Stockholms universitet

Forecasting on the spread between two stocks in pairs trading

Daniella Zhou

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:13
Matematisk statistik
Juni 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:13**
http://www.math.su.se

Forecasting on the spread between two stocks in pairs trading

Daniella Zhou*

June 2020

Abstract

This thesis is about to forecast the spread between the two stocks traded in pairs with 1- and 2-step-ahead in two different ways: directly forecast the spread series using an univariate AR model; and indirectlyforecast the spread series through the forecasting of the prices of the associated stocks using a multivariate VAR yielded ECM model. The fore-casting process follows a time series cross-validation procedure and the Naive forecasting approach is used as a benchmark. The results indicate that none of the AR or VAR yielded ECM model is outperform the Naive approach, however the ECM model is in overall more accurate than the simple AR model.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: dazh9919@student.su.se. Supervisor: Kristoffer Lindensjö, Mathias Lindholm and Felix Wahl.

Acknowledgements

This is a Bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University. I would like to thank my supervisor Kristofer Lindensjö for his helpful inspirations, patient guidance and valuable suggestions. I would also like to thank Mathias Lindholm and Felix Wahl and all my friends and families for their support and encouragement.

Contents

1 Introduction

Pairs trading is a trading strategy that constructed with a market-neutral portfolio which only involves two highly correlated stocks or other securities. Such market-neutral portfolio gives a return which is uncorrelated with the market return. That is, the return provided by the portfolio has its own performance regardless how market changes. Let p_t denote the price at time t and the return r_t is defined as:

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}$$

In this study we will focus on the statistical arbitrage pairs trading which has an idea based on the relative pricing and the arbitrage pricing theory (APT).

The APT is a multi-factor asset pricing model for which that the expected return of a financial asset can be modeled by a linear combination of the expected risk-free rate and several macroeconomic risk-captured variables. This implies that the expected return of two securities should be the same at each specified time frame if these two securities are exposing to the same risk factors and are having exactly the same sensitivity to the factors. If the expected returns differ, it is likely that one of the security is over-priced and the other one under-priced. When such differ appears, investors may gain profits by putting on a long position in the under-priced security and a short position in the other one in some predetermined ratio with an expectation on that the mispricing will correct itself and revert back in the future. The degree of mutual mispricing, or the scaled price difference is called the spread. "The greater the spread, the higher the magnitude of mispricing and greater the profit potential" (Vidyamurthy 2004, p. 74).

This study is about this spread between the two chosen stocks in pairs trading. The main purpose is to forecast this price gap in order to construct some trading strategies in the future if possible. The forecasting is implemented in two different approaches with two different but related data sets, one is to directly forecast this spread based on already observed data, and the other one is to forecast the prices of the chosen stocks and thereon calculate the associated spread. Two time series models, an univariate AR model and a multivariate VAR yielded ECM model are used and compared in terms of the forecasting accuracy.

To fully understand the study, an introduction about some helpful background theories within the univariate as well as multivariate time series analysis and the study related models - AR, VAR and VAR yielded ECM are given in advance in section 2 and 3. Thereafter some useful tests and tools in the time series analysis and forecast are presented in section 4. The main study is in the section 5.

2 Univariate Time Series Model

In this section, we introduce first some basic and important concepts related to the univariate time series analysis, and thereafter the model that is used in the study.

2.1 Stationarity

The stationarity is the central concept within the time series analysis. An univariate time series x_t is said to be *strictly stationary* if the jointly distribution of $\{x_{t_1}, \ldots, x_{t_n}\}$ is invariant regardless how time indexes shift. That is, for any k, the jointly distribution of $\{x_{t_1}, \ldots, x_{t_n}\}$ and $\{x_{t_1+k}, \ldots, x_{t_n+k}\}$ is the same. This strong condition is hard to verify empirically, hence a weaker version is often assumed. An univariate time series x_t is said to be *weakly stationary* if both the mean of x_t and the covariance between x_t and x_{t-l}, for any arbitrary integer l, do not depending on time (Tsay 2010, p. 30). That is:

$$\begin{aligned} E[x_t] &= \mu \\ Cov(x_t, x_{t-l}) &= \gamma_l \end{aligned}$$

Note that the variance of a weakly stationary time series, i.e. $Var(x_t) = Cov(x_t, x_t) = \gamma_0$, is a time invariant constant. The covariance γ_l is often called the lag-l autocovariance.

Any further mention of stationary is refer to weakly stationary assumptions.

2.2 Autocorrelation Function, ACF

The autocorrelation is simply the correlation between the time series x_t and its past values x_{t-l}, for any arbitrary non-negative integer l. Under the assumption of weakly stationary, the so called lag-l autocorrelation is defined as (Tsay 2010, p. 31):

$$\rho_l = \frac{Cov(x_t, x_{t-l})}{\sqrt{Var(x_t)Var(x_{t-l})}} = \frac{\gamma_l}{\gamma_0}$$

A stationary time series x_t is not serially correlated if and only if $\rho_l = 0$ for all $l > 0$.

Sample Autocorrelation Fuction

The lag-l sample ACF can be estimated by:

$$\hat{\rho}_l = \frac{\sum_{t=l+1}^{T}(x_t - \bar{x})(x_{t-l} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \qquad 0 \le l \le T-1 \tag{2.1}$$

where $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ is the sample mean.

2.3 White noise

A white noise is a series of independent and identically distributed (i.i.d.) random variables with finite mean and variance. The ACFs of the white noise series is close to zero duo to the independent characteristics. If the white noise is normally distributed with mean zero and variance σ^2, i.e. $\sim N(0, \sigma^2)$, then the series is called as Gaussian white noise.

2.4 Simple Autoregressive Model, AR

The autoregressive (AR) model presents a linear combination between the current value of an variable and its own previous values. An autoregressive model of order p, short as AR(p) model, is defined as (Tsay 2010, p. 37-46):

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t \tag{2.2}$$

where p is a positive integer, ϕ_0 is the constant term/intercept, $\phi_1, ..., \phi_p$ are the parameters of the model and a_t is a white noise series with zero mean and variance σ_a^2. An AR(p) model is simply saying that the current value of the variable is determined jointly by a linear combination of the latest p previous values and a white noise series. Under the stationary assumptions, the expected value of x_t and x_{t-l} is the same and as well as the variance of x_t and x_{t-l} for any positive integer l. With these two conditions one can obtain the mean value and the variance of the stationary AR(p) model as:

$$E[x_t] = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p} = \mu \tag{2.3}$$

$$Var(x_t) = \frac{\sigma_a^2}{1 - \phi_1^2 - \cdots - \phi_p^2} \tag{2.4}$$

implying that $\phi_1 + \cdots + \phi_p < 1$ and $\phi_1^2 + \cdots + \phi_p^2 < 1$. Using $\phi_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$ and $Cov(a_t, x_{t-l}) = 0$, the ACFs of x_t can be written as:

$$\rho_l = \phi_1 \rho_{l-1} + \cdots + \phi_p \rho_{l-p} \quad \text{for } l > 0$$

Use a so called back-shift operator B such that $B\rho_l = \rho_{l-1}$ one can obtain the following property:

$$(1 - \phi_1 B - \cdots - \phi_p B^p)\rho_l = 0 \quad \text{for } l > 0$$

The equation $(1 - \phi_1 B - \cdots - \phi_p B^p) = 0$ is called the p-order difference equation or the characteristic equation for the AR(p) model and the inverse of the solutions to this equation are called characteristic roots. It can be shown that an stationary AR(p) model satisfies that the absolute value of the characteristic roots are less than 1 and the ACFs of x_t have a exponential decays.

With the back-shift operator B and the characteristic roots ω, the AR(p) model can be rewritten into the form as:

$$(1 - \omega_1 B)(1 - \omega_2 B) \cdots (1 - \omega_p B) = \phi_0 + a_t \tag{2.5}$$

Forecasting for AR

Forecast of a time series is equivalent to compute the expected value of the variable at a future time point of interest with conditional on all the collected available information at current time.

Suppose we are at time h and interested in the forecast value at time $h+l$ (both h and l are arbitrary positive integers), where the h is called as forecasting origin and l the forecasting horizon. Consider an AR(p) model, the value of x_{h+l} is:

$$x_{h+l} = \phi_0 + \sum_{i=1}^{p} \phi_i x_{h+l-i} + a_{h+l}$$

the l-step-ahead forecasting of x_{h+l}, denote as $\hat{x}_h(l)$, is equivalent to the expected value of the x_{h+l} given all the available information at time h which denote as $\mathcal{F}_h$ (Tsay 2010, p. 54-56):

$$\hat{x}_h(l) = E[x_{h+l}|\mathcal{F}_h] = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{x}_h(l-i)$$

Note that $\hat{x}_h(l-i)$ on the right hand is the same as x_{h+l-i} for any integer $l-i \leq 0$. This multiple forecast can be computed recursively through $\hat{x}_h(1)$, $\hat{x}_h(2)$ until the interested step. The associated forecasting error is:

$$e_h(l) = x_{h+l} - \hat{x}_h(l)$$

It can be shown that for a stationary AR(p) model, $\hat{x}_h(l)$ converges to the unconditional expected value $E[x_t] = \frac{\phi_0}{1-\phi_1-\cdots-\phi_p}$, and the variance of the forecast errors, $Var(e_h(l))$ converges to the sample variance $Var(x_t) = \frac{\sigma_a^2}{1-\phi_1^2-\cdots-\phi_p^2}$ as $l \to \infty$ (Tsay 2010, p. 56). These limitations on the forecasting mean and variance makes a stationary AR(p) model predictable. Such property refers to the mean-reverting in finance literature.

2.5 Unit-root Nonstationary and Random Walk

In general, a price series is normally tend to be nonstationary, namely has time dependent mean and/or variance due to the absence of the fixed price level. Such nonstationary series x_t is called as unit-root nonstationary time series and is usually following a random walk process such defined as (Tsay 2010, p. 71-73):

$$x_t = x_{t-1} + a_t \tag{2.6}$$

where a_t is a white noise series with mean zero and variance σ_a^2. The expression is simply saying that with conditional on previous value x_{t-1} and a white noise series, the probability for the current value of x_t to go up or down is the same. This simple random walk model can be seen as a special case of an AR(1) model with constant ϕ_0 equals zero and parameter ϕ_1 equals one.

A random walk process is not mean-reverting, this can be shown through a long-term forecast. Given a random walk process x_t, the l-step-ahead forecast and the corresponding variance of forecast error are:

$$\begin{aligned}\hat{x}_h(l) &= x_h \\ Var\left(e_h(l)\right) &= l\sigma_a^2\end{aligned}$$

which implies that the forecast value is always the same as the value at the origin and the variance of the forecast error goes to infinity as forecasting horizon goes to infinity. Thus a random walk process is not predictable and need to be transformed into stationary in order to do forecast.

Differencing and Integrating

A common way to overcome with the non-stationarity is to construct a series that express the changing in the original series with a specified time interval, such operation is called differencing (Tsay 2010, p. 76). For instance, consider an unit-root nonstationary random walk series x_t in Eq. 2.6, the corresponding first differenced series is then the one time-unit changing of the x_t, i.e:

$$\Delta x_t = x_t - x_{t-1} = a_t$$

The mean and variance of the white noise series a_t is time invariant and hence is stationary.

Series that can be transformed into stationary through d times differencing is said to be integrated of order d, denote as $I(d)$.

3 Multivariate Time Series Model

Same as the univariate time series section, we begin with some related concepts within multivariate time series analysis and thereafter introduce the models that are involved in this study. Since the time series analysis in this study is only regarding two or less times series, all further introduction about the multivariate times series are restricted to the bivariate version.

3.1 Stationarity

The stationarity for a multivariate time series is straightforward. A weakly stationary bivariate time series $\boldsymbol{x}_t$ satisfies following assumptions (Tsay 2010, p. 390-391):

$$\begin{aligned}\boldsymbol{\mu} &= E[\boldsymbol{x}_t] \\ \boldsymbol{\Gamma}_0 &= E[(\boldsymbol{x}_t - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})^T]\end{aligned}$$

where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Gamma}_0$ is the 2×2 symmetric cross-covariance matrix of $\boldsymbol{x}_t$. Note that the matrix and vector variables are denote in bold. The diagonal elements in $\mathbf{\Gamma}_0$ are the variances of x_{1t} and x_{2t} respectively, and the off-diagonal elements are the covariance between x_{1t} and x_{2t}. Analogously, the lag-l cross-covariance matrix is defined as:

$$\mathbf{\Gamma}_l = E[(\boldsymbol{x}_t - \boldsymbol{\mu})(\boldsymbol{x}_{t-l} - \boldsymbol{\mu})^T]$$

Let $\Gamma_{ij}(l)$ denote the ij:th element in $\mathbf{\Gamma}_l$, and note that $\Gamma_{12}(l)$ is the covariance between x_{1t} and $x_{2,t-l}$ and $\Gamma_{21}(l)$ is the covariance between x_{2t} and $x_{1,t-l}$, which measures different relationships. Thus the lag-l cross-covariance matrix is asymmetric.

3.2 Cross-Correlation Matrix, CCM

The multivariate version of autocorrelation is called cross-correlation and the so called lag-l cross-correlation matrix is defined as (Tsay 2010, p. 391-392):

$$\boldsymbol{\rho}_l = \boldsymbol{D}^{-1}\mathbf{\Gamma}_l\boldsymbol{D}^{-1}$$

where $\boldsymbol{D}$ is a diagonal matrix for which diagonal elements equal to the square root of the diagonal elements in $\mathbf{\Gamma}_0$, that is, the standard deviation of the components in $\boldsymbol{x}_t$,. Let $\rho_{ij}(l)$ denote the ij:th element in $\boldsymbol{\rho}_l$, which is the correlation coefficient between x_{it} and $x_{j,t-l}$, can be written as:

$$\rho_{ij}(l) = \frac{\Gamma_{ij}(l)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{Cov(x_{it}, x_{j,t-l})}{sd(x_{it})sd(x_{jt})}$$

where $\Gamma_{ij}(l)$ is the ij:th element in $\mathbf{\Gamma}_l$. For the same reason as $\mathbf{\Gamma}_l$, the lag-l cross-correlation matrix $\boldsymbol{\rho}_l$ is asymmetric. The diagonal elements in $\boldsymbol{\rho}_l$ are actually the lag-l autocorrelation coefficients of the components in $\boldsymbol{x_t}$.

Sample Cross-Correlation Matrix

The lag-l sample cross-correlation matrix can be estimated through the lag-l sample cross-covariance matrix defined as follows (Tsay 2010, p. 392-393):

$$\hat{\mathbf{\Gamma}}_l = \frac{1}{T}\sum_{t=l+1}^{T}(\boldsymbol{x}_t - \bar{\boldsymbol{x}})(\boldsymbol{x}_{t-l} - \bar{\boldsymbol{x}})^T, \qquad l \geq 0 \tag{3.1}$$

where $\bar{\boldsymbol{x}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t$ is the sample mean vector. Let $\hat{\boldsymbol{D}}$ denote the sample standard deviation of the components in $\boldsymbol{x}_t$, the lag-l sample CCM is:

$$\hat{\boldsymbol{\rho}}_l = \hat{\boldsymbol{D}}^{-1}\hat{\mathbf{\Gamma}}_l\hat{\boldsymbol{D}}^{-1}, \qquad l \geq 0 \tag{3.2}$$

3.3 Vector Autoregressive Models, VAR

The vector autoregressive (VAR) model is useful in modeling several time series jointly. Similar to the simple autoregressive model, the VAR model is just a multi-dimensional version of the simple AR model. Consider a 2 time series x_{1t} and x_{2t}, the VAR model of order 1, shortly VAR(1) model of the two series is (Tsay 2010, p. 399-405):

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} \phi_{10} \\ \phi_{20} \end{bmatrix} + \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}$$

which can be conveniently written as:

$$\boldsymbol{x}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{x}_{t-1} + \boldsymbol{a}_t \tag{3.3}$$

where $\boldsymbol{\phi}_0$ is a vector of the constant terms/intercepts, $\boldsymbol{\Phi}_1$ is the parameter matrix and $\boldsymbol{a}_t$ is a vector of serially uncorrelated random white noise with mean zero and a 2×2 symmetric covariance matrix $\boldsymbol{\Sigma}$. The diagonal elements in $\boldsymbol{\Sigma}$-matrix measure the components' variance and the off-diagonal elements measure the concurrent covariance relationship between the components x_{1t} and x_{2t}. The elements in $\boldsymbol{\Phi}_1$-matrix give the dynamic linear relationship within and between the components. For instance, Φ_{11} measures the linear dependence of x_{1t} and $x_{1,t-1}$ which can be seen as the ϕ_1 parameter in the univariate AR model and Φ_{12} measures the linear dependence of x_{1t} and $x_{2,t-1}$. The same with Φ_{22} and Φ_{21}, but note that Φ_{21} measures the linear relationship between x_{2t} and $x_{1,t-1}$ which is different from Φ_{12}. Hence $\boldsymbol{\Phi}_1$-matrix is not symmetric.

More generally, a VAR(p) model is as follows:

$$\boldsymbol{x}_t = \boldsymbol{\phi_0} + \boldsymbol{\Phi}_1 \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{x}_{t-p} + \boldsymbol{a}_t \tag{3.4}$$

where $\boldsymbol{\phi}_0$ and $\boldsymbol{a}_t$ are the same as in Eq. 3.3, and $\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_p$ are the parameter matrix of the model. Using back-shift operator, the model can be rewritten into:

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p)\boldsymbol{x}_t = \boldsymbol{\phi}_0 + \boldsymbol{a}_t$$

where $\boldsymbol{I}$ is a 2×2 identity matrix in bivariate case.

A weakly stationary VAR(p) model satisfies:

- $E[\boldsymbol{x}_t] = E[\boldsymbol{x}_{t-l}]$, for any $l > 0$.
- $E[\boldsymbol{x}_t] = (\boldsymbol{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p)^{-1}\boldsymbol{\phi}_0$, where $\boldsymbol{I}$ is a 2×2 identity matrix in the case of bivariate time series.
- $Cov(\boldsymbol{a}_t, \boldsymbol{x}_t) = \boldsymbol{\Sigma}$, which is the cov-matrix of $\boldsymbol{a}_t$.
- $Cov(\boldsymbol{a}_t, \boldsymbol{x}_{t-l}) = 0$ for any $l > 0$.
- $\boldsymbol{\Gamma}_l = \boldsymbol{\Phi}_1 \boldsymbol{\Gamma}_{l-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{\Gamma}_{l-p}$, for any $l > 0$.
- $\boldsymbol{\rho}_l = \boldsymbol{\Upsilon}_1 \boldsymbol{\rho}_{l-1} + \cdots + \boldsymbol{\Upsilon}_p \boldsymbol{\rho}_{l-p}$, for any $l > 0$, where $\boldsymbol{\Upsilon}_i = \boldsymbol{D}^{-1/2} \boldsymbol{\Phi}_i \boldsymbol{D}^{1/2}$ and $\boldsymbol{D}$ is the diagonal matrix of the standard deviations of $\boldsymbol{x}_t$.
- All zeros of the determinant $|\boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p|$ are located outside the unit circle.

Forecasting for VAR

The forecasting of VAR model can be derived analogously as the forecasting of AR model (Tsay 2010, p. 409).

Consider an VAR(p) model, the value of $\boldsymbol{x}_{h+l}$ is:

$$\boldsymbol{x}_{h+l} = \boldsymbol{\phi}_{\mathbf{0}} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{x}_{h+l-i} + \boldsymbol{a}_{h+l}$$

The l-step-ahead forecast for $\boldsymbol{x}_{h+l}$, conditional on $\mathcal{F}_h$ is then:

$$\hat{\boldsymbol{x}}_h(l) = \boldsymbol{\phi}_{\mathbf{0}} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \hat{\boldsymbol{x}}_h(l-i)$$

the value of $\hat{\boldsymbol{x}}_h(l-i)$ is equal to the $\boldsymbol{x}_{h+l-i}$ for any integer $l - i \leq 0$. And the associated forecasting error is:

$$\boldsymbol{e}_h(l) = \boldsymbol{x}_{h+l} - \hat{\boldsymbol{x}}_h(l)$$

The forecasting for VAR has the same property as the forecasting for AR, that the long-term point forecasting will approaches the unconditional mean of the seires. In other words, a mean-reverting will arise for large l.

3.4 Cointegration

As mentioned earlier that most price series are unit-root nonstationary as well as the stock prices for which is the study object in this thesis. An interesting phenomenon called *cointegration* arises when modelling two unit-root nonstationary time series jointly and is the key characteristics involving pairs trading. The term cointegration was coined by two econometricians Engle & Granger (1987) and they had given a definition of cointegration as follow:

DEFINITION*(cointegration)*: The components of the vector $\boldsymbol{x}_t$ said to be co-integrated of order (d,b), denoted $\boldsymbol{x}_t \sim CI(d,b)$, if (i) all components of $\boldsymbol{x}_t$ are $I(d)$; (ii) there exists a vector $\boldsymbol{\beta}(\neq 0)$ so that $\boldsymbol{w}_t = \boldsymbol{\beta}^T \boldsymbol{x}_t \sim I(d-b), 0 < b \leq d$. The vector $\boldsymbol{\beta}$ is called the cointegrating vector.

In other words, consider two unit-root nonstationary time series x_{1t} and x_{2t}, both are integrated of order 1. If there exist a certain vector $\boldsymbol{\beta} = \begin{bmatrix} 1 & -\beta_1 \end{bmatrix}^T$, such that the series $w_t = x_{1t} - \beta_1 x_{2t}$ is stationary, i.e. $I(0)$, then x_{1t} and x_{2t} is said to be cointegrated.

The stationary linear combination $x_{1t} - \beta_1 x_{2t}$ refers to the spread for which is interested in pairs trading. A presence of the time invariant mean and variance in the spread implies that the components series in the system will adjust themselves to restore this mean if a deviation ever appears. This is the mean-reverting characteristics that requires for the spread between the two stocks in pairs trading. Additionally, a comovement is always appeared between the two cointegrated series (Vidyamurthy 2004, p. 75-76).

3.5 Error Correction Model, ECM

Since there exits 2 unit-root nonstationary time series, while the number of stationary series is only 1 in the cointegration system in bivariate case, therefore the implementation of differencing on each components to achieve stationarity will leads to over-differencing (Tsay 2010, p. 431). To overcome with this over-differencing, an error correction representation is introduced.

3.5.1 Error Correction Model for VAR

Consider a 2-dimensional time series $\boldsymbol{x}_t$ with a VAR(p) model in Eq. 3.4 and let $\Delta\boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$ be the differenced series. Using back-shift operator one can write that $\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p$ and the error correction model for the bivariate VAR(p) process is (Tsay 2010, p. 432-434):

$$\Delta\boldsymbol{x}_t = \boldsymbol{\mu}_t + \boldsymbol{\Pi}\boldsymbol{x}_{t-1} + \boldsymbol{\Phi}_1^*\Delta\boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}^*\Delta\boldsymbol{x}_{t-p+1} + \boldsymbol{a}_t \tag{3.5}$$

where $\boldsymbol{\Pi} = -\boldsymbol{\Phi}(1)$, and $\boldsymbol{\Phi}_j^* = -\sum_{i=j+1}^{p}\boldsymbol{\Phi}_i$ (detailed derivation see appendix C). $\boldsymbol{\mu}_t$ is the 2-dimensional deterministic vector and can be written as $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t$. The role of this deterministic vector will be described in section 3.5.2. We use ECMvar to denote this conversion model of VAR.

The rank of $\boldsymbol{\Pi}$ is the number of cointegrating vectors. If the rank of $\boldsymbol{\Pi}$ is zero implies that $\boldsymbol{x}_t$ is not cointegrated, and if $\boldsymbol{\Pi}$ has full rank implies that no unit-roots are consisted in $\boldsymbol{x}_t$ and ECM becomes meaningless. Therefore the rank of $\boldsymbol{\Pi}$ can be used to determine the existence of cointegration. In the bivariate case, if the rank of $\boldsymbol{\Pi}$ is 1, then the Eq. 3.5 can be rewritten as:

$$\Delta\boldsymbol{x}_t = \boldsymbol{\mu}_t + \boldsymbol{\alpha}\boldsymbol{\beta}^T\boldsymbol{x}_{t-1} + \boldsymbol{\Phi}_1^*\Delta\boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}^*\Delta\boldsymbol{x}_{t-p+1} + \boldsymbol{a}_t \tag{3.6}$$

where both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are (2×1)-matrix and with rank 1, such that $\boldsymbol{\alpha}\boldsymbol{\beta}^T = \boldsymbol{\Pi}$.

Note that the $\boldsymbol{\beta}$-matrix need to has a form as $\begin{bmatrix}1 & -\beta_1\end{bmatrix}^T$ in order to give a stationary series $w_t = \boldsymbol{\beta}^T\boldsymbol{x}_t = x_{1t} - \beta_1 x_{2t}$. This w_t is the spread series as we have already discussed in section 3.4.

3.5.2 Deterministic Function of ECMVar

The deterministic function $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t$ tells if there exist any time trend/drift in a time series. It has the same structure as a simple linear regression. $\boldsymbol{\mu}_0$ is the constant/intercept term. $\boldsymbol{\mu}_1 t$ can be seen as the time dependent slope. If there exist a time trend, then the time series will have values distributed surround this slope.

There are five specific deterministic trend cases considered by Johansen (1995, p. 80-84) and are summarized by Tsay (2010, p. 434-435) as follows:

1. $\boldsymbol{\mu}_t = 0$:

$$\Delta\boldsymbol{x}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^T\boldsymbol{x}_{t-1} + \boldsymbol{\Phi}_1^*\Delta\boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}^*\Delta\boldsymbol{x}_{t-p+1} + \boldsymbol{a}_t$$

Components of $\boldsymbol{x}_t$ are $I(1)$ but has no trends, w_t has no intercept.

2. $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 = \boldsymbol{\alpha} c_0$:

$$\Delta \boldsymbol{x}_t = \boldsymbol{\alpha}(\boldsymbol{\beta}^T \boldsymbol{x}_{t-1} + \boldsymbol{c}_0) + \boldsymbol{\Phi}_1^* \Delta \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}^* \Delta \boldsymbol{x}_{t-p+1} + \boldsymbol{a}_t$$

where c_0 is a non-zero constant in the bivariate case. Components of $\boldsymbol{x}_t$ are $I(1)$ but has no trends, and w_t has an intercept at $-c_0$.

3. $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0$ (non-zero):

Components of $\boldsymbol{x}_t$ are $I(1)$ and has linear trends, and w_t have an intercept.

4. $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\alpha} c_1 t$:

$$\Delta \boldsymbol{x}_t = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}(c_1 t + \boldsymbol{\beta}^T \boldsymbol{x}_{t-1}) + \boldsymbol{\Phi}_1^* \Delta \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_{p-1}^* \Delta \boldsymbol{x}_{t-p+1} + \boldsymbol{a}_t$$

where c_1 is a non-zero constant in the bivariate case. Both components of $\boldsymbol{x}_t$ and w_t have linear trends.

5. $\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t$:

Components of $\boldsymbol{x}_t$ are $I(1)$ and has quadratic trends, w_t has a linear trend.

The first and fifth case are very rare and not common in empirical work. The first case may presents some log-price series and is therefore only used when all components in $\boldsymbol{x}_t$ have zero mean.

3.5.3 Forecasting for ECMVar

The ECMVar model can be used to obtain the forecasting of $\boldsymbol{x}_t$ through the forecast of the differenced series $\Delta \boldsymbol{x}_t$. The advantage with ECMvar forecast is that it imposes the cointegration relationships among the series during the prediction process (Tsay 2010, p. 437).

Consider a ECMvar model with order $p-1$, where p is the corresponding order of the VAR model. With a forecasting origin h and a forecasting horizon l, we obtain:

$$\Delta \boldsymbol{x}_{h+l} = \boldsymbol{\mu}_{h+l} + \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{x}_{h+l-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^* \Delta \boldsymbol{x}_{h+l-i} + \boldsymbol{a}_{h+l}$$
$$\Delta \hat{\boldsymbol{x}}_h(l) = \boldsymbol{\mu}_{h+l} + \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{x}_{h+l-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^* \Delta \hat{\boldsymbol{x}}_h(l-i)$$

where $\Delta \hat{\boldsymbol{x}}_h(l-i) = \Delta \boldsymbol{x}_{h+l-i}$ for any integer $l - i \leq 0$. The associated forecast error is then:

$$\Delta \boldsymbol{e}_h(l) = \Delta \boldsymbol{x}_{h+l} - \Delta \hat{\boldsymbol{x}}_h(l)$$

4 Model Selection, Examination and Evaluation

There are many things that need to be concerned when construct a model. Did the series satisfy the assumptions? Which order should be chosen? How to evaluate the model? And etc. In this section we introduce several useful approaches to determine the stationarity, cointegration, order of the model and the accuracy of the prediction.

4.1 Augmented Dicky-Fuller Test, ADF

There are several ways to determine the stationarity of an univariate time series, here we introduce the augmented Dicky-Fuller test which determine the stationarity by testing the presence of unit-root in the time series.

Consider an AR(p) model in form of Eq. 2.2, the ADF-test is to conduct a t-test on the least square estimated parameter $\hat{\gamma}$ which provided by the equation (Tsay 2010, p. 77):

$$x_t = c_t + \beta x_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta x_{t-i} + a_t$$

where c_t is the deterministic function which can be written as $c_t = c_0 + c_1 t$, $\Delta x_t = x_t - x_{t-1}$ is the first differenced series of x_t. If the coefficient for the x_{t-1} term equals 1 then an unit-root is appeared and the series x_t is then said to be nonstationary. The test hypothesis is then $H_0 : \beta = 1$ versus $H_1 : \beta < 1$ and the corresponding t-statistic is:

$$\text{ADF-test} = \frac{\hat{\beta} - 1}{std(\hat{\beta})}$$

The null hypothesis is rejected if the test statistic exceeds the critical value at an appropriate level.

4.2 Johansen's Cointegration Test

As mentioned in section 3.5.1 that the rank of $\mathbf{\Pi}$ is the number of cointegrating vectors, thus one can examine the rank of $\mathbf{\Pi}$ to determine the existence of cointegration. This is the approach taken by Johansen (1988, 1995). There are two type of Johansen's cointegration test on VAR model:

Trace

Consider the hypothesis

$$H_0 : \text{Rank}(\mathbf{\Pi}) = m \quad \text{versus} \quad H_a : \text{Rank}(\mathbf{\Pi}) > m$$

The associated likelihood ratio (LR) statistic:

$$\mathrm{LR}_{tr}(m) = -(T-p)\sum_{i=m+1}^{k} ln(1-\hat{\lambda}_i)$$

where T is the sample size, p is the order of the model, k is the dimension of the time series and $\hat{\lambda}_i$ is the corresponding eigenvalues of the $\mathbf{\Pi}$-matrix.

Maximum eigenvalue

Consider the hypothesis

$$H_0 : \mathrm{Rank}(\mathbf{\Pi}) = m \quad \text{versus} \quad H_a : \mathrm{Rank}(\mathbf{\Pi}) = m+1$$

The associated LR statistic:

$$\mathrm{LR}_{max}(m) = -(T-p)ln(1-\hat{\lambda}_{m+1})$$

where T, p and $\hat{\gamma}_i$ defines the same as in trace test.

Both trace and max-eigenvalue test start with $m = 0$ until $m = k-1$ where k is the dimension of the $\mathbf{\Pi}$-matrix. At each m the null-hypothesis is rejected if the test statistic exceeds the critical value at an appropriate level, implying that the number of cointegrating vectors is larger than the current m. Because of the presence of unit-root, both $\mathrm{LR}_{tr}(m)$ and $\mathrm{LR}_{max}(m)$ statistics are not chi-square distributed but a function of standard Brownian motions which makes the critical values nonstandard. Thus the corresponding critical values must be evaluated via simulation (Tsay 2010, p. 436-437). Osterwald-Lenum (1992) had proposed a suggestion of the critical values at 1% and 5% levels. MacKinnon, Haug & Michelis (1999) had based on these proposed a more complete and accurate critical p-values.

4.3 Portmanteau Test

All fitted model must be examined for adequacy. An adequate model has residuals behave like white noise series and all AFCs/CCMs of the residuals is close to zero. The model checking is conducted trough the Portmanteau test in terms of testing if the autocorrelations or the cross-correlation matrices in the first m lags of the times series is jointly equal to zero.

Univariate

The test hypothesis is simply $H_0 : \rho_0 = \cdots = \rho_m = 0$ versus $H_1 : \rho_i \neq 0$, for $0 \leq i \leq m$. The Portmanteau statistic proposed by Ljung & Box (1978), also known as the Ljung-Box test statistic is (Tsay 2010, p. 32-33):

$$Q(m) = T(T+2)\sum_{l=1}^{m} \frac{\hat{\rho}_l^2}{T-l} \tag{4.1}$$

where T is the number of observations, m is the interested lag and $\hat{\rho}_l$ is the sample autocorrelation in Eq. 2.1. The $Q(m)$ statistic is asymptotically chi-square distributed with m degrees of freedom. However when testing the residuals, the test statistic is then asymptotically chi-square distributed with $m - q$ degrees of freedom, where q is the number of AR coefficients/parameters used in the model (Tsay 2010, p. 50-51):

$$Q(m) \sim \chi^2(m - q)$$

If the Ljung-Box test statistic exceeds the critical value at an appropriate level, the null hypothesis is rejected and the model is conclude to be inadequate.

Multivariate

The multivariate Portmanteau test has the same test hypothesis as in univariate, $H_0 : \boldsymbol{\rho}_0 = \cdots = \boldsymbol{\rho}_m = 0$ versus $H_1 : \boldsymbol{\rho}_i \neq 0$, for $0 \leq i \leq m$. Thus the statistic is to test the existence of auto- and cross-correlations in the first m lags of the vector time series. The statistic is (Tsay 2010, p. 397-398):

$$Q_k(m) = T^2 \sum_{l=1}^{m} \frac{1}{T-l} tr(\hat{\mathbf{\Gamma}}_l^T \hat{\mathbf{\Gamma}}_0^{-1} \hat{\mathbf{\Gamma}}_l^T \hat{\mathbf{\Gamma}}_0^{-1}) \tag{4.2}$$

where T is the number of observations, k is the dimension for which is fixed to 2 in this study, $tr(\cdot)$ is the trace of the matrix, i.e. the sum of the diagonal elements in the matrix. $\hat{\mathbf{\Gamma}}_l$ is the sample lag-l cross-covariance matrix, see Eq. 3.1. This statistic follows asymptotically a chi-square distribution with k^2m degrees of freedom and when testing the residuals, the degree of freedom becomes $k^2m - q$ (Tsay 2010, p. 407):

$$Q_k(m) \sim \chi^2(k^2m - q)$$

Reject the null hypothesis if the test statistic exceeds the critical value.

4.4 Order Selection

There are several useful tools that are available in determining the order of a time series model. Here we introduce two of them: Akaike information criterion (AIC) and Schwarz–Bayesian information criterion (BIC).

AIC

$$AIC(l) = ln(\tilde{\sigma}_l^2) + \frac{2l}{T}$$

where T is the sample size, and $\tilde{\sigma}_l^2$ is the maximum-likelihood estimate of σ_a^2 with a model of order l. The first term measures the goodness of fit and second term is the penalty function of the criterion because it penalizes a candidate model by the number of parameters used (Tsay 2010, p. 48).

BIC

BIC is closely related to the AIC, the only different is the penalty function:

$$BIC(l) = ln(\tilde{\sigma}_l^2) + \frac{l * ln(T)}{T}$$

It is obviously that BIC is going to select a lower order than AIC when the sample size is sufficiently large.

Since both AIC and BIC estimate the errors, the selected rule is then choose the order that provide the minimum value of the AIC or BIC. "There is no evidence to suggest that one approach outperforms the other in a real application"(Tsay 2010, p. 49).

4.5 Error Measurement

What we are most concerned about in forecasting is the accuracy, thus our model evaluation is performed in the aspect of prediction errors. To give a comparable intuitive reflection of the total prediction errors that generated through the process, three error measurements for which are suggested by Hyndman & Athanasopoulos (2018) are introduced.

Mean Absolute Error, MAE

$$MAE = \frac{1}{T} \sum_{t=1}^{T} |F_t - A_t|$$

Root Mean Square Error, RMSE

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (F_t - A_t)^2}{T}}$$

Mean Absolute Percentage Error, MAPE

$$MAPE = \frac{100\%}{T} \sum_{t=1}^{T} \left| \frac{F_t - A_t}{A_t} \right|$$

where T is the number of predictions, F_t is the predicted value at time t and A_t is the actual value at time t in all three measurements.

4.6 Naive Approach

A naive approach in prediction is to expect that the future prices is equal to present price. More mathematically, consider a time series x_t, the naive forecasting approach is meaning that the expected value of x_{t+l} given all available data at time t is equal to the expected value of x_t, i.e.

$$E[x_{t+l}|\mathcal{F}_t] = E[x_t|\mathcal{F}_t] = x_t$$

for any positive integer l. This forecasting approach gives the same result as the forecasting of random walk process and is also called the random walk forecasts (Hyndman & Athanasopoulos 2018, ch. 3.1). The naive forecasting approach is used as a benchmark in model evaluation.

5 Case Study

The study consists of three part: preprocess, analysis and prediction. In the preprocess part the raw data will be processed and divided into groups. The analysis part will start with several tests to verify if the data satisfies the assumption for pairs trading. An univariate and a multivariate time series model is then fit to the data. Before send the fitted models to the prediction part, a model adequacy test is conducted and all unsatisfactory models are thrown away. The prediction process follows a so called time series cross validation procedure which will be introduced later, and a model evaluation is conducted through a comparison of errors between the forecasting values and the actual values.

5.1 Data

Unfortunately it is impossible to find securities that are facing exactly the same risk factors and are equally sensitive to the factor which in term gives exactly the same expected return in reality. However, the risk factors and corresponding sensitivities for companies that are providing the same service or product within the same industry and market may be highly similar. A cointegration relationship which is the main assumption that need to be satisfied in further analysis may also exist among such companies.

A dual-listed company(DLC) is composed by two corporations which have their respective legal identities and stock exchange listings, while they conduct business as a single operation. Theoretically the risk factors exposure and sensitivity are almost the same due to the DLCs characteristics, hence the stock prices of DLCs should move in a lockstep and give an expected return at same level and thereon give an opportunity of trade in pairs. As a result, we chose one of the world's largest mining company - BHP Billiton - which is also a DLC. The two corporations are the Australian BHP Group Limited, denote as BHP, and the British BHP Group plc, denote as BBL. The stock price of BHP and BBL on the New York Stock Exchange are downloaded from Yahoo Finance. All available historical data after merging until 2020 are used, that is, data from 2003-06-25 to 2019-12-31. After removing all unavailable values there are total 4158 observations.

Figure 1 shows the time plot of the logarithm of the adjusted closing price for both BHP and BBL, for convenience we will use "log price" instead of saying the whole name of "logarithm of the adjusted closing price". As shown, the log price of BHP always lies above the BBL and the the two stocks are moving together. Though the observed price gap between the two stocks are somewhat enlarged after 2010, but still holding a steady state. It is plausible to believe that stocks of a dual-listed company is tradable in pair, however there are some exceptions in the history and it is therefore necessary to test the tradability among these two stocks.

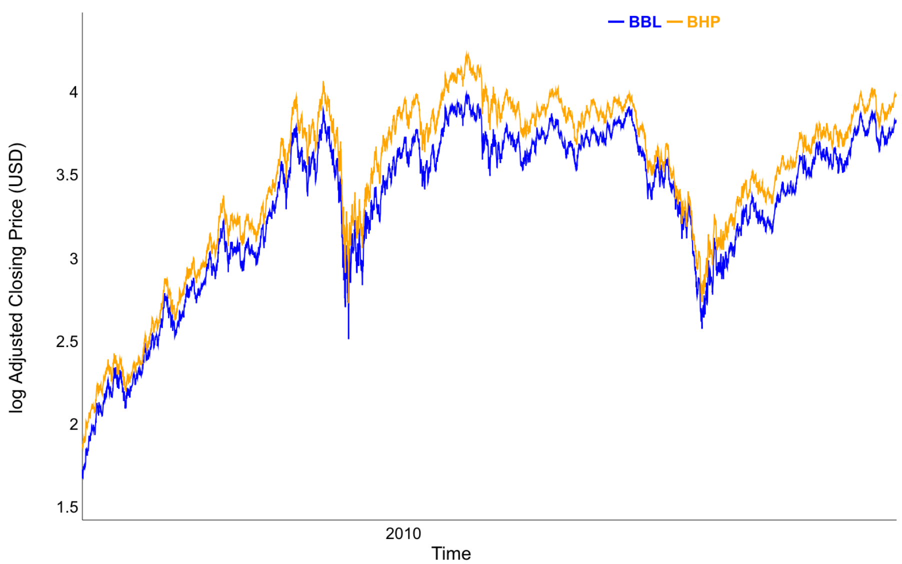


Figure 1: log adjusted closing price of BHP and BBL, from 2013-06-25 to 2019-12-31

5.2 Training and Testing Group

Before data testing and further analysis we divide the data into two groups, a training and a testing group. The purpose with the training and testing group is to conduct value comparison and model evaluation. Training data are used to conduct model fitting and parameter estimating, and thereafter a prediction of the exogenous data, i.e. the data in the testing group. By comparing the predicted and actual values, a model valuation based on error analysis is then taking place.

The testing group contains the latest one year's data, i.e. data from 2019-01-01 to 2019-12-31. Remainders, i.e. data from 2003-06-25 to 2018-12-31, belong to the training group.

5.3 Test of Tradability

A pair is said to be tradable if the pair-making stocks are cointegrated (Vidyamurthy 2004, p. 104). To check whether there exist a cointegration relationship between BHP and BBL, two approaches are used: (1) check if the AR modeled linear combination between the log prices of BHP and BBL is stationary; (2) determine the existence of cointegration through the Johansen's cointegration test on VAR modeled log prices of BHP and BBL.

5.3.1 Linear Combination

First of all, we need to determine which series ought to be the independent variable when constructing a linear regression on the BHP-BBL pair. According to Vidyamurthy (2004, p. 108) that the stock with a lower variance on returns indicate a lower volatility and is therefore appropriate to be treated as the independent variable. It can be shown that:

$$log(p_t) - log(p_{t-1}) \approx \frac{p_t - p_{t-1}}{p_{t-1}}$$

where p_t is the price at time t, hence the difference in the log price can be construed to be the return (Vidyamurthy 2004, p. 30).

The variance on returns of BHP and BBL is 0.0005760112 and 0.0006700049 respectively and hence the BHP series is chosen to be the independent variable. Construct a simple linear regression in the form of $BBL_t = \beta_0 + \beta_1 * BHP_t$, following relationship is obtained by using ordinary least square (OLS) estimation:

$$BBL_t = 0.0156 + 0.9492 * BHP_t + \hat{\varepsilon}_t$$

where BBL_t and BHP_t is the respective log price at time t and $\hat{\varepsilon}_t$ is the estimated residual at time t (details see Table 8 in Appendix). As discussed earlier in section 3.4 and 3.5 that the associated fitted spread can be calculated by:

$$\hat{w}_t = BBL_t - 0.9492 * BHP_t = 0.0156 + \hat{\varepsilon}_t \tag{5.1}$$

The expected value of fitted spread is then simply the same as the intercept of the linear combination since the residuals $\hat{\varepsilon}_t$ is asymptotically normal distributed with mean zero (see Figure 5 in Appendix), i.e. $\mu_{\hat{w}} = E[\hat{w}_t] = E[0.0156 + \hat{\varepsilon}_t] = 0.0156$.

The fitted spread, $\hat{w}_t$, is plotted in Figure 2. The horizontal solid line represents the mean value of the fitted spread regarding the training group. The actual spread regarding the testing group which provided by the same linear combination is referring to the line on the right of the vertical dashed line. A cursory look at this spread series seems to suggest that this par is not stationary and thereby not tradable. But pay attention on the scale of y-axis, that this tiny scale may deceive our eyes. Therefore it is necessary to conduct the ADF-test on this spread series.

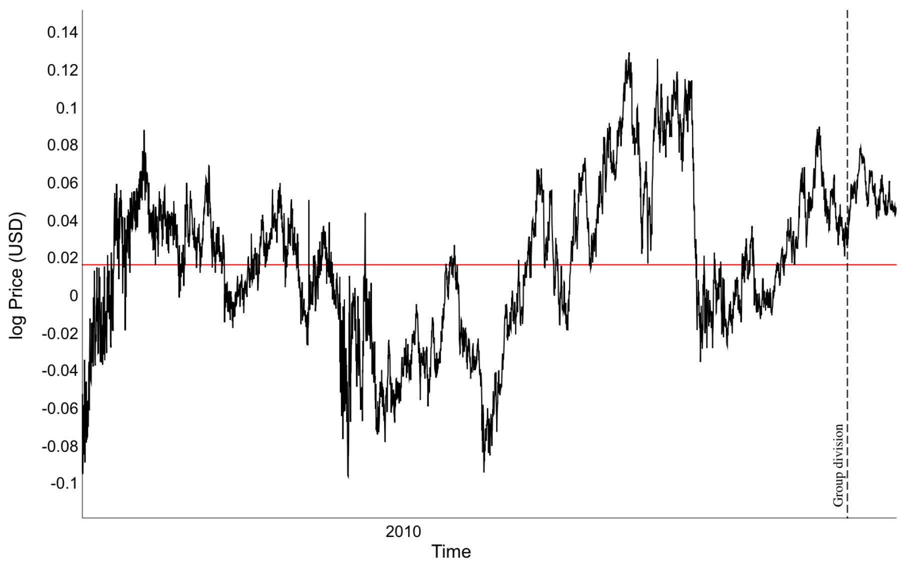


Figure 2: Fitted spread between BHP and BBL. Horizontal solid line represents the mean value of the fitted spread regarding the training group.

According to preceding discuss on ADF-test that the order of the model need to be determined as well as the model parameters. Thus a construction of AR model on the spread series is conducted start by order determination. Large order can increase the computational complexity which is what we want to avoid. For this reason we only search suitable orders below 13.

Conduct first a maximum-likelihood estimation on the variance of the white noise term in an AR(p) model, for which $p = 0, ..., 13$. With this estimation one can compute the corresponding AIC and BIC using the formula given in section 4.4. The order of the model is decided by the lowest value of the information criteria. Values of AIC and BIC for AR model on the spread series are presented in Table 1.

The AIC suggests an AR(5) model on fitted spread series $\hat{w}_t$, while the BIC suggests an AR(4) model. This is agreed with the expectation that BIC will

Table 1: AIC and BIC for AR model on spread series

p	AIC	BIC	p	AIC	BIC
0	-6.3307	-6.3307	7	-9.7832	-9.7720
1	-9.7124	-9.7108	8	-9.7833	-9.7705
2	-9.7649	-9.7617	9	-9.7828	-9.7683
3	-9.7729	-9.7680	10	-9.7831	-9.7670
4	-9.7821	**-9.7757**	11	-9.7815	-9.7639
5	**-9.7837**	-9.7757	12	-9.7812	-9.7620
6	-9.7836	-9.7740	13	-9.7807	-9.7598

Note: selected order marked in bold

select a lower order when sample size is large which mentioned in section 4.4. An ADF-test on $\hat{w}_t$ for both models is conducted for confirmation of stationarity.

The ADF-test gives a test statistic at -3.9573 with a p-value at 0.0111 for AR(4) and a test statistic at -3.8621 with a p-value at 0.01586 for AR(5). Both models have a p-value lower than 5 % which implies that we can reject the null-hypothesis and conclude that the spread series is stationary for both models. A lower p-value implies that the probability for $\hat{w}_t$ to be stationary is higher which may suggest that the AR(4) model is fitting the spread series better regarding the property of the cointegration.

5.3.2 Cointegration

As mentioned in section 4.2 that the order of the VAR model is involved in Johansen's cointegration test, hence AIC and BIC are used to determine the order of the VAR model in the same way as AR model. Again, AIC and BIC gives different results. AIC suggests a VAR(5) and BIC suggests a VAR(4) model on the log prices of BHP and BBL; see Table 2.

Table 2: AIC and BIC for VAR model on log price series

p	AIC	BIC	p	AIC	BIC
0	-7.6773	-7.6773	7	-17.3450	-17.3000
1	-17.2486	-17.2422	8	-17.3438	-17.2924
2	-17.3251	-17.3122	9	-17.3420	-17.2842
3	-17.3338	-17.3145	10	-17.3422	-17.2780
4	-17.3431	**-17.3174**	11	-17.3411	-17.2705
5	**-17.3468**	-17.3147	12	-17.3400	-17.2629
6	-17.3462	-17.3077	13	-17.3385	-17.2550

Note: selected order marked in bold

Since there is no reason to believe a presence of trend in the log price of the two stocks and the spread series is likely to have a non-zero mean when looking at Figure 1, thus a constant restricted (case 2 in section 3.5.2) Johansen's cointegration test is conducted for both VAR models with both trace and max-eigenvalue test type. The test results using MacKinnon-Haug-Michelis p-values are presented in Table 3. The test using Osterwald-Lenums p-values gives the same results and is presented in Table 9 in Appendix. For VAR(4) model, the trace statistics and max-eigenvalue statistics indicate a presence of 1 cointegrating vector at 1 % respective 5 % levels. While for VAR(5) model, both trace and max-eigenvalue statistics indicate a presence of 1 cointegrating vector at 1 % level. This implies that the cointegration relationship between the log prices of BHP and BBL is stronger with a VAR(5) model and might produce a better prediction result in further analysis.

Table 3: Johansen's cointegration test for VAR(4) and VAR(5), case 2

			VAR(4)		
Eigenvalues			0.005020311 0.001904570		
	Hypothesis	Test	5pct	1pct	p-value*
Trace	m=0	27.08425	20.26184	25.07811	0.0049
	m≤1	7.440625	9.164546	12.76076	0.1050
Max-eigen	m=0	19.64362	15.89210	20.16121	0.0122
	m≤1	7.440625	9.164546	12.76076	0.1050
			VAR(5)		
Eigenvalues			0.005195894 0.001652006		
	Hypothesis	Test	5pct	1pct	p-value*
Trace	m=0	26.77869	20.6184	25.07811	0.0055
	m ≤ 1	6.451459	9.164546	12.76076	0.1586
Max-eigen	m=0	20.32723	15.89210	20.16121	0.0094
	m≤1	6.451459	9.164546	12.76076	0.1586

*MacKinnon-Haug-Michelis(1999) p-values

5.4 Model Fitting and Checking

It has been shown that AR(4) modeled spread and VAR(5) modeled log prices are better in line with the requirement of the pairs trading. However, there is no evidence to support that AIC selected order is better than BIC's and "substantive information of the problem under study and simplicity" are also important when choosing the model (Tsay 2010, p. 49). Therefore, both AR(4) and AR(5) for spread series, and both VAR yielded ECMvar(3) and ECMvar(4) for log prices are modeled. Bear in mind that ECMvar models have order of $(p-1)$ where p is the order of the corresponding VAR model. The reason to modelling ECMvar instead of VAR is that the ECMvar model consists the cointegration relationship which is important in further prediction.

5.4.1 Fit AR Model on Spread

The model parameters in AR(5) and AR(6) are estimated by the maximum-likelihood methods (Tsay 2010, p. 49-50) and the constant terms are computed by the Eq. 2.3 respectively. The estimated AR models are as follows:

AR(4) model

$$\begin{aligned}\hat{w}_t = &\ 0.0001615757 + 0.7298x_{t-1} \\ &+ 0.1421x_{t-2} + 0.0218x_{t-3} \\ &+ 0.0960x_{t-4} + a_t\end{aligned}$$

$$\sigma_a^2 = 0.00005634$$

AR(5) model

$$\begin{aligned}\hat{w}_t = &\ 0.0001578812 + 0.7261x_{t-1} \\ &+ 0.1424x_{t-2} + 0.0198x_{t-3} \\ &+ 0.0606x_{t-4} + 0.0409x_{t-5} \\ &+ a_t\end{aligned}$$

$$\sigma_a^2 = 0.00005622$$

Following the discussion in section 2.4, the associated characteristic roots are obtained and the fitted AR models can be rewritten into the form av Eq. 2.5 as follows:

AR(4) model

$$\begin{aligned}&(1 - 0.9929928B)(1 - 0.4602519B) \\ &(1 - 0.4582326B)(1 - 0.4582326B)\hat{w}_t = 0.0001615757 + a_t\end{aligned}$$

AR(5) model

$$\begin{aligned}&(1 - 0.9933985B)(1 - 0.4062324B) \\ &(1 - 0.4062324B)(1 - 0.4996340B) \\ &\qquad\qquad (1 - 0.4996340B)\hat{w}_t = 0.0001578812 + a_t\end{aligned}$$

Since all the characteristic roots are less than 1 in modulus for both models respectively, the spread series $\hat{w}_t$ is again verified to be stationary with both models.

5.4.2 AR Model Checking

The univariate version of Portmanteau test in section 4.3 is used to check the adequacy of AR(4) and AR(5) models. Note that the tested object is the residual series of the AR model, which is eqivalent to the a_t terms in the model. Use Eq. 2.1 and Eq. 4.1, the associated Portmanteau statistics are obtained. The corresponding critical values are provided by $\chi^2(m-q)$, where q is the number of AR coefficients used in the model which is 4 for AR(4) and 5 for AR(5). Tests are conducted with $m = 10, 15, 20$. The results presented in Table 4 suggest to reject the H_0 with all three m for AR(4) but not for AR(5). This implies that there exist a serial correlation in the first 20:th lags of the residuals for AR(4) and no serial correlations for AR(5) at 5 % level. Since the assumption for a_t is that it is a Gaussian white noise series, which means that there is no existence of serial correlations. Thus, we can conclude that the AR(4) is not adequate and all information given by it can not be trusted. Though the spread series do have a higher probability to be stationary with an AR(4) model, we still throw away this model duo to its inadequacy. Only **AR(5)** is send to the prediction process.

Table 4: AR models' Portmanteau test

	m	Statistic	$\chi^2_{0.05}(m-q)$	p-value
AR(4)	10	23.619	12.59159	0.0006135
	15	26.905	19.67514	0.00475
	20	29.6	26.29623	0.02019
AR(5)	10	10.104	11.0705	0.07234
	15	13.496	18.30704	0.1973
	20	15.35	24.99579	0.4265

5.4.3 Fit ECMvar

Since both log prices of the BHP and BBL are unit-root nonstationary and is now verified to be cointegrated, the ECM is then needed in order to transform the series into stationary jointly and thereby predictable.

Consider a bivariant time series $\boldsymbol{x}_t = \begin{bmatrix} BBL_t \\ BHP_t \end{bmatrix}$ and use Eq. 3.6 with a fixed $\boldsymbol{\beta}$-matrix: $\boldsymbol{\beta} = \begin{bmatrix} 1 & -0.9492 \end{bmatrix}^T$ to obtain the maximum-likelihood estimated ECMvar models, that is the model in Eq. 3.6. The reason to fix the $\boldsymbol{\beta}$-matrix like this is to ensure that the linear relationship $BBL_t - 0.9492BHP_t$ (Eq. 5.1) is hold all along the system. Estimated ECMvar model are in the follow:

ECMvar(3)

$$\begin{aligned}\boldsymbol{\Delta x}_t = \begin{bmatrix}0.0009\\0.0007\end{bmatrix} &+ \begin{bmatrix}-0.0256\\-0.0149\end{bmatrix}\begin{bmatrix}1 & -0.9492\end{bmatrix}\boldsymbol{x}_{t-1}\\
&+ \begin{bmatrix}-0.0116 & -0.0056\\0.2812 & -0.3052\end{bmatrix}_1^* \boldsymbol{\Delta x}_{t-1}\\
&+ \begin{bmatrix}-0.1136 & 0.1054\\0.0223 & -0.0334\end{bmatrix}_2^* \boldsymbol{\Delta x}_{t-2}\\
&+ \begin{bmatrix}-0.0477 & 0.0364\\0.0575 & -0.0710\end{bmatrix}_3^* \boldsymbol{\Delta x}_{t-3}\\
&+ \boldsymbol{a}_t\end{aligned}$$

$$\boldsymbol{\Sigma}_a = \begin{bmatrix}0.0006678725 & 0.0005935363\\0.0005935363 & 0.0005715220\end{bmatrix}$$

ECMvar(4)

$$\begin{aligned}\boldsymbol{\Delta x}_t = \begin{bmatrix}0.0009\\0.0007\end{bmatrix} &+ \begin{bmatrix}-0.0254\\-0.0150\end{bmatrix}\begin{bmatrix}1 & -0.9492\end{bmatrix}\boldsymbol{x}_{t-1}\\
&+ \begin{bmatrix}-0.0145 & -0.0023\\0.2814 & -0.3053\end{bmatrix}_1^* \boldsymbol{\Delta x}_{t-1}\\
&+ \begin{bmatrix}-0.1227 & 0.1145\\0.0209 & -0.0320\end{bmatrix}_2^* \boldsymbol{\Delta x}_{t-2}\\
&+ \begin{bmatrix}-0.0720 & 0.0630\\0.0529 & -0.0661\end{bmatrix}_3^* \boldsymbol{\Delta x}_{t-3}\\
&+ \begin{bmatrix}-0.0730 & 0.0571\\-0.0137 & 0.0114\end{bmatrix}_4^* \boldsymbol{\Delta x}_{t-4}\\
&+ \boldsymbol{a}_t\end{aligned}$$

$$\boldsymbol{\Sigma}_a = \begin{bmatrix}0.0006668510 & 0.0005933817\\0.0005933817 & 0.0005715842\end{bmatrix}$$

5.4.4 ECMvar Model Checking

The multivariate Portmanteau test on the residual series is used here to check the adequacy. The associated sample cross-covariance matrix is then the $\boldsymbol{\Sigma}_a$-matrix for each model and the required q value is then 3 for ECMvar(3) and 4 for ECMvar(4). Use Eq. 3.1 and Eq. 4.2 to obtain the corresponding test statistics and compare with the critical values provided by $\chi^2(2^2m - q)$ at 5 % level for $m = 5, 10, 15$. Results are presented in Table 5.

It has shown that H_0 is rejected when $m = 5$ for ECMvar(3) model at 5 % level. Though the p-values at $m = 10$ and $m = 15$ are higher than 0.05, but are

Table 5: ECMvar models' Portmanteau test

	m	Statistic	$\chi^2_{0.05}(2^2m - q)$	p-value
ECMvar(3)	5	30.6210	27.58711	0.02
	10	46.5131	52.19232	0.14
	15	71.2096	75.62375	0.10
ECMvar(4)	5	7.1229	26.29623	0.97
	10	23.2649	50.99846	0.95
	15	49.6793	74.46832	0.71

still very small compare with the corresponding p-values in ECMvar(4) model. Remember that the goal is to accept the model which do not reject the H_0 to achieve a residual series that behaves like a white noise series. Therefore, the ECMvar(3) model is identified as inadequacy at $m = 5$ and relatively inadequacy at $m = 10$ and 15 and hence is thrown away. The **ECMvar(4)** is also send to the prediction process for further analysis.

5.5 Prediction

As mentioned in section 2.4 that the predicted values of a long-term forecasting for a stationary AR model will converge to its unconditional mean. Figure 4 in Appendix is supporting this assertion. What we concerned in pairs trading is the existence and the magnitude of deviation, such convergence is making the long-term forecasting meaningless regarding the pairs trading. Thus, this study is focusing on short-term forecast: 1-step and 2-step forecasting.

The 1-step and 2-step forecasting is conducted by following a procedure called times series cross-validation. In this procedure, the forecasting origin is rolling forward with time, and all historical data prior to the current forecasting origin are put into the training group. Only one observation is predicted with conditional on the associated training set at each time step (Hyndman & Athanasopoulos 2018, Ch. 3.4). In other words, consider a time series $\{x_t\}$, one may conduct a l-step-ahead forecast by using the equation:

$$\hat{x}_{h+i}(l) = E[x_{h+i+l}|\mathcal{F}_{h+i}], \qquad i = 0, 1, 2, ... \tag{5.2}$$

where h is the original forecasting origin.

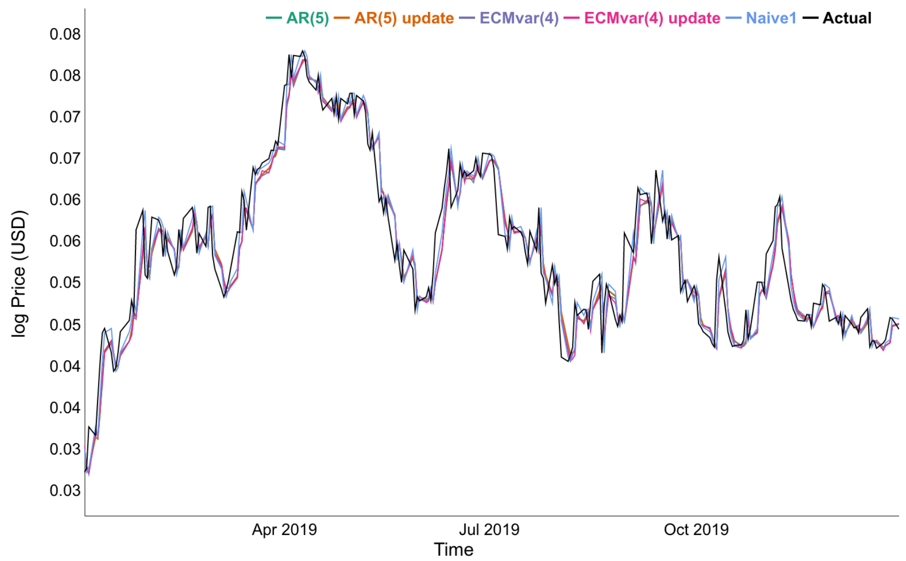
Two types of forecasting is used: (1) only updated historical data to forecast the future values; (2) update both historical data and model parameters at each time step to forecast the future values. Both type 1 and type 2 forecasting are applied to the model AR(5) and ECMvar(4) models. The naive approaches are used only with type 1 forecasting. All predicted values are compared with the actual values which are obtained by the observed log prices of the two stocks through the predetermined linear combination. Model errors measured in MAPE, RMSE and MAE are calculated and analysed.

Use the Eq. 5.2 and previous introduced forecasting formula for each model, one can obtain the forecasting values. The forecasting values of the spread series can be obtained directly by using the AR models. While using the ECMvar models, the forecasting values of the spread series are computed via the predetermined linear combination (Eq. 5.1) between the forecasting value of the log pries of the two stocks.

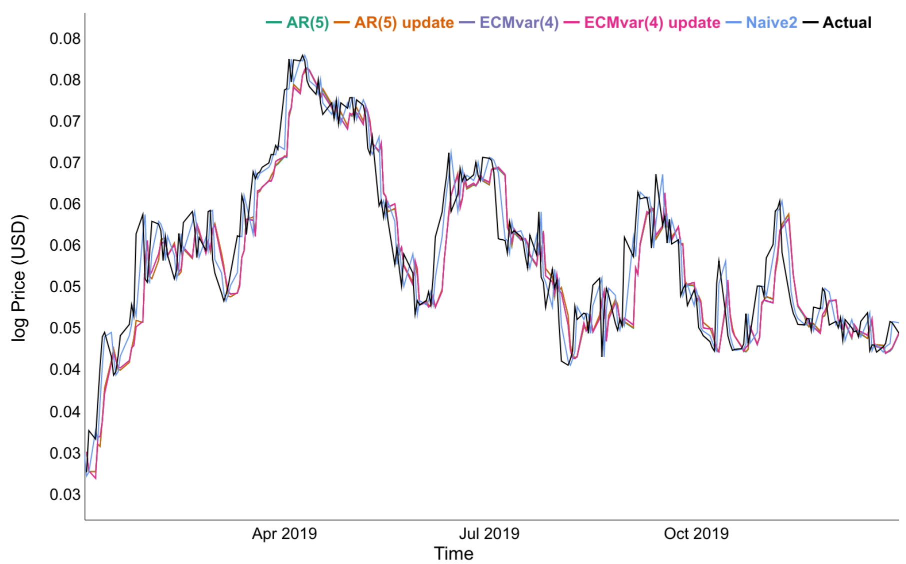
The predicted values are small and the differences between the actual and predicted values are tiny which make it difficult to draw any conclusion about the models intuitively through the images (see Figure 3). However, the gap between the actual and predicts in the 2-step-ahead forecasting is larger which implies that the accuracy is lower in such forecasting. This is in line with the theory that larger steps create larger errors.

Take a closer look at the model errors obtained by the different error measurements, in Table 6 and Table 7. Unfortunately none of our models is outperforming the naive approach either in 1-step nor 2-step forecasts. However type 2 ECMvar(4) is outperformed all other models. In overall, the ECMvar models have a lower error than AR models with all three error measurements and type 2 forecasts are better than type 1 in both 1-step and 2-step forecasts. Although both AR and ECMvar provided forecast spread series are directly or indirectly related to the predetermined linear combination, but ECMvar model do take along this cointegration information with it during every step in the forecasting which AR model do not. The reason that type 2 forecasts are more accurate in prediction is that it use the latest historical data to adjust the parameters which reflect to the relationships between the previous data and current data. This relationship might be changed by some business activities or social events and it is therefore necessary to catch such change and update the available information in order to give a more accurate and believable prediction.

It has been shown in Table 6 that type 1 AR(5) is better than type 2 AR(5) in MAPE while a contrary result in RMSE. Such problem can be caused by the characteristics of the error measurements. MAPE is widely used as a percentage measures in forecast performances, but it has a disadvantage that it put a heavier penalty when handle the negative errors (Flores 1986, Hyndman & Athanasopoulos 2018). A forecast method that minimises the MAE will lead to forecasts of the median, while minimising the RMSE will lead to forecasts of the mean (Hyndman & Athanasopoulos 2018, Ch. 3.4). On the basis of this, one can say that a higher MAPE might implies an existence of more negative errors, a higher RMSE might implies that there exist some relative extreme errors, and a higher MAE might implies an average higher errors. In the case of pairs trading, price changing, no matter if it goes up or down, is equally important to the investors. Hence, the usage of MAPE here is not as suitable as the other two duo to its disadvantage. Regarding RMSE and MAE, an aggressive investor might chose the model with a lower RMSE, with aspiration to seek more profits from those relatively extreme minor errors. While a conservative investor might prefer models with lower MAE.



(a) 1-step-ahead forecasting



(b) 2-step-ahead forecasting

Figure 3: Forecasting for the spread between BHP and BBL, with forecasting origin at 2018-12-31

Summarize the results:

- None of our models is outperforming the naive approach in both 1-step and 2-step forecasts.
- Type 2 ECMvar(4) forecast is outperforming all other models in both 1-step and 2-step forecast.
- ECMvar model provided predictions of the spread between BHP and BBL is better than the AR model provided predictions.
- Rolling forecasting origin based time series cross-validation with update in both historical data and model parameters can improve the forecasting accuracy.

Table 6: Model Errors for 1-step-ahead Forecast

	MAPE	RMSE	MAE
AR(5) *type 1*	4.461833	0.003081	0.002345
AR(5) *type 2*	4.462820	0.003077	0.002341
ECMvar(4) *type 1*	4.424975	0.003070	0.002324
ECMvar(4) *type 2*	4.421785	0.003067	0.002321
Naive	4.304111	0.003026	0.002249

Table 7: Model Errors for 2-step-ahead Forecast

	MAPE	RMSE	MAE
AR(5) *type 1*	6.074521	0.004133	0.003193
AR(5) *type 2*	6.066236	0.004125	0.003186
ECMvar(4) *type 1*	5.991964	0.004092	0.003155
ECMvar(4) *type 2*	5.989897	0.004087	0.003151
Naive	5.948054	0.004035	0.003080

5.6 Discussion

The purpose with this study is to compare the forecast accuracy between AR and ECMvar model on the spread series which provided by two cointegrated time series. The study start with a verification of the feasibility for pairs trading though a cointegration test. Directly with the Johansen's cointegration test and indirectly with the ADF test on the associated linear combination. The order determination of the model is implemented in order to conduct the cointegration tests. In this part, four models pass the tradability test: AR(4) and AR(5) which present the spread series, and VAR(4) and VAR(5) which present the log prices of the two associated stocks. It makes sense that the spread series correspond AR model and the log prices correspond VAR model have the same order. If the current log prices are related to the past 5 days log prices, then the spread between the log prices is also related to the past 5 days spread, and vice versa. The AIC and BIC values used to determine the order of the model are very small, which is mainly due to the small variance of the maximum-likelihood estimated white noise series. These small variance indicate that the volatility in the price of the two stocks is extremely small. However, in the subsequent model examination, AR(4) and VAR(4) yielded ECMvar(3) show an inadequacy in terms of serially- or cross- correlated residuals. As a consequence, only AR(5) and VAR(5) yielded ECMvar(4) are send to the final prediction part. Disappointingly that none of the AR(5) or ECMvar(4) is outperform the naive forecast approach. But in overall, the ECMvar model is better than a simple AR model on the spread series analysis. It has higher prediction accuracy because it comprise the cointegration information and this information is exist in every predicted log prices pairs, which then delivered to the associated predicted spread. The AR predictions is lack of this cointegration information and only concern about the existing information about the spread. It is the log prices that decides how large is the spread, and the stationarity with the spread decides how log prices need to adjust in the future. Thus the AR predictions may deviate more from the actual spread.

There are some restrictions in this study, which can be improved in any further studies. The deterministic function has been restricted to case 2, as there is no significant time trend when looking throughout the whole period. But if concerning the seasonality, especially the business cycle, there is a certain time tend during certain periods. Moreover, the chosen models in this study are the most simple univariate and multivariate time series models. Therefore, using STL decomposition (Seasonal and Trend decomposition using Loess) and some other more advanced time series models like ARIMA or GARCH which concerning the moving average and volatility, the forecasting results can be improved and may have a chance to beat the naive forecasting approach.

References

Engle, R. F. & Granger, C. W. J. (1987), 'Co-integration and error correction: Representation, estimation, and testing', *Econometrica* **55**(2), 251–276.
URL: *https://doi.org/10.2307/1913236*

Flores, B. E. (1986), 'A pragmatic view of accuracy measurement in forecasting', *Omega* **14**(2), 93–98.
URL: *https://doi.org/10.1016/0305-0483(86)90013-7*

Hyndman, R. J. & Athanasopoulos, G. (2018), *Forecasting: principles and practice*, Business Planning and Forecasting, OTexts.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* **12**(2-3), 231–254.
URL: *https://doi.org/10.1016/0165-1889(88)90041-3*

Johansen, S. (1995), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Advanced Texts in Econometrics, Oxford University Press.

Ljung, G. M. & Box, G. E. P. (1978), 'On a measure of lack of fit in time series models', *Biometrika* **65**(2), 297—-303.
URL: *https://doi.org/10.1093/biomet/65.2.297*

MacKinnon, J., Haug, A. & Michelis, L. (1999), 'Numerical distribution functions of likelihood ratio tests for cointegration', *Journal of Applied Econometrics* **14**(5), 563–577.
URL: *https://doi.org/10.1002/(SICI)1099-1255(199909/10)14:5¡563::AID-JAE530¿3.0.CO;2-R*

Osterwald-Lenum, M. (1992), 'A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics', *Oxford Bulletin of Economics and Statistics* **54**(3), 461–472.
URL: *https://doi.org/10.1111/j.1468-0084.1992.tb00013.x*

Tsay, R. S. (2010), *Analysis of Financial Time Sries, Third Edition*, Applied Probability and Statistics, John Wiley Sons Inc.

Vidyamurthy, G. (2004), *Pairs Trading: Quantitative Methods and Analysis*, Finance and investment, John Wiley Sons Inc.

A Tables

Table 8: Linear regression on BHP and BBL

	Dependent variable:
	BBL
BHP	0.9492***
	(0.0013)
Constant	0.0156***
	(0.0045)
Observations	3,907
R^2	0.9928
Adjusted R^2	0.9928
Residual Std. Error	0.0422 (df = 3905)
F Statistic	535,051.0000*** (df = 1; 3905)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Johansen's cointegration test for VAR(4) and VAR(5), case 2

	VAR(4)			
Eigenvalues	0.005020311 0.001904570			
	Hypothesis	Test	5pct*	1pct*
Trace	m=0	27.08	19.96	24.60
	m<=1	7.44	9.24	12.97
Max-eigen	m=0	19.64	15.67	20.20
	m<=1	7.44	9.24	12.97
	VAR(5)			
Eigenvalues	0.005195894 0.001652006			
	Hypothesis	Test	5pct*	1pct*
Trace	m=0	26.78	19.96	24.60
	m<=1	6.45	9.24	12.97
Max-eigen	m=0	20.33	15.67	20.20
	m<=1	6.45	9.24	12.97

*Osterwald-Lenum(1992) critical values

B Figures

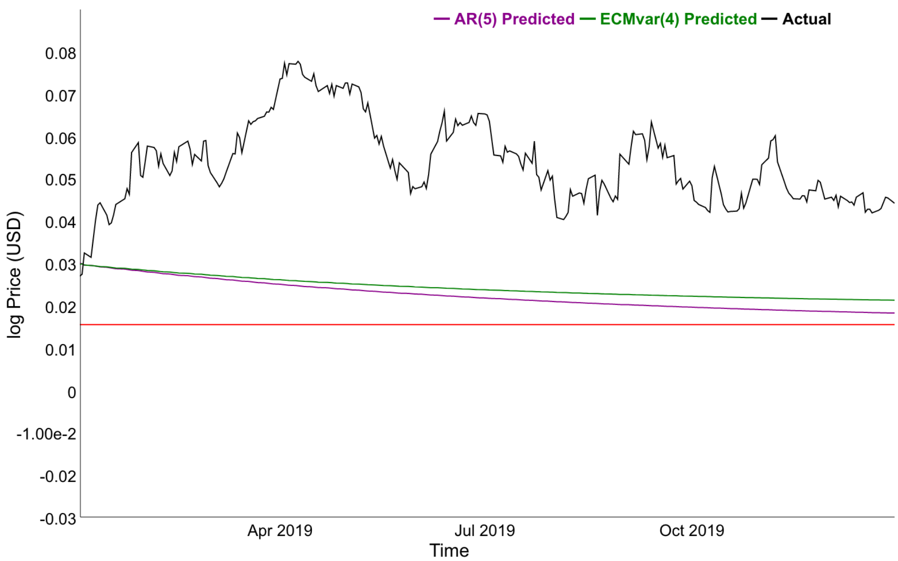
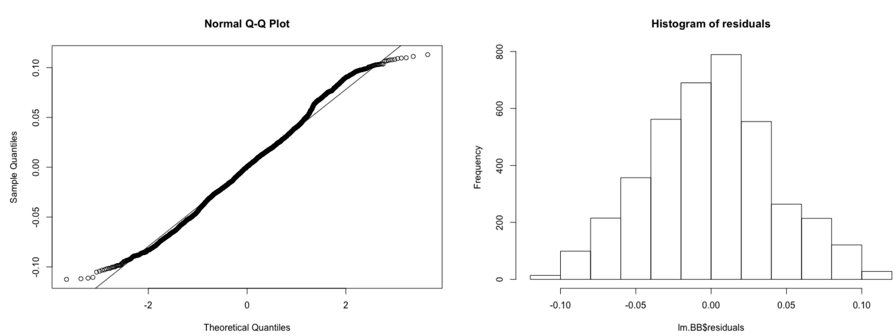


Figure 4: Multi-step-ahead forecasting for the spread between BHP and BBL regarding AR(5) and ECMvar(4) models. Forecasting origin at 2018-12-31 and forecasting horizon reflects the whole 2019. Horizontal line denote the unconditional mean of the spread.



(a) QQ-plot for residuals

(b) Histogram for residuals

Figure 5: Plot for residuals from fitted linear combination between BHP and BBL

C Derivations

From VAR to ECMvar

Consider a VAR(p) model with deterministic function $\boldsymbol{\mu}_t$ in the follows:

$$\boldsymbol{x}_t = \boldsymbol{\mu}_t + \boldsymbol{\Phi}_1 \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{x}_{t-p} + \boldsymbol{a}_t$$

subtract $\boldsymbol{x}_{t-1}$ in both side and obtain:

$$\begin{aligned}
\boldsymbol{x}_t - \boldsymbol{x}_{t-1} &= \boldsymbol{\mu}_t + \boldsymbol{\Phi}_1 \boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{x}_{t-p} + \boldsymbol{a}_t \\
\boldsymbol{\Delta x}_t &= \boldsymbol{\mu}_t + (\boldsymbol{\Phi}_1 - \mathbf{I}) \boldsymbol{x}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{x}_{t-p} + \boldsymbol{a}_t
\end{aligned}$$

Let $\boldsymbol{\Delta x}_{t-1} = \boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}$, $\boldsymbol{\Delta x}_{t-2} = \boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-3}$ and etc, the right hand side can rewrite as:

$$\begin{aligned}
&\boldsymbol{\mu}_t + (\boldsymbol{\Phi}_1 - \mathbf{I}) \boldsymbol{x}_{t-1} + \boldsymbol{\Phi}_2 (\boldsymbol{x}_{t-1} - \boldsymbol{\Delta x}_{t-1}) + \cdots + \boldsymbol{\Phi}_p (\boldsymbol{x}_{t-1} - \boldsymbol{\Delta x}_{t-p+1}) + \boldsymbol{a}_t \\
=&\boldsymbol{\mu}_t + (\boldsymbol{\Phi}_1 - \mathbf{I} + \boldsymbol{\Phi}_2 + \cdots + \boldsymbol{\Phi}_p) \boldsymbol{x}_{t-1} - \boldsymbol{\Phi}_2 \boldsymbol{\Delta x}_{t-1} - \cdots - \boldsymbol{\Phi}_p \boldsymbol{\Delta x}_{t-p+1} + \boldsymbol{a}_t \\
=&\boldsymbol{\mu}_t + (-\mathbf{I} + \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_2 + \cdots + \boldsymbol{\Phi}_p) \boldsymbol{x}_{t-1} - \sum_{i=2}^{p} \boldsymbol{\Phi}_i \boldsymbol{\Delta x}_{t-i+1} + \boldsymbol{a}_t
\end{aligned}$$

Let

$$\begin{aligned}
\boldsymbol{\Pi} &= (-\mathbf{I} + \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_2 + \cdots + \boldsymbol{\Phi}_p) \\
\boldsymbol{\Phi}_j^* &= -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i
\end{aligned}$$

the right hand side becomes:

$$\boldsymbol{\mu}_t + \boldsymbol{\Pi} \boldsymbol{x}_{t-1} + \sum_{j=i-1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{\Delta x}_{t-i+1} + \boldsymbol{a}_t$$

Then we have the ECM for VAR as follows:

$$\boldsymbol{\Delta x}_t = \boldsymbol{\mu}_t + \boldsymbol{\Pi} \boldsymbol{x}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{\Delta x}_{t-j} + \boldsymbol{a}_t$$