Ornstein Uhlenbeck Process Statistical Arbitrage and Cointegration on Equity Markets

by Filip Toth

A Mathematical Exploration for IB Math: Applications & Interpretations HL

1 Brownian Motion and the Wiener Process

Consider a stochastic process, where per each iteration, we add a Δ to the previous value, this Δ is distributed by a normal distribution with a *mu* of zero, and these are independent stochastic events this is called the Wiener process. In discrete time, we can define the Wiener process with the following equation

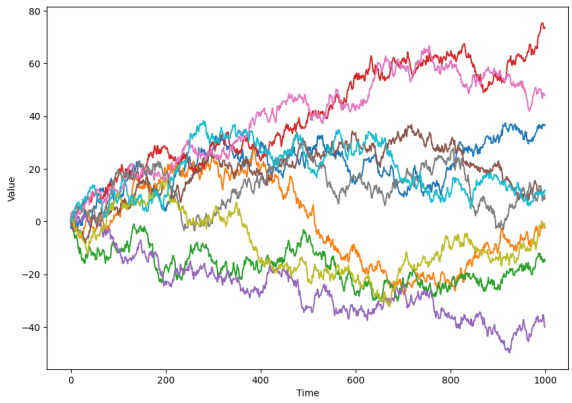
$$W_{t+1} = W_t + \mathcal{N}(0, \sigma^2)$$

Where W_{t+1} is the next value of the process (under discrete time), $\mathcal{N}(0, \sigma^2)$ is a normally distributed drift variable with the variance σ^2.

Intuitively, the expected value, i.e. the limit of this process as it approaches infinity, should be zero, since the normal distribution is symmetric. (maybe prove this? Using a limits on the $\mathbb{E}[W]$ and the normal dist)

The spread of the Wiener process grows at a rate of $\sigma_t = \sigma\sqrt{t}$, due to each timestep adding additional compounded uncertainty.

Here's an example of a Wiener process with $W_0 = 0$ and $\sigma^2 = 1$:



In the diagram, we can clearly observe that the the variance of the Wiener process gets larger as the time progresses.

The wiener process is crucial and has many financial applications, for example, equity prices under the assumption of the efficient market hypothesis follow a purely-stochastic Wiener process. This although disregards stochastic drift, which applies directional pressure on equity

prices, i.e., the Wiener process is market-agnostic, and applies to equity which do not follow the upward trend of the market due to the positive risk-free rate.

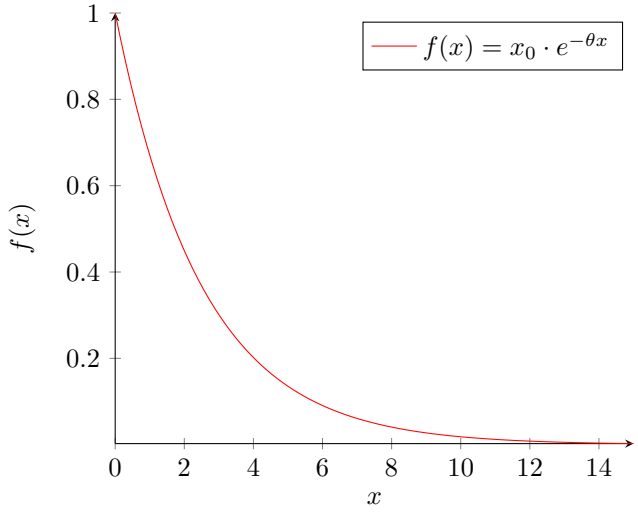
2 Ornstein-Uhlenbeck Process

2.1 Deriving a Mean Reversion Mechanism

But what if we don't want the variance of the process to get larger as time progresses? We call these processes mean-reverting, as in they tend to revert their values back towards the mean μ. This has numerous financial applications (not to mention its applications outside of finance), for example: statstical arbitrage, interest rate models, heston model, and mean reverting equities.

So how do we implement mean reversion? We can start by thinking about what mean reversion actually is? Mean reversion is the tendency of a variable (or usually a stochastic process) to revert back to its long-term mean. If we consider the roots of brownian motion - physics -, we can consider a mean-reverting term to be similar to a force pushing a particle back towards its mean; note that this doesn't happen instantaneuously, it happens over a period of time, where the steadyness of this decay is dictated by the rate of mean-reversion, denoted as θ.

We know that an exponential function with a negative coefficient applied to the exponent will produce a function that will slowly decay.

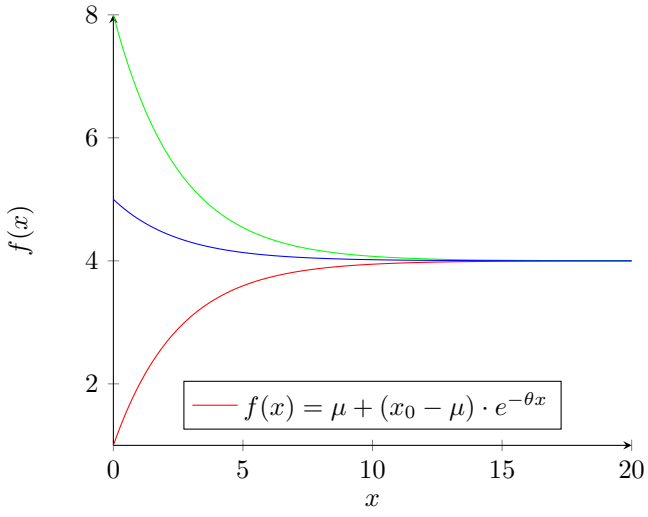


But this is not exactly mean reverting, we need to adjust the function to not only have the

tendency to approach zero, but rather, the function must approach the mean. And the function must be applied from any starting point x_0. Considering that $f(0) = x_0$, we can adjust the function and introduce a base term: $f(x) = x_0 e^{-\theta x}$, this ensures the initial value is x_0, since $e^0 = 1$. Now we have to ensure that the function converges at the mean, not at zero. Initially, we can try to add a the mean as a constant term, such that $f(x) = \mu + x_0 e^{-\theta x}$, but when graphing this function, we notice that we break our first condition - the initial value of the function $f(0)$ is not equal to x_0. We have to adjust for this in the coefficient of the exponential term. We notice that the difference between x_0 and $f(0)$ is equal to the mean, thus modifying the function to the following will solve our problem:

$$f(x) = \mu + (x_0 - \theta)e^{-\theta x}$$

The function produces the following graph, assuming $\mu = 4$, $\theta = 0.4$ and varying the initial value x_0 to be $x_0 \in \{8, 5, -1\}$ for each of the separate lines respectivelly:



2.2 Differentiating for the Mean-reverting Force

We can now think about creating a stochastic process that will take the Wiener process as a basis, but add mean-reverting properties in order to ensure the variance stays within a certain range. Just adding a term for the brownian motion will not be enough, this would be misinterpreting the

mean-reverting effect and would be mostly useless for our purposses, thus we can't do something like:

$$X_t = \mu + (X_0 - \mu) \cdot e^{-\theta t} + \sigma W_t$$

We must now calculate the tug or the actual amount by which the function changes per unit t. This will be necessary for our stochastic process, as they are inherently defined in differential form (in descrete time), where the next value depends on the previous value and we simply add a modifying term, e.g. $M_{t+1} = M_t + \sigma$. We can do this by taking a simple derivative of the function. This derivative can then be thought of as the per-unit t amount of tug applied for each value above or below the mean.

$$\frac{d}{dt}\left(\mu + (X_0 - \mu) \cdot e^{-\theta t}\right)$$
$$\frac{d}{dt}\mu + \frac{d}{dt}\left(X_0 - \mu\right) e^{-\theta t}$$

We know that the derivative of the constant μ is zero, thus we can eliminate the term. We are left with the derivative of a product, for which we can use the product rule of differentiation: $\frac{d}{dt} x \cdot y = x'y + xy'$. In our case, the two terms are $(X_0 - \mu)$ and $e^{-\theta t}$. The first term is a constant that does not chage with respect to t, thus it will be zero. The second term can be differentiated using the chain rule.

We need to define an outer and an inner function in order to apply the chain rule, the inner function will be e^x and the outer function will be $-\theta t$. The $\frac{d}{dx} e^x = e^x$ is a standard differential. And the $\frac{d}{dx} - \theta x$ will simply collapse down to $-\theta$, since it's a constant that is applied to our differentiating variable. The chain rule states:

$$\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$$
$$\frac{d}{dt} e^{-\theta x} = e^{-\theta x} \cdot (-\theta) = -\theta e^{-\theta x}$$

We can now revisit the prudct rule and apply it, we know that the derivative of our first term

is zero, thus we can ignore the first term of the product rule. Only leaving the following: $\frac{d}{dt}x \cdot y = xy'$, we have already calculated $y' = -\theta e^{-\theta x}$ in the previous step. Thus we can Just multiply $-\theta e^{-\theta x}$ with $(X_0 - \mu)$, and we get the following result for our whole derivative:

$$\frac{d}{dt}\left(\mu + (X_0 - \mu) \cdot e^{-\theta t}\right) = (-\theta e^{-\theta x}) \cdot (X_0 - \mu)$$

We can rearrange this to:

$$\frac{d}{dt}\left(\mu + (X_0 - \mu) \cdot e^{-\theta t}\right) = -\theta(X_0 - \mu)e^{-\theta t}$$

This differential desribes the mean-reverting tug applied to our brownian motion model.

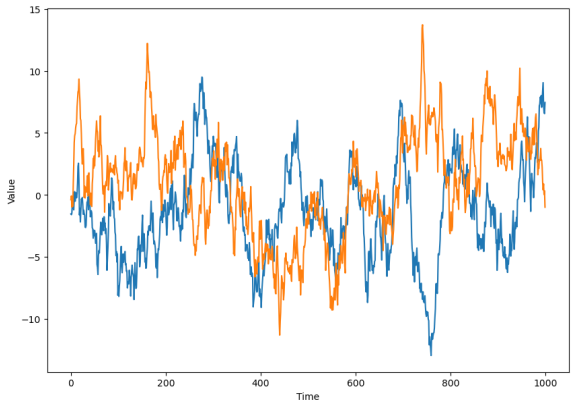
2.3 Full Naive Derivation of the OU Process

We can now try to perform a few experiments on a naively-formulated OU (Ornstein Uhlenbeck) process. We can begin by simply adding a brownian motion to our stochastic process, and assuming (naive step) $X_0 = X_t$, since at each point, the differential will get us tug, this will mean that our t is now zero. We are left with the following stochastic process:

$$X_{t+1} = -\theta(X_t - \mu)e^{-\theta \cdot 0} + W_t$$

$$X_{t+1} = -\theta(X_t - \mu)e + \mathcal{N}(0, \sigma^2)$$

We can not create stochastic paths along the process and they should be mean reverting. A stochastic path (more formally called a sample path of a stochastic process) is simply a possible outcome of running the process, esentially a random path that conforms to the process.



Here we can see that the stochastic process is much more contained within a range, and is thus mean-reverting.

2.4 Formal Derivation of the OU Process

The Ornstein Uhlenbeck process is formally defined as a stochastic differential equation (SDE). An SDE is just a differential equation with a stochastic term.

We begin with adding a brownian motion component. A standard wiener process in its differential form is: $dX_t = dW_t$, this is pretty straight forward if you think about it, the change in X_t is just the change in W_t, this being a stochastic component, dW_t will always be different, and for the sake of simplicity, we can visualize this as if the term follows the normal distribution $dW_t \sim \mathcal{N}(0, \sigma^2)$, this isn't entire true because the variance of the normal distribution itself depends on dt, but we will get to that later.

We now add the drift term (which we derived in a previous subchapter) $\theta(\mu - X_t)$ (we derived it as $-\theta(X_t - \mu)$, but these are functionally equivalent), we can say that this drift happens - continuously - over time t, thus can be written as $\theta(\mu - X_t)dt$, recall that dt here represents an infinitesmal (very small) change in time t, thus the whole term represents the change in the process value X_t as we change time t by an infinitesmal amount and can thus be written as $\frac{dX_t}{dt} = \theta(\mu - X_t)$, but this is just basic differential equations. We end up with:

$$dX_t = \theta(\mu - X_t)dt + dW_t$$

We also add a variable σ to our process, not to be confused with the standard deviation, in our context, θ represents the intensity of the brownian motion (stochastic) component. We end up with the final formal definition of the OU process:

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t$$

2.5 Analytical Solution to the Ornstein-Uhlenbeck SDE

3 Long-term Equilibrium Cointegration in Equity Markets

We perform the **Two Step Engle-Granger Test** for cointegration of time series to get the cointegration coefficient. This method is widely accepted in quantitative finance and econometrics.

But what is cointegration? Cointegration represents the level to which a linear combination of two (or more) time series $X_t = \beta Y_t$ follows a stationary time series $I(0)$. Or in other words, whether the linear combination has a zero long-term mean. We can take any linear combination, such as the linear regression of X_t on Y_t, in the next sub-chapter, we form this regression and add an error term, the error term is esentially formed from a linear combination of the two time series. A high cointegration coefficient means that the two time series move in tandem with each other and that deviations in their spread revert to the long-term mean (since the mean of the spread is zero), divergence of spreads resulting in a zero long-term mean spread is impossible because the function is continuous.

3.1 Static Regression of Equity Time Series

First, we need to determine the residuals $\hat{\epsilon}_t$ (similar to the error) of a regression between the two time series of the prices of the equities we're analyzing. We define a linear regression $Y_t = \alpha + \beta X_t + \epsilon_t$. Where Y_t and X_t are the t-th elements of the time series of the equity prices of X and Y. And where α is the intercept term (which is the mean level of Y_t where X_t is zero) and where β is mean slope coefficient, i.e. the change in the value of Y_t per unit X_t. We define

the following intuitively or from the definition of the linear regression:

$$\beta = \frac{\sum_{i=0}^{n}(X_i - \mu_X)\cdot(Y_i - \mu_Y)}{\sum_{x\in X}(x-\mu_X)^2}$$

$$\alpha = \mu_Y - \beta\cdot\mu_X$$

Now we calculate the residuals $\hat{\epsilon}_t$ using the formula $\hat{\epsilon}_t = Y_t - (\alpha + \beta X_t)$, which we obtain from rearranging the terms of the linear regression equation with resduals. We can interpret these residuals as the difference (for each timestep t) between the actual value X_t and the predicted value - the value of the regression solution for the parameter t.

3.2 Dickey Fuller Test

The Dickey Fuller Test or (DF Test for short) is a statistical test that tests for the stationarity of a time series, or in other words, it tests for if the order of integration of the time series is $I(0)$. The non-augmented version of the Dickey Fuller Test is only applicable for an auto-regressive AR(1) process. Equity spreads are widely considered to be AR(1) processes. We can define an AR(1) time series as $y_t = \alpha + \phi y_{t-1} + \epsilon_t$, where α is the constant shift term, ϕy_{t-1} is the autoregressive term where ϕ is a coefficient denoting how strongly the previous value of the time series affects the current value at time t, and ϵ_t (not to be confused with the residuals from our regression model) denotes a random error term, which IS assumed to be white noise $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Now we can define the hypotheses for the DF Test.

$$H_0 : \phi = 1$$

$$H_1 : \phi < 1$$

Where the null hypothesis denotes that the process is non-stationary and the alternative hypothesis indicates that the process is stationary. We could simply use a t-test to see if a $\hat{\phi}$ estimator is statistically significant, and based on that reject or fail to reject the null hypothesis of the DF Test. But the t-test is inapplicable in models where the central limit theorem doesn't

hold true (because the ϵ values of the process are not independent and identically distributed), thus we need to use the DF test. First, we transform our original time series expression in as a differencing equation. Note that we also rewrite the $y_t - y_{t-1}$ as Δy_t.

$$y_t = \alpha + \phi y_{t-1} + \epsilon_t$$

$$y_t - y_{t-1} = \alpha + \phi y_{t-1} - y_{t-1} + \epsilon_t$$

$$\Delta y_t = \alpha + y_{t-1}(\phi - 1) + \epsilon_t$$

For the sake of simplicity, we introduce δ instead of $(\phi - 1)$. We can simplify our express to $\Delta y_t = \alpha + \delta y_{t-1} + \epsilon_t$, which is the standard form you'll find in most literature on the DF Test. Our hypotheses remain and can be rewritten in terms of δ instead of ϕ.

$$H_0 : \delta = 0$$

$$H_1 : \delta < 0$$

We apply an Ordinary Least Squares linear regression to find the estimator $\hat{\delta}$, this is the p-value for our test and we compare it against our hypotheses. But in our case, we this won't be rigorous enough since we didn't account for autocorrelation (In our case, this means that previous values have some sort of effect on the current value y_t in our time series) properties of financial time series, to remedy this and correctly account for autocorrelation, we use the Augmented Dickey Fuller Test or ADF for short.

3.3 Augmented Dickey Fuller Test

This is an extension of the Dickey Fuller Test that correctly accounts for autocorrelation between consecutive values of the time series and is also extended to the autoregressive AR(p) process; it tests whether a time series of the order of integration $I(p)$ is stationary and for the presence of a unit root (which is a prerequisite of non-stationarity). In the ADF test, our hypotheses stay

the same.

$$H_0 : \delta = 0 \rightarrow \text{non-stationary, unit root}$$

$$H_1 : \delta < 0 \rightarrow \text{stationary, no unit root}$$

We need to add lags to our expression to account for autocorrelation. We can model autocorrelation as some coefficient applied on some previous value of the time series also being accounted in the value of the time series at t. Considering we add two lags to account for autocorrelation between the current value y_t and values y_{t-1} and y_{t-2}, our expression will intuitively become $\Delta y_t = \alpha + y_{t-1}(\phi - 1) + \beta_1(\Delta y_{t-1}) + \beta_2(\Delta y_{t-2}) + \epsilon_t$. Where the betas β_1 are our coefficients onf the lagged differences. We also define p to be the number of lags we include, this value is automatically optimized in various statistical application programming interfaces (APIs) and one can use various Bayesian estimation methods to get the optimal p value, or one can perform a t-test (or an F-test when testing for the significance of multiple variables) and keep adding lagged differences (thus increasing p) until the addition is no longer statistically significant. Our expression can be generalized in the following form:

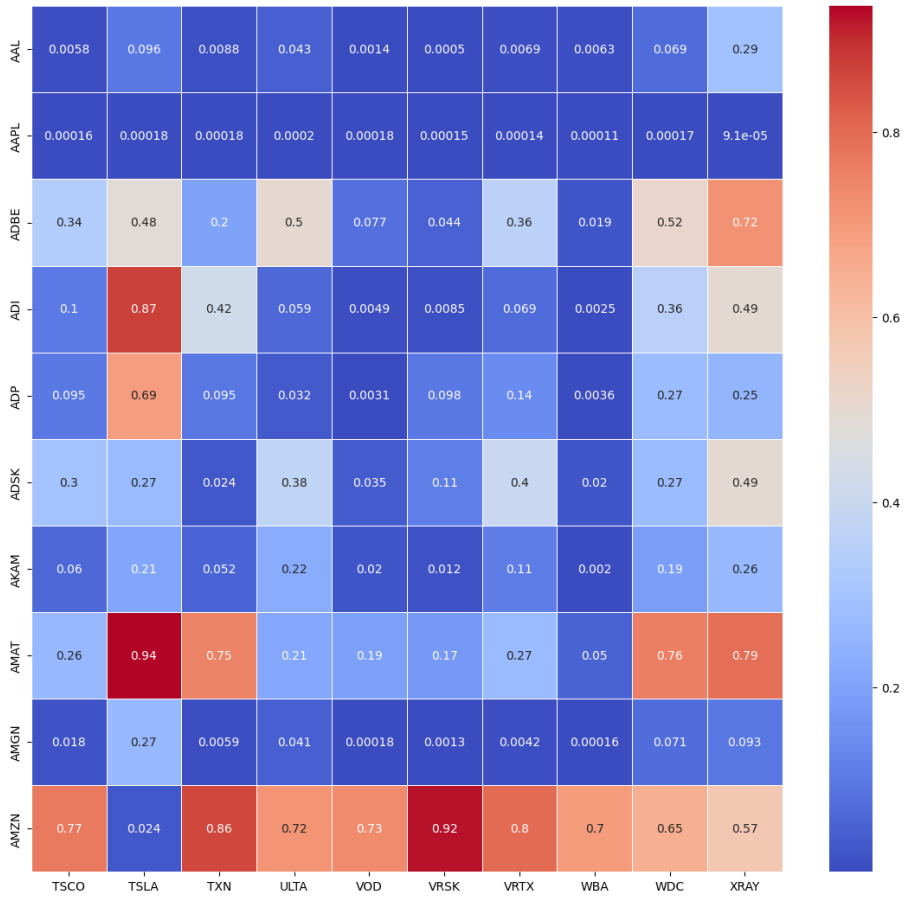
$$\Delta y_t = \alpha\delta y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-1} + \epsilon_t$$

Now that we have this equation, we can do an Ordinary Least Squares regression (in its multiple linear regression form, since we have multiple regressors) to find $\alpha, \delta, \beta_1, \beta_2, \ldots, \beta_p$, we then use the δ value as the final p-value for the Augmented Dickey Fuller Test.

3.4 Performing the Two-Step Engle Granger Test on Equity Prices

We now perform the two-step engle granger test on 78 mature stocks from the selected from NADSAQ-100 index from Janurary 1st 2014 to January 1st 2024. We pair every stock with every other stock resulting in $78^2 = 6084$ pairs, for each of these pairs we perform an OLS regression and get the residual values (as described in subsection 1) and then perform the Augmented

Dickey Fuller Test using the 'statsmodels' library. We extract the p-values of the test and end up with the following heatmap:



Only 100 pairs are displayed as the complete 78x78 heatmap is too large to present here.

We now select 10 of the most cointegrated (lowest p-values) equity pairs to further examine, the ten most cointegrated pairs were, in descending order of cointegration, we observe that most of the equity pairs with very heavy cointegration have AAPL (Apple Inc.) as the base stock, this is interesting as AAPL has heavily outperformed even the growth-heavy NASDAQ, thus this should create a larger spread between most other stocks, suggesting limitations of our test.

Stock 1 Ticker	Stock 2 Ticker	ADF P-Value
AAPL	XRAY	9.104296439449678e-05
AAPL	WBA	0.00011254543577701014
AAPL	VRTX	0.00014280647279306093
AAPL	VRSK	0.00015429781044445213
AMGN	WBA	0.00016469559752946386
AAPL	TSCO	0.00016489122300362627
AAPL	WDC	0.0001652537550668745
AAPL	TSLA	0.00017722866444703876
AMGN	VOD	0.00018081206670114813
AAPL	TXN	0.00018223283943511726

For reference, we can plot the residuals of the initial regression of a select heavily cointegrated pair and a non-cointegrated pair to better visualize the idea of cointegration.

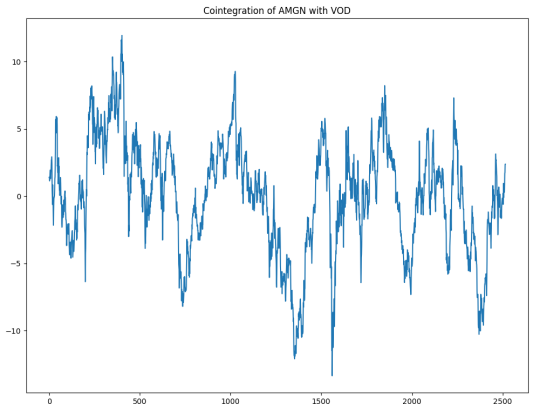


Figure 1: AMGN, VOD, p-value: 0.000181

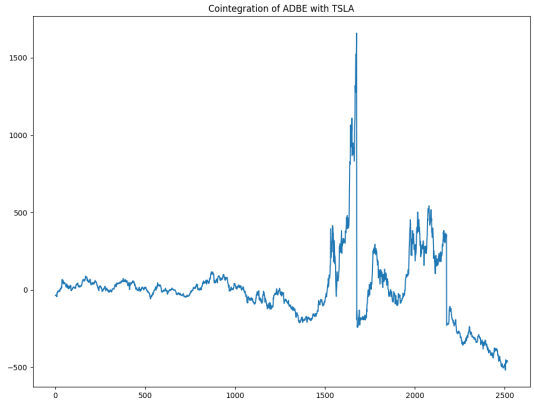


Figure 2: ADBE, TSLA, p-value: 0.484251

3.5 Engle Granger Residuals vs Equity Spreads

If we plot the residuals of AMGN and VOD, we can observe that the resultant time series clearly follows an OU process (or a time-discrete AR(1) time series), as is also indicated by the test we performed, but if we plot the equity spreads (values of stock 1 minues values of stock 2), we

observe a different pattern, indicating that our model is still incomplete and that OLS residuals might not give us the full picture. Let's examine the difference between the graph of the residuals and the graph of the spreads:

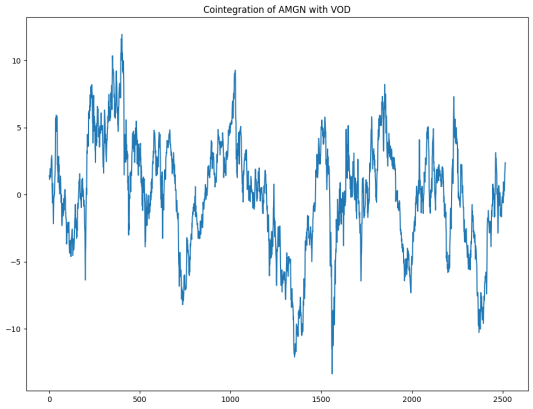


Figure 3: AMGN-VOD cointegration regression residuals

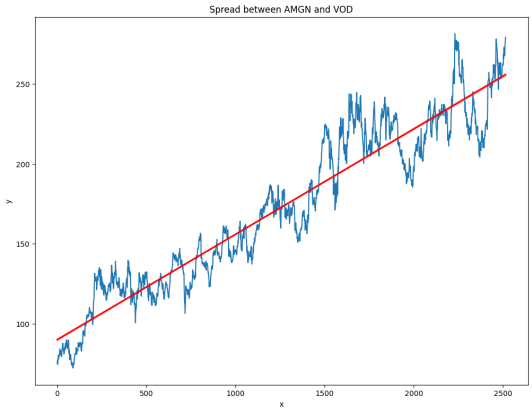


Figure 4: AMGN-VOD spread with line of best fit

We can see that the cointegration esentially disregards deterministic drift in the spreads (deterministic drift is esentially the non-stochastic trend). This makes sense if you think about our test methodology, we defined an OLS linear regression between two stock price time series and then we used to residuals to perform the stationarity test, our graph just shows the residuals at each time step. A residual is just the difference between the what the regression model predicts (using a standard linear equation model $y = \beta x + \alpha$) and the actual value of the time series.

Our Ornstein Uhlenbeck model resembles the residuals and not the actual spreads, it doesn't yet account for deterministic drift.

4 Optimal Execution Region and OU Trading Strategy

We define an Ornstein Uhlenbeck process with the following Stochastic Differential Equation and we have an analytical solution:

$$dX_t = \theta(\mu - X_t)dt - \sigma dW_t$$

$$X_t = Y_t + \mu = \mu + e^{-\theta(t-s)}(X_s - \mu) + \sigma \int_s^t e^{-\theta(t-u)} dW_u$$

5 Ajne Street

Pls hire me any internships pls

6 Derivation of the Gradient for Non-linear Programming

We optimize for the variable θ. We get following base function.

$$X_{t+1} = X_t + -\theta(X_t - \mu)e + \mathcal{N}(0, \sigma^2)$$

We must get the objective function which is thre mean squared error in terms of θ. We consider the the values to be in a pre-computed list $\mathcal{X}$, where $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n \in \mathcal{X}$

$$MSE = \frac{1}{T}\sum_{i=1}^{T} \mathcal{X}_i = \frac{1}{T}\sum_{i=1}^{T} X_i + -\theta(X_t - \mu)e + \mathcal{N}(0, \sigma^2)$$

7 Least Squares Regression for Mean-reverting Parameter

$$\mathcal{X}_{t+1} = \mathcal{X}_t - \theta(X_t - \mu)e + \mathcal{N}(0, \sigma^2)$$

$$\mathcal{X}_{t+1} - \mathcal{X}_t - \mathcal{N}(0, \sigma^2) = -\theta(X_t - \mu)e$$

$$\frac{-\mathcal{X}_{t+1} + \mathcal{X}_t + \mathcal{N}(0, \sigma^2)}{e(X_t - \mu)} = \theta$$

For now, we can disregard the normally-distributed Wiener process term. Have to use Ito's Lemma.