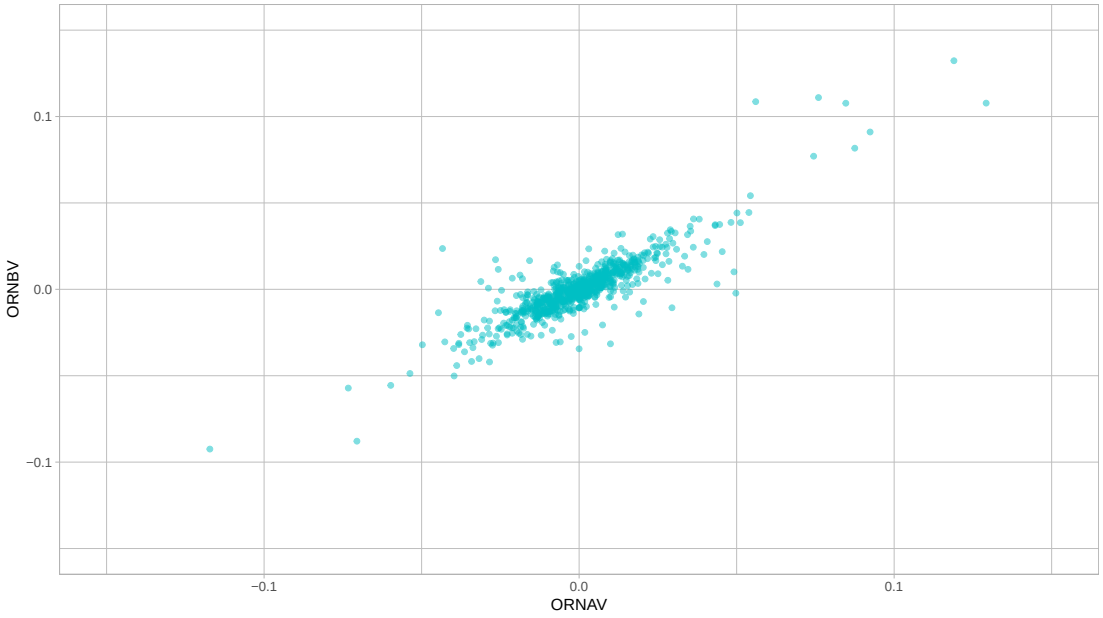


Figure 15. Scatter plot of log returns

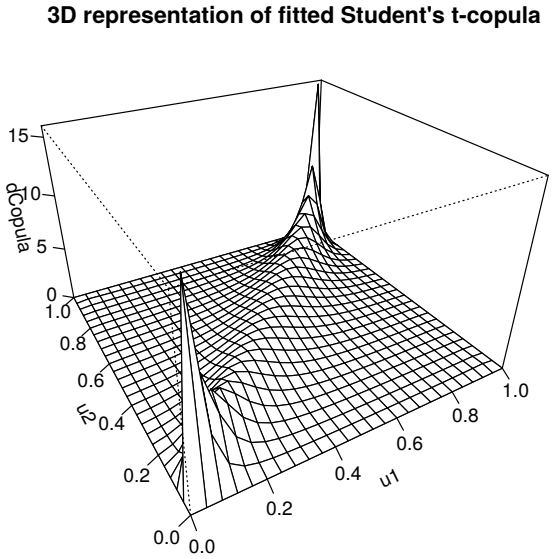


Figure 16. Density plot of the fitted Student's t-copula

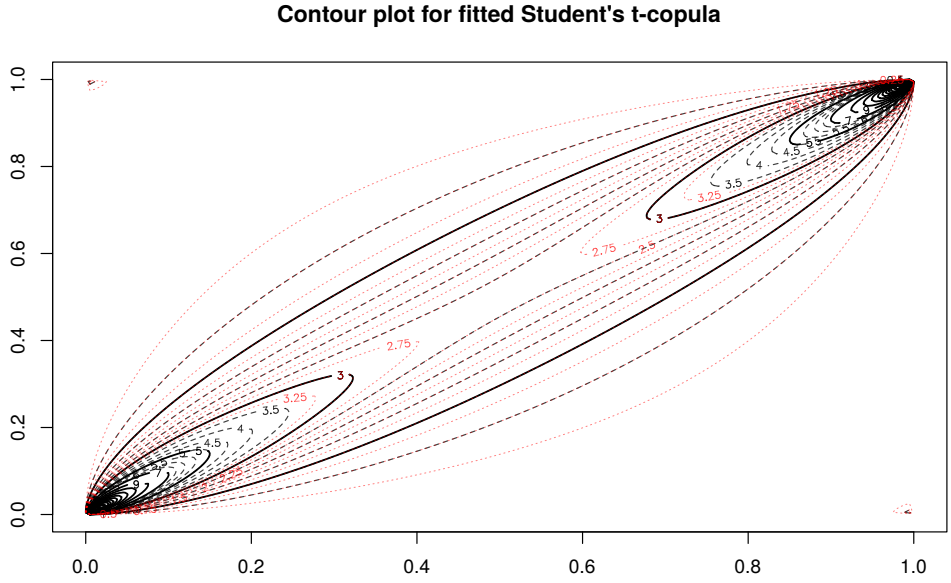


Figure 17. Contour plot of the fitted Student's t-copula

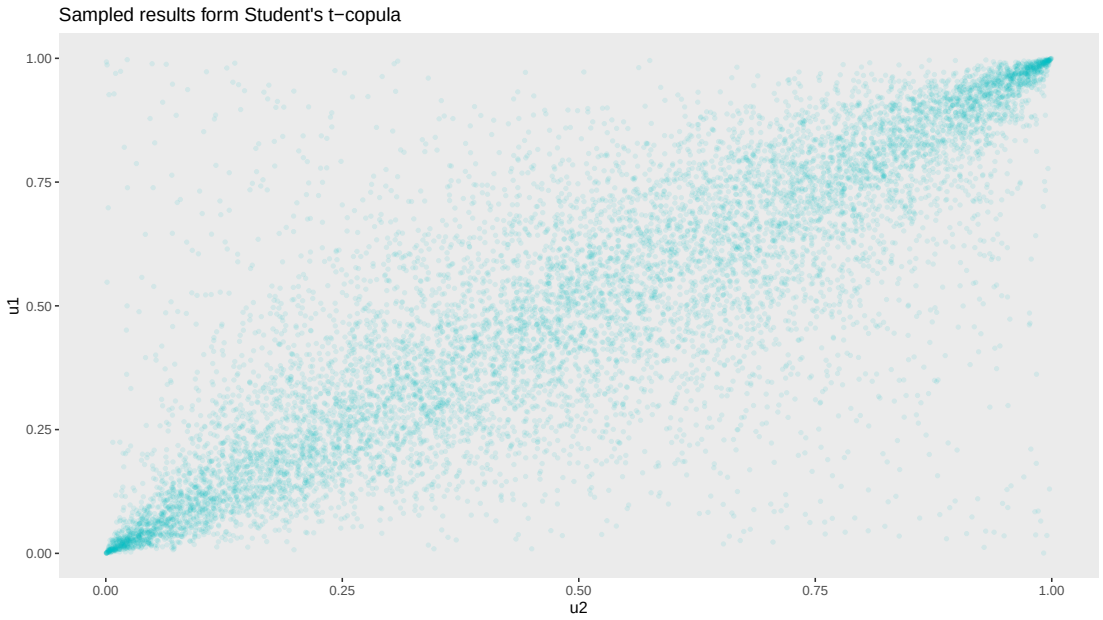


Figure 18. Sampled values from fitted Student's t-copula

After converting back to the original scale, these sampled values can be plotted to Figure 15 and compared to the original observed values to illustrate how copula captures the dependency structure between those two share classes. Figure 19 shows that the fitted copula yields rather similar observations, but fails to capture the most extreme events. Visually, trading signals would be indicated by those independent dots that deviate significantly from the dependency structure indicated by the dense cloud.

Like Krauss and Stübinger (2017), copulas were modelled in R. Copulas can be modelled in R using package *VineCopula* written by Technical University of Münich mathematical statistician Thomas Nagler et al. Nagler's research interests include applying vine copulas to portfolio optimization problems, discussed in Nagler et al. (2019a). Fundamentally, vine copulas are a collection of joint bivariate copulas, each of which may be parameterized differently. This allows modelling the dependency structure of a large portfolio decomposed to dependencies between each pair of securities in the portfolio. Because bivariate copulas are the building blocks of vine copulas, this package contains all of the required methods for selecting, fitting and sampling bivariate copulas. For example, it provides a function for selecting the appropriate family of copulas based on AIC, BIC and log-likelihood. This method tests for all common copula families, including the Gaussian, Student (t-copula), Clayton, Gumbel, Frank and Joe copula families. Functions provided by the package are described in detail in Nagler et al. (2019b).

Series were first forced inside the unit square by applying the empirical cumulative distribution

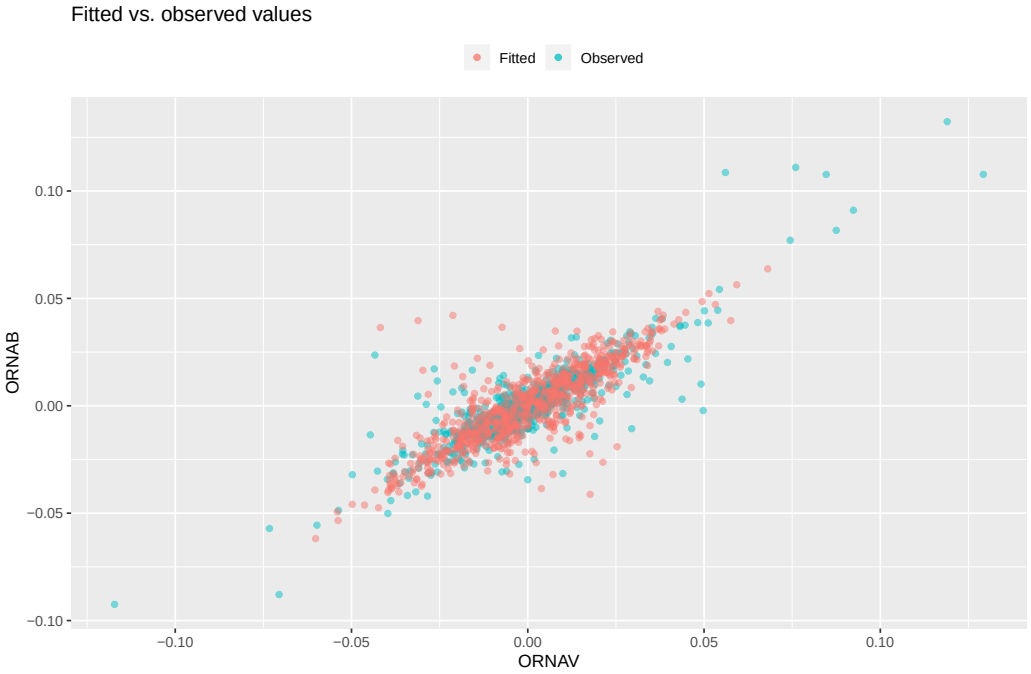


Figure 19. Fitted vs. observed values

function (CDF) to each series. This yields values with uniform distribution $U(0,1)$. These values could then be converted to any distribution, including the original, by applying the inverse CDF of that distributution. Once inside the unit square, best fitting copula can be selected.

R function `BiCopSelect()` for Selection and Maximum Likelihood Estimation of Bivariate Copula Families from package *VineCopula* is defined as

```
BiCopSelect(u1,u2, familyset = NA, selectioncrit = "AIC",
            indeptest = FALSE, level = 0.05, weights = NA,
            rotations = TRUE, se = FALSE, presel = TRUE,
            method = "mle"
),
```

where `u1` and `u2` are data vectors of equal length with values in [0,1], `familyset` is a vector of bivariate copula families to select from. The vector has to include at least one bivariate copula family that allows for positive and one that allows for negative dependence. Default value `NA` indicates that selection among all possible families is performed. The most relevant encodings are listed in Table 8.

`selectioncrit` accepts values `"AIC"` for Akaike information criterion, `"BIC"` for Bayesian

Table 8. List of copula encodings in package VineCopula

Encoding	Copula
0	Independence
1	Gaussian
2	Student t
3	Clayton
4	Gumbel
5	Frank
6	Joe

information criterion and `"logLik"` for log-likelihood. Pairs were selected based on Akaike information criterion. `method` defaults to maximum likelihood estimation `"mle"`, but inversion of Kendall's tau `"itau"` could also be used. Copulas were selected based on maximum likelihood estimation.

To generate trading signals, partial derivatives of the copula functions are required. Nowadays those are rather easy to obtain, for example, with package `Deriv`, thanks to recent advances in symbolic computation.

Package `VineCopula` contains first and second order derivatives for Copula functions defined in the package. Function `BiCopHfunc()` can be used to evaluate partial derivatives with respect to data vectors u and v for previously selected `BiCop` object. For a previously trained `BiCop` object `obj`, partial derivative with respect to normalized current trading data series u at point (u_t, v_t) is given as `BiCopHfunc1(u, v, obj)` and trading signal is generated if the returned value is greater than 0.95 or smaller than 0.05.

As there are very few known universal goodness-of-fit tests for copulas, pairs were chosen based on the selections in previous two methods. This serves as a validation criterion for pair selection in those two methods, as seeing independence copula emerge in copula selection is often an indication of a flaw in pair selection when pairs were chosen by assumed statistical dependency.

Copulas were fitted on the selected pairs and the profitability of copula-based trading rules was evaluated and compared to the profitability of non-copula approaches. Huang and Prokhorov (2014) have proposed a new rank-based goodness-of-fit test built on White's information criterion, but its reliability has not been validated and the procedure has not yet implemented by any of the common statistical software.

Trades were performed using the returns-based approach of Liew and Wu (2013), meaning

that each trading signal was created purely based on the preceding return. Trigger values for opening trades were above 0.95 and below 0.05, and positions were closed when both of the partial derivatives crossed 0.5.

4 Results

This chapter discusses observations in pair selection and profitability of different trading strategies. It compares the number of times each pair was selected in distance and cointegration-based selection as well as discusses the amount of trading opportunities created by each of the four signal creation methods and compares the magnitude of returns provided by each of the methods.

4.1 Selected pairs

The distance method identified 423 unique pairs within the best 20 pairs of each trading period suitable for trading. All theoretical pairs composed of two different share classes in the underlying company are represented among these pairs. A visualization of typical pairs selected by distance method is presented in Appendix A4.1 and A4.2. In Appendix A4.1 there are four pairs that consist of different classes of shares in the same companies among the the top five pairs with the lowest sums of squared differences. These are Orion, Oriola, Stockmann and Kesko. At positions 7, 8 and 9 there are three more of such companies - Ålandsbanken, Stora Enso and SSAB.

Table 9. Number of times in top 20 (distance)

Stockmann Oyj Abp A - Stockmann Oyj Abp B	56
Orion Oyj A - Orion Oyj B	53
Kesko Oyj B - Kesko Oyj A	50
Oriola Oyj A - Oriola Oyj B	39
Stora Enso Oyj A - Stora Enso Oyj R	39
Metsä Board Oyj A - Metsä Board Oyj B	30
Stora Enso Oyj R - UPM-Kymmene Oyj	29
Ålandsbanken Abp A - Ålandsbanken Abp B	22
Stora Enso Oyj A - UPM-Kymmene Oyj	19
Sampo Oyj A - Nordea Bank Abp	17
SSAB A - SSAB B	17
Elisa Oyj - Telia Company	15
Ilkka-Yhtymä Oyj 2 - Keskisuomalainen Oyj A	13
Citycon Oyj - Technopolis Oyj	11
Apetit Oyj - Rapala VMC Oyj	11
Citycon Oyj - Sampo Oyj A	10

Table 9 lists number of times each pair ranked within the best 20 pairs. During the 66 periods, Stockmann was selected 56 times, Orion 53 times and Kesko 50 times. The most reliably selected pair not composed of different share classes of the same company was Stora Enso - UPM-Kymmene, which was selected 19 times. Pairs that were selected fewer than 10 times were excluded from the table.

Table 10. Number of times in top 20 (cointegration)

Orion Oyj A - Orion Oyj B	31
Stora Enso Oyj A - Stora Enso Oyj R	29
Oriola Oyj A - Oriola Oyj B	28
Metsä Board Oyj A - Metsä Board Oyj B	22
Ålandsbanken Abp A - Ålandsbanken Abp B	18
Stockmann Oyj Abp A - Stockmann Oyj Abp B	16
Kesko Oyj B - Kesko Oyj A	14
Elecster Oyj A - Uponor Oyj	9
Citycon Oyj - Technopolis Oyj	8
Lassila & Tikanoja Oyj - SRV Yhtiöt Oyj	7
Konecranes Oyj - Vaisala Oyj A	7
Sampo Oyj A - Nordea Bank Abp	7
Elecster Oyj A - Vaisala Oyj A	7
Stockmann Oyj Abp A - Viking Line Abp	7
Wulff-Yhtiöt Oyj - Lassila & Tikanoja Oyj	6
Elecster Oyj A - Tikkurila Oyj	6

Similar observations were found in cointegration based selection. Table 10 lists the most commonly selected pairs based on MacKinnon's p-value. The order at the top changes slightly, and no single pair was selected in more than 50% of the periods. This time, pairs that were selected less than 6 times were excluded from the table. In cointegration based selection, all 20 pairs were on average, significantly cointegrated at the significance level of 0.1%. However, the cointegration-based selection resulted in 641 unique pairs among the 20 selected pairs during each of the 66 periods. This is an increase of 52% to the distance-based selection.

The variability seems to be much higher when using the cointegration method. In total, 382 unique pairs made it to the top-10 in cointegration method. This is significantly more than the 212 pairs selected in the top-10 based on distance.

Tables 11 and 12 list pairs with most trades by selection method. Although pair Stockmann A - Stockmann B was selected on more periods than any other pair based on distance criterion, it did not create as many trading opportunities than pairs that were selected 2nd and 5th most

often. On the other hand, the most often selected cointegrated pairs provided the greatest number of trading opportunities.

By cointegration method, Orion A - Orion B is the most suitable pair for trading. In distance method it ranks second. However, when only the first places are considered, Orion A - Orion B ranks significantly more often as the most cointegrated or most closely matching pair than any other pair in either of the selection methods. (Tables 13 and 14).

Orion A - Orion B provided the greatest number of trading opportunities using both trading strategies. Interestingly, this pair provided only 8% less trading opportunities using cointegration method although it was selected 42% less often. For this pair, the cointegration method seems to provide more trading opportunities per period.

Table 11. Pairs with most trades (distance)

Orion Oyj A - Orion Oyj B	342
Stora Enso Oyj A - Stora Enso Oyj R	313
Stockmann Oyj Abp A - Stockmann Oyj Abp B	266
Oriola Oyj A - Oriola Oyj B	229
Kesko Oyj B - Kesko Oyj A	196
Metsä Board Oyj A - Metsä Board Oyj B	142
Ålandsbanken Abp A - Ålandsbanken Abp B	112
SSAB A - SSAB B	104
Stora Enso Oyj R - UPM-Kymmene Oyj	64
Ilkka-Yhtymä Oyj 2 - Keskisuomalainen Oyj A	53
Stora Enso Oyj A - UPM-Kymmene Oyj	52
Elecster Oyj A - Tikkurila Oyj	46
Elisa Oyj - Telia Company	36
Citycon Oyj - Technopolis Oyj	34
Viking Line Abp - Keskisuomalainen Oyj A	34
Sampo Oyj A - Nordea Bank Abp	34
Fiskars Oyj Abp - Rapala VMC Oyj	30
Citycon Oyj - Sampo Oyj A	28
Citycon Oyj - Nordea Bank Abp	28
Apetit Oyj - Rapala VMC Oyj	28
Rapala VMC Oyj - Marimekko Oyj	24
Ålandsbanken Abp B - eQ Oyj	22
Elecster Oyj A - Vaisala Oyj A	22
Wärtsilä Oyj Abp - Aspo Oyj	22
Cramo Oyj - Ramirent Oyj	20
Ponsse Oyj 1 - Yleiselektroniikka Oyj	20
Finnair Oyj - Viking Line Abp	20
Ilkka-Yhtymä Oyj 2 - Pohjois-Karjalan Kirjapaino	20
Finnair Oyj - Sanoma Oyj	20
Sampo Oyj A - Aktia Bank Abp	20
Pohjois-Karjalan Kirjapaino - Keskisuomalainen ...	20

Table 12. Pairs with most trades (cointegration)

Orion Oyj A - Orion Oyj B	315
Stora Enso Oyj A - Stora Enso Oyj R	274
Oriola Oyj A - Oriola Oyj B	250
Stockmann Oyj Abp A - Stockmann Oyj Abp B	172
Metsä Board Oyj A - Metsä Board Oyj B	134
Ålandsbanken Abp A - Ålandsbanken Abp B	74
Kesko Oyj B - Kesko Oyj A	52
Lassila & Tikanoja Oyj - SRV Yhtiöt Oyj	40
Stockmann Oyj Abp A - Viking Line Abp	40
Ramirent Oyj - Wulff-Yhtiöt Oyj	32
Konecranes Oyj - Vaisala Oyj A	32
Elecster Oyj A - Uponor Oyj	30
Citycon Oyj - Technopolis Oyj	26
Elecster Oyj A - Vaisala Oyj A	26
Sampo Oyj A - Nordea Bank Abp	24
Ålandsbanken Abp A - CapMan Oyj	24
Apetit Oyj - Rapala VMC Oyj	24
Finnair Oyj - Viking Line Abp	24
Wulff-Yhtiöt Oyj - Glaston Oyj Abp	23
Elecster Oyj A - Tikkurila Oyj	22
Nurminen Logistics Oyj - Lehto Group Oyj	21
YIT Oyj - SRV Yhtiöt Oyj	20
Aspo Oyj - Wulff-Yhtiöt Oyj	20
Technopolis Oyj - eQ Oyj	20

Table 13. Number of times with the lowest SSD (distance)

Orion Oyj A - Orion Oyj B	39
Stora Enso Oyj A - Stora Enso Oyj R	14
Oriola Oyj A - Oriola Oyj B	4
SSAB A - SSAB B	4
Stockmann Oyj Abp A - Stockmann Oyj Abp B	3
Kesko Oyj B - Kesko Oyj A	2

Table 14. Number of times with the lowest MacKinnon p-value

Orion Oyj A - Orion Oyj B	18
Stora Enso Oyj A - Stora Enso Oyj R	11
Oriola Oyj A - Oriola Oyj B	9
Ålandsbanken Abp A - Ålandsbanken Abp B	3
Stockmann Oyj Abp A - Stockmann Oyj Abp B	3
Keskisuomalainen Oyj A - Sanoma Oyj	2

4.2 Returns of the distance method

The distance method generated 2349 unique trades during the 16-year period. Table 15 lists summary statistics for distance-based trades. On average, each trade provided a return of 3%. Worst trade provided a loss of 53% and best trade yielded a gain of 91%. These are simple returns indicating the average profit or loss made when position is closed. The returns are not annualized here because some positions stay open for longer and some positions are realized sooner. The average duration of an open position is discussed later in this chapter.

Table 15. Summary statistics for distance trades

	long profit	short profit	total profit
count	2349	2349	2349
mean	0.02	0.02	0.03
std	0.12	0.12	0.11
min	-0.60	-0.47	-0.53
25%	-0.02	-0.03	-0.00
50%	0.02	0.02	0.04
75%	0.07	0.07	0.10
max	0.74	1.13	0.91

Table 16 lists win ratios of distance trades per position determined at the end of the fitting period. The pair which has the lowest sum of squared differences ranks first, the second lowest SSD ranks second and so on. The best two pairs per each period seem to provide more consistent profit than pairs that have, on average, more distance between them. A win ratio of 94% is stunningly good and only 17 bad trades during 66 trading periods, or 33 years of trading, is extraordinary. The average quality of pairs dwindles after best two pairs, but all 20 pairs win more frequently than they lose.

The hightest ranked pairs provide an average a return of 3% and win 94% of trades. Returns of the best two pairs are very consistent, the worst loss was -11% and the greatest gain was 14%. The best pair opens, on average, 4.68 times per trading period, indicating a return of 32% p.a. ignoring the trading costs and rent paid for shorted assets. The second best pair provides an average return of 4% and opens, on average, 3.2 times per period for a total return of 28% p.a. The 20^{th} pair provides an average return of 1% per trade and opens 1.5 times per period, yielding only 3% p.a. before subtracting the costs associated with trading. Only the first 7 pairs appear suitable for profitable trading using the distance method. (Table 18)

Table 16. Distance win ratio by position

rank	loss	win	ratio
1	17	289	.94
2	13	195	.94
3	25	108	.81
4	29	119	.80
5	31	92	.75
6	28	93	.77
7	33	77	.70
8	31	64	.67
9	30	61	.67
10	34	51	.60
11	26	68	.72
12	35	56	.62
13	28	84	.75
14	28	59	.68
15	32	50	.61
16	37	57	.61
17	37	42	.53
18	32	59	.65
19	32	53	.62
20	41	57	.58

Table 17. Duration of distance trades before convergence

	all pairs	best 5 pairs	pairs 6-10	pairs 11-20
count	2349	925	503	921
mean	52 days	30 days	63 days	67 days
std	52 days	42 days	53 days	53 days
25%	7 days	3 days	16 days	19 days
median	30 days	10 days	45 days	56 days
75%	90 days	37 days	114 days	114 days

Better pairs converge more quickly. 50% of the best 5 pairs converge within 10 days, while for pairs 6 to 10 median is 45 days and for pairs 11 to 20 it is 56 days. Mean duration is high for all pairs because some trades do not converge, which results in duration of 170 days or longer. Converge within 10 days is in line with observations made in Rinne and Suominen (2017).

Table 18. Summary statistics of absolute returns from an individual distance trade by position

rank	count	mean	std	min	25%	50%	75%	max
1	309	0.03	0.02	-0.10	0.02	0.02	0.04	0.13
2	212	0.04	0.03	-0.11	0.02	0.04	0.05	0.14
3	133	0.04	0.07	-0.32	0.02	0.04	0.07	0.25
4	148	0.04	0.07	-0.33	0.01	0.04	0.08	0.23
5	123	0.03	0.08	-0.25		0.05	0.09	0.18
6	122	0.05	0.14	-0.40	0.01	0.07	0.11	0.91
7	110	0.03	0.11	-0.37	-0.01	0.05	0.12	0.32
8	95	0.03	0.11	-0.30	-0.03	0.05	0.12	0.23
9	91	0.02	0.16	-0.53	-0.05	0.05	0.14	0.30
10	85	0.02	0.11	-0.42	-0.03	0.03	0.11	0.22
11	95	0.04	0.14	-0.41	-0.01	0.07	0.14	0.31
12	92	0.04	0.12	-0.36	-0.04	0.07	0.13	0.31
13	113	0.06	0.12	-0.39		0.08	0.13	0.44
14	89	0.03	0.13	-0.37	-0.04	0.05	0.12	0.26
15	83	0.04	0.14	-0.29	-0.04	0.07	0.13	0.37
16	94	0.03	0.16	-0.44	-0.09	0.06	0.15	0.35
17	79	0.01	0.16	-0.44	-0.06	0.02	0.14	0.27
18	91	0.04	0.15	-0.42	-0.04	0.04	0.15	0.35
19	86	0.03	0.13	-0.27	-0.07	0.06	0.12	0.25
20	99	0.01	0.14	-0.31	-0.07	0.02	0.10	0.36

4.3 Returns of the cointegration method

In total, 2 456 trades were made for a mean profit of 4% per trade. The worst trade lost 84% of the initial value and the best trade yielded 199%. (Table 19).

The win ratio of the most cointegrated pair is almost as good as the win ratio of the pair with smallest SSD. After best five pairs, win ratio quickly decays to $\approx$ 60%. As with the distance method, all 20 pairs win on more trades than they lose. (Table 20).

Table 21 lists cointegration trades by position. It is apparent that the most cointegrated pairs

Table 19. Summary statistics for cointegration trades

	long profit	short profit	total profit
count	2456	2456	2456
mean	0.02	0.03	0.04
std	0.13	0.15	0.16
min	-0.85	-0.83	-0.84
25%	-0.02	-0.02	-0.02
50%	0.01	0.02	0.05
75%	0.06	0.08	0.11
max	0.94	1.99	1.99

open more often and provide more reliable returns. Compared to the distance method though, returns of the most cointegrated pairs vary a lot more than returns of the pairs with the least sum of squared differences between them.

Table 20. Cointegration win ratio by position

rank	loss	win	ratio
1	20	272	.93
2	26	174	.87
3	27	133	.83
4	34	106	.76
5	30	88	.75
6	36	68	.65
7	35	66	.65
8	36	65	.64
9	37	54	.59
10	33	90	.73
11	41	69	.63
12	29	83	.74
13	34	53	.61
14	36	69	.66
15	31	70	.69
16	31	59	.66
17	36	64	.64
18	37	54	.59
19	41	61	.60
20	36	80	.69

The most cointegrated pair provides an average return of 3% and opens 4.44 times per period. This means an annualized return of 30% before subtraction of costs. The 2nd most cointegrated pair provides a mean return of 4%, opens 3.1 times per period and yields 27%

p.a. Average annualized returns of pairs 10 to 20 vary between 3% and 30%.

Table 21. Summary statistics of absolute returns from a cointegration trade by position

	count	mean	std	min	25%	50%	75%	max
1	293	0.03	0.06	-0.37	0.02	0.03	0.05	0.63
2	203	0.04	0.11	-0.84	0.02	0.04	0.06	0.59
3	162	0.04	0.09	-0.50	0.02	0.04	0.08	0.27
4	143	0.05	0.14	-0.55		0.05	0.11	0.70
5	118	0.06	0.22	-0.49		0.06	0.12	1.99
6	105	0.04	0.18	-0.73	-0.04	0.05	0.11	0.66
7	101	0.06	0.19	-0.46	-0.05	0.08	0.14	0.61
8	101	0.07	0.20	-0.30	-0.04	0.07	0.13	0.82
9	91	0.01	0.16	-0.46	-0.08	0.06	0.11	0.31
10	123	0.04	0.14	-0.38	-0.01	0.05	0.12	0.42
11	110	0.02	0.18	-0.75	-0.07	0.07	0.13	0.36
12	112	0.08	0.15	-0.34	-0.01	0.08	0.12	0.56
13	88	0.06	0.23	-0.55	-0.06	0.05	0.16	0.82
14	105	0.03	0.16	-0.81	-0.07	0.05	0.13	0.45
15	101	0.04	0.18	-0.60	-0.03	0.08	0.14	0.59
16	90	0.02	0.15	-0.63	-0.04	0.05	0.11	0.38
17	100	0.05	0.26	-0.60	-0.07	0.07	0.13	1.61
18	91	0.01	0.15	-0.40	-0.09	0.03	0.11	0.41
19	103	0.01	0.15	-0.34	-0.07	0.05	0.10	0.55
20	116	0.05	0.17	-0.45	-0.03	0.05	0.13	0.68

Table 22. Duration of cointegration trades before convergence

	all pairs	best 5 pairs	pairs 6-10	pairs 11-20
count	2456	919	521	1016
mean	58 days	33 days	71 days	73 days
std	57 days	46 days	59 days	57 days
25%	8 days	3 days	18 days	22 days
median	34 days	10 days	50 days	60 days
75%	104 days	45 days	128 days	123 days

4.4 Returns of copula method

Tables 23 and 24 list selected copulas per pair formation method. The t-copula proved to be the best fitting copula in most cases using either of the formation methods. This is consistent with results for Cramér-von Mises goodness-of-fit test on the DAX 30 constituents from January 2005 until December 2014 listed in Table 2 where most favorable values are highlighted in bold. If one were to fit a single copula type to all pairs, t-copula would be the best choice. In this thesis though, copula trades were performed using the best fitting copula.

Table 23. Copula selections in distance pairs

Copula family	**n**	**%**	**cum%**
t	531	40.2	40.2
Gumbel	155	11.7	52.0
Gaussian	150	11.4	63.3
Survival Gumbel	141	10.7	74.0
Survival Joe	82	6.2	80.2
Joe	64	4.8	85.1
Clayton	48	3.6	88.7
Rotated Tawn type 1 180 degrees	46	3.5	92.2
Tawn type 1	26	2.0	94.2
Survival Clayton	25	1.9	96.1
Rotated Joe 270 degrees	12	0.9	97.0
Rotated Joe 90 degrees	8	0.6	97.6
Rotated Clayton 90 degrees	7	0.5	98.1
Rotated Gumbel 270 degrees	7	0.5	98.6
Rotated Tawn type 1 270 degrees	6	0.5	99.1
Rotated Clayton 270 degrees	5	0.4	99.5
Rotated Tawn type 1 90 degrees	5	0.4	99.8
Rotated Gumbel 90 degrees	2	0.2	100.0
Total	1320	100.0	100.0

Tables 25 and 26 list summary statistics of the returns for individual copula trades in distance-based and cointegration-based selection strategies. Because per-trade returns of copula approach are much smaller than the returns of previous trading strategies, copula strategy should create significantly more trading opportunities to be able to compete with profitability of distance and cointegration strategies.

However, the copula method combined with either of the pair selection criteria results in fewer trades than cointegration- and distance-based trading signal generation. The annualized returns of the best distance-selected pairs reach only 6% before deducting trading costs. For the best cointegration-selected pairs, the annualized return is only 4%.

Table 24. Copula selections in cointegrated pairs

Copula family	**n**	**%**	**cum%**
t	359	27.2	27.2
Gaussian	201	15.2	42.4
Survival Gumbel	169	12.8	55.2
Survival Joe	128	9.7	64.9
Gumbel	105	8.0	72.9
Clayton	70	5.3	78.2
Tawn type 1	55	4.2	82.3
Joe	53	4.0	86.4
Rotated Tawn type 1 180 degrees	51	3.9	90.2
Survival Clayton	44	3.3	93.6
Rotated Joe 90 degrees	24	1.8	95.4
Rotated Clayton 270 degrees	13	1.0	96.4
Rotated Joe 270 degrees	11	0.8	97.2
Rotated Tawn type 1 90 degrees	11	0.8	98.0
Rotated Gumbel 270 degrees	8	0.6	98.6
Rotated Gumbel 90 degrees	7	0.5	99.2
Rotated Tawn type 1 270 degrees	6	0.5	99.6
Rotated Clayton 90 degrees	5	0.4	100.0
Total	1320	100.0	100.0

Table 25. Absolute returns from copula trades (distance) by position

rank	count	mean	std	min	25%	50%	75%	max
1	188	0.01	0.01	-0.05		0.01	0.01	0.07
2	161	0.01	0.02	-0.06		0.01	0.02	0.09
3	158	0.01	0.03	-0.05			0.02	0.10
4	108	0.01	0.03	-0.05		0.01	0.03	0.18
5	145	0.01	0.03	-0.08			0.02	0.15
6	89	0.01	0.05	-0.17		0.01	0.03	0.18
7	81	0.01	0.03	-0.06	-0.01	0.01	0.02	0.08
8	81	0.01	0.03	-0.10	-0.01		0.02	0.10
9	86		0.04	-0.13	-0.01		0.02	0.19
10	83	0.01	0.04	-0.23	-0.01		0.03	0.15
11	85	0.01	0.05	-0.10	-0.01		0.02	0.27
12	80		0.05	-0.23	-0.01		0.02	0.18
13	91	0.01	0.05	-0.18	-0.02		0.03	0.11
14	74	0.02	0.05	-0.05		0.01	0.03	0.26
15	95		0.05	-0.24	-0.01		0.02	0.14
16	61		0.04	-0.11	-0.01		0.03	0.18
17	71	0.01	0.03	-0.08			0.02	0.15
18	71	0.01	0.04	-0.05		0.01	0.03	0.17
19	50	0.02	0.06	-0.10		0.02	0.04	0.34
20	64		0.04	-0.07	-0.02		0.02	0.16

Table 26. Absolute returns from copula trades (cointegration) by position

rank	count	mean	std	min	25%	50%	75%	max
1	131	0.01	0.03	-0.06		0.01	0.02	0.15
2	103		0.04	-0.15		0.01	0.02	0.16
3	102	0.01	0.04	-0.14		0.01	0.03	0.11
4	73	0.01	0.06	-0.19	-0.01	0.01	0.03	0.30
5	74	0.01	0.06	-0.11	-0.01		0.04	0.25
6	72		0.05	-0.18	-0.02		0.03	0.18
7	61	0.01	0.06	-0.13	-0.02	0.01	0.04	0.18
8	67	0.01	0.06	-0.14	-0.02	0.01	0.03	0.18
9	74	-0.01	0.10	-0.70	-0.02		0.02	0.18
10	94		0.07	-0.46	-0.01		0.03	0.17
11	63	0.01	0.04	-0.09	-0.02	0.01	0.03	0.12
12	53	0.01	0.05	-0.12	-0.02		0.03	0.14
13	64		0.06	-0.19	-0.01		0.03	0.14
14	73	0.01	0.04	-0.16		0.01	0.03	0.12
15	77		0.05	-0.24	-0.01	0.01	0.02	0.13
16	57		0.04	-0.10	-0.01		0.03	0.11
17	69	0.01	0.12	-0.57	-0.01	0.01	0.03	0.66
18	97	0.01	0.06	-0.18	-0.01		0.03	0.31
19	85	0.01	0.05	-0.11	-0.02		0.02	0.17
20	72	0.01	0.04	-0.07	-0.01	0.01	0.03	0.15

Win ratios for both selection methods are presented in Table 27 by position. As with other approaches, copula trades seem to win more often than they lose. However, in copula method the win ratio is less dependent on the *quality* of the pair and averages to $\approx 60\%$ independent of the ranking in pair selection.

Because copula trades used a returns-based signal creation method as opposed to value-based method in distance and cointegration strategies, positions are more responsive to noise and tend to stay open for much shorter period of time. (Tables 28 and 29). This creates a smaller window for deviations to occur, and limits the potential of gains and losses. Therefore, the minimum and maximum returns per trade are significantly smaller than in distance and cointegration strategies.

Table 27. Copula win ratios

(a) Distance

rank	loss	win	ratio
1	44	143	.76
2	42	114	.73
3	55	99	.64
4	36	71	.66
5	45	91	.67
6	31	57	.65
7	29	50	.63
8	29	50	.63
9	35	49	.58
10	39	43	.52
11	31	49	.61
12	33	44	.57
13	39	49	.56
14	24	49	.67
15	42	50	.54
16	27	33	.55
17	27	42	.61
18	22	48	.69
19	11	38	.78
20	32	30	.48

(b) Cointegration

rank	loss	win	ratio
1	36	93	.72
2	32	68	.68
3	29	71	.71
4	24	46	.66
5	29	44	.60
6	29	42	.59
7	24	35	.59
8	29	37	.56
9	34	37	.52
10	38	53	.58
11	24	37	.61
12	24	29	.55
13	28	34	.55
14	21	49	.70
15	33	44	.57
16	27	30	.53
17	21	46	.69
18	35	60	.63
19	35	48	.58
20	27	43	.61

Table 28. Duration of copula trades (distance) before convergence

	all pairs	best 5 pairs	pairs 6-10	pairs 11-20
count	1922	760	420	742
mean	7 days	8 days	7 days	6 days
std	7 days	7 days	7 days	6 days
25%	2 days	2 days	2 days	2 days
median	5 days	6 days	5 days	5 days
75%	10 days	11 days	9 days	8 days

Table 29. Duration of copula trades (cointegration) before convergence

	all pairs	best 5 pairs	pairs 6-10	pairs 11-20
count	1561	483	368	710
mean	6 days	7 days	7 days	6 days
std	6 days	6 days	7 days	6 days
25%	2 days	2 days	3 days	2 days
median	5 days	6 days	5 days	5 days
75%	9 days	10 days	9 days	9 days

4.5 Empirical testing

To test the real-life performance of pairs trading, simulations were run with the most commonly selected pair Orion Class A - Orion Class B. This pair was the best suitable pair for pairs trading by both the cointegration criterion and the minimum distance criterion. Trading strategy was formed based on rolling windows – parameters of the trading system were adjusted each day based on the values of previous days. Both trading signal generation methods provided approximately equal number of trading opportunities, about 115 during the 12-year period for which data was available for both of the stocks.

The distance method provided a bit higher returns than the cointegration method. Often the trades opened on the same day, but the closing dates differed. Ignoring the transaction costs, the distance method yielded 1 400% and the cointegration method returned 900% during the 12 years. For that period, the market benchmark OMX Helsinki 25 yielded approximately 50%. (Figure 20). Annualized Sharpe ratio was 1.36 (distance) and 1.16 (cointegration).

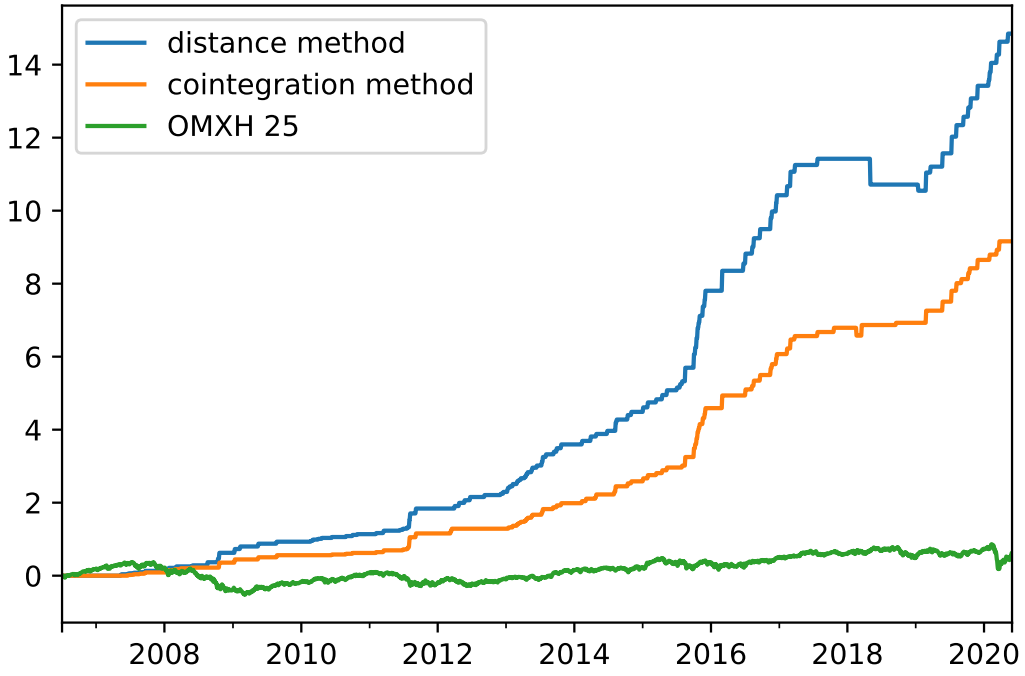


Figure 20. Trading Orion A - Orion B with no transaction costs.

When capital gains taxes are subtracted from the returns, previously reported 1 400% decreases to only 500% for the distance method and 900% reduces to 350%. (Figure 21). This is much closer to the reality, and shows how much capital gains taxes reduce the returns of frequent trading.

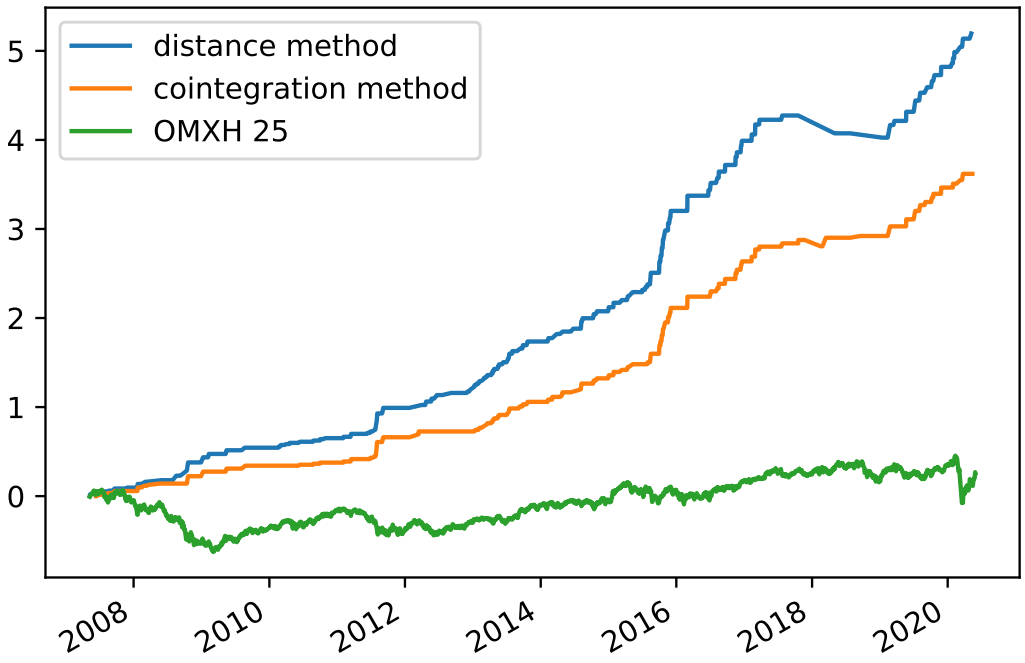


Figure 21. Trading Orion A - Orion B with 34 % capital gains tax.

When trading costs are added to the already high tax burden, returns decline quite sharply. Just 1% of transaction costs per opened position reduces the 12-year gains of the distance-method to 180% and to 140%. (Figure 22).

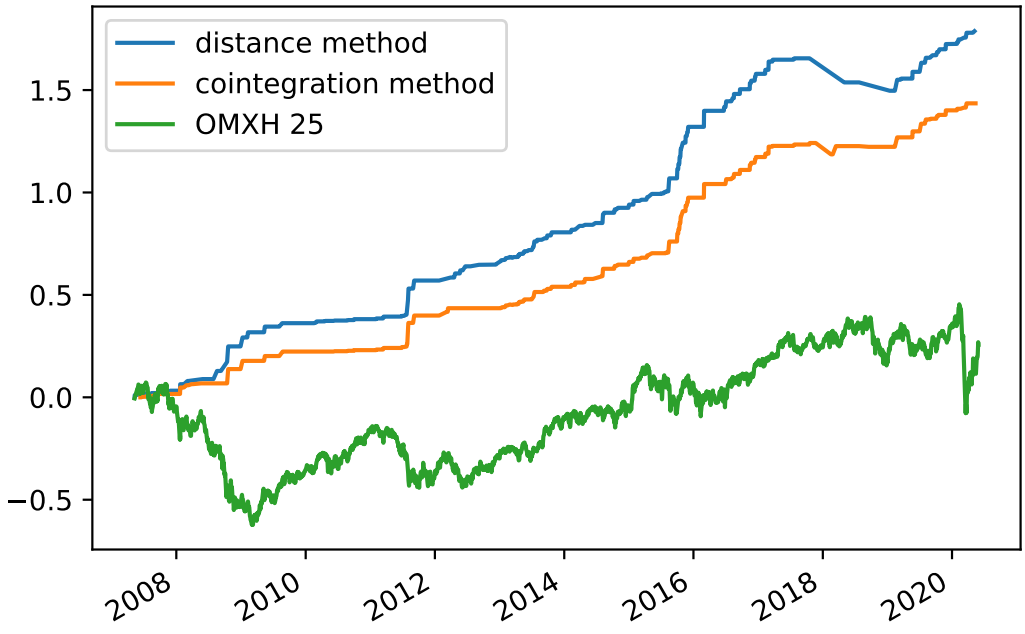


Figure 22. Trading Orion A - Orion B with 34 % capital gains tax and 1% transaction costs per opened position.

4.6 Summary

Table 30 presents annualized return of an equally weighted portfolio composed of the best five pairs per trading period per trading strategy with reference to corresponding OMXH 25 index return. The most significant periods are 12-19, from July 2007 to August 2009. Those periods contain the financial crisis of 2007-2008 and periods 12-15 show significant −30% p.a. market decline, periods 16 and 17 experience a −70% p.a. market crash and period 19 a 107% market gain. All trading strategies did protect the investor from those steep declines, often yielding double-digit gains period.

Table 30. Annualized returns of a portfolio consisting the best five pairs per period per method compared to the market returns (OMXH 25)

period	cointcop	distcop	cointegration	distance	omxh25
1	0.14	0.06	0.32	0.13	0.25
2		0.28	0.03	0.02	0.32
3	0.17	0.06	0.25	0.15	0.24
4	0.06	0.15	0.16	0.10	0.45
5	0.12	0.54	0.30	0.12	0.24
6	-0.05	0.96	0.67	0.09	-0.05
7	0.15	0.77	1.04	0.10	0.46
8	0.06	0.11	0.37	0.09	0.50
9	0.02	0.05	-0.01	-0.04	0.37
10	0.10	0.14	0.29	0.05	0.09
11	0.06	0.33	0.09	0.07	0.04
12	0.08	0.42	0.44	0.08	-0.32
13	0.09	0.60	0.34	0.16	-0.31
14	0.11	4.66	0.47	0.11	-0.32
15	0.15	0.55	0.97	0.08	-0.29
16	0.34	0.98	2.13	0.05	-0.69
17	0.10	1.38	1.09	0.08	-0.73
18	0.50	1.45	0.59	0.20	0.22
19	0.09	0.29	0.04	0.09	1.07
20	0.03	0.57	0.04	0.08	0.41
21	-0.07	-0.02	0.06	0.04	0.25
22	0.10	0.31	0.28	-0.01	0.47
23	0.14	0.20	0.17	-0.04	0.20

Continues on next page

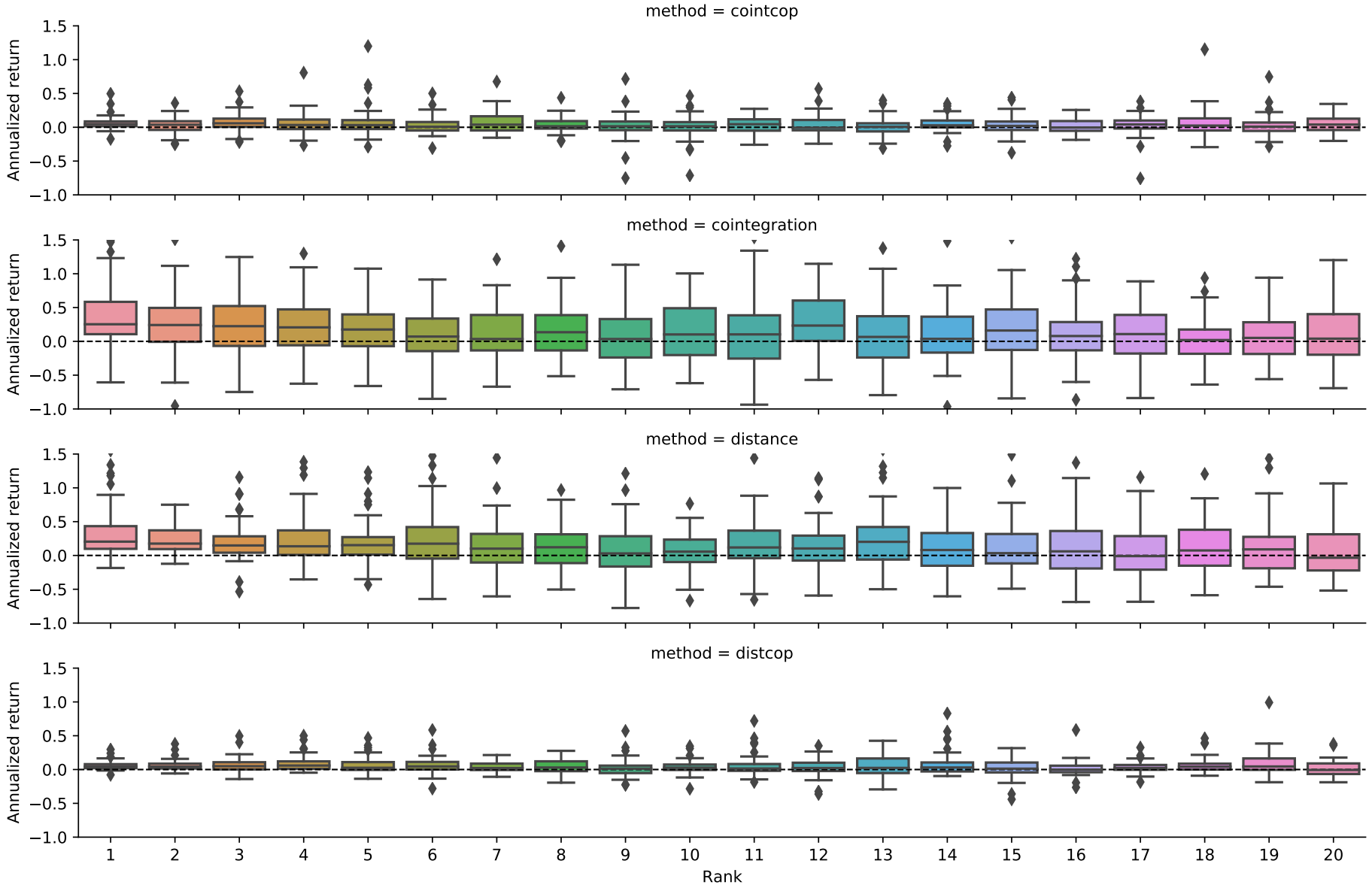
period	cointcop	distcop	cointegration	distance	omxh25
24	0.01	-0.00	0.06	0.09	0.20
25	-0.05	0.34	0.21	0.11	0.53
26	-0.08	-0.13	0.01	0.03	0.23
27	-0.00	0.31	-0.00	0.09	-0.26
28	0.02	0.14	0.52	0.03	-0.43
29	-0.11	0.53	0.38	0.11	-0.35
30	-0.04	0.21	0.16	0.20	0.43
31	0.13	0.21	0.09	0.11	-0.03
32	-0.09	-0.03	-0.01	0.19	-0.21
33	-0.04	0.09	0.37	0.04	0.09
34	-0.01	0.35	0.03	0.11	0.37
35	0.01	-0.10	0.14	-0.02	0.31
36	0.11	0.19	0.13	-0.00	0.00
37	0.06	0.43	0.07	0.03	0.34
38	0.01	0.03	0.06	0.09	0.51
39	-0.01	0.49	0.04	-0.00	0.16
40	0.04	0.38	0.28	0.03	0.15
41	0.06	0.29	0.39	0.02	0.09
42	-0.01	0.09	0.13	-0.01	0.05
43	0.11	0.12	0.19	0.03	0.38
44	0.03	0.10	0.18	0.03	0.28
45	0.00	0.29	0.14	0.02	-0.01
46	0.06	0.26	0.54	0.04	0.01
47	0.06	0.67	0.39	0.04	-0.11
48	0.11	0.29	0.23	0.10	-0.05
49	0.08	0.23	0.34	0.07	0.18
50	0.24	-0.23	0.34	0.05	0.25
51	0.30	0.18	0.09	0.06	0.29
52	0.04	0.14	0.06	0.06	0.20
53	-0.04	0.05	0.25	0.00	0.25
54	-0.01	0.16	0.26	0.00	0.07
55	-0.07	0.20	0.12	0.02	-0.08
56	0.06	0.19	-0.01	0.06	0.16
57	0.04	0.59	0.11	0.05	0.17
58	0.00	0.80	0.11	0.14	0.13

Continues on next page

period	cointcop	distcop	cointegration	distance	omxh25
59	0.05	-0.14	0.06	0.10	-0.08
60	0.06	0.71	0.19	0.04	-0.09
61	0.03	0.51	0.02	0.05	0.10
62	0.02	0.13	0.18	0.07	0.05
63	0.12	0.70	0.25	0.03	-0.15
64	0.04	0.19	0.08	-0.01	0.13
65	0.08	0.32	0.22	0.04	-0.42
66	0.03	-0.02	0.52	0.12	-0.03

Figure 23 displays annualized returns of each trading strategy per ranking in pair selection. The best two pairs in the distance method and the cointegration method provide consistent profit, but all copula strategies provide insufficient or negative expected returns.

Figure 23. Annualized returns per ranking per method



4.7 Future Work

This thesis establishes a foundation on which profitable pairs trading strategies can be constructed using either distance or cointegration based pair selection. Yet it failed to create a feasible trading strategy using the *returns-based* copula method with thresholds $P(U \leq u|V = v) > 0.95$ and $P(V \leq v|U = u) < 0.05$ or vice-versa. Further research is required on parametrization of the copula method to determine if a higher signal creation threshold would yield better trades and on how *level-based* strategy discussed in Liew and Wu (2013) would perform.

The real-life feasibility of trading strategies presented in this thesis remains to be tested. Generated pairs are theoretical and assume that any security can be sold short. Such assumption may not be realistic, as most brokers have a relatively small list of the most liquid assets that they let investors to short. Real-life returns may also be hindered by the realization of the execution risk - investor may not be able to find a buyer or seller for the asset when the position ought to be opened or closed.

5 Conclusions

This thesis proves that the observations made in Rinne and Suominen (2017) using a single pair (Stora Enso - UPM) are more widespread in the OMX Helsinki stock exchange and that the distance-method based transaction returns for a holding-period of approximately one week are indeed in the range of 2.4% or higher, with return volatility of 4.2% or lower, for a more optimal pair than Stora Enso - UPM. It also concludes that similar returns are achievable using cointegration-based trading, but the volatility of those returns is much higher, as is the risk of a stochastic change leading to the disappearance of mean-reversion.

Findings regarding the amount of trading opportunities generated by the copula method are in line with Harju (2016). While Harju used only cointegration criterion for selecting the pairs, this thesis combined copula method also with distance-based pair selection. Distance-based selection seems to create more trading opportunities, resulting in slightly better win ratio, and providing more consistent returns than cointegration-based selection. However, the magnitude of returns achieved through copula-based trading signal generation is not sufficient enough to compete with distance-based or cointegration-based trading signal creation.

This thesis confirms the assumptions made in previous literature about what types of pairs are suitable for pairs trading. When pair formation is limited to allow only pairs consisting of companies from the same industry, distance-based and cointegration-based selection favors pairs formed of different share classes of one company. An example of this is Orion Class A - Orion Class B. Such pairs provide more trading opportunities than other types of pairs (e.g. Stora Enso - UPM).

A trading system was simulated using the most frequently suggested pair Orion Class A - Orion Class B. From late 2007 to mid 2020, this pair yielded significant positive returns using both the cointegration and the distance method. However, the returns are highly sensitive to the costs associated with trading.

REFERENCES

Aas, Kjersti, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken (2009). "Pair-Copula Constructions of Multiple Dependence". In: *Insurance: Mathematics and Economics* vol. 44 (2), pp. 182–198. Link (visited on 05/01/2020) (cit. on p. 27).

Ané, Thierry and Cécile Kharoubi (2003). "Dependence Structure and Risk Measure". In: *The Journal of Business* vol. 76 (3), pp. 411–438 (cit. on pp. 25, 26, 28).

Blázquez, Mario Carrasco, Carmen De la Orden De la Cruz, and Camilo Prado Román (2018). "Pairs Trading Techniques: An Empirical Contrast". In: *European Research on Management and Business Economics; Madrid* vol. 24 (3), pp. 160–167. Link (visited on 04/06/2019) (cit. on p. 16).

Bolgün, Kaan Evren, Engin Kurun, and Serhat Güven (2012). "Adaptive Pairs Trading Strategy Performance in Turkish Derivatives Exchange with the Companies Listed on Istanbul Stock Exchange". In: *Journal of Derivatives & Hedge Funds; Basingstoke* vol. 18 (2), pp. 113–126. Link (visited on 05/11/2019) (cit. on p. 16).

Chen, Cathy W.S., Zona Wang, Songsak Sriboonchitta, and Sangyeol Lee (2017). "Pair Trading Based on Quantile Forecasting of Smooth Transition GARCH Models". In: *The North American Journal of Economics and Finance* vol. 39, pp. 38–55. Link (visited on 04/06/2019) (cit. on p. 15).

Chen, Danni, Jing Cui, Yan Gao, and Leilei Wu (2017). "Pairs Trading in Chinese Commodity Futures Markets: An Adaptive Cointegration Approach". In: *Accounting & Finance* vol. 57 (5), pp. 1237–1264. Link (visited on 08/06/2019) (cit. on pp. 21, 22, 39, 45, 47).

Chiu, Mei Choi and Hoi Ying Wong (2015). "Dynamic Cointegrated Pairs Trading: Mean–Variance Time-Consistent Strategies". In: *Journal of Computational and Applied Mathematics* vol. 290, pp. 516–534. Link (visited on 04/06/2019) (cit. on p. 15).

Clegg, Matthew and Christopher Krauss (2018). "Pairs Trading with Partial Cointegration". In: *Quantitative Finance* vol. 18 (1), pp. 121–138. Link (visited on 08/06/2019) (cit. on pp. 17, 24, 37, 39).

Cryer, Jonathan D. and Kung-sik Chan (2008). *Time Series Analysis: With Applications in R*. 2nd ed. Springer Texts in Statistics. New York: Springer. 491 pp. (cit. on p. 38).

De Moura, Carlos Eduardo, Adrian Pizzinga, and Jorge Zubelli (2016). "A Pairs Trading Strategy Based on Linear State Space Models and the Kalman Filter". In: *Quantitative Finance* vol. 16 (10), pp. 1559–1573. Link (visited on 08/06/2019) (cit. on p. 32).

Dickey, David A. and Wayne A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series With a Unit Root". In: *Journal of the American Statistical Association* vol. 74 (366), pp. 427–431 (cit. on p. 22).

Do, Binh and Robert Faff (2010). "Does Simple Pairs Trading Still Work?" In: *Financial Analysts Journal; Charlottesville* vol. 66 (4), pp. 83–95. Link (visited on 05/26/2019) (cit. on pp. 16, 38).

Durbin, J. and G. S. Watson (1950). "Testing for Serial Correlation in Least Squares Regression: I". In: *Biometrika* vol. 37 (3/4), pp. 409–428 (cit. on p. 22).

— (1951). "Testing for Serial Correlation in Least Squares Regression. II". In: *Biometrika* vol. 38 (1/2), pp. 159–177 (cit. on p. 22).

Economist, The (2019). "March of the Machines". In: *The Economist; London* vol. 433 (9163), pp. 23–24, 26. Link (visited on 05/03/2020) (cit. on p. 11).

Ehrman, Douglas (2006). *The Handbook of Pairs Trading: Strategies Using Equities, Options, and Futures | Wiley*. Hoboken, NJ: John Wiley & Sons. 272 pp. (cit. on pp. 10, 16, 18).

Engle, Robert F. and C. W. J. Granger (1987). "Co-Integration and Error Correction: Representation, Estimation, and Testing". In: *Econometrica* vol. 55 (2), pp. 251–276 (cit. on pp. 21, 22).

Farago, Adam and Erik Hjalmarsson (2019). "Stock Price Co-Movement and the Foundations of Pairs Trading". In: *Journal of Financial & Quantitative Analysis* vol. 54 (2), pp. 629–665. Link (visited on 08/06/2019) (cit. on p. 25).

Ferreira, Luan (2008). "New Tools for Spread Trading". In: *Futures; Chicago* vol. 37 (12), pp. 38–41. Link (visited on 12/30/2019) (cit. on p. 28).

Figuerola-Ferretti, Isabel, Ioannis Paraskevopoulos, and Tao Tang (2018). "Pairs-Trading and Spread Persistence in the European Stock Market". In: *Journal of Futures Markets* vol. 38 (9), pp. 998–1023. Link (visited on 05/24/2020) (cit. on pp. 37, 45).

Focardi, Sergio M., Frank J. Fabozzi, and Ivan K. Mitov (2016). "A New Approach to Statistical Arbitrage: Strategies Based on Dynamic Factor Models of Prices and Their Performance". In: *Journal of Banking & Finance* vol. 65, pp. 134–155. Link (visited on 02/10/2019) (cit. on p. 14).

Gatev, Evan, William N. Goetzmann, and K. Geert Rouwenhorst (1999). "Pairs Trading: Performance of a Relative Value Arbitrage Rule". In: *NBER Working Paper Series; Cambridge* vol. P. 7032. Link (visited on 05/11/2019) (cit. on pp. 12, 14, 18, 39, 45).

— (2006). "Pairs Trading: Performance of a Relative-Value Arbitrage Rule". In: *The Review of Financial Studies* vol. 19 (3), pp. 797–827. Link (visited on 02/10/2019) (cit. on pp. 12, 15, 16, 19, 39).

Genest, Christian and Louis-Paul Rivest (1993). "Statistical Inference Procedures for Bivariate Archimedean Copulas". In: *Journal of the American Statistical Association* vol. 88 (423), pp. 1034–1043 (cit. on p. 26).

Grau-Carles, Pilar, Luis Miguel Doncel, and Jorge Sainz (2019). "Stability in Mutual Fund Performance Rankings: A New Proposal". In: *International Review of Economics & Finance* vol. 61, pp. 337–346. Link (visited on 02/16/2020) (cit. on p. 42).

Göncü, Ahmet and Erdinc Akyildirim (2016). ”A Stochastic Model for Commodity Pairs Trading”. In: *Quantitative Finance* vol. 16 (12), pp. 1843–1857. Link (visited on 08/06/2019) (cit. on pp. 15, 32).

Hain, Martin, Julian Hess, and Marliese Uhrig-Homburg (2018). ”Relative Value Arbitrage in European Commodity Markets”. In: *Energy Economics* vol. 69, pp. 140–154. Link (visited on 08/05/2019) (cit. on pp. 17, 24).

Hansen, Peter Reinhard (2005). ”A Test for Superior Predictive Ability”. In: *Journal of Business & Economic Statistics; Alexandria* vol. 23 (4), pp. 365–371 (cit. on p. 40).

Harju, Jens (2016). ”Pairs Trading Profitability in the Finnish Stock Market: A Comparison between Three Methods”. Master’s thesis. Lappeenranta: Lappeenranta University of Technology. Link (visited on 06/20/2020) (cit. on pp. 12, 13, 79).

Haug, Espen Gaarder and Nassim Nicholas Taleb (2011). ”Option Traders Use (Very) Sophisticated Heuristics, Never the Black–Scholes–Merton Formula”. In: *Journal of Economic Behavior & Organization* vol. 77 (2), pp. 97–106. Link (visited on 04/19/2020) (cit. on p. 28).

HE 88/2020, vp (2020). *Hallituksen esitys eduskunnalle vuoden 2020 neljänneksi lisätalousarvioksi*. Link (visited on 06/12/2020) (cit. on p. 10).

Hogan, Steve, Robert Jarrow, Melvyn Teo, and Mitch Warachka (2004). ”Testing Market Efficiency Using Statistical Arbitrage with Applications to Momentum and Value Strategies”. In: *Journal of Financial Economics* vol. 73 (3), pp. 525–565. Link (visited on 04/14/2019) (cit. on p. 14).

Hsu, Po-Hsuan, Yu-Chin Hsu, and Chung-Ming Kuan (2010). ”Testing the Predictive Ability of Technical Analysis Using a New Stepwise Test without Data Snooping Bias”. In: *Journal of Empirical Finance* vol. 17 (3), pp. 471–484. Link (visited on 05/12/2019) (cit. on p. 40).

Huang, Chin-Wen, Chun-Pin Hsu, and Wan-Jiun Paul Chiou (2015). ”Can Time-Varying Copulas Improve the Mean-Variance Portfolio?” In: *Handbook of Financial Econometrics and Statistics*. Ed. by Cheng-Few Lee and John C. Lee. New York, NY: Springer, pp. 233–251. Link (visited on 12/29/2019) (cit. on pp. 26, 27).

Huang, Wanling and Artem Prokhorov (2014). ”A Goodness-of-Fit Test for Copulas”. In: *Econometric Reviews* vol. 33 (7), pp. 751–771. Link (visited on 02/16/2020) (cit. on p. 56).

Huck, Nicolas (2010). ”Pairs Trading and Outranking: The Multi-Step-Ahead Forecasting Case”. In: *European Journal of Operational Research* vol. 207 (3), pp. 1702–1716. Link (visited on 02/10/2019) (cit. on pp. 15, 31).

— (2013). ”The High Sensitivity of Pairs Trading Returns”. In: *Applied Economics Letters* vol. 20 (14), pp. 1301–1304. Link (visited on 05/11/2019) (cit. on pp. 19–21).

Huck, Nicolas (2015). "Pairs Trading: Does Volatility Timing Matter?" In: *Applied Economics* vol. 47 (57), pp. 6239–6256. Link (visited on 05/11/2019) (cit. on pp. 15, 17, 20).

Huck, Nicolas and Komivi Afawubo (2015). "Pairs Trading and Selection Methods: Is Cointegration Superior?" In: *Applied Economics* vol. 47 (6), pp. 599–613. Link (visited on 08/06/2019) (cit. on p. 14).

Jacobs, Heiko and Martin Weber (2015). "On the Determinants of Pairs Trading Profitability". In: *Journal of Financial Markets* vol. 23, pp. 75–97. Link (visited on 02/10/2019) (cit. on p. 16).

Jensen, Michael C. (1968). "The Performance of Mutual Funds in the Period 1945-1964". In: *The Journal of Finance* vol. 23 (2), pp. 389–416 (cit. on p. 41).

Joe, Harry (1996). *Families of m-Variate Distributions with given Margins and $m(m-1)/2$ Bivariate Dependence Parameters*. Institute of Mathematical Statistics, pp. 120–141. Link (visited on 05/02/2020) (cit. on p. 27).

Joe, Harry and Taizhong Hu (1996). "Multivariate Distributions from Mixtures of Max-Infinitely Divisible Distributions". In: *Journal of Multivariate Analysis* vol. 57 (2), pp. 240–265. Link (visited on 05/02/2020) (cit. on pp. 27, 28).

Keating, C. and W. Shadwick (2002). "A Universal Performance Measure". In: *Journal of Performance Measurement* vol. 6 (3), pp. 59–84 (cit. on p. 41).

Kendall, M. G. (1938). "A New Measure of Rank Correlation". In: *Biometrika* vol. 30 (1/2), pp. 81–93 (cit. on p. 27).

Kharoubi-Rakotomalala, Cécile and Frantz Maurer (2013). "Copulas In Finance Ten Years Later". In: *Journal of Applied Business Research; Laramie* vol. 29 (5), n/a. Link (visited on 05/15/2019) (cit. on pp. 26, 28).

Kim, Saejoon and Jun Heo (2017). "Time Series Regression-Based Pairs Trading in the Korean Equities Market". In: *Journal of Experimental & Theoretical Artificial Intelligence* vol. 29 (4), pp. 755–768. Link (visited on 08/06/2019) (cit. on p. 15).

Kirchgässner, Gebhard, Jürgen Wolters, and Uwe Hassler (2013). *Introduction to Modern Time Series Analysis*. Second edition. Springer Texts in Business and Economics. Heidelberg: Springer. 319 pp. (cit. on pp. 22, 38).

Krauss, Christopher (2017). "Statistical Arbitrage Pairs Trading Strategies: Review and Outlook". In: *Journal of Economic Surveys* vol. 31 (2), pp. 513–545. Link (visited on 04/06/2019) (cit. on p. 15).

Krauss, Christopher, Xuan Anh Do, and Nicolas Huck (2017). "Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500". In: *European Journal of Operational Research* vol. 259 (2), pp. 689–702. Link (visited on 04/06/2019) (cit. on p. 11).

Krauss, Christopher and Johannes Stübinger (2017). "Non-Linear Dependence Modelling with Bivariate Copulas: Statistical Arbitrage Pairs Trading on the S&P 100". In: *Applied*

Economics vol. 49 (52), pp. 5352–5369. Link (visited on 08/06/2019) (cit. on pp. 27, 28, 31, 54).

Kuang, P., M. Schröder, and Q. Wang (2014). "Illusory Profitability of Technical Analysis in Emerging Foreign Exchange Markets". In: *International Journal of Forecasting* vol. 30 (2), pp. 192–205. Link (visited on 05/12/2019) (cit. on p. 40).

Kupiainen, Jukka (2008). "Pairs Trading -strategia Suomen osakemarkkinoilla". Master's thesis. Lappeenranta: Lappeenranta University of Technology. Link (visited on 06/20/2020) (cit. on p. 12).

Law, K.F., W.K. Li, and Philip L.H. Yu (2018). "A Single-Stage Approach for Cointegration-Based Pairs Trading". In: *Finance Research Letters* vol. 26, pp. 177–184. Link (visited on 04/06/2019) (cit. on p. 24).

Lei, Yaoting and Jing Xu (2015). "Costly Arbitrage through Pairs Trading". In: *Journal of Economic Dynamics and Control* vol. 56, pp. 1–19. Link (visited on 02/10/2019) (cit. on p. 17).

Leybourne, S. J. and B. P. M. McCabe (1994). "A Simple Test for Cointegration". In: *Oxford Bulletin of Economics & Statistics* vol. 56 (1), pp. 97–103. Link (visited on 05/05/2019) (cit. on p. 22).

Liew, Rong Qi and Yuan Wu (2013). "Pairs Trading: A Copula Approach". In: *Journal of Derivatives & Hedge Funds; Basingstoke* vol. 19 (1), pp. 12–30. Link (visited on 05/11/2019) (cit. on pp. 15, 27, 31, 56, 78).

Lintilhac, Paul Sopher and Agnès Tourin (2017). "Model-Based Pairs Trading in the Bitcoin Markets". In: *Quantitative Finance* vol. 17 (5), pp. 703–716. Link (visited on 08/06/2019) (cit. on p. 16).

Litterman, Robert (2011). "Who Should Hedge Tail Risk?" In: *Financial Analysts Journal; Charlottesville* vol. 67 (3), pp. 6, 9–11. Link (visited on 06/14/2020) (cit. on p. 10).

MacKinnon, James G. (1994). "Approximate Asymptotic Distribution Functions for Unit-Root and Cointegration Tests". In: *Journal of Business & Economic Statistics* vol. 12 (2), pp. 167–176. Link (visited on 12/01/2019) (cit. on p. 47).

Mikkelsen, Andreas (2018). "Pairs Trading: The Case of Norwegian Seafood Companies". In: *Applied Economics* vol. 50 (3), pp. 303–318. Link (visited on 08/06/2019) (cit. on pp. 17, 38).

Montana, Giovanni and Francesco Parrella (2009). "Data Mining for Algorithmic Asset Management". In: *Data Mining for Business Applications*. Ed. by Longbing Cao, Philip S. Yu, Chengqi Zhang, and Huaifeng Zhang. Boston, MA: Springer US, pp. 283–295. Link (visited on 08/05/2019) (cit. on pp. 15, 32).

Murthi, B. P. S., Yoon K. Choi, and Preyas Desai (1997). "Efficiency of Mutual Funds and Portfolio Performance Measurement: A Non-Parametric Approach". In: *European*

Journal of Operational Research vol. 98 (2), pp. 408–418. Link (visited on 02/16/2020) (cit. on p. 41).

Nagler, Thomas, C. Bumann, and C. Czado (2019a). "Model Selection in Sparse High-Dimensional Vine Copula Models with an Application to Portfolio Risk". In: *Journal of Multivariate Analysis* vol. 172, pp. 180–192. Link (visited on 08/01/2019) (cit. on p. 54).

Nagler, Thomas, Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, and Tobias Erhardt (2019b). *VineCopula: Statistical Inference of Vine Copulas*. Version of R package 2.3.0. Link (cit. on p. 54).

Neal, Robert (1996). "Direct Tests of Index Arbitrage Models". In: *The Journal of Financial and Quantitative Analysis* vol. 31 (4), pp. 541–562 (cit. on p. 14).

Nelsen, Roger B. (2006). *An Introduction to Copulas*. 2nd ed. Springer Series in Statistics. New York: Springer. 269 pp. (cit. on pp. 26, 27).

Nikoloulopoulos, Aristidis K., Harry Joe, and Haijun Li (2012). "Vine Copulas with Asymmetric Tail Dependence and Applications to Financial Return Data". In: *Computational Statistics & Data Analysis*. 1st Issue of the Annals of Computational and Financial Econometrics vol. 56 (11), pp. 3659–3673. Link (visited on 05/02/2020) (cit. on p. 27).

Oriola (2020). *Share Capital and Shares - Oriola*. Link (visited on 05/30/2020) (cit. on p. 48).

Osakekeisari (2018). *Pörssistä Poistuneet Yhtiöt*. Link (visited on 06/03/2020) (cit. on p. 34).

Rad, Hossein, Rand Kwong Yew Low, and Robert Faff (2016). "The Profitability of Pairs Trading Strategies: Distance, Cointegration and Copula Methods". In: *Quantitative Finance* vol. 16 (10), pp. 1541–1558. Link (visited on 05/11/2019) (cit. on pp. 16, 20, 31, 38).

Rinne, Kalle and Matti Suominen (2017). "How Some Bankers Made a Million by Trading Just Two Securities?" In: *Journal of Empirical Finance* vol. 44, pp. 304–315. Link (visited on 04/06/2019) (cit. on pp. 12, 13, 17, 65, 79).

Romano, Joseph P. and Michael Wolf (2005). "Stepwise Multiple Testing as Formalized Data Snooping". In: *Econometrica* vol. 73 (4), pp. 1237–1282 (cit. on p. 40).

Roy, Roch (1977). "On the Asymptotic Behaviour of the Sample Autocovariance Function for an Integrated Moving Average Process". In: *Biometrika* vol. 64 (2), pp. 419–421 (cit. on p. 21).

Ruxanda, Gheorghe and Sorin Opincariu (2018). "Bayesian Neural Networks with Dependent Dirichlet Process Priors. Application to Pairs Trading". In: *Economic Computation & Economic Cybernetics Studies & Research* vol. 52 (4), pp. 5–18. Link (visited on 08/06/2019) (cit. on p. 32).

Sharpe, William F. (1966). "Mutual Fund Performance". In: *The Journal of Business* vol. 39 (1), pp. 119–138 (cit. on p. 42).

— (1994). "The Sharpe Ratio". In: *Journal of Portfolio Management; New York* vol. 21 (1), p. 49. Link (visited on 06/02/2019) (cit. on p. 42).

Shen, Shih-yu and Andrew Minglong Wang (2001). ”On Stop-Loss Strategies for Stock Investments”. In: *Applied Mathematics and Computation* vol. 119 (2-3), pp. 317–337. Link (visited on 05/03/2020) (cit. on p. 18).

Smith, R. Todd and Xun Xu (2017). ”A Good Pair: Alternative Pairs-Trading Strategies”. In: *Financial Markets and Portfolio Management; New York* vol. 31 (1), pp. 1–26. Link (visited on 04/06/2019) (cit. on p. 17).

Song, Qingshuo and Qing Zhang (2013). ”An Optimal Pairs-Trading Rule”. In: *Automatica* vol. 49 (10), pp. 3007–3014. Link (visited on 02/10/2019) (cit. on p. 32).

SSAB (2020). *Tietoja osakkeesta*. Link (visited on 01/05/2020) (cit. on p. 43).

Stander, Yolanda, Daniel Marais, and Ilse Botha (2013). ”Trading Strategies with Copulas”. In: *Journal of Economic and Financial Sciences* vol. 6 (1), pp. 83–107. Link (visited on 05/11/2019) (cit. on pp. 25, 26).

Stübinger, Johannes and Jens Bredthauer (2017). ”Statistical Arbitrage Pairs Trading with High-Frequency Data”. In: *International Journal of Economics and Financial Issues; Mersin* vol. 7 (4). Link (visited on 04/06/2019) (cit. on pp. 17, 21).

Sullivan, Ryan, Allan Timmermann, and Halbert White (1999). ”Data-Snooping, Technical Trading Rule Performance, and the Bootstrap”. In: *The Journal of Finance* vol. 54 (5), pp. 1647–1691 (cit. on p. 40).

Taleb, Nassim Nicholas and Mark Blyth (2011). ”The Black Swan of Cairo”. In: *Foreign Affairs* vol. 90 (3), pp. 33–39. Link (visited on 06/12/2020) (cit. on p. 10).

Tianyong, H., D. Ming, and W. Liang (2013). ”Profitability of Pairs Trading Tactics in China’s Stock Market”. In: *2013 Suzhou-Silicon Valley-Beijing International Innovation Conference*. 2013 Suzhou-Silicon Valley-Beijing International Innovation Conference, pp. 111–115 (cit. on p. 16).

Tourin, Agnès and Raphael Yan (2013). ”Dynamic Pairs Trading Using the Stochastic Control Approach”. In: *Journal of Economic Dynamics and Control* vol. 37 (10), pp. 1972–1981. Link (visited on 02/10/2019) (cit. on p. 22).

Triantafyllopoulos, Kostas and Giovanni Montana (2011). ”Dynamic Modeling of Mean-Reverting Spreads for Statistical Arbitrage”. In: *Computational Management Science; Dordrecht* vol. 8 (1-2), pp. 23–49. Link (visited on 04/06/2019) (cit. on pp. 15, 32).

Watsham, Terry and Keith Parramore (1997). *Quantitative Methods in Finance*. 1st ed. London: Thomson Learning. 395 pp. (cit. on p. 21).

White, Halbert (2000). ”A Reality Check for Data Snooping”. In: *Econometrica* vol. 68 (5), pp. 1097–1126 (cit. on p. 40).

Vidyamurthy, Ganapathy (2004). *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons. 230 pp. (cit. on pp. 11, 23, 24, 45).

Xie, Wenjun, Rong Qi Liew, Yuan Wu, and Xi Zou (2016). ”Pairs Trading with Copulas”. In: *Journal of Trading; New York* vol. 11 (3), pp. 41–52. Link (visited on 05/11/2019) (cit. on pp. 15, 28, 39).

Yan-Xia Lin, Michael McCrae, and Chandra Gulati (2006). ”Loss Protection in Pairs Trading Through Minimum Profit Bounds: A Cointegration Approach”. In: *Journal of Applied Mathematics & Decision Sciences* vol. 2006 (4), pp. 1–14. Link (visited on 03/21/2020) (cit. on p. 21).

Yang, Jian, Juan Cabrera, and Tao Wang (2010). ”Nonlinearity, Data-Snooping, and Stock Index ETF Return Predictability”. In: *European Journal of Operational Research* vol. 200 (2), pp. 498–507. Link (visited on 05/12/2019) (cit. on p. 40).

Yu, Philip L.H. and Renjie Lu (2017). ”Cointegrated Market-Neutral Strategy for Basket Trading”. In: *International Review of Economics & Finance* vol. 49, pp. 112–124. Link (visited on 04/06/2019) (cit. on pp. 15, 24).

Appendix 1. Companies

Table A1.1. Listed companies on OMX Helsinki

Symbol	Name	Sector	Data available from
AFAGR	Afarak Group Oyj	Basic Materials	1997-01-02
ALBAV	Ålandsbanken Abp A	Financials	1997-01-02
ALBBV	Ålandsbanken Abp B	Financials	1997-01-02
AMEAS	Amer Sports Oyj	Consumer Goods	1997-01-02
APETIT	Apetit Oyj	Consumer Goods	1997-01-02
CRA1V	Cramo Oyj	Industrials	1997-01-02
CTY1S	Citycon Oyj	Financials	1997-01-02
EFO1V	Efore Oyj	Industrials	1997-01-02
ELEAV	Elecster Oyj A	Industrials	1997-01-02
FIA1S	Finnair Oyj	Consumer Services	1997-01-02
FSKRS	Fiskars Oyj Abp	Consumer Goods	1997-01-02
HONBS	Honkarakenne Oyj B	Consumer Goods	1997-01-02
HUH1V	Huhtamäki Oyj	Industrials	1997-01-02
ILK2S	Ilkka-Yhtymä Oyj 2	Consumer Services	1997-01-02
INVEST	Investors House Oyj	Financials	1997-01-02
KCR	Konecranes Oyj	Industrials	1997-01-02
KELAS	Kesla Oyj A	Industrials	1997-01-02
KEMIRA	Kemira Oyj	Basic Materials	1997-01-02
KESKOB	Kesko Oyj B	Consumer Services	1997-01-02
MARAS	Martela Oyj A	Consumer Goods	1997-01-02
METSA	Metsä Board Oyj A	Basic Materials	1997-01-02
METSB	Metsä Board Oyj B	Basic Materials	1997-01-02
NEO1V	Neo Industrial Oyj	Industrials	1997-01-02
NLG1V	Nurminen Logistics Oyj	Industrials	1997-01-02
NOKIA	Nokia Oyj	Technology	1997-01-02
NRE1V	Nokian Renkaat Oyj	Consumer Goods	1997-01-02
OLVAS	Olvi Oyj A	Consumer Goods	1997-01-02
OUT1V	Outokumpu Oyj	Basic Materials	1997-01-02
PKK1V	Pohjois-Karjalan Kirjapaino	Consumer Services	1997-01-02
PNA1V	Panostaja Oyj	Financials	1997-01-02
PON1V	Ponsse Oyj 1	Industrials	1997-01-02
RAIVV	Raisio Oyj Vaihto-osake	Consumer Goods	1997-01-02
RAUTE	Raute Oyj A	Industrials	1997-01-02
SAGCV	Saga Furs Oyj C	Consumer Goods	1997-01-02
SAMPO	Sampo Oyj A	Financials	1997-01-02

Continues on next page

Appendix 1. (continued)

Table A1.1 – *Continued from previous page*

Symbol	Name	Sector	Data available from
STCAS	Stockmann Oyj Abp A	Consumer Services	1997-01-02
STCBV	Stockmann Oyj Abp B	Consumer Services	1997-01-02
STEAV	Stora Enso Oyj A	Basic Materials	1997-01-02
STERV	Stora Enso Oyj R	Basic Materials	1997-01-02
TIETO	Tieto Oyj	Technology	1997-01-02
TULAV	Tulikivi Oyj A	Industrials	1997-01-02
UPM	UPM-Kymmene Oyj	Basic Materials	1997-01-02
UPONOR	Uponor Oyj	Industrials	1997-01-02
UUTEC	Plc Uutechnic Group Oyj	Industrials	1997-01-02
VAIAS	Vaisala Oyj A	Industrials	1997-01-02
VIK1V	Viking Line Abp	Consumer Services	1997-01-02
WRT1V	Wärtsilä Oyj Abp	Industrials	1997-01-02
YEINT	Yleiselektroniikka Oyj	Industrials	1997-01-02
YIT	YIT Oyj	Industrials	1997-01-02
HKSAV	HKScan Oyj A	Consumer Goods	1997-02-06
ICP1V	Incap Oyj	Industrials	1997-05-05
ATRAV	Atria Oyj A	Consumer Goods	1997-06-06
POY1V	Pöyry Oyj	Industrials	1997-12-02
RAMI	Ramirent Oyj	Industrials	1998-04-30
VALOE	Valoe Oyj	Industrials	1998-05-15
BITTI	Bittium Oyj	Technology	1998-09-15
EXL1V	Exel Composites Oyj	Industrials	1998-10-19
RAP1V	Rapala VMC Oyj	Consumer Goods	1998-12-04
FORTUM	Fortum Oyj	Utilities	1998-12-18
MMO1V	Marimekko Oyj	Consumer Goods	1999-03-12
IFA1V	Innofactor Plc	Technology	1999-03-15
TLT1V	Teleste Oyj	Technology	1999-03-30
KSLAV	Keskisuomalainen Oyj A	Consumer Services	1999-04-19
SAA1V	Sanoma Oyj	Consumer Services	1999-05-03
KESKOA	Kesko Oyj A	Consumer Services	1999-06-01
TPS1V	Technopolis Oyj	Financials	1999-06-08
BIOBV	Biohit Oyj B	Health Care	1999-06-18
ELISA	Elisa Oyj	Telecommunications	1999-07-01
METSO	Metso Oyj	Industrials	1999-07-01
SOLTEQ	Solteq Oyj	Technology	1999-09-06

Continues on next page

Appendix 1. (continued)

Table A1.1 – *Continued from previous page*

Symbol	Name	Sector	Data available from
DIGIA	Digia Oyj	Technology	1999-09-27
DIGIGR	Digitalist Group Oyj	Technology	1999-09-28
ASPO	Aspo Oyj	Industrials	1999-10-01
ACG1V	Aspocomp Group Oyj	Industrials	1999-10-04
DOV1V	Dovre Group Oyj	Industrials	1999-10-15
FSC1V	F-Secure Oyj	Technology	1999-11-05
NDA FI	Nordea Bank Abp	Financials	2000-01-31
BAS1V	Basware Oyj	Technology	2000-02-29
TRH1V	Trainers' House Oyj	Technology	2000-03-15
ETTE	Etteplan Oyj	Industrials	2000-04-27
SIEVI	Sievi Capital Oyj	Financials	2000-05-22
WUF1V	Wulff-Yhtiöt Oyj	Industrials	2000-10-09
EQV1V	eQ Oyj	Financials	2000-11-01
SSH1V	SSH Communications Security	Technology	2000-12-20
CTH1V	Componenta Oyj	Industrials	2001-03-20
CAPMAN	CapMan Oyj	Financials	2001-04-02
GLA1V	Glaston Oyj Abp	Industrials	2001-04-02
TEM1V	Tecnotree Oyj	Technology	2001-04-02
LAT1V	Lassila & Tikanoja Oyj	Industrials	2001-10-01
REG1V	Revenio Group Oyj	Health Care	2001-10-01
SUY1V	Suominen Oyj	Consumer Goods	2001-10-01
QPR1V	QPR Software Oyj	Technology	2002-03-08
TELIA1	Telia Company	Telecommunications	2002-05-03
NESTE	Neste Oyj	Oil & Gas	2005-04-18
ALMA	Alma Media Oyj	Consumer Services	2005-04-29
CGCBV	Cargotec Oyj	Industrials	2005-06-01
KNEBV	KONE Oyj	Industrials	2005-06-01
OKDAV	Oriola Oyj A	Health Care	2006-07-03
OKDBV	Oriola Oyj B	Health Care	2006-07-03
ORNAV	Orion Oyj A	Health Care	2006-07-03
ORNBV	Orion Oyj B	Health Care	2006-07-03
OTE1V	Outotec Oyj	Industrials	2006-10-10
SRV1V	SRV Yhtiöt Oyj	Industrials	2007-06-12
SOPRA	Soprano Oyj	Technology	2007-11-05
AKTIA	Aktia Bank Abp	Financials	2009-09-29

Continues on next page

Appendix 1. (continued)

Table A1.1 – *Continued from previous page*

Symbol	Name	Sector	Data available from
TIK1V	Tikkurila Oyj	Industrials	2010-03-26
SCANFL	Scanfil Oyj	Industrials	2012-01-02
SOSI1	Sotkamo Silver AB	Basic Materials	2012-07-17
SIILI	Siili Solutions Oyj	Technology	2012-10-15
TAALA	Taaleri Oyj	Financials	2013-04-24
ENDOM	Endomines	Basic Materials	2013-05-14
AM1	Ahlstrom-Munksjö Oyj	Basic Materials	2013-06-07
CAV1V	Caverion Oyj	Industrials	2013-07-01
OVARO	Ovaro Kiinteistösijoitus Oyj	Financials	2013-10-14
NOHO	NoHo Partners Oyj	Consumer Services	2013-11-28
VALMT	Valmet Oyj	Industrials	2014-01-02
SSABAH	SSAB A	Basic Materials	2014-08-01
SSABBH	SSAB B	Basic Materials	2014-08-01
NIXU	Nixu Oyj	Technology	2014-12-05
ATG1V	Asiakastieto Group Oyj	Financials	2015-03-27
ROBIT	Robit Oyj	Industrials	2015-05-21
PIHLIS	Pihlajalinna Oyj	Health Care	2015-06-04
TNOM	Talenom Oyj	Industrials	2015-06-11
PIZZA	Kotipizza Group Oyj	Consumer Services	2015-07-07
EVLI	Evli Pankki Oyj	Financials	2015-12-02
CONSTI	Consti Yhtiöt Oyj	Industrials	2015-12-11
HOIVA	Suomen Hoivatilat Oyj	Financials	2016-03-31
LEHTO	Lehto Group Oyj	Industrials	2016-04-28
TOKMAN	Tokmanni Group Oyj	Consumer Services	2016-04-29
QTCOM	Qt Group Oyj	Technology	2016-05-02
DNA	DNA Oyj	Telecommunications	2016-11-30
KAMUX	Kamux Oyj	Consumer Services	2017-05-12
SILMA	Silmäasema Oyj	Health Care	2017-06-09
ROVIO	Rovio Entertainment Oyj	Consumer Goods	2017-09-29
TTALO	Terveystalo Oyj	Health Care	2017-10-11
HARVIA	Harvia Oyj	Consumer Goods	2018-03-22
ALTIA	Altia Oyj	Consumer Goods	2018-03-23
KOJAMO	Kojamo Oyj	Financials	2018-06-15
OMASP	Oma Säästöpankki Oyj	Financials	2018-11-30
TALLINK	AS Tallink Grupp FDR	Consumer Services	2018-12-03

Appendix 2. Removed companies

Table A2.1. List of companies removed from OMX Helsinki

Date	Company	Reason
2018-03-15	Ahtium Oyj	bankruptcy
2018-02-21	Affecto Oyj	acquisition (CGI Nordic Investments Limited)
2018-01-31	Lemminkäinen Oyj	merger (YIT)
2017-07-09	PKC Group	acquisition (Motherson Sumi)
2017-06-29	Comptel	acquisition (Nokia)
2017-06-09	Norvestia	merger (CapMan)
2017-05-12	Sponda	acquisition (Polar Bidco)
2017-03-04	Ahlström Oyj	merger (Munksjö)
2016-09-30	Biotie Therapies	acquisition (Acorda Therapeutics)
2016-08-25	Finnlines	acquisition (Grimaldi Group)
2016-08-11	Okmetic	acquisition (NSIG)
2015-05-18	Vacon	acquisition (Danfoss)
2015-04-09	Turvatiimi	acquisition (Atine Group)
2014-12-19	Oral Hammaslääkärit	acquisition (Renideo Group)
2014-11-20	Rautaruukki	acquisition (SSAB)
2014-01-10	Pohjola Pankki	acquisition (OP-Pohjola)
2013-11-21	Stonesoft	acquisition (McAfee)
2013-10-28	GeoSentric	unknown
2013-10-10	Tiimari	bankruptcy
2012-12-21	Interavanti	acquisition
2012-12-15	Nordic Aluminium	acquisition (Lival)
2012-08-08	Aldata Solution	acquisition (Symphony Technology)
2012-02-13	Tekla	acquisition (Trimble Finland)
2011-11-17	Elcoteq SE	bankruptcy
2011-09-23	Salcomp	acquisition (Nordstjernan)
2010-11-06	Larox	acquisition (Outotec)
2010-05-05	Tamfelt	acquisition (Metso)
2010-03-19	Julius Tallberg-Kiinteistöt	unknown
2009-08-26	Terveystalo Healthcare	acquisition (Star Healthcare)
2009-05-14	Rocla	acquisition (Mitsubishi Caterpillar Forklift E...
2009-02-07	Evia	bankruptcy
2008-11-12	Stromsdal	bankruptcy
2008-07-04	Kemira GrowHow	acquisition (Yara Nederland B.V.)
2008-05-29	Birka Line	acquisition (Rederiaktiebolaget Eckerö)

Appendix 2. (continued)

Date	Company	Reason
2008-05-05	OMX	acquisition (NASDAQ OMX Group)
2008-03-17	Perlos	acquisition (Lite-On)
2007-12-21	eQ	acquisition (Straumur-Burdaras)
2007-08-20	FIM	acquisition (Glitnir)
2007-08-05	Kylpyläkasino	acquisition (Restel)
2007-06-27	Puuharyhmä	acquisition (Aspro Ocio S.A.)
2007-04-09	Evox Rifa Group	acquisition (KEMET Electronics)
2006-09-27	Sentera	acquisition (SysOpen Digia)
2006-09-14	Fortum Espoo	acquisition (Fortum)
2006-06-14	Pohjola	acquisition (OKO)
2006-05-22	Suomen Spar	acquisition (SOK)
2006-03-29	Saunalahti Group	acquisition (Elisa)
2006-01-12	Kekkilä	acquisition (Vapo)
2005-09-15	Plandent	acquisition (Planmeca)
2005-07-11	Alma Media	merger (Almanova)
2005-06-21	Chips	acquisition (Orkla ASA)
2005-02-03	Turun Arvokiinteistöt	acquisition (Nordea)
2004-12-31	Yomi	merger (Elisa)
2004-11-05	Hackman	acquisition (Alifin Oy)
2004-11-05	Tamro	acquisition (Phoenix International Beteiligung...
2004-06-28	Polar Kiinteistöt	acquisition (IVG Immobiliére SAS)
2004-04-27	WM-data Novo	acquisition (WM-data AB)
2004-03-29	Janton	acquisition (BACPE Finland Holdings)
2004-01-23	Instrumentarium	acquisition (General Electric Finland)

Appendix 3. Trading periods

Table A3.1. Analyzed periods

	Fitting period	Trading period
1	January 1, 2004 to December 31, 2004	January 1, 2005 to July 2, 2005
2	March 25, 2004 to March 25, 2005	March 26, 2005 to September 24, 2005
3	June 17, 2004 to June 17, 2005	June 18, 2005 to December 17, 2005
4	September 9, 2004 to September 9, 2005	September 10, 2005 to March 11, 2006
5	December 2, 2004 to December 2, 2005	December 3, 2005 to June 3, 2006
6	February 24, 2005 to February 24, 2006	February 25, 2006 to August 26, 2006
7	May 19, 2005 to May 19, 2006	May 20, 2006 to November 18, 2006
8	August 11, 2005 to August 11, 2006	August 12, 2006 to February 10, 2007
9	November 3, 2005 to November 3, 2006	November 4, 2006 to May 5, 2007
10	January 26, 2006 to January 26, 2007	January 27, 2007 to July 28, 2007
11	April 20, 2006 to April 20, 2007	April 21, 2007 to October 20, 2007
12	July 13, 2006 to July 13, 2007	July 14, 2007 to January 12, 2008
13	October 5, 2006 to October 5, 2007	October 6, 2007 to April 5, 2008
14	December 28, 2006 to December 28, 2007	December 29, 2007 to June 28, 2008
15	March 22, 2007 to March 21, 2008	March 22, 2008 to September 20, 2008
16	June 14, 2007 to June 13, 2008	June 14, 2008 to December 13, 2008
17	September 6, 2007 to September 5, 2008	September 6, 2008 to March 7, 2009
18	November 29, 2007 to November 28, 2008	November 29, 2008 to May 30, 2009
19	February 21, 2008 to February 20, 2009	February 21, 2009 to August 22, 2009
20	May 15, 2008 to May 15, 2009	May 16, 2009 to November 14, 2009
21	August 7, 2008 to August 7, 2009	August 8, 2009 to February 6, 2010
22	October 30, 2008 to October 30, 2009	October 31, 2009 to May 1, 2010
23	January 22, 2009 to January 22, 2010	January 23, 2010 to July 24, 2010
24	April 16, 2009 to April 16, 2010	April 17, 2010 to October 16, 2010
25	July 9, 2009 to July 9, 2010	July 10, 2010 to January 8, 2011
26	October 1, 2009 to October 1, 2010	October 2, 2010 to April 2, 2011
27	December 24, 2009 to December 24, 2010	December 25, 2010 to June 25, 2011
28	March 18, 2010 to March 18, 2011	March 19, 2011 to September 17, 2011
29	June 10, 2010 to June 10, 2011	June 11, 2011 to December 10, 2011
30	September 2, 2010 to September 2, 2011	September 3, 2011 to March 3, 2012
31	November 25, 2010 to November 25, 2011	November 26, 2011 to May 26, 2012
32	February 17, 2011 to February 17, 2012	February 18, 2012 to August 18, 2012
33	May 12, 2011 to May 11, 2012	May 12, 2012 to November 10, 2012
34	August 4, 2011 to August 3, 2012	August 4, 2012 to February 2, 2013
35	October 27, 2011 to October 26, 2012	October 27, 2012 to April 27, 2013

Continues on next page

Appendix 3. (continued)

Table A3.1 – *Continued from previous page*

	Fitting period	Trading period
36	January 19, 2012 to January 18, 2013	January 19, 2013 to July 20, 2013
37	April 12, 2012 to April 12, 2013	April 13, 2013 to October 12, 2013
38	July 5, 2012 to July 5, 2013	July 6, 2013 to January 4, 2014
39	September 27, 2012 to September 27, 2013	September 28, 2013 to March 29, 2014
40	December 20, 2012 to December 20, 2013	December 21, 2013 to June 21, 2014
41	March 14, 2013 to March 14, 2014	March 15, 2014 to September 13, 2014
42	June 6, 2013 to June 6, 2014	June 7, 2014 to December 6, 2014
43	August 29, 2013 to August 29, 2014	August 30, 2014 to February 28, 2015
44	November 21, 2013 to November 21, 2014	November 22, 2014 to May 23, 2015
45	February 13, 2014 to February 13, 2015	February 14, 2015 to August 15, 2015
46	May 8, 2014 to May 8, 2015	May 9, 2015 to November 7, 2015
47	July 31, 2014 to July 31, 2015	August 1, 2015 to January 30, 2016
48	October 23, 2014 to October 23, 2015	October 24, 2015 to April 23, 2016
49	January 15, 2015 to January 15, 2016	January 16, 2016 to July 16, 2016
50	April 9, 2015 to April 8, 2016	April 9, 2016 to October 8, 2016
51	July 2, 2015 to July 1, 2016	July 2, 2016 to December 31, 2016
52	September 24, 2015 to September 23, 2016	September 24, 2016 to March 25, 2017
53	December 17, 2015 to December 16, 2016	December 17, 2016 to June 17, 2017
54	March 10, 2016 to March 10, 2017	March 11, 2017 to September 9, 2017
55	June 2, 2016 to June 2, 2017	June 3, 2017 to December 2, 2017
56	August 25, 2016 to August 25, 2017	August 26, 2017 to February 24, 2018
57	November 17, 2016 to November 17, 2017	November 18, 2017 to May 19, 2018
58	February 9, 2017 to February 9, 2018	February 10, 2018 to August 11, 2018
59	May 4, 2017 to May 4, 2018	May 5, 2018 to November 3, 2018
60	July 27, 2017 to July 27, 2018	July 28, 2018 to January 26, 2019
61	October 19, 2017 to October 19, 2018	October 20, 2018 to April 20, 2019
62	January 11, 2018 to January 11, 2019	January 12, 2019 to July 13, 2019
63	April 5, 2018 to April 5, 2019	April 6, 2019 to October 5, 2019
64	June 28, 2018 to June 28, 2019	June 29, 2019 to December 28, 2019
65	September 20, 2018 to September 20, 2019	September 21, 2019 to March 21, 2020
66	December 1, 2018 to December 1, 2019	December 2, 2019 to May 31, 2020

Appendix 4. Example distance pairs

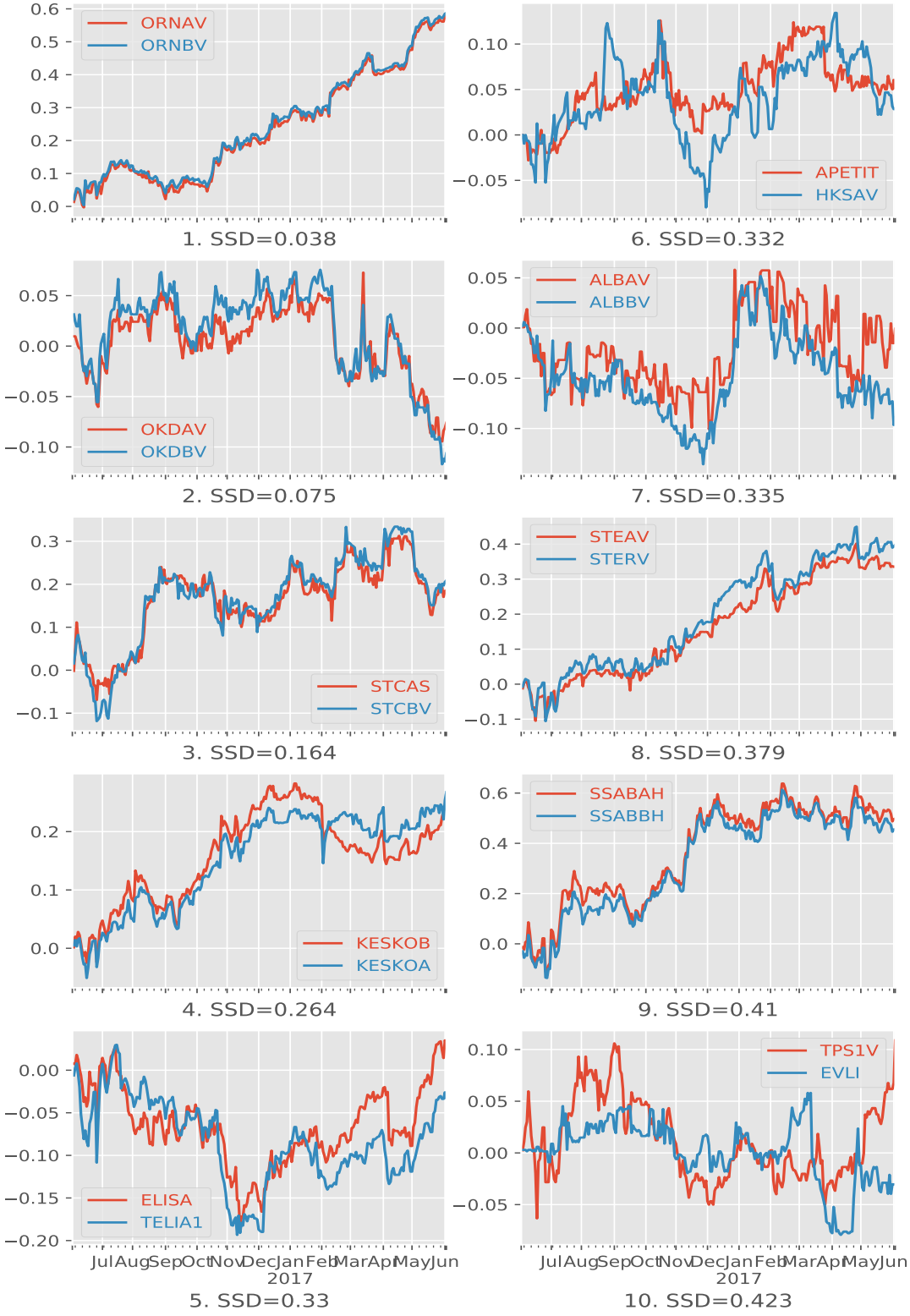


Figure A4.1. Top 10 pairs with the lowest sum of squared differences on trading period 55

Appendix 4. (continued)

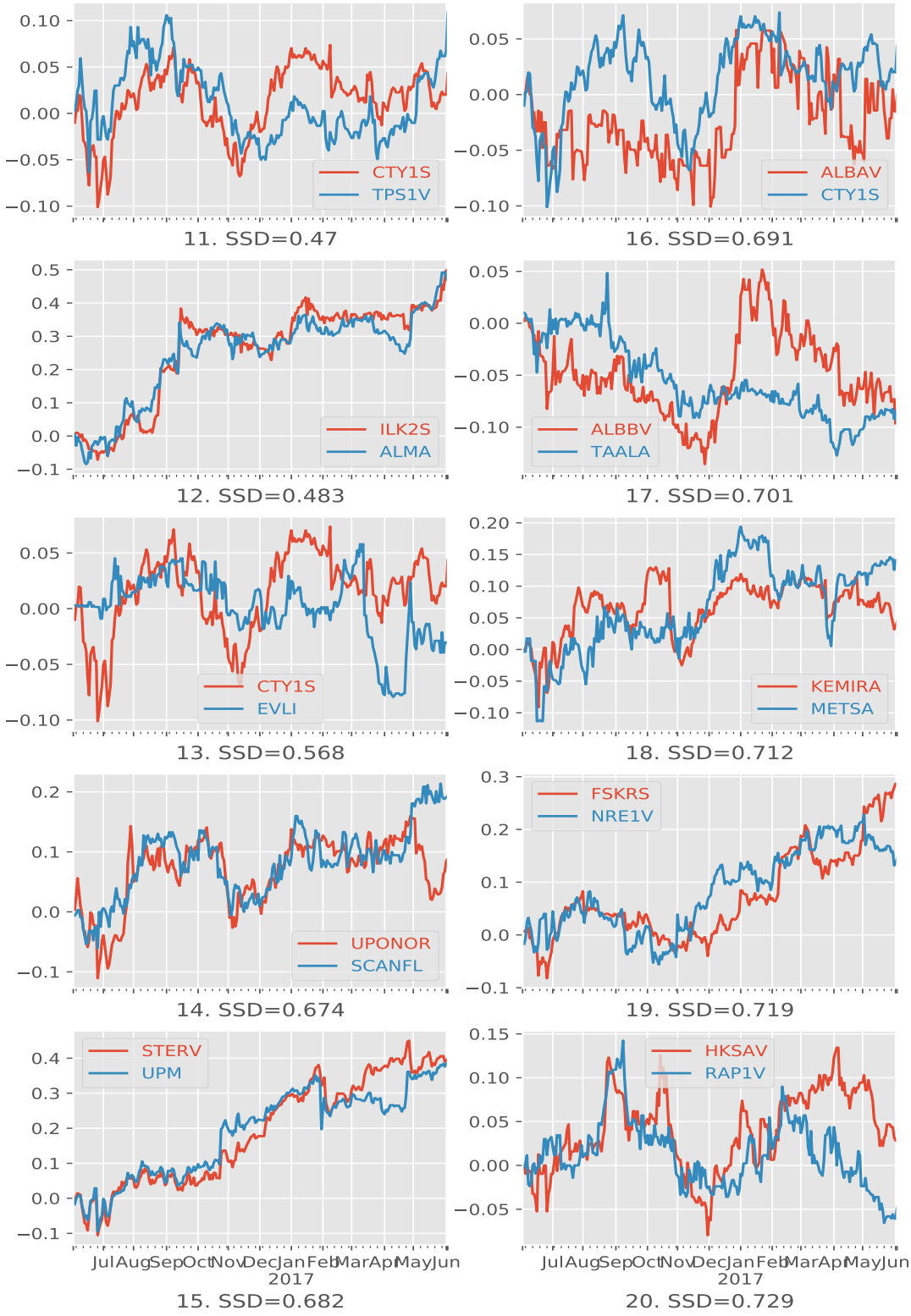


Figure A4.2. Pairs 11-20 with the lowest sum of squared differences on trading period 55

Appendix 5. Example cointegration pairs

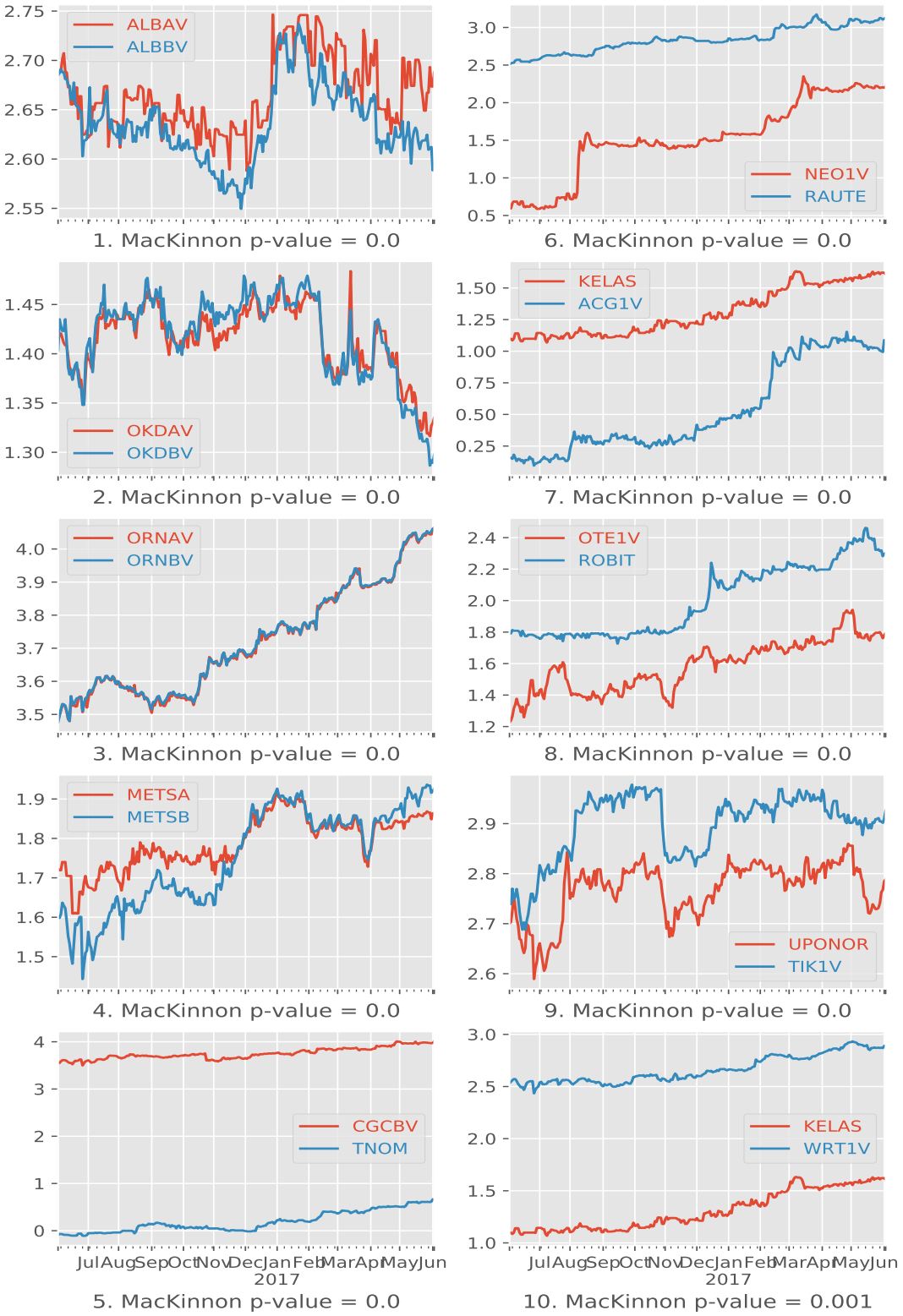


Figure A5.1. Top 10 pairs with the lowest MacKinnon p-value on trading period 55

Appendix 5. (continued)

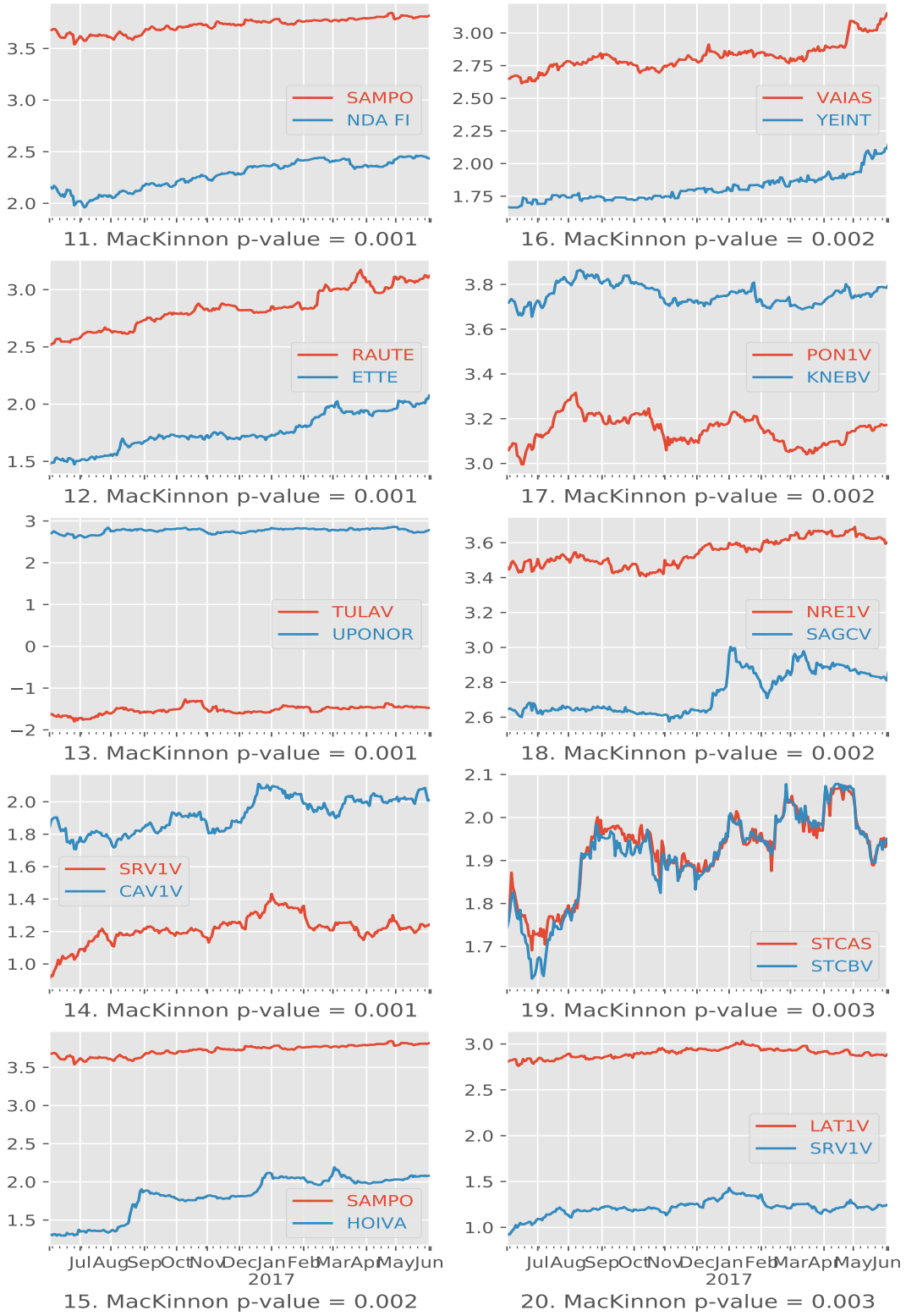


Figure A5.2. Pairs 11-20 with the lowest MacKinnon p-value on trading period 55