Learning Cointegration for Trading Strategies

Tanya Sandoval

July 24, 2016

Abstract

This report introduces the topic of cointegration and its application to trading strategies. By modelling asset prices with an economic link in common, it is sometimes possible to arrive at a stationary spread whose properties can be used to reduce exposure to systematic risk. The essential elements of cointegration such as stationary and mean-reversion are discussed, as well as some of the statistical tests available to detect this relationship. Simulated and real data examples are provided, the latter focusing on a detected cointegration between Brent Crude and Gasoil futures. Lastly, examples of simple trading strategies applying cointegration are discussed, along with aspects that must be considered when trading under market real conditions.

Contents

1 Introduction

Historically *cointegration* as a concept arose from statistical evidence that many US macroeconomic time series (like GDP, wages, employment, etc.) did not follow conventional econometric theory but rather were described by *unit root processes*, also known as "integrated of order 1" I(1). Before the 1980s many economists used linear regressions on non-stationary time series, which Granger and Newbold showed to be a dangerous approach that could lead to *spurious correlation*. For integrated I(1) processes, Granger and Newbold showed that de-trending does not work to eliminate the problem, and that the superior alternative is to check for cointegration, earning them the Nobel prize.

This report summarises some first learnings of this concept and first-steps at applying it to trading strategies. In particular, the focus is on studying two assets from the commodities space to see if similar properties can be detected as in equities.

- Section 2 provides details for the datasets used throughout the report
- Section 3 introduces stationarity and mean-reversion in time series, which are key elements of cointegration
- Section 4 goes into detail about cointegration and how to test for it, as well as assessing its quality
- Section 5 is then dedicated to its application to trading strategies and assessing their performance in terms of profit and loss (P&L)
- Appendix A summarises some of the mathematical methods involved, such as Multivariate Regression, Autoregressive models (AR(p)), Dickey-Fuller test, optimal lag order and stability conditions
- Appendix B starts to examine cointegration in other energy commodities, such as Italian and Dutch gas

All the relevant scripts to arrive at the results can be found in the project repository "*finalProject/TS*" in the attached USB drive. In particular the ipython notebook *Coint_Brent_Gasoil_v2.ipynb* demonstrates how to run the code, which is omitted in this report for brevity. The project repository is also available online on Github `https://github.com/tsando/CQF/tree/master/finalProject`.

2 Datasets

2.1 Simulated Data

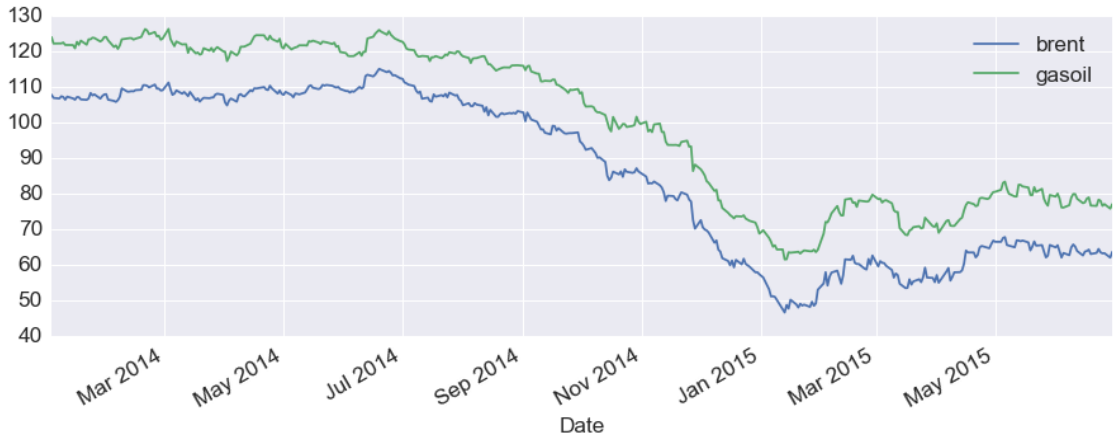
Stochastic processes are used to simulate asset prices. Monte Carlo (MC) techniques are used when relevant and random variables are drawn from the normal distribution $N(\mu = 0, \sigma = 1)$.

2.2 Real Data

- For simplicity, only two financial series are used. As the focus is on commodities, Brent crude and a by-product (Low Sulphur Gasoil) were selected since they were assumed to be good candidates for cointegration given their deep economic link
- The Brent [12] and Gasoil [5] Futures prices traded in the Intercontinental Exchange (ICE) were taken from Quandl's Steven Continuous Series [20] using the *Roll on Last Trading Day with No Price Adjustment* version and the *Settle* field as the closing price on the day

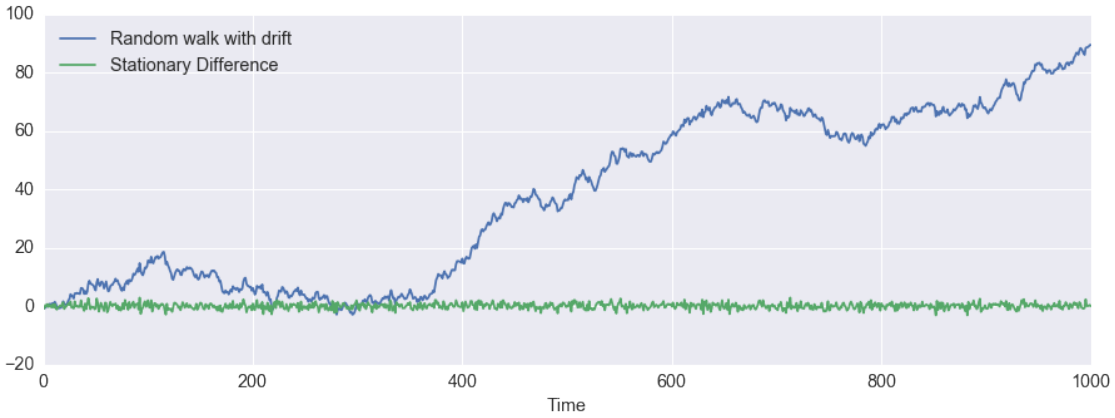
- The two series were then joined to produce a single dataset consisting of daily settlement prices for Brent and Gasoil
- The dataset spans 1.5 'trading years' (equivalent to $\sim$ 252 days). The period selected was Jan-2014 to Dec-2014 for the in-sample testing and Jan-2015 to Jun-2015 for the out-of-sample testing. This was because several sources recommend to use one year of historic data to estimate the cointegration parameters and trade the estimates for a 6-month period, given that the parameters might change over time
- Dates with missing values after joining the two series were removed from the dataset
- Since Gasoil is traded in metric tons and Brent in barrels, the gasoil series was divided by 7.45, which is the ICE conversion factor [1]

The figure below shows the resulting dataset (spanning both in-sample and out-of-sample periods), where the two series indeed seem to be closely related, having very similar trends.



3 Stationarity and Mean-Reversion

Before cointegration is introduced, it is important to understand the concept of *stationarity*. A time series is stationary when the parameters of its generating process do not change over time. In particular, its long-run mean and variance stay constant. This property is fundamental when applying linear regression and forecasting models. Often, processes with a drift or trend, like stock prices, are non-stationary but can be transformed to become stationary. For example, by differencing prices we get returns, which are in general stationary. The figure below shows how a simulated random walk with drift $Y_t = \alpha + Y_{t-1} + \epsilon_t$ can be made stationary by differencing it $\Delta Y_t = Y_t - Y_{t-1} = \alpha + \epsilon_t$



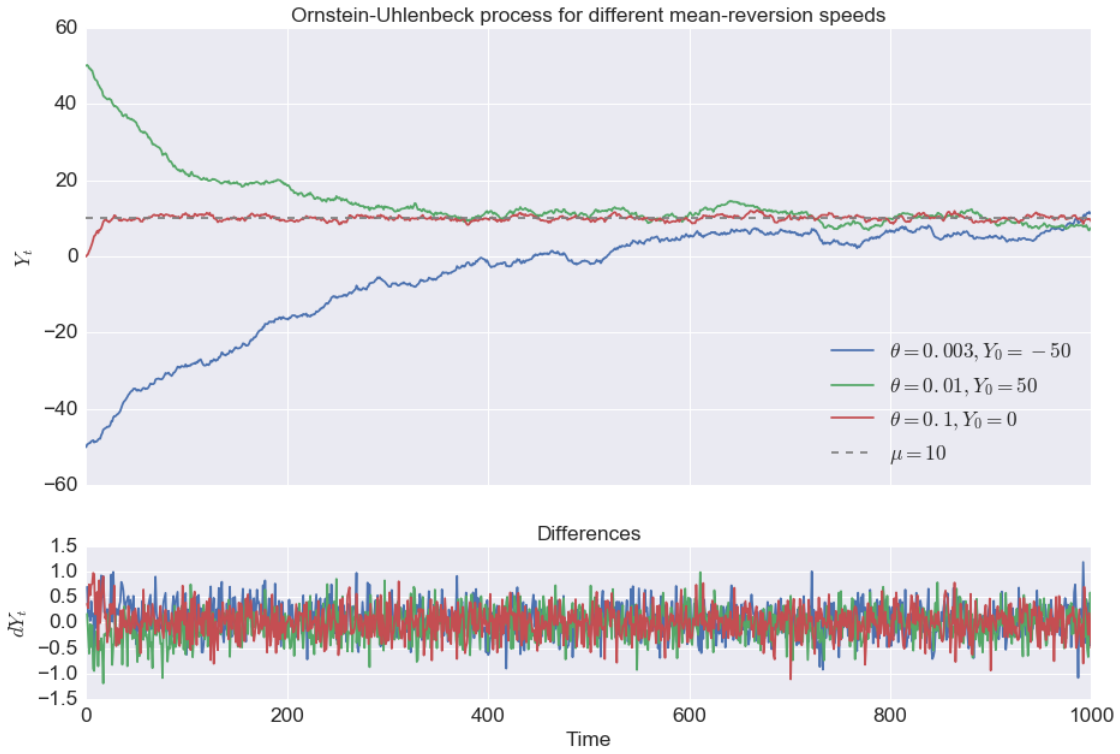
3.1 Mean-Reversion

A stationary series is *mean-reverting* if over time it drifts towards its long-term mean (the historical equilibrium level). A popular model in this category is the Ornstein–Uhlenbeck (OU) process:

$$dY_t = \theta(\mu - Y_t)dt + \sigma dW_t \tag{1}$$

where θ is the speed of reversion, μ is the equilibrium level, σ the variance and W_t is a Wiener Process (Brownian Motion). In a discrete setting this states that the further away the process is from the mean, the greater the 'pull back' to it is. This is in contrast to the random walk above, which has 'no memory' of where it has been at each particular instance of time.

The figure below shows three OU processes with the same mean $\mu = 10$ but different mean-reversion speeds. Indeed it can be noted the one with the highest θ reverts to the mean first. Their differences dY_t are plotted as well and these appear to become stationary significantly faster than the process itself, almost insensitive to the speed θ. Therefore, if we are able to transform a time series to be stationary and mean-reverting, we can design trading strategies using these properties which are more independent of market effects. In a later section we shall see how the OU parameters can be used to design exit/entry thresholds and also assess the *quality* of mean-reversion.



3.2 Tests

We require a more robust method to confirm whether a series is stationary than just by eye. Several statistical tests exist, such as the **Augmented Dickey-Fuller (ADF)** test, Phillips–Perron test, Hurst exponent, Kalman filters, etc. Here we only implement the ADF test and the mathematical details can be found in Appendix A.1.1.

3.2.1 Augmented Dickey-Fuller (ADF)

The ADF test equation implemented was:

$$\Delta Y_t = c_0 + \phi Y_{t-1} + \sum_k^p \phi_k \Delta Y_{t-k} + \epsilon_t \tag{2}$$

where a time-trend term has not been included due to the nature of financial time series [11]. The coefficients are estimated using the familiar linear regression (see Appendix A) whereas the optimal lag order p is discussed below. The python script can be found in *analysis.py*. All the results were validated against the popular python equivalents from the *statsmodels* library [17].

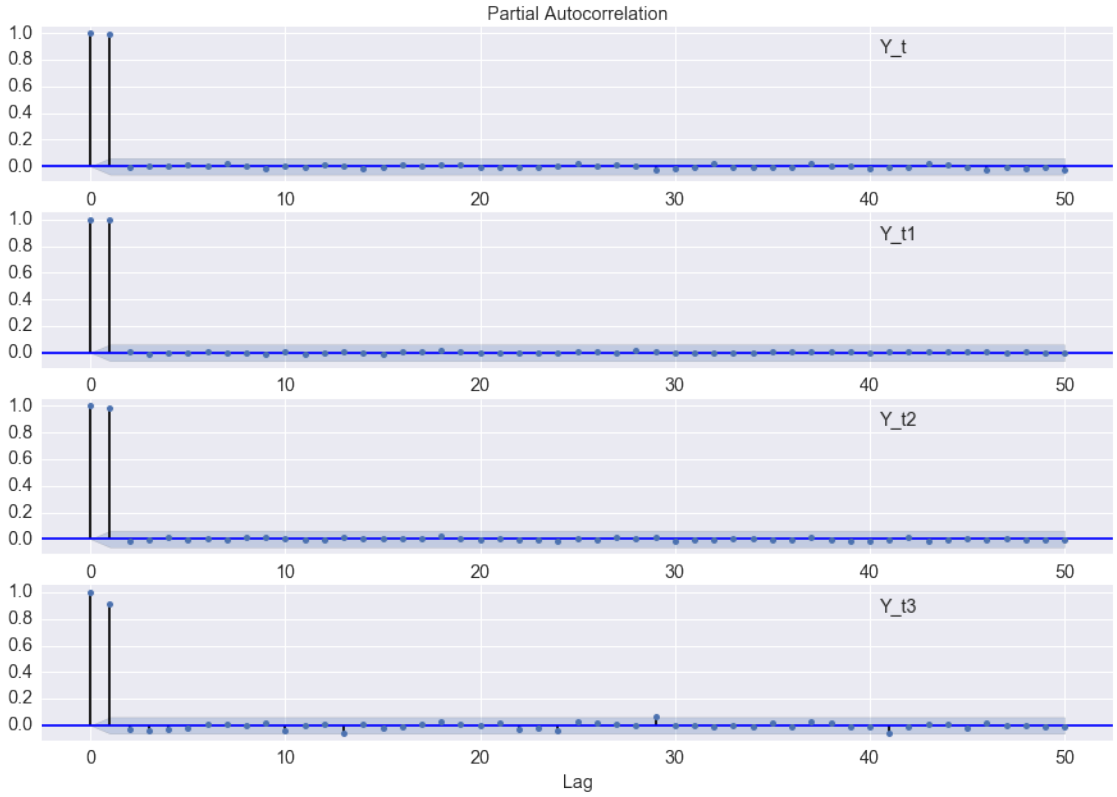
Optimal Lag Selection Choice of lag order can be a difficult problem. Standard approaches use an *information criteria*, such as the Akaike Information Criterion (AIC). However, different methods can lead to different results. Also, keeping more lags can lead to *model overfitting*. In practice, the choice of optimal lag is also evident from the Partial Autocorrelation Function [2] (PACF) since the significant lags would show above confidence limits. Typically for the ADF test

it is enough to take $p=1$, however in the interest of exploring this aspect, here we look at the results using AIC and PACF.

- **AIC:** Iterating over different lag orders, the one yielding the lowest AIC value is taken as the optimal lag. For the simulated random-walk-with-drift and OU processes above the results are summarised in the table below.

Process	OU θ	AIC optimal lag
Y_t	0	22
Y_{t1}	0.003	1
Y_{t2}	0.010	10
Y_{t3}	0.100	2

- **PACF:** Given that the optimal lag order from AIC comes out quite high in some cases, we instead use the empirical results from the PACF plot below, where it can be seen only the first lag is well above the 95% confidence band (the first 'spike' represents $p=0$). Given this, we therefore assume it is 'safe' to take $p=1$ to carry out the ADF test.



ADF Using as optimal lag $p=1$, we run the ADF test and compare the corresponding t-statistic to the critical values (taken from statsmodels, based on MacKinnon(2010) [15]). The results are

summarised in the table below, where we confirm what we expected: the null hypothesis of non-stationary is rejected for all, except for Y_t, which by definition is non-stationary given its drift

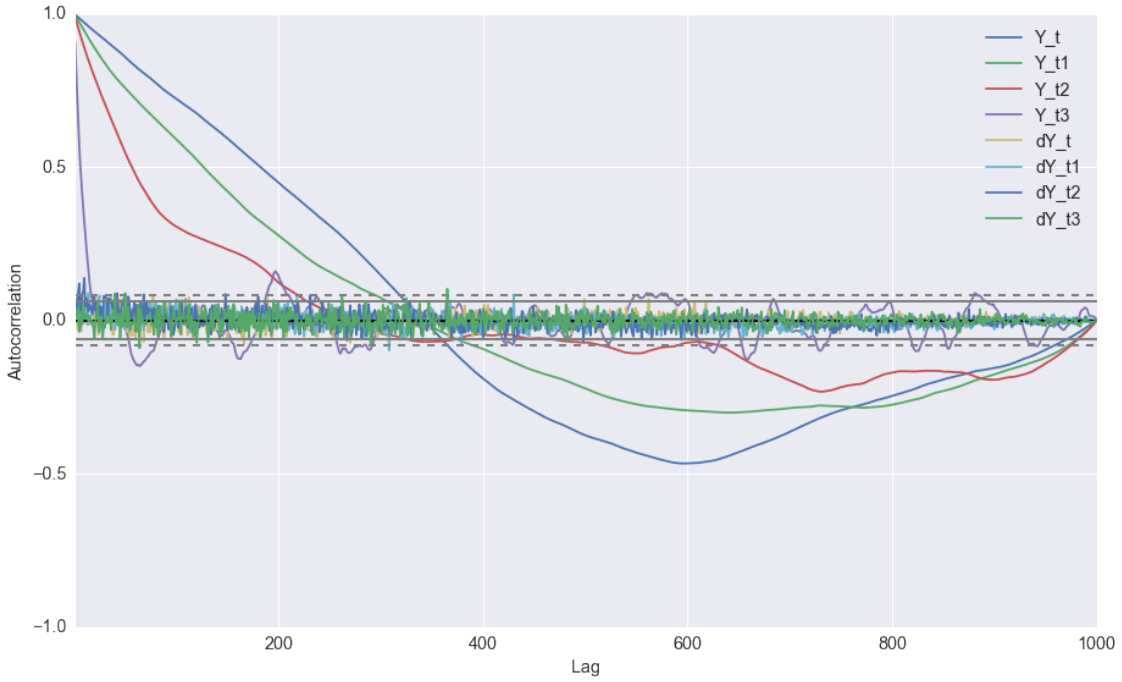
Process	θ	ADF t-stat	5% Crit. Val.	p-value	Stationary	Stable
Y_t	0	0.1093	-2.8644	0.9667	No	Yes
Y_{t1}	0.003	-6.0020	-2.8644	1.65E-07	Yes	Yes
Y_{t2}	0.01	-8.7597	-2.8644	2.69E-14	Yes	Yes
Y_{t3}	0.1	-10.1133	-2.8644	9.87E-18	Yes	Yes
dY_t	0	-31.8476	-2.8644	0.0	Yes	No
dY_{t1}	0.003	-29.4587	-2.8644	0.0	Yes	Yes
dY_{t2}	0.01	-29.3639	-2.8644	0.0	Yes	Yes
dY_{t3}	0.1	-31.3506	-2.8644	0.0	Yes	No

Stability Check To ensure further the reliability of results, a stability check can be done on the estimated coefficients by looking at their eigenvalues within the unit circle (see Appendix A.1.3). The results of the self-implementation are displayed in the table above. All cases were found stable, except dY_t and dY_{t3}. This demonstrates that stationarity does not imply stability. The unstable nature of dY_t may be due to the drift term added $dY_t = \alpha + \epsilon_t$. A close inspection of the problematic root of dY_{t3} shows it is just right on the boundary of the unit circle.

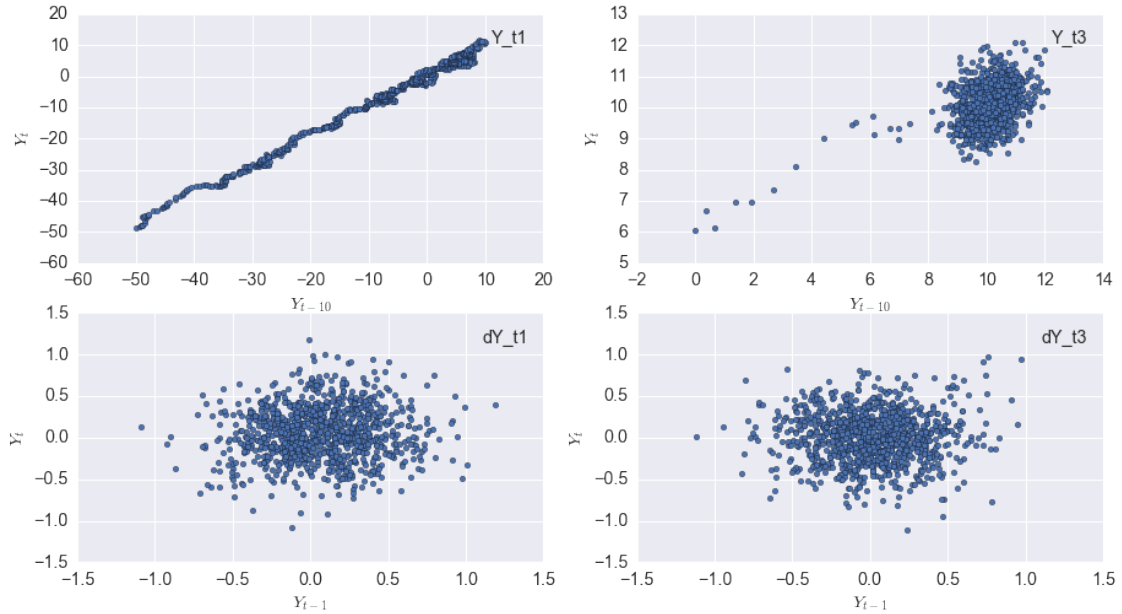
3.3 Other useful plots

In addition to the PACF, when assessing stationarity the below plot types are useful (see references [3] and [4]):

- **Autocorrelation plot:** This shows the autocorrelation function (ACF) at varying time lags. For perfectly stationary series or independent and identically distributed (iid) random variables, the autocorrelations should be near zero for all time-lag separations. The horizontal lines displayed in the plot correspond to 95% and 99% confidence levels. Indeed in the example below none of the dY_t processes show significant autocorrelation. Also, the OU process with the highest mean-reversion speed Y_{t3} rapidly loses autocorrelation to its lags, and looking stationary.



- **Lag plot:** this is a scatter plot between the series Y_t and one of its lags Y_{t-p}. Like in the autocorrelation plot, a stationary series would not exhibit any relationship. Below an example is shown for two of the OU processes and their differences. This confirms the results from the autocorrelation plot - the fastest mean-reverting Y_{t3} has little relation to the 10th lag, unlike Y_{t1}, which has a lower speed. As expected, the stationary differences are insensitive to the lags, even for Y_{t-1}.



4 Cointegration

Having covered the key concepts to understand cointegration, we proceed to explain this topic. Two or more time series $\mathbf{Y}_t = (y_{1t}, \ldots, y_{nt})'$ are said to be cointegrated if a linear combination exists which makes the collection 'integrated of order zero' I(0) i.e stationary:

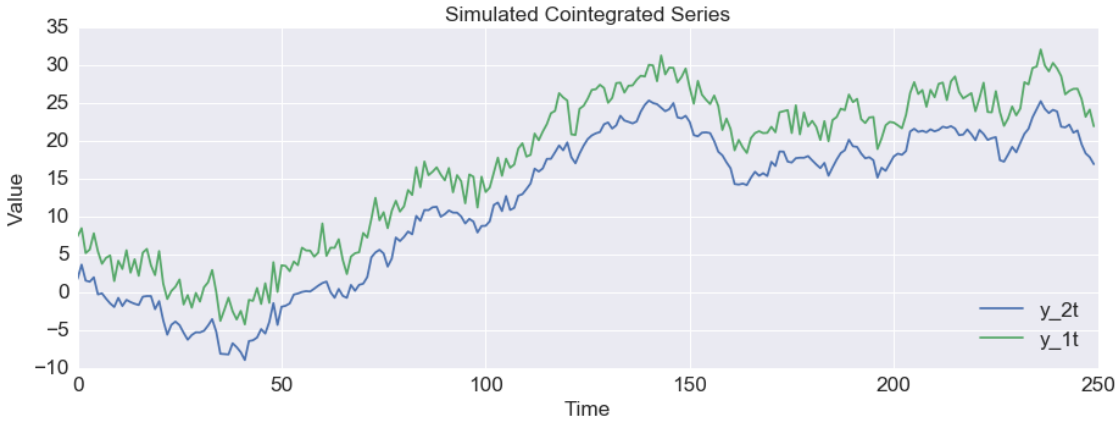
$$\beta' \mathbf{Y}_t = \beta_1 y_{1t} + \cdots + \beta_n y_{nt} \sim I(0) \tag{3}$$

This is known as the *long-run (static) equilibrium* model and is expressed in normalised form as:

$$y_{1t} = \beta_2 y_{2t} + \cdots + \beta_n y_{nt} + e_t \tag{4}$$

where the residual $e_t \sim I(0)$ is referred to as the *cointegrating residual/spread* and $\beta = (1, -\beta_2, \ldots, -\beta_n)'$ is the *cointegrating vector*. Due to time constraints, this report covers only a pair of cointegrated time series, however the concept can be extended to more series and higher orders of integration. In real data, cointegration usually exists when there is a deep economic link between the assets and hence these cannot drift too far apart because economic forces will act to restore the long-run equilibrium. The figure below shows a simulated pair of cointegrated assets y_{1t} and y_{2t}, where y_{2t} is defined as an I(1) process. If y_{1t} is supposed to have a strong link to y_{2t}, the price of y_{1t} should vary similarly. This is simulated by shifting up y_{2t} and adding some noise (residual) drawn from a normal distribution, so y_{1t} is defined as the dependent variable and y_{2t} as the independent variable:

$$y_{1t} = y_{2t} + 5.0 + \epsilon_t \tag{5}$$



4.1 Tests

Again, we need a robust statistical test to confirm cointegration. The three main approaches are:

- **Engle–Granger two-step method:** This can only be used to test a *single* cointegrating relationship. The steps are:

(i) Estimate the cointegrating residual $\hat{e}_t = \hat{\beta}'\mathbf{Y}_t$, e.g. using linear regression

(ii) Test $\hat{e}_t$ for stationarity, e.g. using the ADF test, where the hypotheses to be tested are:

$$H_0 : \hat{e}_t = \hat{\beta}'\mathbf{Y}_t \sim I(1) \qquad \text{(no cointegration)} \tag{6}$$

$$H_1 : \hat{e}_t = \hat{\beta}'\mathbf{Y}_t \sim I(0) \qquad \text{(cointegration)} \tag{7}$$

We refer to this use case of the ADF as 'CADF'

- **Johansen test:** Based on maximum likelihood techniques, this allows for more than one cointegrating relationship, but it is subject to asymptotic conditions when the sample size is too small
- **Phillips–Ouliaris test:** Uses a modified version of the Dickey-Fuller distribution to test the cointegrating spread for stationarity. This is a better choice when dealing with small samples

In this project only the Engle-Granger two-step method is considered.

4.1.1 Engle-Granger Two-Step

For our simulated cointegrated series, the cointegration relationship is then represented by the regression:

$$y_{1t} = c + \beta_2' y_{2t} + e_t \tag{8}$$

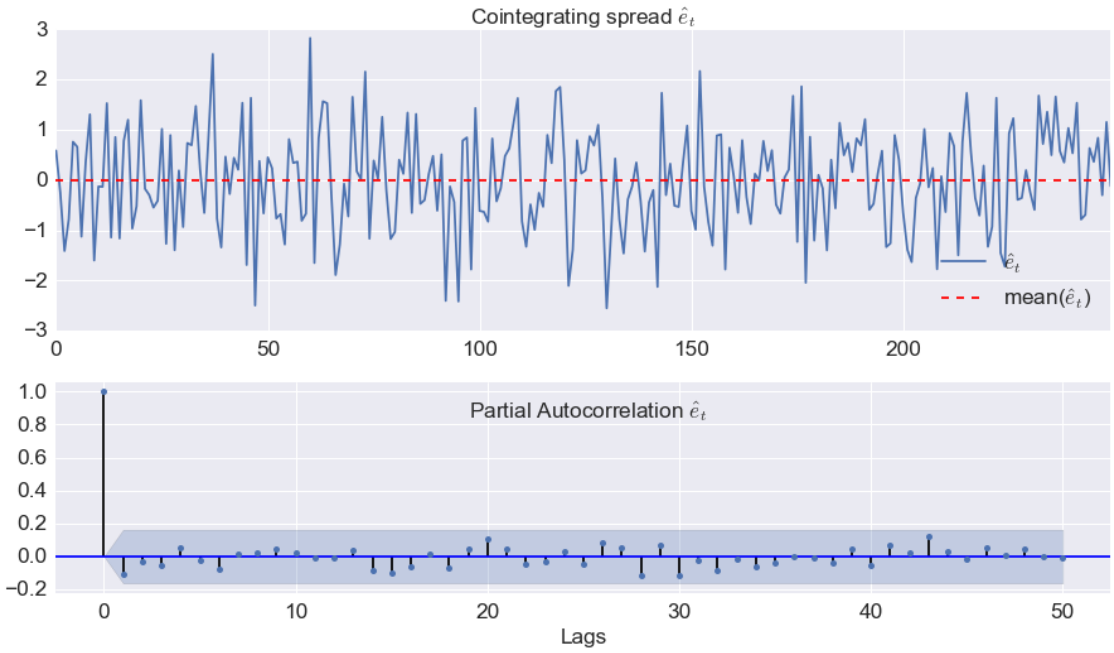
whose parameters are estimated using OLS. Hence, the estimated cointegrating spread is:

$$\hat{e}_t = y_{1t} - \hat{c} - \hat{\beta}_2' y_{2t} \tag{9}$$

We then test $\hat{e}_t$ for stationarity using ADF. Since the mean of $\hat{e}_t$ is zero, the ADF can be implemented without a constant or trend (See reference [16]). Also note that the critical values used are taken as in statsmodels from MacKinnon (2010) [13], whereas other sources suggest to use the Phillips-Ouliaris tabulation in this case. The figure below shows the OLS fit:

$$\hat{y}_{1t} = 5.0330 + 1.0052 y_{2t} + \hat{e}_t \tag{10}$$

which is in good agreement with its true value. Plotted is also the estimated $\hat{e}_t$ and its PACF. The PACF shows the spread is *memoryless*, i.e. no lag order appears significant so it is a random (Markov) process. Although $\hat{e}_t$ appeared memoryless, conservatively the CADF t-statistic was computed using one lag. This was -12.051, which was below the 1% critical value (-2.5748), confirming the stationarity of the spread, which by design we expected.



4.2 Error Correction Model (ECM)

Cointegration implies the existence of an Error Correction Model (ECM), which provides an adjustment to the long-run equilibrium from the short-run dynamics. This is particularly useful when modelling non-stationary series, like market prices, which can lead to spurious regression results. Suppose the cointegrated pair is represented by $\mathbf{Y}_t = (y_t, x_t)'$. One can arrive at the ECM result as follows:

- Consider a dynamic regression model to allow for a wide variety of dynamic patterns in the data. This is done by including lags for both x_t and y_t:

$$y_t = \alpha y_{t-1} + \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \epsilon_t \tag{11}$$

- By knowing the above equation should be consistent with the long-run equilibrium model $y_t = b_0 + b_1 x_t + e_t$, it can be re-written as:

$$\Delta y_t = \beta_1 \Delta x_t - (1-\alpha) e_{t-1} + \epsilon_t \tag{12}$$

  where e_{t-1} is the lagged cointegrating spread from the equilibrium model. The parameter $-(1-\alpha)$ is interpreted as the speed of correction towards the equilibrium level (more details on this in the next section)

- Since all the variables in the ECM are I(0), OLS can be used to estimate the parameters

4.3 Quality of mean-reversion

If the cointegrating spread e_t is stationary, we could use the OU process to model it:

$$de_t = \theta(\mu_e - e_t)dt + \sigma dW \tag{13}$$

In discrete time this can be written as:

$$\Delta e_t = \alpha\mu_e - \alpha e_{t-1} + \epsilon_{t,\tau} \tag{14}$$

where $\alpha = 1 - e^{-\theta\tau}$ and τ is a small period of time. This is in fact the implied ECM representation for e_t. As discussed in the Appendix, the above equation can be written in its AR(1) representation as:

$$e_t = \alpha\mu_e + (1-\alpha)e_{t-1} + \epsilon_{t,\tau} \tag{15}$$

Re-writing in terms of $C = \alpha\mu_e$ and $B = 1 - \alpha$ we see it is a simple regression which can be determined with OLS:

$$e_t = C + Be_{t-1} + \epsilon_{t,\tau} \tag{16}$$

In particular, to assess the quality of the spread for trading strategies the parameters of interest are:

$$\theta = -\frac{\ln B}{\tau} \qquad \mu_e = \frac{C}{1-B} \quad \sigma_{OU} = \sqrt{\frac{2\theta}{1-e^{-2\theta\tau}} Var[\epsilon_{t,\tau}]} \tag{17}$$

- μ_e is the long-run equilibrium level of the OU process
- The mean-reversion speed θ can be translated into a half-life $\tilde{\tau} \propto \ln 2/\theta$, which is the time between equilibrium situations, i.e. when $e_t = \mu_e$. Hence, a high θ (small $\tilde{\tau}$) is desirable to trigger trading signals
- The standard deviation defined as $\sigma_{eq} = \sigma_{OU}/\sqrt{2\theta}$ can be used to plot trading bounds for entry/exit signals as $\mu_e \pm \sigma_{eq}$
- Here we use $\tau = 1/252$ because our data has a daily frequency and the in-sample period covers one trading year, i.e. ~ 252 trading days

4.4 Real Data Example

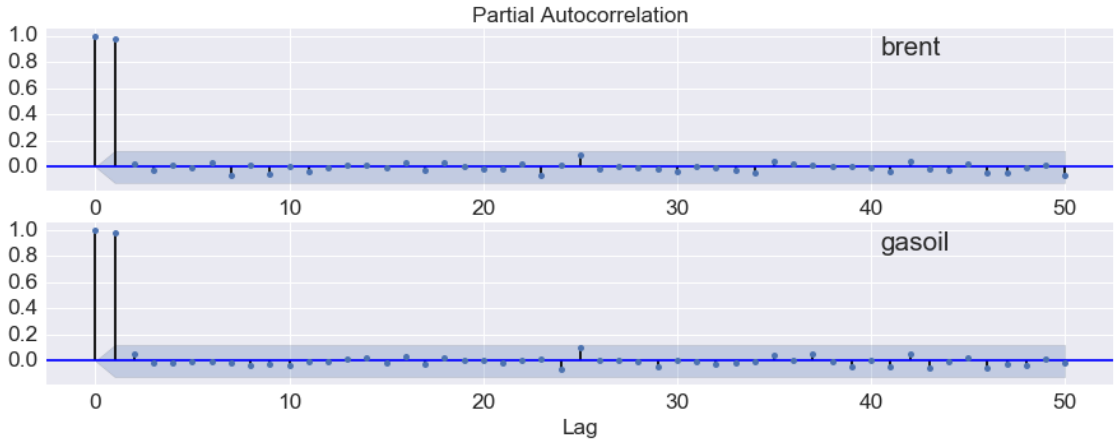
We now proceed to apply the concepts discussed above to real data:

- We use the in-sample dataset for Brent crude and Gasoil (Jan-2014 to Dec-2014). The out-of-sample dataset is later used to backtest trading strategies
- For now we assume Brent is the independent variable x_t and Gasoil the dependent variable y_t, and later check if this is accurate via *Granger's causality test*

4.4.1 Stationarity

First we check the individual price series for unit root I(1). We apply the ADF test using one lag only as recommended by their PACF plot below. A drift term (constant) is also included in the test. The results are summarised in the table below and support the assumption that Brent and Gasoil are I(1) processes whilst their differences are I(0).

Series	ADF t-stat	5% Crit. Val.	p-value	Stationary	Stable
x_t (Brent)	3.5768	-2.8728	1.00	No	No
y_t (Gasoil)	3.9374	-2.8728	1.00	No	No
Δx_t	-18.1437	-2.8728	2.49E-30	Yes	Yes
Δy_t	-18.9247	-2.8728	0.00	Yes	Yes



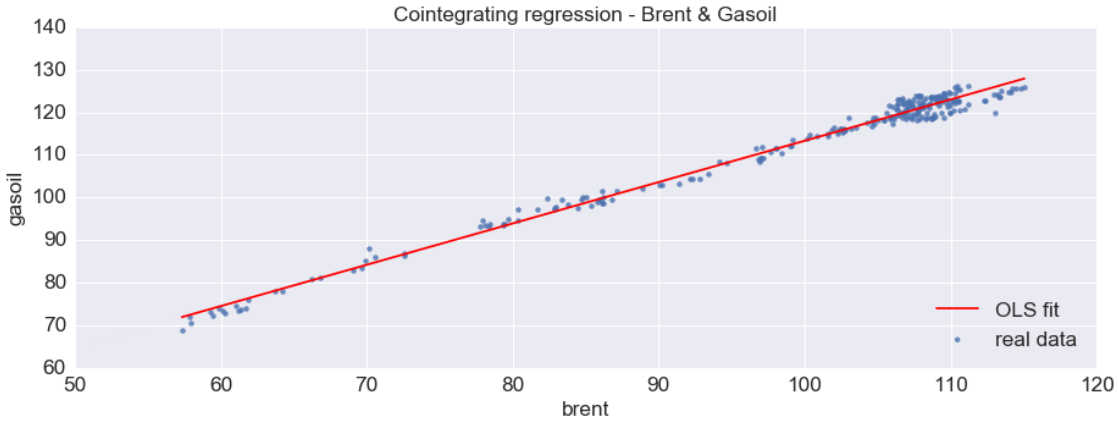
4.4.2 Cointegration

We then proceed to test whether the pair is cointegrated, using the Engle-Granger two-step procedure previously described.

- The long-run equilibrium model is shown in the figure below. The OLS estimate was:

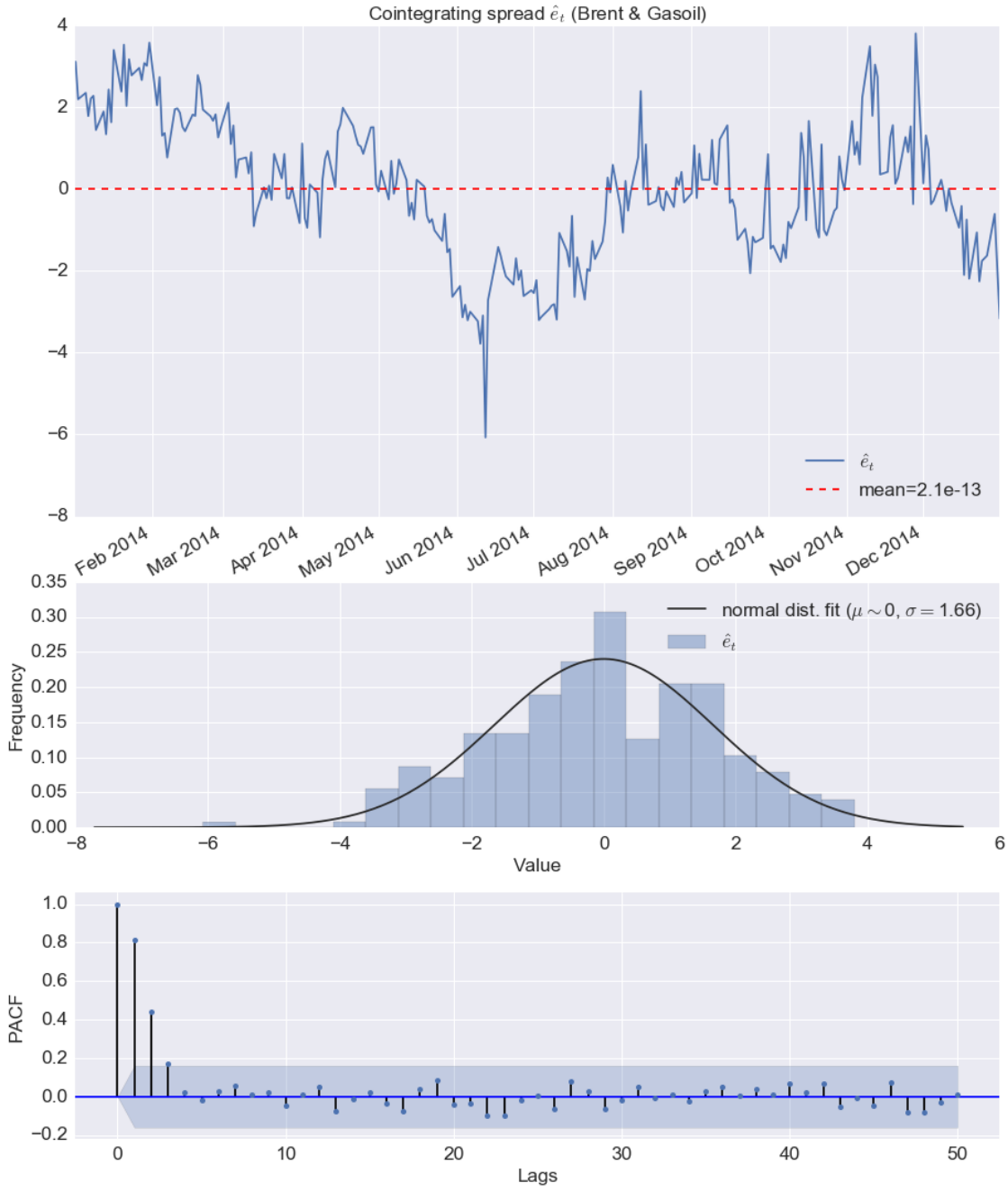
$$\hat{y}_x = 16.3229 + 0.9699 x_t + \hat{e}_t \tag{18}$$

with goodness-of-fit $R^2 = 0.986$.



- The estimated cointegrating spread $\hat{e}_t$ series, histogram and PACF are shown below. First the mean of $\hat{e}_t$ is zero for all practical purposes ($\sim 10^{-13}$).
- We also want $\hat{e}_t$ to be normally distributed. The histogram below shows the spread distribution fitted to a normal distribution with $\mu \sim 0$ and $\sigma = 1.66$. Two tests (*Lilliefors* and *Anderson-Darling*, see references [18] and [8]) were carried out to assess the goodness of fit and neither rejected the null hypothesis that $\hat{e}_t$ is normally distributed
- Next, we note that the PACF plot shows a peculiarity of this spread - it seems to have an AR(3) 'memory', given the first three lags appear significant. Memory in the cointegrating spread is actually *not desirable* because it will interfere with both mean-reversion and diffusion components of the OU process. One hypothesis is that this has an 'economic explanation' in terms of commodity pricing, cyclicality or storage (e.g. asset cannot dramatically drop because there are storage/delivery 'carry' costs borne by the sellers)
- Next, $\hat{e}_t$ was tested for stationarity using the CADF test. Since $\hat{\mu}_e \sim 0$ the test was implemented without a constant or trend. Also, as suggested by the PACF plot, the test was ran with 3 lags. The results below show stationary could only be confirmed at the 95% C.L.

Series	ADF t-stat	5% Crit. Val.	1% Crit. Val.	p-value	Stationary (95% C.L.)	Stable
$\hat{e}_t$	-2.2335	-1.9421	-2.5746	0.0245	Yes	Yes



4.4.3 ECM

The corresponding ECM adjustment was determined to be:

$$\begin{aligned} \Delta y_t = &\quad -0.0861 + \quad 0.6455\Delta x_t - 0.1623 e_{t-1} + \epsilon_t \qquad \text{(with constant, } \mathrm{R}^2 = 0.407) \\ \Delta y_t = &\qquad\qquad\quad 0.6596\Delta x_t - 0.1633 e_{t-1} + \epsilon_t \qquad \text{(no constant, } \mathrm{R}^2 = 0.422) \end{aligned} \tag{19}$$

This represents the second-order adjustment to the price, which as we see is quite significant.

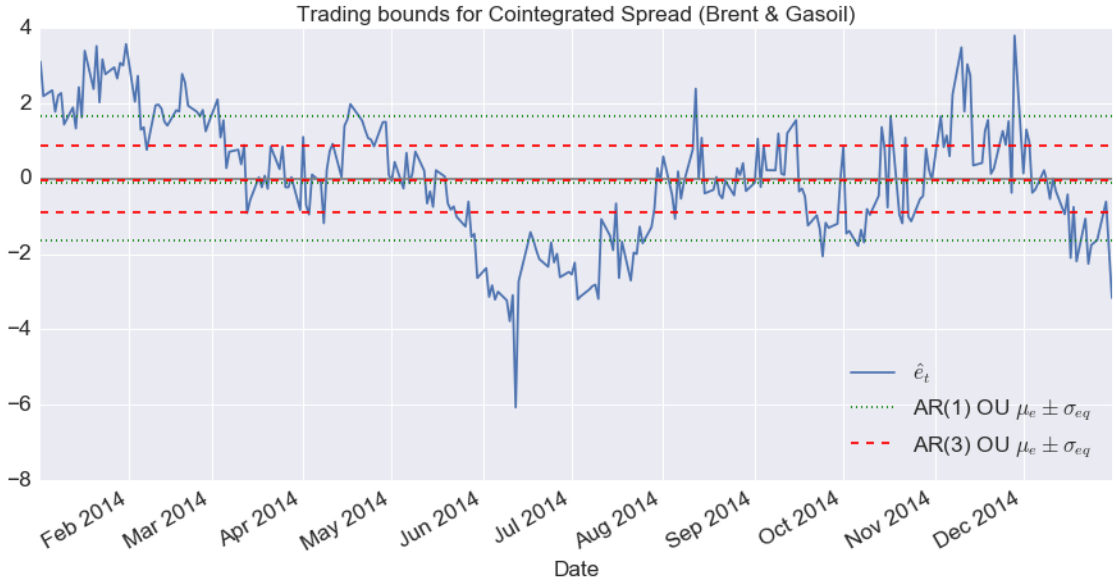
4.4.4 Quality of mean-reversion

Given the peculiar PACF of $\hat{e}_t$, which suggests it is an AR(3) process, fitting to the OU process, which is an AR(1) process, might not be suitable. We proceed nevertheless to compare the results between both fits. For AR(3) we take the OU parameters as for AR(1): the constant term equals $\alpha\mu_e$ and the coefficient of the e_{t-1} term is $\alpha = 1 - e^{-\theta\tau}$. The table below shows the results obtained, where we see the trading bounds are:

$$\mu_e \pm \sigma_{eq} = \begin{cases} -0.1255 \pm 1.6541 & \text{AR(1)} \\ -0.0529 \pm 0.8820 & \text{AR(3)} \end{cases} \tag{20}$$

The AR(3) fit yields a lower standard error in θ than AR(1), although not by much and yet the θ values are quite different. This perhaps indicates the OU fit is not suitable regardless of the lag order.

Process	θ	$\tilde{\tau}$	μ_e	σ_{OU}	σ_{eq}	$s.e._\theta$
AR(1)	49.1253	0.0141	-0.1255	16.3957	1.6541	7.3850
AR(3)	258.9327	0.0027	-0.0529	20.0723	0.8820	5.6638



4.5 Granger Causality

When implementing the cointegrating regression to estimate e_t, an assumption must be made about which the dependent variable is. It is actually important to determine the best choice for the independent variable and the dependent variable, as this can largely influence the parameters estimates and test results.

- Optimal lag determined from the PACF plot of Brent and Gasoil, which shows only maxlag=1 is necessary
- Only do this test in the in-sample, assuming the relationship holds in the out-of-sample

4.6 Issues with Cointegration

- Johansen procedure - MLE for multivariate cointegration on asset price data (levels, not returns)

5 Trading Strategies

Below we describe a few strategies which exploit the cointegration properties. It should be noted that it is actually very difficult to find a directly tradable asset that possesses mean-reverting behaviour - in practice this is instead achieved by creating a *stationary portfolio* of separately interrelated assets. For the demonstrations below however, we remain with a the 2-asset portfolio from before: Brent and Gasoil.

5.1 Regime changes

Tests for cointegration assume that the cointegrating vector is constant during the period of study. In reality, it is possible that the long-run relationship between the underlying variables change (shifts in the cointegrating vector can occur), e.g from changes in their *fundamental factors*. This is especially likely to be the case if the sample period is long. Hence, it should not be assumed that because the assets have passed a cointegration test historically, they will continue to remain cointegrated. Practitioners' advice is to estimate using one year of historic data and trade the estimates for a 6-month period. This is then a good introduction to the next topic, which is backtesting.

5.2 Backtesting

Backtesting is the process of testing a trading strategy or model on historical data to gauge its effectiveness. When we backtest a model, the results achieved are highly dependent on the tested period and this can cause the strategy to fail in the future, as regime changes or *model overfitting* could have taken place. To alleviate this, the test data is usually split into *in-sample* and *out-of-sample* (a 'rule-of-thumb' is to use an 80-20 split). The model/strategy is then fitted/tested to the in-sample data *only* and then tested on the out-of-sample data, which was 'unseen'. This process provides a better way to assess the true performance.

5.3 Naive Beta-Hedging Strategy

** Factor models** are a way of explaining the returns of one asset via a linear combination of the returns of other assets. This can be expressed as a simple linear regression:

$$\Delta Y = \alpha + \beta_1 \Delta X_1 + \beta_2 \Delta X_2 + \cdots + \beta_n \Delta X_n + \epsilon_t \tag{21}$$

We can interpret the betas as the *exposure* of asset Y to the other assets and α as the market-neutral excess return:

$$\alpha = \mathrm{E}[\Delta Y - \beta_1 \Delta X_1 + \beta_2 \Delta X_2 + \cdots + \beta_n \Delta X_n] \tag{22}$$

since $\mathrm{E}[\epsilon_t] = 0$. For the case of Brent and Gasoil only this would be:

$$\Delta Y_{gasoil} = \alpha + \beta \Delta X_{brent} + \epsilon_t \tag{23}$$

We could take a short position in Brent equal to $\beta \Delta X_{brent}$ to try to eliminate the *risk* in our Gasoil position. Hence, we would expect our returns to be on average:

$$\mathrm{E}[\Delta Y_{gasoil} - \beta X_{brent}] = \alpha \tag{24}$$

For the **in-sample** data, running the above regression would actually yield negative returns, although less severe than just going long on Gasoil:

$$\begin{aligned} \mathrm{E}[\Delta Y^{is}_{gasoil} - 0.6282 \Delta X^{is}_{brent}] = &\quad \alpha = -0.0915 \qquad (R^2 = 0.354) \\ \mathrm{E}[\Delta Y^{is}_{gasoil}] = &\quad \alpha = -0.2148 \qquad \text{(no fit)} \end{aligned} \tag{25}$$

One problem could be that the estimated beta is not constant as we walk forward in time. As such, the short position we took out in Brent is not perfectly hedging our portfolio. Another is that additional factors or assets should be included into the model. Surprisingly, our performance in the **out-of-sample** seems to be better but that was just mere 'luck' from positive fluctuations in the data, and still worse off than just going long on Gasoil over this period:

$$\begin{aligned} \mathrm{E}[\Delta Y^{os}_{gasoil} - 0.6282 \Delta X^{os}_{brent}] = &\quad \mathrm{E}[\alpha_t] = 0.02231 \qquad \text{(in-sample fit)} \\ \mathrm{E}[\Delta Y^{os}_{gasoil}] = &\quad \mathrm{E}[\alpha_t] = 0.05806 \qquad \text{(no fit)} \end{aligned} \tag{26}$$

5.4 Pairs Trading Strategy

Moving to a more sophisticated approach, we now apply the concepts of cointegration learned. Although it is commonly referred as 'pairs trading', the concept applies to any stationary portfolio of assets. For a pair of assets, if they are cointegrated, e.g. because they belong to the same 'sector', they are likely to be exposed to similar market factors. Occasionally their relative prices will diverge due to certain events, but eventually will revert to the long-run equilibrium. Hence, positions are taken relative to where the **cointegrating spread** $\hat{e}_t$ is with respect to the equilibrium level. The objective is to hedge the position from price level dynamics (market risk). The P&L generated will then be driven by the quality of mean-reversion. For Gasoil and Brent, we estimated the **in-sample** cointegrating relationship as:

$$Y^{is}_{gasoil} = 0.9699 X^{is}_{brent} + 16.3229 + \hat{e}^{is}_t \tag{27}$$

Under the assumption that the relationship holds, we use the same beta and constant to estimate the **out-of-sample** spread:

$$\hat{e}^{os}_t = Y^{os}_{gasoil} - 0.9699 X^{os}_{brent} - 16.3229 \tag{28}$$

This is then what we use to **backtest** the strategy.

5.4.1 Bounds from OU process

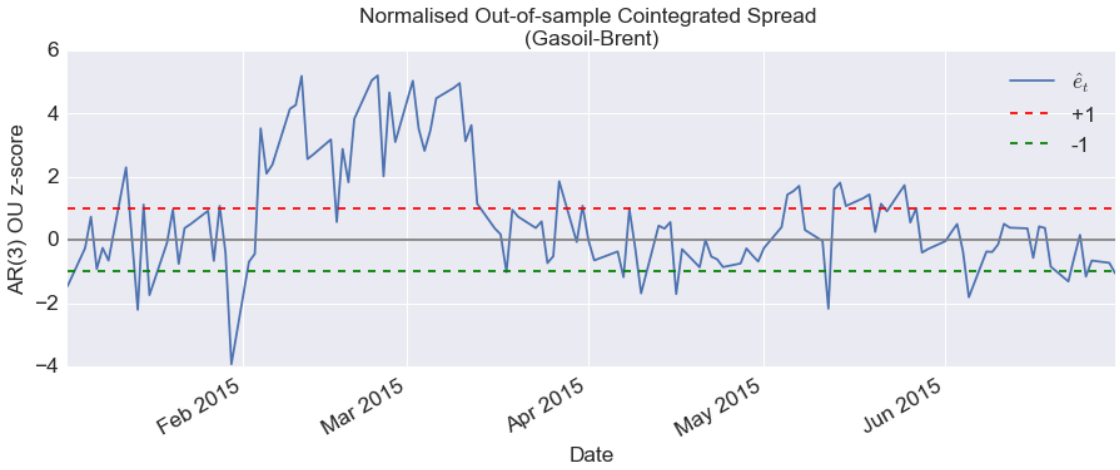
As we saw, we could define thresholds for trading the spread by fitting to the OU process and using the fitted mean μ_e and relative standard deviation σ_{eq}. Assuming normality of the spread $e_{t,\tau \to \infty} \sim N(\mu_e, \sigma^2_{eq})$, a z-score could be used to 'normalise' the spread:

$$z = \frac{\hat{e}_t - \mu_e}{\sigma_{eq}} \tag{29}$$

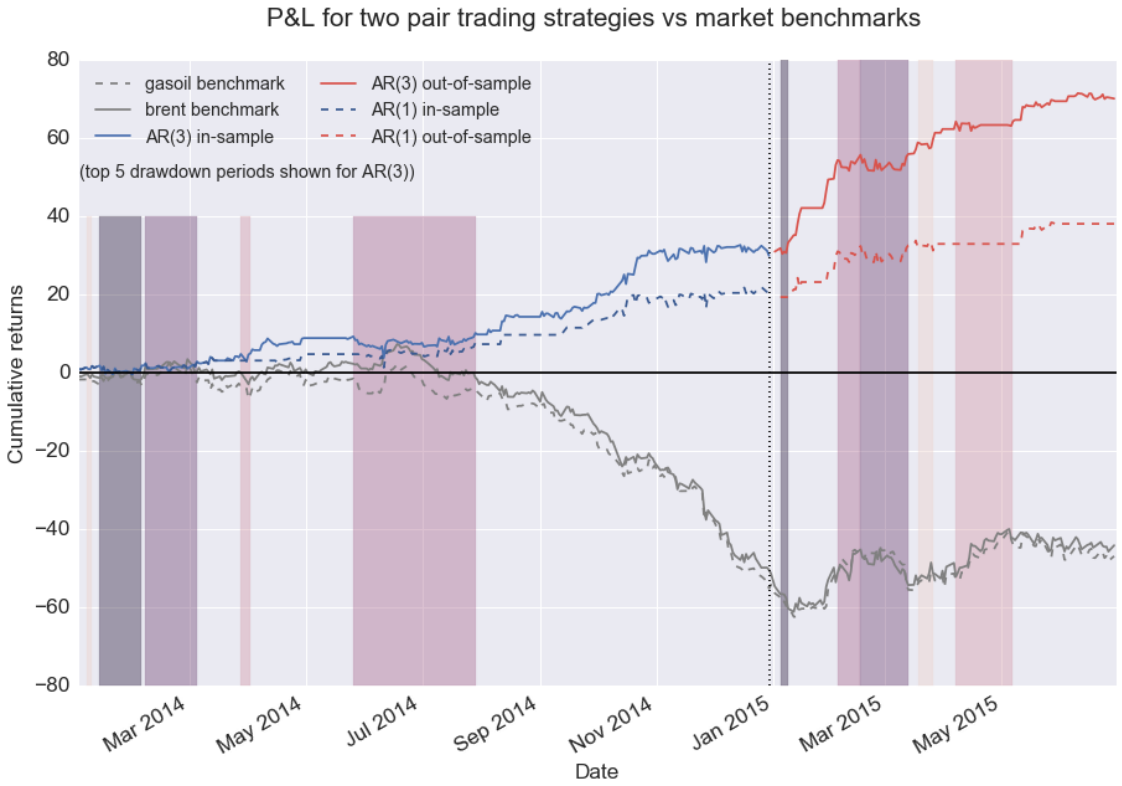
Given this, a simple strategy would then be:

- Go "Long" the spread when the z-score is below -1.0: this means longing Y_{gasoil} and shorting X_{brent}
- Go "Short" the spread when the z-score is above 1.0: this means shorting Y_{gasoil} and longing X_{brent}
- Exit positions when the z-score is within $[-0.1, 0.1]$, i.e. near zero

The figure below shows the out-of-sample z-score spread with the AR(3) OU entry/exit bounds (Eqn. 29).



The corresponding P&L using the AR(3) bounds is shown below for both the in-sample and out-of-sample periods. A positive performance is observed throughout, although with prolonged drawdown periods. In addition, note this does not include **transaction costs** which could certainly add up to an overall loss. The P&L using the AR(1) bounds is also shown, where it is clear it is worse off than AR(3). Both however are better off than having simply gone long in Gasoil or Brent - this is taken as the market 'benchmark'. Given the positive results, it would be interesting to see the P&L from extending to more than 2 assets cointegrated with oil, with a strategy that trades multiple cointegrated spreads at the same time, or a single spread composed of more factors. In this case, the **Johansen test** would need to be implemented instead of CADF.



5.4.2 Bounds from Optimisation

Although earlier the Gasoil-Brent spread passed the Anderson-Darling test for normality, we also saw it had memory AR(3). Hence, it is perhaps not optimal to use an OU fit to define the bounds, as this is only suited for an AR(1) spread. The P&L plot also suggests the performance is sensitive to where bounds are. This motivates the search for the *optimal bound*, which maximises the P&L or any other P&L-related metric, e.g. number of trades. Optimisation techniques can be used to test different combinations of $\mu_e \pm \sigma_{eq}$ in the in-sample data, although again, this can suffer from model overfitting and cause a worse performance in the out-of-sample. Due to time constraints this was not evaluated in this report.

5.4.3 Dynamic Bounds

Another alternative is to use rolling values for $\mu_e \pm \sigma_{eq}$ instead of static. Again, an optimisation could be done to find the optimal period for the rolling window, which could be different for μ_e and σ_{eq} given their different sensitivity to time-scales.

References

[1] https://www.theice.com/publicdocs/futures/ICE_Gas_Oil_Crack.pdf

[2] http://nl.mathworks.com/help/econ/autocorrelation-and-partial-autocorrelation.html

[3] http://pandas.pydata.org/pandas-docs/stable/visualization.html#autocorrelation-plot

[4] https://github.com/pydata/pandas/blob/master/pandas/tools/plotting.py

[5] https://www.theice.com/products/34361119/Low-Sulphur-Gasoil-Futures

[6] http://statsmodels.sourceforge.net/stable/vector_ar.html#lag-order-selection

[7] https://en.wikipedia.org/wiki/Autoregressive_model

[8] http://statsmodels.sourceforge.net/devel/generated/statsmodels.stats.diagnostic.kstest_normal.html

[9] https://en.wikipedia.org/wiki/Vector_autoregression

[10] https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

[11]

[12] https://www.theice.com/products/219/Brent-Crude-Futures

[13] http://statsmodels.sourceforge.net/stable/generated/statsmodels.tsa.stattools.adfuller.html#statsmodels.tsa

[14] https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test

[15] http://statsmodels.sourceforge.net/stable/generated/statsmodels.tsa.stattools.adfuller.html#statsmodels.tsa

[16] file:///C:/Users/Tanya.Sandoval/Downloads/Cointegration%20-%20Book%20Chapter%20-%20UWashington%20E%20Zivot.pdf

[17] http://statsmodels.sourceforge.net/

[18] http://statsmodels.sourceforge.net/notebooks/generated/statsmodels.stats.diagnostic.normal_ad.html

[19] http://matthieustigler.github.io/Lectures/Lect2ARMA.pdf

[20] https://www.quandl.com/data/SCF/documentation/about

[21] https://en.wikipedia.org/wiki/Information_criterion

A Multivariate Regression

Also known as 'generalised linear model', it generalises linear regression to multiple input variables (regressors) and n observations. It is best expressed in matrix form as:

$$Y = X\beta + \epsilon \tag{30}$$

where Y is a vector representing the endogenous (dependent) variables, X is a matrix representing the exogenous (independent) variables, β is the coefficients vector and ϵ the residuals vector. The

OLS method, which minimises the sum of squared residuals via the Maximum Likelihood Estimation method (MLE), is used to estimate the parameters:

$$\hat{\beta} = (X'X)^{-1}X'Y\hat{\epsilon} = Y - X\hat{\beta} \tag{31}$$

The covariance matrix of the residuals estimate is:

$$\hat{\Sigma} = scale \times \sum \epsilon\epsilon' \tag{32}$$

where $scale = 1/n$ using the MLE estimator or $scale = 1/(n - kp)$ using the OLS estimator for a model with k variables and p lags. The covariance matrix for the coefficients is:

$$(XX')^{-1} \otimes \hat{\Sigma} \tag{33}$$

where $\otimes$ is the *Kronecker product*. The Log Likelihood Function for OLS is:

$$\log(L) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{n}{2} \tag{34}$$

The variance of the residuals and parameters are therefore the diagonal elements of the corresponding covariance matrices, from which the standard errors can be calculated. These are the conventions used in the script *analysis.py*. Additional mathematical details can be found in reference [9]. The are several assumptions about the nature of the variables in the model to hold. In particular, these should be *stationary* and the ϵ homoscedastic (with finite variance) and normally distributed. The main applications of the multivariate regression are: * Vector Autoregression Models - also known as VAR(p), these can be used to forecast, test for stationarity (e.g. ADF test) or model stationary series like returns * Error Correction Model - also known as ECM, these are used to model series which aren't stationary or that have stochastic trends, like prices

A.1 Autoregression Models - AR(p)

Also referred to as VAR(p) where p is the lag order, it is simply a linear regression on a time series and its lagged (past) values:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \tag{35}$$

where c is a constant (also known as the drift term), ϕ_i are the parameters of the model and ϵ_t is the error term. Whether to exclude the constant c or not depends on the nature of what we are trying to model. Computationally, the model can be fitted in one go by using the OLS method described above with a special matrix formulation. Example code for this can be found in *analysis.py* in the project repository. As shown in Section X, the VAR(p) system can be re-written in terms of differences and lagged differences, which is how it is commonly expressed in some cases, for example in the ADF test. See more details in references [9] and [7].

A.1.1 Dickey-Fuller Test and ADF

The *Dickey-Fuller test* examines the null hypothesis of whether a unit root is present in the autoregressive model (also known as AR(p), see Appendix) of the time series. For example, a simple AR(1) model is:

$$Y_t = \beta Y_{t-1} + \epsilon_t \tag{36}$$

If $\beta = 1$ the series is said to have a 'unit root' and hence is non-stationary. The equation can be re-written as:

$$\Delta Y_t = (\beta - 1)Y_{t-1} + \epsilon_t = \phi Y_{t-1} + \epsilon_t \tag{37}$$

where $\phi = \beta - 1$. Hence, testing for unit root is equivalent to testing $\phi = 0$. The value of the test statistic$(\hat{\phi})$/std.err$(\hat{\phi})$ is then compared to the relevant critical values for the Dickey-Fuller distribution. If found lower, then the null hypothesis $\phi = 0$ is rejected and the series can be considered stationary. There are three main versions of the test depending on whether drift and/or time-dependent terms are included:

- Test for a unit root: $\Delta Y_t = \phi Y_{t-1} + \epsilon_t$
- Test for a unit root with drift: $\Delta Y_t = c_0 + \phi Y_{t-1} + \epsilon_t$
- Test for a unit root with drift and deterministic time-trend: $\Delta Y_t = c_0 + c_1 t + \phi Y_{t-1} + \epsilon_t$

Each version of the test has its own critical value which depends on the size of the sample. Which version to use is not straightforward and the wrong choice can lead to wrong conclusions. In general, financial time series exclude the time-trend. There is an extension to the test called the **Augmented Dickey–Fuller test (ADF)**, which removes autocorrelation effects by including lagged difference terms $\phi_p \Delta Y_{t-p}$. The optimal lag order could then be determined from the information criteria (see above). Clearly the above equations belong to the family of generalised linear models, which means the parameters can be estimated using the familiar linear regression described above. Additional details on this topic can be found in references [10], [14] and the Appendix.

A.1.2 Optimal Lag Order

To select the optimal lag order, one approach uses the Akaike Information Criterion (AIC). Iterating over different lag orders, the one yielding the lowest value of AIC is selected. Statsmodels suggests to try up to a maximum lag order of $12 * (n/100)^{1/4}$ where n is the number of observations. There are different definitions of AIC used - we use the same as in statstools [6], which has different definitions for AR(p) and the ADF test as:

$$AIC = \log|\hat{\Sigma}| + 2\frac{1+k}{n} \qquad \text{(AR(p) model)} \tag{38}$$

$$AIC = -2\log(L) + 2k \qquad \text{(ADF test)} \tag{39}$$

where k is the number of estimated parameters. Other information criteria can be used, see for example reference [21].

A.1.3 Stability Condition

It is required for the eigenvalues of the estimated coefficients matrix or vector to be inside the unit circle (<1):

$$|\lambda I - \hat{\beta}| = 0 \tag{40}$$

This is equivalent to requiring the roots of the characteristic polynomial of the AR(p) system to be outside the unit circle - see reference [21].

B Cointegration between Italian and Dutch Gas