Swedish Equities; Casanovas or committed Cointegrated partners

By Alexander Fors & Ossian Markiewicz

Bachelor’s Thesis
Department of Statistics
Uppsala University

Supervisor: Lars Forsberg

Spring 2016
Abstract

This thesis investigates the long-run stability of Cointegrated pairs in the Swedish Equity Market. Stability is evaluated by estimating pairs in an in-sample period then rolling the window forward. A Pairs Trading strategy is then applied to the estimated pairs and traded out-of-sample. The relationships are found to diminish over time and most break off. Negative compound annual growth rates are obtained for the period. However there are enough lasting cointegrating relationships for the strategy to be applicable but the returns are highly dependent on the complexity of the trading rules.

Keywords: Cointegration, Stability, Mean Reversion, Pairs Trading.
Contents

1 Introduction 4

2 Theory 6
   2.1 Statistical Theory ............................................. 6
      2.1.1 Stationarity and Integration ............................. 6
      2.1.2 Augmented Dickey-Fuller’s test ......................... 7
      2.1.3 Cointegration and Error Correction mechanism .......... 8
      2.1.4 VAR .................................................. 10
      2.1.5 Test for Cointegration .................................. 12
      2.1.6 Johansen’s Approach .................................... 13
   2.2 Financial Theory ............................................... 14
      2.2.1 Pairs Trading ........................................... 14
      2.2.2 Calculate Return ....................................... 15
      2.2.3 Sharpe Ratio ........................................... 16
      2.2.4 Constructing Confidence Intervals ...................... 17

3 Method 18
   3.1 Data ...................................................... 18
   3.2 Finding Cointegrated Pairs .................................. 18
   3.3 Stability of Cointegrating Relationships ................... 19
   3.4 Trading the two Portfolios .................................. 20

4 Results 21
   4.1 Evaluating Stability ......................................... 21
   4.2 Portfolio Performance ....................................... 23

5 Discussion 24
   5.1 Conclusions .................................................. 25
   5.2 Future research ............................................... 25

6 References 26

7 Appendix 28
1 Introduction

Cointegration can be thought of as a stationary relationship between two non-stationary series (Marmol & Velasco, 2004). When two variables share the same stochastic and deterministic trends, there is a linear combination between the two variables that cancels both trends. The resulting cointegrating relation is not trending despite the fact that the variables by themselves are.

The most defining property of a cointegrated series is that of mean reversion (Vidya-murthy). A mean reverting process oscillates around a steady-state value over time. If it were to depart from the long-run level of equilibrium it would eventually snap back. Cointegration and the mean reverting property have proven to be a useful tool for econometricians and time-series analysts.

In the mid 80’s a group of mathematicians, computer scientists and physicists convened at Morgan Stanley to develop arbitrage strategies using statistical techniques (Ibid). One such group of techniques came to be known as ”Pairs Trading”. This is in its essence an umbrella term of trading strategies designed to find pairs of securities whose prices moved together. If the prices were to depart from the equilibrium, positions would be taken based on the idea that they eventually correct themselves.

Cointegration theory is a very fitting concept for Pairs Trading (Alexander & Dimitriu, 2002). In Pairs Trading, the existence of a cointegrating relationship between two equities’ prices is a proof of such a long-term relationship. This implies that the two prices follow the same long-term trend and that the price differentials are cancelled leaving a stationary series (Hendry & Juselius, 2001). Alexander et al. showed in their study from 2001 that the profit potential is highly dependent on the presence of the long-term equilibrium, or in other words, the presence of cointegrated pairs of equities. Buncic & Roca (2005) finds that the relationships are more or less stable depending on the time period that is examined. A pair that is cointegrated in one period doesn’t necessarily have to continue its relationship in another time frame.

The purpose of this thesis is two fold and aims to answer the following questions: Are the cointegrating relationships stable over time? And if they are, is Pairs Trading a profitable strategy? We study how stable cointegrating relationships are between equities in a real trading environment. Three portfolios are constructed from equities on Large Cap
in the Swedish Equity market between 2005 to 2015. A market portfolio containing all cointegrated pairs and two portfolios of the 20 and 10 highest correlated and cointegrated pairs. A Pairs Trading strategy is then applied to the 20 and 10 portfolios and compared to the OMXS30 index.

The outline of this thesis is as follows: In chapter two the theoretical framework needed to answer our research questions is presented. Chapter three explains the method that has been used. In the fourth chapter the results are presented. And lastly, chapter five where the results are discussed, conclusions are made and suggestions for future research proposed. The R-code that we used can be seen in the Appendix.
2 Theory

In this section the theoretical framework is divided into a statistical and financial part. The first, giving an overview of the statistical tools necessary to answer our research question and purpose and the second to give a theoretical background to Pairs Trading.

2.1 Statistical Theory

2.1.1 Stationarity and Integration

The covariance-stationary time series is characterised by a constant mean and constant covariance, only dependent on lag length $k$ (Asteriou & Hall). This form of stationarity only considers the first two moments and will from now on be termed as just stationarity. For a time series $Y_t$ the two first moments are

$$E(Y_t) = \mu, \forall t, \quad (2.1)$$

$$\text{Cov}(Y_t, Y_{t+k}) = \gamma_k, \forall t \text{ and } k. \quad (2.2)$$

Covariance only dependent on lag length $k$, implies a constant variance

$$\text{Var}(Y_t) = \sigma^2, \forall t. \quad (2.3)$$

When these three conditions are fulfilled, a series possess what is called a mean reverting property (Ibid.). If the series were to wander off, the property will eventually cause it to revert back to the long-run equilibrium. As a result the time series oscillate around a constant mean value over time. This is shown mathematically by considering an autoregressive (AR) model of the first order

$$Y_t = \phi Y_{t-1} + u_t, \quad (2.4)$$

where $|\phi| < 1$ and $u_t$ is iid $(0, \sigma_u^2)$. Here, the present value of variable $Y_t$ is affected by its own lag, the subsequent period $Y_{t-1}$ with the coefficient $\phi$, and by the stochastic error term $u_t$. Since both mean and variance for $u_t$ is constant its distribution is identical at all points in time. Such a process is known as "white-noise" (Cryer & Chan). Shown by
sequentially expanding $Y_{t-k}$ in (2.4) to an infinite MA($\infty$) process

$$Y_t = u_t + \phi u_{t-1} + \phi^2 u_{t-2} + \phi^3 u_{t-3} + ...$$

(2.5)

For $|\phi| < 1$ the effects of the error terms will dissipate over time and the series will revert back to its long-run equilibrium. If $\phi = 1$ it would in turn imply a non-stationary time series as the effect of each $u_{t-k}$ never dies out. This is called a unit root process.

A non-stationary time series can be transformed into a stationary one. This is achieved by taking the first difference of the series until it becomes stationary. A series that is stationary after differencing it $d$ times is said to be integrated of order $d$. The notation for an integrated series of order $d$ is, $I(d)$.

### 2.1.2 Augmented Dickey-Fuller’s test

Dickey & Fuller devised a formal test to detect non-stationarity among time series, the Dickey-Fuller (DF) test. Their key insight was that proving non-stationarity is equivalent to testing for the presence of a unit root. Because the error terms in the original test-model rarely was white-noise Dickey & Fuller suggested an augmented version of the test i.e. the Augmented Dickey-Fuller (ADF) test. This model meant to eliminate the autocorrelation by including extra lagged terms of the dependent variable

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t,$$

(2.6)

where $\gamma = (\theta - 1)$ with the null hypothesis $H_0 : \gamma = 0$ tested against the alternative $H_a : \gamma < 0$. If $\gamma < 0$ it implies that $|\theta| < 1$, as a result $y_t$ would be stationary. However if $\gamma = 0$ then $y_t$ would be a random walk. The $\sum_{i=1}^{p} \beta_i \Delta y_{t-i}$ component represents the sum of all extra lagged terms with their corresponding coefficients $\beta_i$ (Dickey & Fuller, 1979). Two additional regression equations can be used to test for the existence of a unit root

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t,$$

(2.7)

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + a_2 t + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t.$$

(2.8)
Equation (2.7) contains a constant $\alpha_0$, allowing for an intercept in the time series. In (2.8) the constant is still included with an additional term $a_2 t$ representing a trend in the series. In addition to choosing the appropriate model, the correct number of autoregressive lags needs to be estimated. For a correctly specified model the test statistic is obtained by performing a t-test,

$$ADF_{obs} = \frac{\hat{\gamma}}{\hat{\sigma}_\gamma},$$

the estimated coefficient $\gamma$ is compared to the critical value from Dickey-Fuller’s distribution.

According to Shiller & Perron (1985) the power of the unit root test depend on the time span of the data rather than the frequency of the observations. However it has been shown in later research that the power of the ADF test is improved with higher sampling frequencies for stock data i.e. daily data are better than monthly and so on (Choi & Chung, 1995).

### 2.1.3 Cointegration and Error Correction mechanism

Two time series $Y_t$ and $X_t$ are said to be cointegrated if they contain a unit root while a linear combination between them exists that eliminates the non-stationarity. In this case $Y_t$ and $X_t$ are cointegrated of order $d$ and $b$, $Y_t \sim I(d)$ and $X_t \sim I(b)$, where $d \geq b > 0$. If $d = b$ the linear combination of the two is integrated of order $d - b$ hence equal to zero (Asteriou & Hall)

$$\theta_1 Y_t + \theta_2 X_t = u_t \sim I(0).$$

The coefficients $\theta_1$ and $\theta_2$ are called the coefficients vector and represent the linear combination of $Y_t$ and $X_t$ that generates a stationary time series. In the presence of cointegration, interest generally lies in the long-run relationship between the two time series, also called the equilibrium value. Through transformation of equation (2.10) to

$$Y_t = -\frac{\theta_2}{\theta_1} X_t + \epsilon_t,$$

the equilibrium value of the relationship is obtained.
If cointegration is present any shock that would put it in disequilibrium will be corrected through an error correction mechanism (ECM). The ECM handles both short- and long-term shocks to the model. To understand the mechanism, ECM can be mathematically derived, from \( Y_t \sim I(1) \) and \( X_t \sim I(1) \). First \( Y_t \) is regressed upon \( X_t \) (Ibid.)

\[
Y_t = \beta_0 + \beta_1 X_t + u_t. \tag{2.12}
\]

Such a regression could explain a significant amount of the variation in \( Y_t \) but may be completely spurious. This can be solved by taking first difference of both series resulting in \( \Delta Y_t \sim I(0) \) and \( \Delta X_t \sim I(0) \) which ensures stationarity in model

\[
\Delta Y_t = a_0 + a_1 \Delta X_t + \Delta u_t. \tag{2.13}
\]

Now the spurious issue is resolved and the coefficients \( a_0, a_1 \) can be estimated. Both coefficients are correct but equation (2.13) only explains the short-term relationship. In the case of \( Y_t \) and \( X_t \) being cointegrated,

\[
Y_t + \hat{\beta}_0 + \hat{\beta}_1 X_t = u_t \sim I(0), \tag{2.14}
\]

is the linear combination between the two. This model is not spurious and describes the long-run relationship. Thus combining equation (2.13) and (2.14),

\[
\Delta Y_t = a_0 + a_1 \Delta X_t - \pi(Y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1 X_{t-1}) + e_t, \tag{2.15}
\]

both the short- and long-run behaviour is included in the model, where \( \Delta u_t = \pi(Y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1 X_{t-1}) + e_t \). Here \( a_1 \) is the multiplier of any short-run impact on \( X_t \) and its effects on the change in \( Y_t \). The speed of correction in the case of disequilibrium is represented by \( \pi \) equal to \( (1 - a_0) \). When the model is in equilibrium \( (Y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1 X_{t-1}) \) will be zero. Through reparametrization and rearrangement of equation (2.15),

\[
Y_t - Y_{t-1} = a_0 + a_1 X_t - a_1 X_{t-1} - Y_{t-1} + a_0 Y_{t-1} - \pi \hat{\beta}_0 + \pi \hat{\beta}_1 X_{t-1} + e_t, \tag{2.16}
\]

\[
Y_t = a_0 - \pi \hat{\beta}_0 + a_0 Yt - 1 + a_1 X_t + (\pi \hat{\beta}_1 - a_1) X_{t-1} + e_t, \tag{2.17}
\]
and by simplifying notation \( a_0 - \pi \hat{\beta}_0 = \delta_0 \) and \( \pi \hat{\beta}_1 - a_1 = a_2, \)

\[
Y_t = \delta_0 + a_0 Y_{t-1} + a_1 X_t + a_2 X_{t-1} + e_t.
\] (2.18)

We now have an autoregressive distributed lag (ARDL) model (2.18) that describes the relationship between \( Y_t \) and \( X_t \). With further reparametrization and the simplification of \( Y^*_t = Y_{t-1} = Y_{t-2} = \ldots = Y_{t-k} \) for both parameters \( Y_t \) and \( X_t \)

\[
Y^*_t - a_0 Y^*_t = \delta_0 + a_1 X^*_t + a_2 X^*_t + e_t,
\] (2.19)

\[
Y^*_t(1 - a_0) = \delta_0 + (a_1 + a_2) X^*_t + e_t,
\] (2.20)

\[
Y^*_t = \frac{\delta_0}{(1 - a_0)} + \frac{(a_1 + a_2) X^*_t}{(1 - a_0)} + e_t,
\] (2.21)

\[
Y^*_t = \beta_0 + \beta_1 X^*_t + e_t.
\] (2.22)

Equation (2.22) now describes the long-run relationship between \( Y_t \) and \( X_t \) through the coefficient \( \beta_1 = \frac{a_1 + a_2}{1 - a_0} \), with the crucial assumption \( a_0 < 1 \).

### 2.1.4 VAR

A vector autoregressive (VAR) model is the multivariate version of the regular AR model. The VAR model is built upon the idea that no distinction is made between endogenous and exogenous variables (Sims, 1980). First consider the bivariate VAR model derived from two time series \( y_t \) and \( x_t \) (Asteriou & Hall)

\[
y_t = \beta_{10} - \beta_{12} x_t + \gamma_{11} y_{t-1} + \gamma_{12} x_{t-1} + u_{yt},
\] (2.23)

\[
x_t = \beta_{20} - \beta_{21} y_t + \gamma_{21} y_{t-1} + \gamma_{22} x_{t-1} + u_{xt},
\] (2.24)

where \( y_t \) and \( x_t \) are assumed to be stationary and the error terms uncorrelated. Equation (2.23) and (2.24) can in turn be written as

\[
\begin{bmatrix}
1 & \beta_{12} \\
\beta_{12} & 1
\end{bmatrix}
\begin{bmatrix}
Y_t \\
X_t
\end{bmatrix} =
\begin{bmatrix}
\beta_{10} \\
\beta_{20}
\end{bmatrix}
+ \begin{bmatrix}
\gamma_{11} & \gamma_{12} \\
\gamma_{21} & \gamma_{22}
\end{bmatrix}
\begin{bmatrix}
Y_{t-1} \\
X_{t-1}
\end{bmatrix}
+ \begin{bmatrix}
u_{yt} \\
u_{xt}\end{bmatrix},
\] (2.25)
and reparametrized
\[ Bz_t = \Gamma_0 - \Gamma_1 z_{t-1} + u_t, \quad (2.26) \]

where \( B = \begin{bmatrix} 1 & \beta_{12} \\ \beta_{12} & 1 \end{bmatrix}, \quad z_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix}, \quad \Gamma_0 = \begin{bmatrix} \beta_{10} \\ \beta_{20} \end{bmatrix}, \quad \Gamma_1 = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}, \quad \text{and} \quad u_t = \begin{bmatrix} u_{yt} \\ u_{xt} \end{bmatrix}. \]

By multiplying equation (2.26) by the inverse of \( B 

\[ z_t = A_0 + A_1 Z_{t-1} + e_t, \quad (2.27) \]

is obtained. Here \( A_0 = B^{-1} \Gamma_0, \quad A_1 = B^{-1} \Gamma_1 \) and \( e_t = B^{-1} u_t \). The resulting VAR models are represented by

\[ y_t = a_{10} + a_{12} y_{t-1} + a_{12} x_{t-1} + e_{1t}, \quad (2.28) \]

\[ x_t = a_{20} + a_{21} y_{t-1} + a_{22} x_{t-1} + e_{2t}, \quad (2.29) \]

where, \( e_t = B^{-1} u_t \rightarrow \begin{bmatrix} e_{1t} = (u_{yt} + \beta_{12} u_{xt})/(1 - \beta_{12} \beta_{21}) \\ e_{2t} = (u_{xt} + \beta_{12} u_{yt})/(1 - \beta_{12} \beta_{21}) \end{bmatrix}. \]

The error terms are here composites of \( u_{yt} \) and \( u_{xt} \) representing white-noise processes. The models explain the variation in \( x_t \) and \( y_t \) as the function of lagged values (Hendry & Juselius, 2001). Each equation is estimated separately with regular OLS with no need to distinguish between endogenous and exogenous variables.

The \( \alpha \)-coefficients describe the changes that restore the model to its level of equilibrium. The long-run relationships between the variables are in turn characterised by the \( \beta \)-coefficients. Lastly, \( \gamma \) represent the short-term changes from previous lags.

Since every variable in VAR is explained by its own lagged values it is important to decide on the number of lags that (Juselius). If too many are included, the degrees of freedom would be consumed and the possibility of multicollinearity introduced, too few will tend to lead to specification errors. When specifying the lag length, \( k \) VAR models are estimated with up to \( k \) lags. The \( k \):th VAR models are

\[ y_t = \beta_{10} - \beta_{12} x_t + \gamma_{11} y_{t-1} + \gamma_{12} x_{t-1} + ... + \gamma_{1k} y_{t-k} + \gamma_{1k+1} x_{t-k} + u_{yt}, \quad (2.30) \]

\[ x_t = \beta_{20} - \beta_{21} y_t + \gamma_{21} y_{t-1} + \gamma_{22} x_{t-1} + ... + \gamma_{2k} y_{t-k} + \gamma_{2k+1} x_{t-k} + u_{xt}, \quad (2.31) \]

...
\[ y_t = \alpha_{10} + \alpha_{11}y_{t-1} + \alpha_{12}x_{t-1} + \ldots + \alpha_{1k}y_{t-k} + \alpha_{1k+1}x_{t-k} + \epsilon_{1t}, \quad (2.32) \]

\[ x_t = \alpha_{20} + \alpha_{21}y_{t-1} + \alpha_{22}x_{t-1} + \ldots + \alpha_{2k}y_{t-k} + \alpha_{2k+1}x_{t-k} + \epsilon_{2t}, \quad (2.33) \]

and an information criterion is estimated for each model. Two criterions that can be considered are Akaike- and Scwharz information criterion. These are measures of relative quality for each model and is used to select \( k \). The best\(^1\) model is the one that minimises the value of these measures (Ibid.).

### 2.1.5 Test for Cointegration

Engle & Granger devised a test in 1987 for identifying possible cointegrated relationships between two time series. The test had a few shortcomings. For instance it was found to yield different results depending on what variable that was regressed upon the other (Asteriou & Hall).

Søren Johansen developed another approach based on the works of Engle & Granger to solve the issues in the original test. Johansen’s test is capable of estimating the \( n - 1 \) cointegrating vectors, where \( n \) is the number of variables. In the case of \( n = 2 \), only one cointegrating relationship is possible. Johansen’s test for cointegration can be mathematically derived by extending a single-equational ECM to a multivariate ECM

\[ Z_t = A_1 Z_{t-1} + A_2 Z_{t-2} + \ldots + A_k Z_{t-k} + u_t, \quad (2.34) \]

where \( Z_t = [X_t, Y_t] \) is a matrix of \( I(1) \) variables (Ibid.). Equation (2.34) can be rewritten as a vector error-correction model (VECM)

\[ \Delta Z_t = \Gamma_1 \Delta Z_{t-1} + \Gamma_2 \Delta Z_{t-2} + \ldots + \Gamma_{t-k} \Delta Z_{t-k-1} + \Pi Z_{t-1} + u_t, \quad (2.35) \]

where \( \Gamma_i = (I - A_1 - A_2 - \ldots - A_k) \) and \((i = 1, 2, \ldots, k - 1)\). The long-run equilibrium is described by \( \Pi \) equal to \(-(I - A_1 - A_2 - \ldots - A_k)\) and may in turn be rewritten as the product of \( \alpha \) and \( \beta' \). Here \( \alpha \) represent the speed of adjustment to equilibrium and \( \beta' \) represent the matrix of long-run coefficients. By expanding \( \Pi \), \( \beta' Z_{t-1} \) is obtained and equal to \((Y_{t-1} - \beta_0 - \beta_1 X_{t-1})\) that form the error-correction term in the model. Through reparametrisation of (2.35)

\(^1\)Under the assumption of a correctly specified model.
\[
\begin{pmatrix}
\Delta Y_t \\
\Delta X_t
\end{pmatrix} = \Gamma_1 \begin{pmatrix}
\Delta Y_{t-1} \\
\Delta X_{t-1}
\end{pmatrix} + \Pi \begin{pmatrix}
Y_{t-1} \\
X_{t-1}
\end{pmatrix} + e_t,
\] (2.36)

is obtained (Ibid.). By expanding \( \Pi \)
\[
\begin{pmatrix}
\Delta Y_t \\
\Delta X_t
\end{pmatrix} = \Gamma_1 \begin{pmatrix}
\Delta Y_{t-1} \\
\Delta X_{t-1}
\end{pmatrix} + \begin{pmatrix}
\alpha_{11} & \alpha_{12} \\
\alpha_{21} & \alpha_{22}
\end{pmatrix} \begin{pmatrix}
\beta_{11} & \beta_{21} \\
\beta_{12} & \beta_{22}
\end{pmatrix} \begin{pmatrix}
Y_{t-1} \\
X_{t-1}
\end{pmatrix} + e_t.
\] (2.37)

Now, if only the error-correction part (i.e. the last three matrices in equation (2.37)) is considered the following is retrieved
\[
\Pi_1 Z_{t-1} = (\alpha_{11}\beta_{11})(\alpha_{11}\beta_{21}) \begin{pmatrix}
Y_{t-1} \\
X_{t-1}
\end{pmatrix},
\] (2.38)

where \( \Pi_1 \) represent the first row of the \( \Pi \) matrix. This equation can further be derived into
\[
\Pi_{t-1} = \alpha_{11}(\beta_{11}Y_{t-1} + \beta_{21}X_{t-1}),
\] (2.39)

which is the final cointegrating vector with \( \alpha_{11} \) representing its speed of adjustment to equilibrium.

2.1.6 Johansen’s Approach

Johansen suggested an approach along with his test for cointegration. In the first step each variables’ order of integration is determined. Then the lag length that yields the best VAR model is chosen with Akaike information criterion (AIC),

\[
AIC = -log(\text{maximum likelihood}) + \\
+2(\text{numbers of independently adjusted parameters within the model}).
\] (2.40)

The model with the lowest AIC value from (2.40) is then chosen to represent the most appropriate lag length for the possible pairs \( Y_t \) and \( X_t \). As the third step, the appropriate model with respect to the deterministic components in the system is determined. Maximum eigenvalue statistic is then used to find the rank of the cointegrating matrix. With null hypothesis of \( r \) cointegrating vectors tested against the alternative of \( r + 1 \) vectors.
The procedure works by ordering the estimated eigenvalues \((\lambda_1 > \lambda_2 > ... > \lambda_n)\) in a descending order and evaluating each to find the ones significantly different from zero. Likelihood Ratio-test statistic,

\[
\lambda_{\text{max}} = -T \times \ln(1 - \hat{\lambda}_{r+1}),
\]

is then used to test how many of the eigenvalues are significantly different from zero (Asteriou & Hall).

2.2 Financial Theory

2.2.1 Pairs Trading

In the market the general idea is to buy undervalued securities and sell overvalued ones. This would however only be possible if the true price of each security was known (Elliott et al., 2005). Pairs Trading tries to resolve this by using the idea of relative pricing between two securities. If the two securities are cointegrated their prices should move together around a long-term equilibrium. If they diverge from their equilibrium the spread either increases or decreases. If the spread were to move too far and past a pre-set threshold point positions are taken. A trade is opened by buying one of the two equities (long position, earning money when equity prices increase) and selling the other one (short position, earning money when equity prices fall). When the equities converge to their long-term equilibrium both positions are closed and a profit is made.

Consider two securities \(A\) with price \(P^A\) and \(B\) with price \(P^B\) both prices are logarithmized

\[
Log(P_t^A) = n_{A_t} + \epsilon_{A_t},
\]

\[
Log(P_t^B) = n_{B_t} + \epsilon_{B_t},
\]

where \(n_{A_t}, n_{B_t}\) represents two unit root processes (random walks) and components \(\epsilon_{A_t}, \epsilon_{B_t}\) represent the stationary part. If the two securities are cointegrated, a constant \(\gamma\) exist so that \(n_{A_t} = \gamma n_{B_t}\) and as a result the linear combination of the two securities,

\[
Log(P_t^A) - \gamma Log(P_t^B) = (n_{A_t} - \gamma n_{B_t}) + (\epsilon_{B_t} - \epsilon_{A_t}),
\]

is stationary when \((n_{A_t} - \gamma n_{B_t}) = 0\) (Vidyamurthy). The linear relationship of both
securities can then be reinterpreted as

\[ \log(P_t^A) - \gamma \log(P_t^B) = \mu + \epsilon_t, \quad (2.45) \]

here \( \mu \) is a constant and represent the long-term distance between both securities and \( \epsilon_t \) is a stationary time series with mean zero. For each time period \( t \) the spread of both securities can be constructed by first regressing security \( A \) on security \( B \) with OLS,

\[ \log(P_t^A) = \mu + \gamma \log(P_t^B) + \epsilon_t. \quad (2.46) \]

The spread of both securities would then be equal to

\[ \text{Spread}_t = \log(P_t^A) - [\mu + \gamma \log(P_t^B)], \quad (2.47) \]

where \( \text{Spread}_t \) is \( I(0) \) with a mean zero and standard deviation \( \sigma_{\text{spread}} \). Therefore the spread between both oscillates around zero as a result of the mean reverting property.

### 2.2.2 Calculate Return

Each pair in the portfolio is assigned 1 SEK of capital beforehand. As a result even though a pair is not traded during the period capital is still held ready for trading if the opportunity would arise. The invested capital for both the long and short position is assumed to be 0.5 SEK each. Hence the portfolio does not hold any leverage.

Consider a portfolio consisting of a long position of one share in security \( A \) and a short position of \( \gamma \) shares in security \( B \). The position is constructed so as \( \gamma \) shares in \( B \) is equivalent to one share of \( A \). The return of such a trade is calculated as

\[ [\log(P_{t+i}^A) - \log(P_t^A)] - \gamma [\log(P_{t+i}^B) - \log(P_t^B)]. \quad (2.48) \]

A higher price of \( P^A \) at time \( t + i \) relative to \( t \) yields a positive return. For \( P^B \) a higher price in \( t + i \) relative to time \( t \) instead yields negative return due to the short position in the security, hence the negative sign in front of \( \gamma \). By rearranging the terms in equation (2.48) to

\[ [\gamma \log(P_t^B) - \log(P_t^A)] - [\gamma \log(P_{t+i}^B) - \log(P_{t+i}^A)] = \text{Spread}_t - \text{Spread}_{t+i}, \quad (2.49) \]

15
the return is viewed in terms of the spread. Measured as the difference in the spread between time point $t$ and $t + i$. If both equities diverge from their long-term equilibrium and the spread increase to three percent in time $t$ followed by conversion of both assets back to equilibrium in $t + i$, the return is $3 - 0 = 3$ percent. Positions are marked to market daily and have a daily return equal to

$$Spread_{t-1} - Spread_t.$$ (2.50)

Return of an open position is calculated as the movement in the spread from time $t - 1$ to $t$. A portfolio of $k$ assets will have the daily return of

$$\sum_{i=1}^{k} \left( Spread_{i,t-1} - Spread_{i,t} \right) * C_{i,t-1} \over \sum_{i=1}^{k} C_{i,t-1},$$ (2.51)

where $C_{i,t-1}$ is the committed capital for pair $i$ at time $t - 1$.

Any excess return from previous trades in a pair is reinvested for new trades in the same pair. All open pairs are closed on the last day of trading.

Transaction costs are not included in the return calculations. The cost for performing a pairs trade are numerous and some of the commonly discussed are transaction commissions, margins for short positions and bid-ask spreads.

### 2.2.3 Sharpe Ratio

One way of measuring return of a portfolio in relation to the risk taken is the Sharpe ratio, first introduced by William Sharpe in 1966. The differential return was defined as the chosen portfolios return, $R_p$, subtracted by the benchmark portfolio, $R_f$. In modern times the praxis for approximating the benchmark return has been the risk free rate. The differential return is then adjusted by dividing it with the standard deviation, $\sigma_p$, for the chosen portfolio. (Sharpe, 1994) The formula is given by

$$S_h = {R_p - R_f \over \sigma_p}.$$ (2.52)

A higher Sharpe ratio, as a rule of thumb, is interpreted as the additional return an investor is obtaining for the added risk. In other words the higher Sharpe ratio the better.
2.2.4 Constructing Confidence Intervals

The confidence intervals serve to give an indication of the current price of an asset in relation to its long-term equilibrium. Consider the spread between two securities with mean, $\mu$, and standard deviation, $\sigma$, calculated as

$$\mu_{\text{spread}_i} = \frac{1}{T} \sum_{\forall t} \text{spread}_i,$$  \hspace{1cm} (2.53)

$$\sigma_{\text{spread}_i} = \sqrt{\frac{\sum_{\forall t} (\text{spread}_i - \mu_{\text{spread}_i})^2}{T - 1}},$$  \hspace{1cm} (2.54)

where $i$ is the specific pair of securities in the portfolio and $t$ represents number of trading days. The intervals are constructed with a predetermined scalar, $M$, multiplied with the standard deviation subtracted and added to the mean,

$$\text{UpperBoundary}_{\text{spread}_i} = \mu_{\text{spread}_i} + M \times \sigma_{\text{spread}_i},$$  \hspace{1cm} (2.55)

$$\text{LowerBoundary}_{\text{spread}_i} = \mu_{\text{spread}_i} - M \times \sigma_{\text{spread}_i}.$$  \hspace{1cm} (2.56)
3 Method

The outline of this section is as follows: First a part about the data followed by the process of selecting tradable pairs. Thirdly, how the stability is tested and lastly a section on how the trading was performed. All tests were performed using $R^2$.

3.1 Data

The dataset used in this study was retrieved from Factset$^3$. The data consists of daily adjusted closing prices for all companies on the Stockholm Stock Exchange (OMXS) Large Cap between the periods 2005-2015. There are 103 equities included in the study and restrictions were made to only include those companies listed on Large Cap due to the liquidity of the equities. It is important that they have a fairly high turnover to avoid abruptions in the series. Among class shares only the one with the highest turnover was included.

In accordance with Gatev et al. (2006) a rolling window is set up with a ratio of 2:1. An estimation window of two years where tests for cointegration are performed. We use a rolling window of two years to estimate stability and a one year out-of-sample to trade. There are approximately 260 trading days for any given year. Equities that did not trade for the whole estimation period were excluded. Further, the data was transformed by the natural logarithm in order to scale it and still keep its overall characteristics.

3.2 Finding Cointegrated Pairs

For all time series, $Y_t$, the order of integration is determined with ADF-test, allowing for an intercept, $\alpha_0$ and a non-stochastic trend, $\alpha_2t$. The observed value of $\gamma$ from (2.8) is compared to the critical value of the Dickey Fuller’s distribution. The order of integration $d$ is obtained from the ADF-test and assigned to each security $Y_{t,t} \sim I(d)$. Only pairs with same order of integration are allowed, $Y_t \sim I(d)$ and $X_t \sim I(d)$.

From this a matrix of all possible pairs from each industry is formed. Depending on the time period this usually yields up to 400 possible pairs. Further, to decide on the appropriate lag length for Johansen’s test, ”VARselect”$^4$ was used. The function

---

$^2$R is an open source statistical program.

$^3$A company offering financial data and software.

$^4$Built in ”vars” package.
is designed to estimate $k$ VAR-models including different lag lengths of $k$, see equation (2.30 to 2.33). The maximal number of $k$ was set to 20, giving height for each model’s most appropriate lag length.

The final model chosen for Johansen’s test is represented in equation (2.35). Here $\delta_1 t$ represents a linear trend in the cointegrated equation (CE) allowing for exogenous growth. Intercepts are also used for both CE, $\mu_1$ and the VAR, $\mu_2$. When the model is formed the number of cointegrated vectors are determined by the rank of $\Pi$ in equation (2.36). To determine the rank of $\Pi$ the maximal eigenvalue statistic is used, denoted by $\lambda_{max}$, seen in equation (2.41). From this the number of cointegrated relationships is obtained for each pair of equities and now only cointegrated pairs remain.

All these pairs form a ”Market Portfolio” of cointegrated pairs. Next, all possible cointegrating pairs are sorted on correlation. And now we form our two correlated portfolios. One portfolio containing the 20 highest correlated and cointegrated pairs, which we name ”20 Portfolio” and the second portfolio containing the 10 highest correlated and cointegrated pairs, named ”10 Portfolio”.

### 3.3 Stability of Cointegrating Relationships

To test how the cointegrated relationships evolve over time the estimation window is shifted forward 6, 12, 18 and 24 months. Creating four subsequent periods. This is done for all cointegrated pairs on the market for every year and for the 20 and 10 Portfolios.

Consider the following example, initial estimation window: 2005-01-01 to 2006-12-31. Where the first of the subsequent periods are created by shifting it forward 6 months i.e. 2005-06-01 to 2007-05-31, the second, shifted 12 months; 2006-01-01 to 2007-12-31, third, 18 months; 2006-06-01 to 2008-05-31 and lastly 24 months; 2007-01-01 to 2008-12-31.

If we were to have 50 pairs that are cointegrated in the initial estimation window. Then those 50 pairs are tested separately for a cointegrated relationship in each subsequent period for the Market Portfolio. The same is done for the correlated portfolios (i.e. 20 and 10 Portfolio). Continuing the example: if the 50 pairs are still cointegrated as the window is shifted forward we say that the relationships are stable over time and that Pairs Trading can be profitable. If none of the pairs still maintain their relationship they are all declared unstable and the strategy is not applicable. In the case where some relationships remain Pairs Trading might be profitable. But to distinguish between stable
cointegrated pairs and unstable in such a scenario is near impossible.

3.4 Trading the two Portfolios

As mentioned in section 2.2.1 we want to take a position in equities $A$ and $B$ when the spread move beyond the threshold point. We define it as two standard deviations. Such a movements are due to shocks to the underlying assets and should only be temporary. The mean reversion property of the linear combination of equities $A$ and $B$ result in the spread returning back to zero. If both equities diverge and as a result the spread is positive, a long position in $A$ and a short position in $B$ is taken, vice versa if the spread is negative. Both positions will be held until the spread return to zero. However if a position is taken and the deviation of both assets continue and the spread pass three standard deviations a stop loss is activated and the position is exited.

In practice this is done by first estimating $\mu$ and $\gamma$ with equation (2.47). With these each spread is calculated and constructed for the trade window. From the estimation window $\mu_{\text{spread}}$ and $\sigma_{\text{spread}}$ is calculated with equations (2.53 & 2.54). Lastly intervals are constructed by equations (2.55 & 2.56), see Figure 1 for example.

![Figure 1: An example of trading two equities $A$ and $B$ during a period of 50 days. On day nine they diverge and spread increase to more than two standard deviations (represented by blue line). Hence a long position of $A$ and a short position of $B$ is opened. Both assets start to converge and spread returns to zero (red line) on day 31 and the position is closed. Two days later (i.e. day 33) they diverge again and the spread is yet again beyond two standard deviations and the same position is opened. Now however they continue to diverge and the spread moves past three standard deviations (yellow line) and the stop loss kicks in on day 39.](image-url)
4 Results

This chapter is divided into two subsections: Stability of the cointegrated relationships and Portfolio Performance. First, stability results are presented for our three portfolios: Market, 20 and 10 where the estimation window is shifted forward in time as explained in section 3.3. The two portfolios 20 and 10 are compared to each other and to the Market Portfolio. In the second subsection the returns and Sharpe ratios for our two 20 and 10 Portfolios are presented and compared against the OMXS30 benchmark index.

4.1 Evaluating Stability

First we examine all cointegrated relationships in the Market Portfolio to decide if there are enough stable relationships for Pairs Trading to be profitable.

Table 1: Presents the cointegrated pairs of the Market Portfolio. The left part shows how many cointegrated pairs there were among all possible pairs for each time period. And on the right, the number of still cointegrated pairs in the four subsequent periods 6, 12, 18 and 24 months.

<table>
<thead>
<tr>
<th>Years</th>
<th>Possible pairs</th>
<th>Cointegrated pairs</th>
<th>6 months</th>
<th>12 months</th>
<th>18 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-2006</td>
<td>293</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006-2007</td>
<td>364</td>
<td>38</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2007-2008</td>
<td>345</td>
<td>37</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2008-2009</td>
<td>408</td>
<td>37</td>
<td>18</td>
<td>14</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>2009-2010</td>
<td>385</td>
<td>67</td>
<td>28</td>
<td>22</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>2010-2011</td>
<td>418</td>
<td>29</td>
<td>16</td>
<td>10</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>2011-2012</td>
<td>375</td>
<td>72</td>
<td>13</td>
<td>9</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>2012-2013</td>
<td>387</td>
<td>35</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>2013-2014</td>
<td>441</td>
<td>66</td>
<td>15</td>
<td>17</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Average</td>
<td>372</td>
<td>42</td>
<td>16</td>
<td>11</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

When all equities are considered an average of 391 possible pairs are found, 48 of those were cointegrated during the two year estimation window. Among those, 17, 11, 7 and 5 pairs were still cointegrated for the four subsequent periods respectively.

For a twelve month shift, on average about a fifth of the cointegrated pairs still maintain their relationship. This would in turn cover the trading window. For this period the number of stable pairs ranges from 4 to 22 depending on the estimation window. We deem it enough to use our 20 and 10 Portfolios to try to capture these pairs.
Table 2: Presents the cointegrated pairs of the 20 Portfolio. The numbers of still cointegrated pairs are shown for the 20 Portfolio in the four subsequent periods.

<table>
<thead>
<tr>
<th>Cointegration in estimation window</th>
<th>Cointegration in subsequent periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years</td>
<td>Possible pairs in the data</td>
</tr>
<tr>
<td>2005-2006</td>
<td>293</td>
</tr>
<tr>
<td>2006-2007</td>
<td>364</td>
</tr>
<tr>
<td>2007-2008</td>
<td>345</td>
</tr>
<tr>
<td>2008-2009</td>
<td>408</td>
</tr>
<tr>
<td>2009-2010</td>
<td>385</td>
</tr>
<tr>
<td>2010-2011</td>
<td>418</td>
</tr>
<tr>
<td>2011-2012</td>
<td>375</td>
</tr>
<tr>
<td>2012-2013</td>
<td>387</td>
</tr>
<tr>
<td>2013-2014</td>
<td>441</td>
</tr>
<tr>
<td>Average</td>
<td>372</td>
</tr>
</tbody>
</table>

The 20 Portfolio has on average 7, 6, 3, 2 still cointegrated pairs in the subsequent periods. When sorting by highest correlation we find that a higher proportion of cointegrated pairs remain after twelve months compared to the Market Portfolio.

Table 3: Presents the cointegrated pairs of the 10 Portfolio. The numbers of still cointegrated pairs are shown for the 10 Portfolio in the four subsequent periods.

<table>
<thead>
<tr>
<th>Cointegration in estimation window</th>
<th>Cointegration in subsequent periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years</td>
<td>Possible pairs in the data</td>
</tr>
<tr>
<td>2005-2006</td>
<td>293</td>
</tr>
<tr>
<td>2006-2007</td>
<td>364</td>
</tr>
<tr>
<td>2007-2008</td>
<td>345</td>
</tr>
<tr>
<td>2008-2009</td>
<td>408</td>
</tr>
<tr>
<td>2009-2010</td>
<td>385</td>
</tr>
<tr>
<td>2010-2011</td>
<td>418</td>
</tr>
<tr>
<td>2011-2012</td>
<td>376</td>
</tr>
<tr>
<td>2012-2013</td>
<td>387</td>
</tr>
<tr>
<td>2013-2014</td>
<td>441</td>
</tr>
<tr>
<td>Average</td>
<td>372</td>
</tr>
</tbody>
</table>

For the 10 Portfolio, on average; 4, 3, 2 and 1 pair(s) are still cointegrated in the subsequent periods. When the 10 highest correlated pairs are selected also a higher proportion of cointegrated pairs remain after twelve months compared to the Market Portfolio.

For both the 20 and 10 Portfolios we observe the same proportions of stable pairs remaining after twelve months. This could be due to the selection process of the highest correlated pairs. But the majority of the pairs still does not hold over time. However, we deem it to be enough stable pairs to be worthwhile to consider the profitability of trading the portfolios.
4.2 Portfolio Performance

The results of our Pairs Trading strategy with the 20 and 10 Portfolios are summarised in Table 4.

Table 4: Presents the returns and Sharpe ratios of 20 and 10 Portfolios and for the benchmark index OMXS30.

<table>
<thead>
<tr>
<th>Return (%)</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>Mean</th>
<th>Std</th>
<th>Sharpe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portfolio 20</td>
<td>6.1</td>
<td>1.2</td>
<td>-9.3</td>
<td>-2.7</td>
<td>8.0</td>
<td>-2.2</td>
<td>-9.1</td>
<td>-5.3</td>
<td>-5.1</td>
<td>-2.0</td>
<td>6.1</td>
<td>neg.</td>
</tr>
<tr>
<td>Portfolio 10</td>
<td>6.4</td>
<td>4.7</td>
<td>-13.2</td>
<td>-3.8</td>
<td>10.3</td>
<td>-5.6</td>
<td>-9.4</td>
<td>-5.3</td>
<td>-5.9</td>
<td>-2.4</td>
<td>7.8</td>
<td>neg.</td>
</tr>
<tr>
<td>OMXS30</td>
<td>-7.1</td>
<td>-37.4</td>
<td>37.2</td>
<td>20.0</td>
<td>-16.1</td>
<td>10.5</td>
<td>18.0</td>
<td>10.6</td>
<td>-1.2</td>
<td>3.8</td>
<td>22.1</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Negative returns are obtained for both portfolios. For the 20 Portfolio the return is −2.0 and −2.4 for the 10 Portfolio. For the same period the OMXS30 index had an annual growth of 3.8 percent with a Sharpe ratio of 0.17.

There is no relationship between the number of still cointegrated pairs and the obtained return. Take the 10 Portfolio for example: the pairs traded in 2011 yield a return of 10.3 percent. These were estimated in 2009-2010 and only three pairs remained cointegrated in the trading window. Compared to the pairs traded in 2013 with a return of −9.4 percent. During that estimation window (2011-2012) six pairs were still cointegrated for the trading window. If the 20 and 10 Portfolios are compared to the OMXS30 index, we note that when the market had a positive growth the 20 and 10 Portfolios generated negative returns and for years of negative growth the portfolios yielded positive returns (with the exception of 2015).
5 Discussion

Are the cointegrating relationships stable over time? Theory suggests that the presence of cointegrating relationships are only possible if there is a true link between two time series (Asteriou & Hall). However for all time periods we see the number of cointegrated pairs diminish over time. The relationships are unstable and most break off. Some could be attributed to Type-II error, with an \( \alpha \) of five percent, we should have five percent Type-II errors. In this case, that would result in about 20 pairs, of our 48, being accepted as cointegrated when they are in fact not. That leaves 28 pairs, correctly estimated with only 17 still cointegrated six months forward in time. The difference (11 pairs) are more difficult to explain, but structural breaks in the cointegrated relationship between pairs of equities is a probable cause (Buncic & Roca, 2005). The number of cointegrated pairs for the Market Portfolio differ greatly from year to year, which in turn could imply that pairs are falling in and out of their cointegrated relationships.

If they are stable over time, is Pairs Trading a profitable strategy? The profitability of Pairs Trading is highly dependent on the ability to find the presence of long-term equilibrium (Alexander et al., 2001). When we sort on high correlation the 20 and 10 Portfolios find a higher proportion of stable cointegrated relationships compared to the Market Portfolio. However these portfolios yielded negative return and underperformed the benchmark index OMXS30.

The poor performance can be attributed to two factors. First, pairs of equities diverge and start to converge but do not reach their equilibrium before the trade window runs out. This may be the result of a trade window being too short. It is also possible that these pairs are not cointegrated and will not converge in the future. Which of the two we are dealing with is impossible to determine at that point in time.

Secondly, other pairs are oscillating around the stop loss threshold. As a result a position is opened below the threshold and then closed above it. These trades therefor result in continuous losses throughout the trade window. Some of these losses could be avoided with more complex trading rules.

We therefore argue that the returns of the portfolios are not only dependent on the ability of finding cointegrated pairs but also rely on the complexity of the trading rules.

\[5\text{Given that the true probability distribution of our population follow the chosen test distribution.}\]
5.1 Conclusions

We did not manage to find portfolios where the majority of the cointegrated pairs lasted throughout the trade window.

The Market Portfolio consisted of a fifth of the pairs that were desirable to trade in comparison the 20 and 10 Portfolios consist of a third. The 10 Portfolio did not contain a higher proportion of cointegrated pairs than the 20 Portfolio. Neither did it outperform in returns.

Finally to sum up our two research questions: Are the cointegrating relationships stable over time? If they are is Pairs Trading a profitable strategy? There is instability, but there are enough relationships on the market for Pairs Trading to be applicable. Whether the strategy is profitable or not is a function of sorting out stable cointegrated pairs and applying sophisticated trading rules.

5.2 Future research

We find it interesting to further investigate complimentary ways of determining whether equities are stable. As a suggestion; one could combine multiple estimation windows to draw conclusions on how stable the cointegrating relationships are.

Additionally, we only tested the number of cointegrated pairs in each period as the estimation window was shifted forward. It would therefore be interesting to see if it is the same pairs that stay cointegrated or if there are different pairs falling in and out of their cointegrated relationships.

Another aspect to examine, would be how our results fare with different lengths of the estimation and trade window.
6 References


#Appendix#

```r
#Packages
install.packages("xts") #Packed for transforming dates
install.packages("tseries") #For ADF-test
install.packages("data.table")
install.packages("urca") #For cointegration test
install.packages("vars")
library(xts)
library(tseries)
require(data.table)
library(urca)
library(vars)

#Input: Set dates, Correlation & Cut-off
StartdateTestwindow <- "2009-01-01"
EnddateTestwindow <- "2010-12-31"
StartdateTradewindow <- "2011-01-01"
EnddateTradewindow <- "2011-12-30"
CorrCutOff <- 0.8

file <- FINALBOOK.no.class.shares.2 #Set name of imported dataset
industry <- Industri.klar.no.class.share
Significance.level.adf <- 0.05
Significance.level.cointegration.test <- "5pct" #choices "10pct", "5pct" or "1pct"
number.of.pairs.in.portfolio <- 10

#Create Test & Trade period for later use
RangeTestW <- paste(StartdateTestwindow,"::",EnddateTestwindow,sep="")
RangeTradewindow <- paste(StartdateTradewindow,"::",EnddateTradewindow,sep="")
RangeWholeW <- paste(StartdateTestwindow,"::",EnddateTradewindow,sep="")

#Change dates to row names
file.xts <- xts(x=file[,,-1],order.by=as.POSIXct(file$Date))

#Create the Test & Trade window
WholeW <- file.xts[RangeWholeW] #The two Test & Trade windows in one matrix
Filteredsample<- WholeW[,colSums(is.na(WholeW)) == 0]
#Used to filter out Stocks with missing values in either Test or Trade window

#Filter industry by na in sample
industry <- industry[1,]
dim.WholeW <- dim(WholeW)
for (i in 1:dim.WholeW[2]){
  if (colSums(is.na(WholeW[,i])) != 0){
    industry[1,i] <- NA
  }
}
industry <- industry[,colSums(is.na(industry)) == 0]
```

28
# From the filtered sample we divide it into Test window to get dimensione
TestW <- Filteredsample[RangeTestW]

# Logaritm of prices
Filteredsample <- log(Filteredsample)

# Exclude Stationary time series these are not of interest
TestW.dim <- dim(TestW)
Filteredsample.dim <- dim(Filteredsample)
for(i in 1:TestW.dim[2]){
  if(adf.test(Filteredsample[c(1:TestW.dim[2]),i])$p.value > Significance.level.adf){
    Filteredsample[,i] <- matrix(Filteredsample[,i])
  } else {
    Filteredsample[,i] <- NA
  }
}

# Exclude stock from industry vector if they are stationary in TestW
dim.Filteredsample <- dim(Filteredsample)
for (i in 1:dim.Filteredsample[2]){  
  if (colSums(is.na(Filteredsample[,i])) != 0){
    industry[1,i] <- NA
  }
}
industry <- industry[,colSums(is.na(industry)) == 0]

# For Loop: Fills matrix with only the non stationary series the rest is NA values
non.stationary.WholeW <- Filteredsample[,colSums(is.na(Filteredsample)) == 0]

# Exclude the columns containing NA values i.e only non stationary time series left

# From the filtered sample we divide it into Test and Trade window
TestW <- non.stationary.WholeW[RangeTestW]
TradeW <- non.stationary.WholeW[RangeTradeW]

# Create all possible pairs for the non stationary time series
TestW.new.dim <- dim(TestW)
possible.pairs <- t(combn(TestW.new.dim[2],2))

# Exclude all pairs from different industries
possible.pairs.dim <- dim(possible.pairs)
for (i in 1:possible.pairs.dim[1]){  
  if (industry[1,possible.pairs[i,1]] != industry[1,possible.pairs[i,2]]){
    possible.pairs[i,] <- NA
  }
}
possible.pairs <- possible.pairs[rowSums(is.na(possible.pairs)) == 0 ,]

# Choose correlation
possible.pairs.dim <- dim(possible.pairs)
all.corr.matrix <- matrix(NA, nrow=possible.pairs.dim[1], ncol=1)
for (i in 1:possible.pairs.dim[1]){ 
  all.corr.matrix[i,] <- abs(cor(TestW[,c(possible.pairs[i,1],possible.pairs[i,2])])[2,1])
  #Creates matrix with the correlation for each pair
}
corr.possible.pairs <- cbind(possible.pairs, all.corr.matrix)
#Combine matrices possible.pairs with corresponding correaltion
sorted.corr.possible.pairs <- data.table(corr.possible.pairs, key="V3")
#Sorts the new matrix on correlation
final.pairs.for.coint.test <- sorted.corr.possible.pairs
final.pairs.for.coint.test <- (data.frame(final.pairs.for.coint.test))
#Transform data to data frame and remove correlation column

#Calculate the lag lenght with AIC for each possible pair
final.pairs.for.coint.test.dim <- dim(final.pairs.for.coint.test)
VAR.result.matrix <- matrix(NA, nrow=final.pairs.for.coint.test.dim[1], ncol=1)
for (i in 1:final.pairs.for.coint.test.dim[1]){ 
  VAR.result.matrix[i,1] <- matrix((VARselect(TestW[,c(final.pairs.for.coint.test[i,1],final.pairs.for.coint.test[i,2])], lag.max=20, type="both"))$selection)[1,1]
}
#Choose only the cointegrated pairs from all possible pairs
final.pairs.for.coint.test <- as.matrix(final.pairs.for.coint.test)
cointegrated.possible.pairs <- matrix(NA, nrow=final.pairs.for.coint.test.dim[1], ncol=3)
for (i in 1:final.pairs.for.coint.test.dim[1]){ 
  johansen.test <-ca.jo(TestW[,c(final.pairs.for.coint.test[i,1],final.pairs.for.coint.test[i,2])],
    ecdet = "none", type="eigen", K=(max(2,VAR.result.matrix[i,1])))
  if((johansen.test@teststat)[2] > johansen.test@cval[2,
    Significance.level.cointegration.test]){ 
    cointegrated.possible.pairs[i,] <- (final.pairs.for.coint.test[i,])
  }
}

Filtered.cointegrated.possible.pairs <- cointegrated.possible.pairs[
  rowSums(is.na(cointegrated.possible.pairs)) == 0,]
#Filters out non cointegrated pairs (i.e. rows with NA values)

### This section is only used to test the Test Window rolling forward###
#New trade window
StartdateTradewindow <- "2013-06-01"
EnddateTradewindow <- "2015-06-30"
StartdateTradewindow <- "2014-01-01"
EnddateTradewindow <- "2015-12-30"
StartdateTradewindow <- "2013-06-01"
EnddateTradewindow <- "2015-06-30"
StartdateTradewindow <- "2014-01-01"
EnddateTradewindow <- "2015-12-30"

final.pairs.for.trading <- final.pairs.for.trading
dim(final.pairs.for.trading)

RangeTradew <- paste(StartdateTradewindow,"::",EnddateTradewindow,sep="")
TradeW <- non.stationary.WholeW[RangeTradew]

#Calculate the lag length with AIC for each possible pair
final.pairs.for.trading.dim.TradeW <- dim(final.pairs.for.trading)
VAR.result.matrix.TradeW <- matrix(NA, nrow=final.pairs.for.trading.dim.TradeW[1], ncol=1)
for (i in 1:final.pairs.for.trading.dim.TradeW[1]){
  VAR.result.matrix.TradeW[i,1] <- matrix((VARselect(TradeW[,c(final.pairs.for.trading[i,1],final.pairs.for.trading[i,2])], lag.max=20, type="both"))$selection)[1,1]
}

final.pairs.for.trading.TradeW <- as.matrix(final.pairs.for.trading)
cointegrated.pairs.TradeW <- matrix(NA, nrow=final.pairs.for.trading.dim.TradeW[1], ncol=2)
for (i in 1:final.pairs.for.trading.dim.TradeW[1]){
  johansen.test <- ca.jo(TradeW[,c(final.pairs.for.trading.TradeW[i,1],final.pairs.for.trading.TradeW[i,2])], ecdet = "none", type="eigen", K=(max(2,VAR.result.matrix[i,1])))
  if((johansen.test@teststat)[2] > johansen.test@cval[2,Significance.level.cointegration.test]){
    cointegrated.pairs.TradeW[i,] <- (final.pairs.for.trading.TradeW[i,])
  }
}

cointegrated.pairs.TradeW <- cointegrated.pairs.TradeW[rowSums(is.na(cointegrated.pairs.TradeW)) == 0,] #Filters out non cointegrated pairs (i.e. rows with NA values)
dim(cointegrated.pairs.TradeW)

###This section is only used to test the Test Window rolling forward####

#Chose our trade portfolio of coint and highly corr pairs
final.pairs.for.trading <- tail(Filtered.cointegrated.possible.pairs, n=number.of.pairs.in.portfolio)
#Picks the highest correlated and cointegrated pairs for trading
final.pairs.for.trading <- (final.pairs.for.trading)[-3]
#Transform data to data frame and remove correlation column
#Calculate the spread for each selected pairs to Trade
final.pairs.for.trading.dim <- dim(final.pairs.for.trading)
TradeW.dim <- dim(TradeW)
Trade.portfolio.spreads <- matrix(NA, nrow=TradeW.dim[1], ncol=final.pairs.for.trading.dim[1])
TestW.portfolio.spreads <- matrix(NA, nrow= TestW.dim[1], ncol=final.pairs.for.trading.dim[1])
hr.i <- matrix(NA, nrow=final.pairs.for.trading.dim[1])
intercept.i <- matrix(NA, nrow=final.pairs.for.trading.dim[1])
for (i in 1:final.pairs.for.trading.dim[1]){
  hr.i[i,] <- as.numeric((lm(TestW[,final.pairs.for.trading[i,1]] ~ TestW[,final.pairs.for.trading[i,2]]))$coefficients[2])#Extract hedge ratio
  intercept.i[i,] <- as.numeric((lm(TestW[,final.pairs.for.trading[i,1]] ~ TestW[,final.pairs.for.trading[i,2]]))$coefficients[2])#Extract intercept
$coefficients[1])$

Trade.portfolio.spreads[,i] <- TradeW[,final.pairs.for.trading[i,1]] - (intercept.i[i,]
+ hr.i[i,]*TradeW[,final.pairs.for.trading[i,2]])

# Creates portfolio for the spread of all chosen pairs
TestW.portfolio.spreads[,i] <- TestW[,final.pairs.for.trading[i,1]] - (intercept.i[i,]
+ hr.i[i,]*TestW[,final.pairs.for.trading[i,2]])
}

# Name each spread in matrix with the combination of both stock names
matrix.i <- matrix(NA, nrow=TradeW.dim[1], ncol=1)
Trade.portfolio.spreads.dim <- dim(Trade.portfolio.spreads)
Trade.portfolio.spreads.named <- data.frame(Trade.portfolio.spreads)
for (i in 1:number.of.pairs.in.portfolio){
    matrix.i <- matrix(Trade.portfolio.spreads[,i])
    colnames(Trade.portfolio.spreads.named)[i] <- paste(colnames(TestW[,final.pairs.for.trading[i,1]]),
              colnames(TestW[,final.pairs.for.trading[i,2]]), sep = " & ")
}

# Create matrix for each coefficient and each pair
trades.tradew <- matrix(NA, nrow=TradeW.dim[1], ncol=final.pairs.for.trading.dim[1])
meanT.i <- matrix(NA, nrow = final.pairs.for.trading.dim[1], ncol=1)
sdT.i <- matrix(NA, nrow = final.pairs.for.trading.dim[1], ncol=1)
upperThr.i <- matrix(NA, nrow = final.pairs.for.trading.dim[1], ncol=1)
lowerThr.stoploss.i <- matrix(NA, nrow = final.pairs.for.trading.dim[1], ncol=1)
lowerThr.takeprofit.i <- matrix(NA, nrow = final.pairs.for.trading.dim[1], ncol=1)
for (i in 1:final.pairs.for.trading.dim[1]){  
    meanT.i[i,1] <- as.numeric(mean(TestW.portfolio.spreads[,i],na.rm=TRUE))
    sdT.i[i,1] <- as.numeric(sd(TestW.portfolio.spreads[,i],na.rm=TRUE))
    upperThr.i[i,1] <- meanT.i[i,1] + 2 * sdT.i[i,1]
    lowerThr.i[i,1] <- meanT.i[i,1] - 2 * sdT.i[i,1]
    upperThr.stoploss.i[i,1] <- meanT.i[i,1] + 3 * sdT.i[i,1]
    lowerThr.stoploss.i[i,1] <- meanT.i[i,1] - 3 * sdT.i[i,1]
    upperThr.takeprofit.i[i,1] <- meanT.i[i,1] + 0.25 * sdT.i[i,1]
    lowerThr.takeprofit.i[i,1] <- meanT.i[i,1] - 0.25 * sdT.i[i,1]
    for (j in 1:TradeW.dim[1]){  
        if ((Trade.portfolio.spreads.named[j,i] > upperThr.i[i,1]) && (Trade.portfolio.spreads.named[j,i]
            < upperThr.stoploss.i[i,1])){
            trades.tradew[j,i] <- 1
        } else if ((Trade.portfolio.spreads.named[j,i] < lowerThr.i[i,1]) && (Trade.portfolio.spreads.named[j,i]
            > lowerThr.stoploss.i[i,1])){
            trades.tradew[j,i] <<- (-1)
        } else if (Trade.portfolio.spreads.named[j,i] < lowerThr.stoploss.i[i,1]){
    } else if (Trade.portfolio.spreads.named[j,i] > upperThr.takeprofit.i[i,1]){
```r
trades.tradew[j, i] <- (2)
} else if (Trade.portfolio.spreads.named[j, i] > upperThr.stoploss.i[i, 1]){
  trades.tradew[j, i] <- (2)
} else {
  trades.tradew[j, i] <- 0
}
}

# Second iteration to simulate holding the pair when we buy outside of the upper/lower thr. and it correct towards mean
for (i in 1:final.pairs.for.trading.dim[1]){
  for (j in 2:TradeW.dim[1]){
    if ((trades.tradew[j-1, i] == 1) && (Trade.portfolio.spreads.named[j, i] < upperThr.i[i, 1]) &&
        (Trade.portfolio.spreads.named[j, i] > upperThr.takeprofit.i[i, 1])){
      trades.tradew[j, i] <- 1
    } else if ((trades.tradew[j-1, i] == -1) && (Trade.portfolio.spreads.named[j, i] > lowerThr.i[i, 1]) &&
               (Trade.portfolio.spreads.named[j, i] < lowerThr.takeprofit.i[i, 1])){
      trades.tradew[j, i] <- -1
    }
  }
}

# Take exp of log:ed TradeW
TradeW <- exp(TradeW)

# Calculate return
return.portfolio <- matrix(NA, nrow=TradeW.dim[1], ncol=final.pairs.for.trading.dim[1])
return.portfolio[1,] <- 1
for (i in 1:final.pairs.for.trading.dim[1]){
  for (j in 2:(TradeW.dim[1])){
    if ((trades.tradew[j-1, i] == -1)){
      return.portfolio[j, i] <- matrix(matrix(return.portfolio[j-1, i]) + matrix((return.portfolio[j-1, i]) * 0.5 *
                                             ((matrix(TradeW[j-1, final.pairs.for.trading[i, 1]]) - matrix(TradeW[j-1, final.pairs.for.trading[i, 2]]) +
                                             (matrix(TradeW[j-1, final.pairs.for.trading[i, 2]]) - matrix(TradeW[j, final.pairs.for.trading[i, 2]])))/matrix(TradeW[j-1, final.pairs.for.trading[i, 2]])))
    } else if ((trades.tradew[j-1, i] == 1)){
      return.portfolio[j, i] <- matrix(matrix(return.portfolio[j-1, i]) + matrix((return.portfolio[j-1, i]) * 0.5 *
                                             ((matrix(TradeW[j-1, final.pairs.for.trading[i, 2]]) - matrix(TradeW[j-1, final.pairs.for.trading[i, 1]]) +
                                             (matrix(TradeW[j-1, final.pairs.for.trading[i, 1]]) - matrix(TradeW[j, final.pairs.for.trading[i, 1]])))/matrix(TradeW[j-1, final.pairs.for.trading[i, 1]])))
    } else {
      return.portfolio[j, i] <- matrix(return.portfolio[j-1, i])
    }
  }
}

return.portfolio.dim <- dim(return.portfolio)
((sum(return.portfolio[return.portfolio.dim[1],]))-number.of.pairs.in.portfolio)
```
/number.of.pairs.in.portfolio

#Export Excel sheets of return
write.table(return.portfolio,"/Users/Scruffy/Desktop/Statistik C/Return.20.pairs.2010.txt", sep="\t")

#############################################################################
#Plot Test and Trade Window#############################################################################

#Plot to see test window
par(mfrow=c(2,3))#plot 2 by 3 matrix
for (i in 1:final.pairs.for.trading.dim[1]){
  stock.pair.for.plot <- i
  spread.chosen.stock <- matrix( TestW.portfolio.spreads[,stock.pair.for.plot])
  plot(spread.chosen.stock, main="TestW")
  lines(spread.chosen.stock)
  abline(h = meanT.i[stock.pair.for.plot,1], col = "red", lwd =2)
  abline(h = meanT.i[stock.pair.for.plot,1] + 1 * sdT.i[stock.pair.for.plot,1], col = "blue", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] - 1 * sdT.i[stock.pair.for.plot,1], col = "blue", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] + 2 * sdT.i[stock.pair.for.plot,1], col = "yellow", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] - 2 * sdT.i[stock.pair.for.plot,1], col = "yellow", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] + 0.25 * sdT.i[stock.pair.for.plot,1], col = "green", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] - 0.25 * sdT.i[stock.pair.for.plot,1], col = "green", lwd=2)
}

#Plot to see trading
par(mfrow=c(2,3))#
for (i in 1:10){
  stock.pair.for.plot <- i
  spread.chosen.stock <- matrix(Trade.portfolio.spreads.named[,stock.pair.for.plot])
  plot(spread.chosen.stock, yaxt='n', ann=FALSE)
  lines(spread.chosen.stock)
  abline(h = meanT.i[stock.pair.for.plot,1], col = "red", lwd =2)
  abline(h = meanT.i[stock.pair.for.plot,1] + 1 * sdT.i[stock.pair.for.plot,1], col = "blue", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] - 1 * sdT.i[stock.pair.for.plot,1], col = "blue", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] + 2 * sdT.i[stock.pair.for.plot,1], col = "yellow", lwd=2)
  abline(h = meanT.i[stock.pair.for.plot,1] - 2 * sdT.i[stock.pair.for.plot,1], col = "yellow", lwd=2)
  #abline(h = meanT.i[stock.pair.for.plot,1] + 0.25 * sdT.i[stock.pair.for.plot,1], col = "green", lwd=2)
  #abline(h = meanT.i[stock.pair.for.plot,1] - 0.25 * sdT.i[stock.pair.for.plot,1], col = "green", lwd=2)
}

34