Statistical Arbitrage: Pairs Trading - A Brief Introduction

Don't show my code much because most of it is on my work computer with Bloomberg. Hopefully there are enough examples supplied through my personal computer to help you understand some basic statistical arbitrage trading strategies. Conditional Probabilities using copulas are at the second half of the paper.

By: Ryan Finegan

Pairs Trading Using Cointegration (Price Spread)

Main Topics:

Cointegration: Allows us to determine if we are able to form a mean reverting pair of assets. It is a long term relationship between asset prices (I usually look into a specific long term regime such as a bull economy where the 10-2 spread differential month over month (or Q/Q is increasing) where these assets' movements might be more related. (will go over more later)

Correlation: Short term relationship of returns

Arbitrage: Mispricing - Buy low, sell high

Pairs Trading: Assets sharing underlying factors that affect their movements

Stationary Series: Fixed mean and variance (equilibrium) **[weak-sense stationary]**

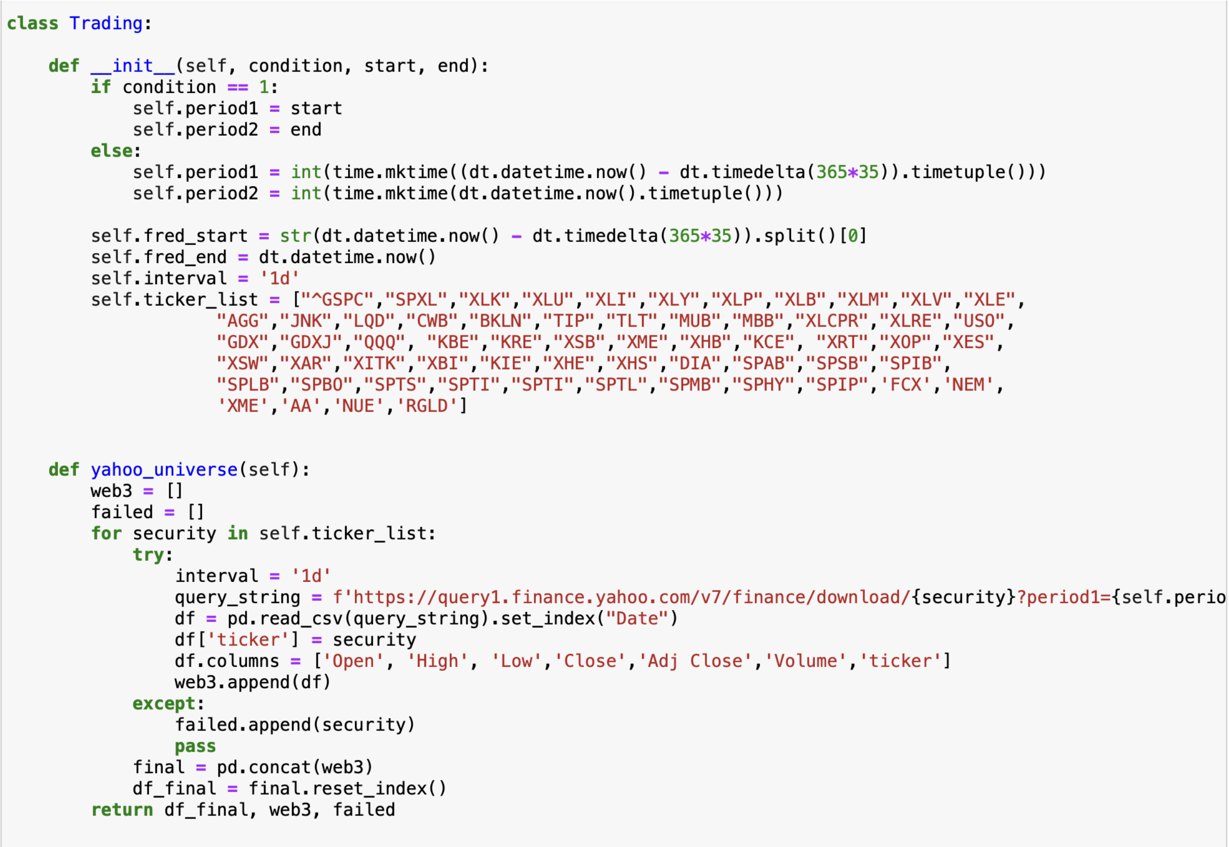
Random Walk vs Stationarity:

Stationarity differs with a random walk as it has a constant mean and variance. Random walks are dependent on time ("drunkard's walk"). You are equally likely to go forward than backwards.

Ultimately we are creating a stationary linear combination between similar moving assets and buying the security that is underpriced and shorting the security that is overpriced.

Caveat, Before Intro:

For ethical reasons, I created scripts for this article and didn't use products I created at work or on my own trading accounts. Here is a brief snippet of the code I used to grab the data. I used python, yahoo finance data, and a SQL database to store all my data. I think this snippet will be enough for someone to replicate the data grab process. If you have a Bloomberg, I can maybe later supply code for that but still grappling with ethical issues since I created most of those scripts while I was working. The tickers below in the ticker list can obviously be changed for your specific security universe.



```python
class Trading:

    def __init__(self, condition, start, end):
        if condition == 1:
            self.period1 = start
            self.period2 = end
        else:
            self.period1 = int(time.mktime((dt.datetime.now() - dt.timedelta(365*35)).timetuple()))
            self.period2 = int(time.mktime(dt.datetime.now().timetuple()))

        self.fred_start = str(dt.datetime.now() - dt.timedelta(365*35)).split()[0]
        self.fred_end = dt.datetime.now()
        self.interval = '1d'
        self.ticker_list = ["^GSPC","SPXL","XLK","XLU","XLI","XLY","XLP","XLB","XLM","XLV","XLE",
                            "AGG","JNK","LQD","CWB","BKLN","TIP","TLT","MUB","MBB","XLCPR","XLRE","USO",
                            "GDX","GDXJ","QQQ", "KBE","KRE","XSB","XME","XHB","KCE", "XRT","XOP","XES",
                            "XSW","XAR","XITK","XBI","KIE","XHE","XHS","DIA","SPAB","SPSB","SPIB",
                            "SPLB","SPBO","SPTS","SPTI","SPTI","SPTL","SPMB","SPHY","SPIP",'FCX','NEM',
                            'XME','AA','NUE','RGLD']


    def yahoo_universe(self):
        web3 = []
        failed = []
        for security in self.ticker_list:
            try:
                interval = '1d'
                query_string = f'https://query1.finance.yahoo.com/v7/finance/download/{security}?period1={self.perio
                df = pd.read_csv(query_string).set_index("Date")
                df['ticker'] = security
                df.columns = ['Open', 'High', 'Low','Close','Adj Close','Volume','ticker']
                web3.append(df)
            except:
                failed.append(security)
                pass
            final = pd.concat(web3)
            df_final = final.reset_index()
        return df_final, web3, failed
```

Brief Intro:

Pairs trading, in its simplest form, is a type of statistical arbitrage where a two security portfolio is created including two similar assets. There is a spread between these two securities and for this strategy to work well, the spread has to be stationary. With this mean reverting spread, profitability is achieved when you short the overpriced security and go long the underpriced security as the spread mean-reverts or converges back to equilibrium.

Buy Low, Sell High:

This strategy is based on the phrase “Buy Low, Sell High”. Trading the spread seeks out mispricings by finding securities that move together or have the same activity in specific market regimes. There are a couple of preliminary steps to find good pairs arbitrage opportunities.

1. Do market research and find ETFs or securities that should be very similar
2. Run a correlation matrix on group of stocks to see similar price movements

3. Run a unit root test on the spread or price ratio to see if the spread is stationary (should do this process every time)
4. Plot the spread (I like using Plotly in Python) and see if there is a specific regime where the spread deviates more from equilibrium (more profitable) or if there is a period where the spread isn't stationary.

Ideally, this could be an easy arbitrage strategy when used correctly. In past experiences, I typically have many pairs that I look at with respect to macroeconomic regimes (10-2 M/M Differences) or market (VIX) / underlying volatility. In specific regimes, these spreads are more likely to be stationary where a market neutral strategy would perform better.

Stationarity:

This is the basis of mean reversion trading strategies. There are two main types of stationarity, strict and weak. Strict is unrealistic and not found in many stochastic processes while weak can be found in many time series.

Strict: The joint distributions have to be the same for all time periods.

Weak: This case is more relaxed where the joint distribution has to exhibit a constant mean, variance, and autocovariance (or have no seasonality). All strict stationarity stochastic processes are weakly stationary.

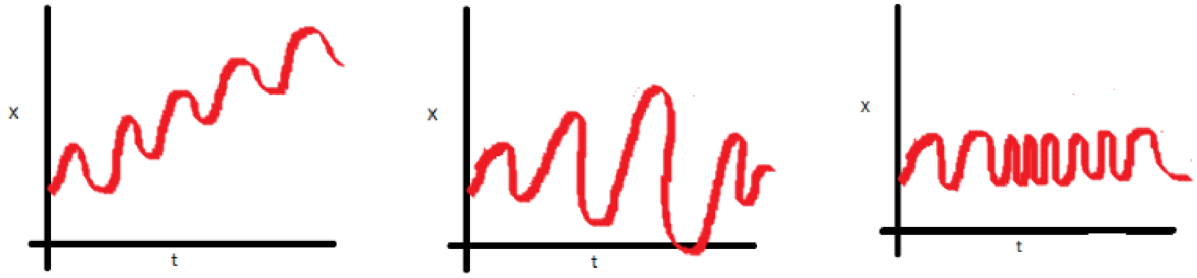
Therefore, stationarity is a stochastic process whose joint distribution doesn't change over time. Any time series is considered a stochastic process because the data points are probabilistic in nature (there is no certainty of future data).

A joint distribution is a combination of more than two distributions. (What's the probability of both time series taking these values again?)

How do you check for stationarity?

Plots:

You can look at a plot and observe the mean, variance, and seasonality of the stochastic process



Picture Credit: Analytics Vidya

None of these time series are weakly stationary. The far left time series doesn't have a constant mean. The middle isn't stationary either because it has a time-dependent variance. The process on the right also doesn't meet the criteria because it has a time-dependent covariance.

Augmented Dickey Fuller Test - A Formal Test:

Without getting too technical, the ADF test basically is a formal, robust test to check if the change in price in the next period (say 10 minute time stamp) is going to be proportional to the difference between the current price and the mean price. If it is not proportional, we fail to reject the null hypothesis and the stochastic process is a random walk. However, if it is proportional to the mean price minus the current price, we can reject the null hypothesis and now can most likely believe that the current price will revert back to the mean price (stationary).

In other words, we are testing the regression coefficient between the next period price change and the mean price and current price differential. If we can reject the hypothesis that this proportion is zero, there is a good likelihood that this price series is stationary.

For example, if the p-value of the ADF test is less than your significance level (0.05 usually - although I've had PM's where the level is 0.10), and the current price is currently lower than the mean, the change in price will most likely be upwards towards equilibrium.

Cointegration:

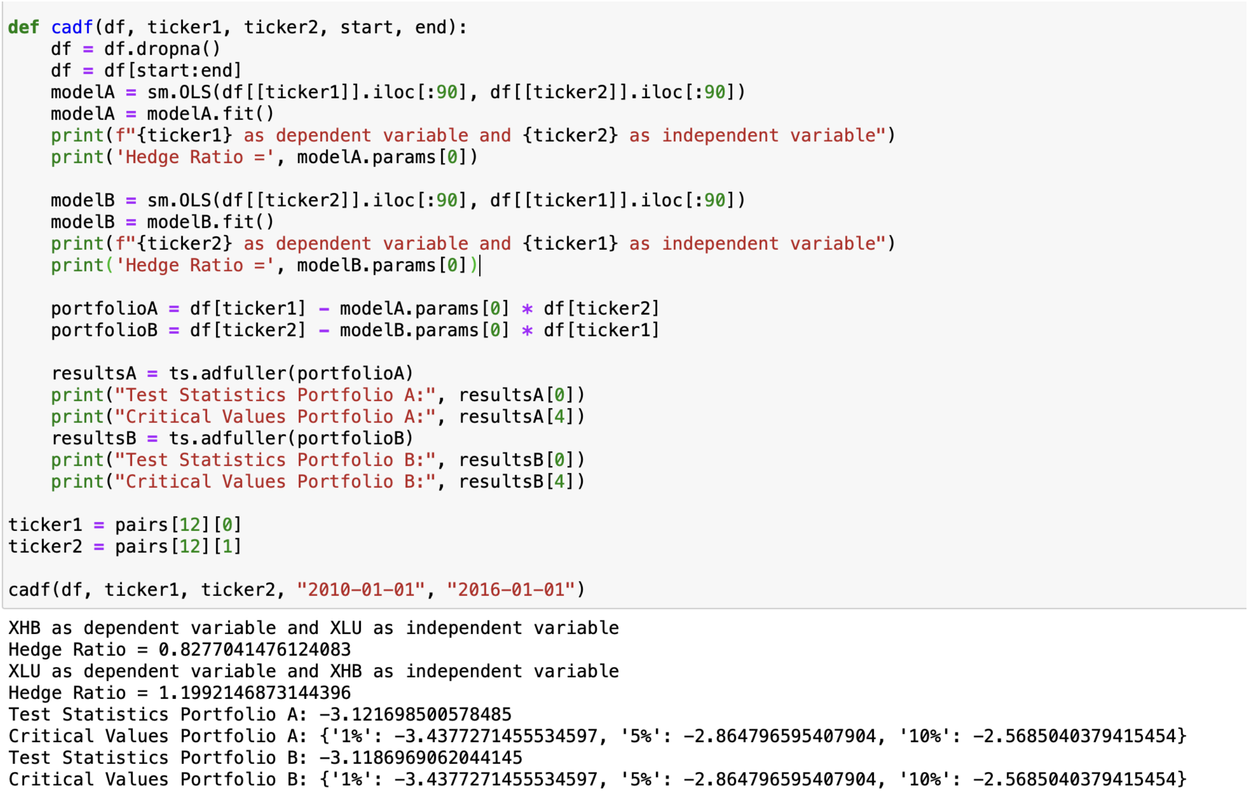
Individual security's price series are usually not stationary. The few that are stationary are rarely tradeable on exchanges, such as the ^VIX Index or other OTC securities like CDX Indices. However, a common practice is holding a portfolio of securities that will result in a stationary spread. If these securities in the portfolio are stationary, they are said to be cointegrated. A CADF test is used to test for this.

CADF Test:

A CADF test is short for a Cointegrated Augmented Dickey Fuller Test. This is the same test as discussed earlier except it is done on the portfolio (the spread). Ideally, we would like to reject the null hypothesis at a 5% alpha. With this rule satisfied, the portfolio is said to be cointegrated.

1. Determine optimal hedge ratio
2. Form portfolio (spread)
3. Run stationarity test on the portfolio (ADF Test)

Example:



```
def cadf(df, ticker1, ticker2, start, end):
    df = df.dropna()
    df = df[start:end]
    modelA = sm.OLS(df[[ticker1]].iloc[:90], df[[ticker2]].iloc[:90])
    modelA = modelA.fit()
    print(f"{ticker1} as dependent variable and {ticker2} as independent variable")
    print('Hedge Ratio =', modelA.params[0])

    modelB = sm.OLS(df[[ticker2]].iloc[:90], df[[ticker1]].iloc[:90])
    modelB = modelB.fit()
    print(f"{ticker2} as dependent variable and {ticker1} as independent variable")
    print('Hedge Ratio =', modelB.params[0])

    portfolioA = df[ticker1] - modelA.params[0] * df[ticker2]
    portfolioB = df[ticker2] - modelB.params[0] * df[ticker1]

    resultsA = ts.adfuller(portfolioA)
    print("Test Statistics Portfolio A:", resultsA[0])
    print("Critical Values Portfolio A:", resultsA[4])
    resultsB = ts.adfuller(portfolioB)
    print("Test Statistics Portfolio B:", resultsB[0])
    print("Critical Values Portfolio B:", resultsB[4])

ticker1 = pairs[12][0]
ticker2 = pairs[12][1]

cadf(df, ticker1, ticker2, "2010-01-01", "2016-01-01")
```

```
XHB as dependent variable and XLU as independent variable
Hedge Ratio = 0.8277041476124083
XLU as dependent variable and XHB as independent variable
Hedge Ratio = 1.1992146873144396
Test Statistics Portfolio A: -3.121698500578485
Critical Values Portfolio A: {'1%': -3.4377271455534597, '5%': -2.864796595407904, '10%': -2.5685040379415454}
Test Statistics Portfolio B: -3.1186969062044145
Critical Values Portfolio B: {'1%': -3.4377271455534597, '5%': -2.864796595407904, '10%': -2.5685040379415454}
```

In this above example, XHB and XLU are cointegrated at the 5% alpha level from 2010 to 2016 using the ADF test. Obviously, the correct methodology would not be to just randomly test different spreads without having a good understanding of the market or securities you are trading within the portfolio. This example was just meant to show the code and results for interpretation. A good process would be to have a hypothesis on which two securities would be cointegrated during which regime and test if that hypothesis produces a stationary portfolio in that specific time period.

Cointegration vs. Correlation:

Correlation is more commonly touched on in probability courses, but it is different from correlation. Many times you will see two securities that are very much correlated in their returns time series but aren't cointegrated when looking at the price series. This is also true the other way around. XLI and XLB show a cointegrated price series, but their returns are relatively not correlated at all.

This can be explained with two main differences between the two. Correlation looks at two return series whether it be 5 minute tick data to monthly data and the main concern is if the two securities are moving in the same direction. Cointegration on the other hand is concerned with two price series and whether they diverge over a long period of time.

Recap:

a. **Cointegration:** long-term and prices
b. **Correlation:** short-term and returns

Steps:

In a simplified example, the three steps for a pairs arbitrage strategy are as follows:

1. Find co-moving securities by cointegration tests
2. Maximize spread / half-life
3. Specify trading rules to optimize profits

Stationarity Applicable to Cointegration Pairs Trading:

Hudson and Thames DefinitivGuidetoPairsTrading is an amazing reference for this:

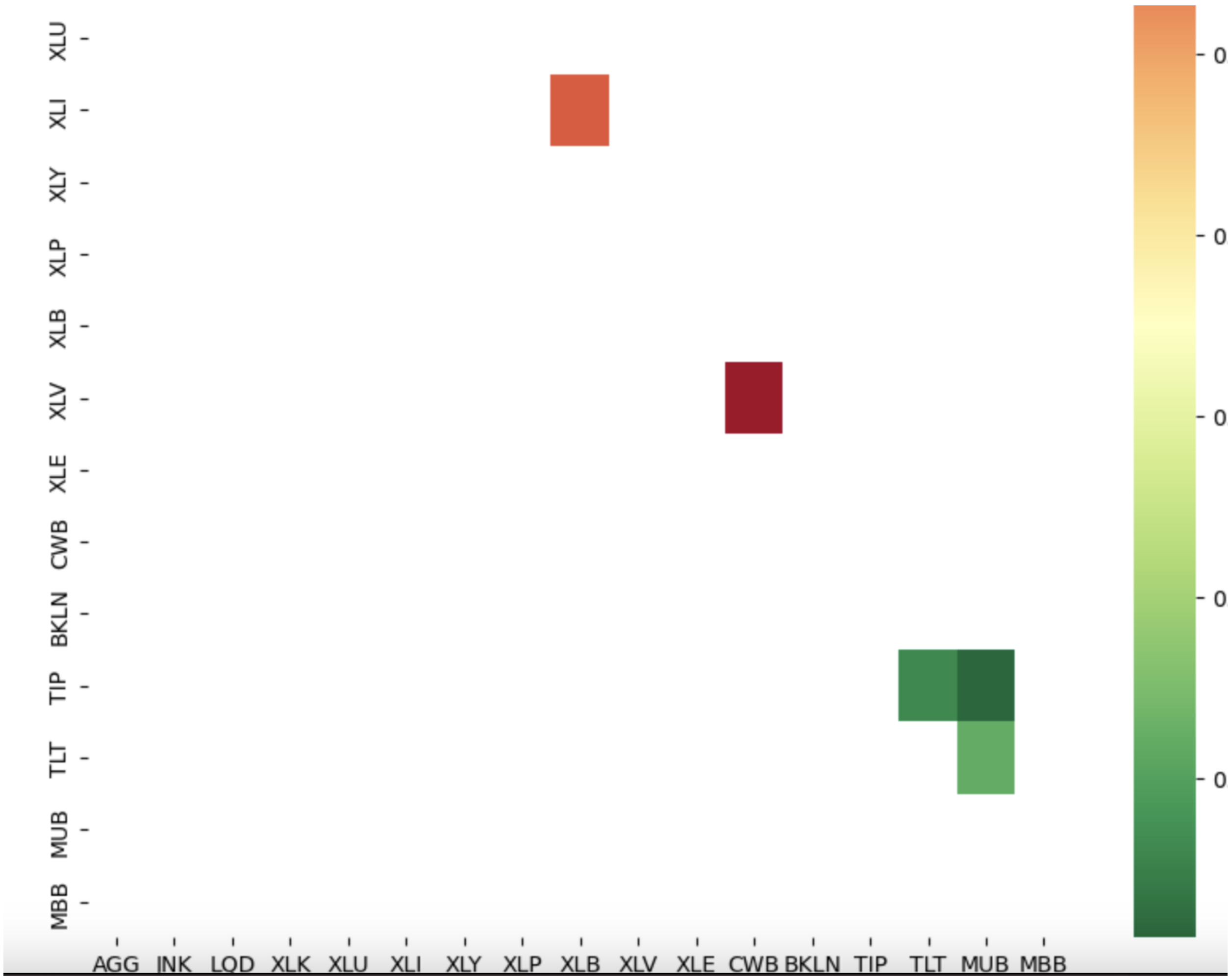
> **Chapter 2:**
>
> "Perform a linear regression between the two asset prices and check if the residual is stationary using the Augmented Dick-Fuller (ADF) test. If the residual is stationary, then the two asset prices are cointegrated. The cointegration coefficient is obtained as the coefficient of the regressor." - Hudson and Thames

You should have to run this twice using the other asset as the dependent variable. Of the two ADF tests, use the one that has the shorter half-life (more on that later) or that is more statistically significant (lower p-value).

Engle-Granger can help find the cointegration coefficient that a cointegrated spread can be constructed.

If we wanted more than two assets, a Johansen test would be run instead.

Cointegrated Pairs



This is the result of my script that finds assets that are heavily cointegrated at a 10% confidence level.

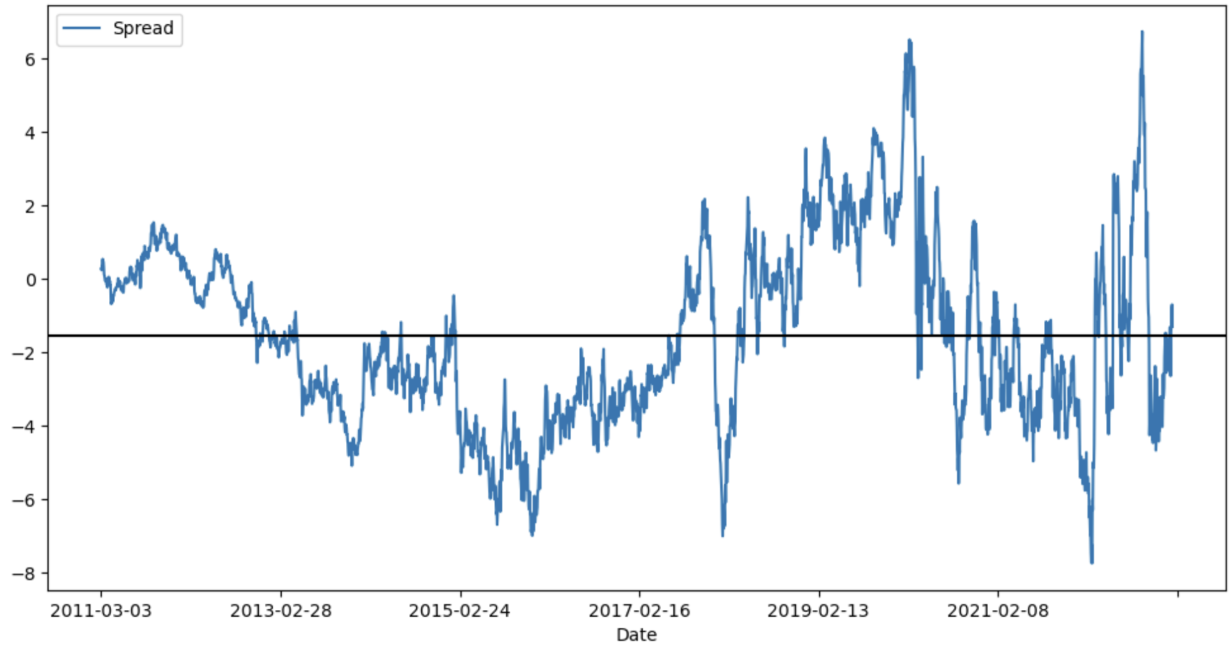
Hedge Ratio:

Number of units of an asset in the portfolio that we should be long or short.

Hedge ratio = XLI - 0.97 XLU

Long one share of SPDR Industrial (XLI) and short 0.97 shares of SPDR Utilities (XLU)

Spread For XLU and XLP



Don’t need to rebalance shares; however, if you use log prices, this will be in market value and rebalancing will need to be touched on

1. Calculate hedge ratio
2. Form portfolio
3. Calculate ADF statistic for the portfolio

Once you have a cointegrated pair, you can trade using the bollinger bands strategy on the optimal spread.

For example, when spread is 0.75 standard deviations from the mean, you go long the spread. [XLI - 0.97 XLU]

When spread is 0.75 standard deviations from the mean, you short the spread.

Exit both positions when the spreads converge back to equilibrium.

Hedge ratio should be computed without look ahead bias (somewhat of a training period)

Intercept of linear regression to compute spread is zero

Dependent = Hedge * Independent + Intercept

XLI = Hedge * XLI + 0 (reduces overfitting)

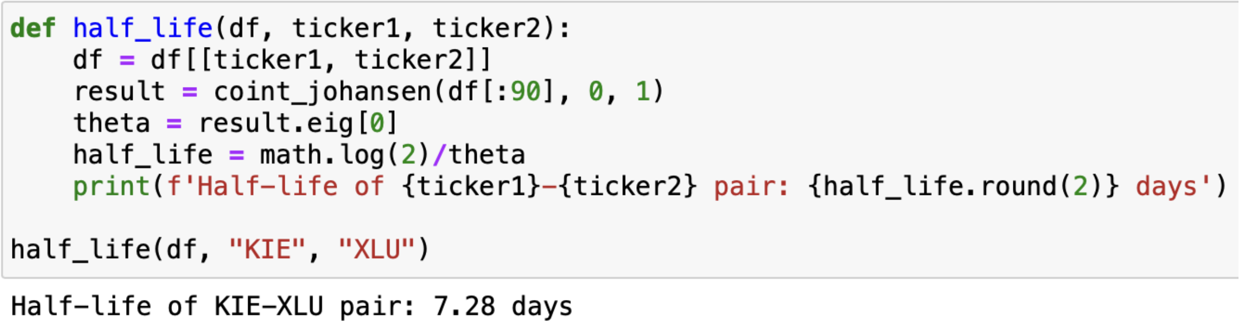
Bollinger Bands:

Bollinger bands are really just an oversimplification of trading rules for these mean reverting strategies. They are used for entering and exiting long and short positions dictated by the spread that was formed through the ADF or Johansen tests. I don't think I will be touching on this much because I feel stationarity is much more important and these bollinger bands can vary depending on the trader and how strict these rules are optimized / created. Bollinger band strategies work great in my experience for my actual scripts in credit portfolio management on assets that behave like Volatility such as the CDX Indices.

Half-Life:

The half-life concept is very similar to the use of bollinger bands. These are varying rules set in place by the specific trader depending on their independent style. The half-life is for actual execution because it is helpful to know when or how often the price series reverts back to equilibrium. Short-term traders might not want a half-life that is too long because that could take away from their alpha.

The half-life basically looks at the expected half life of the trading strategy. The lower the half-life, the more opportunities there's going to be for the trader as the holding period would be much shorter than a portfolio with a longer half-life. Therefore, a shorter half-life is equivalent to saying there are more profitable trading opportunities with this given portfolio. However, the main reason why I will not touch on this in more depth is the relative importance when compared to stationarity tests such as the ADF and Johansen Test. If stationarity is not present within the given portfolio's spread, then there should be no trading strategy in place anyway. I will leave a small function for finding the half-life between pairs, but that is as far as I will go in this article.



```
def half_life(df, ticker1, ticker2):
    df = df[[ticker1, ticker2]]
    result = coint_johansen(df[:90], 0, 1)
    theta = result.eig[0]
    half_life = math.log(2)/theta
    print(f'Half-life of {ticker1}-{ticker2} pair: {half_life.round(2)} days')

half_life(df, "KIE", "XLU")
```

```
Half-life of KIE-XLU pair: 7.28 days
```

More than Two:

Statistical arbitrage pairs trading could work with more than two securities. Instead of a ADF test, the Johansen Test would need to be employed to test for cointegration on the 2 plus securities.

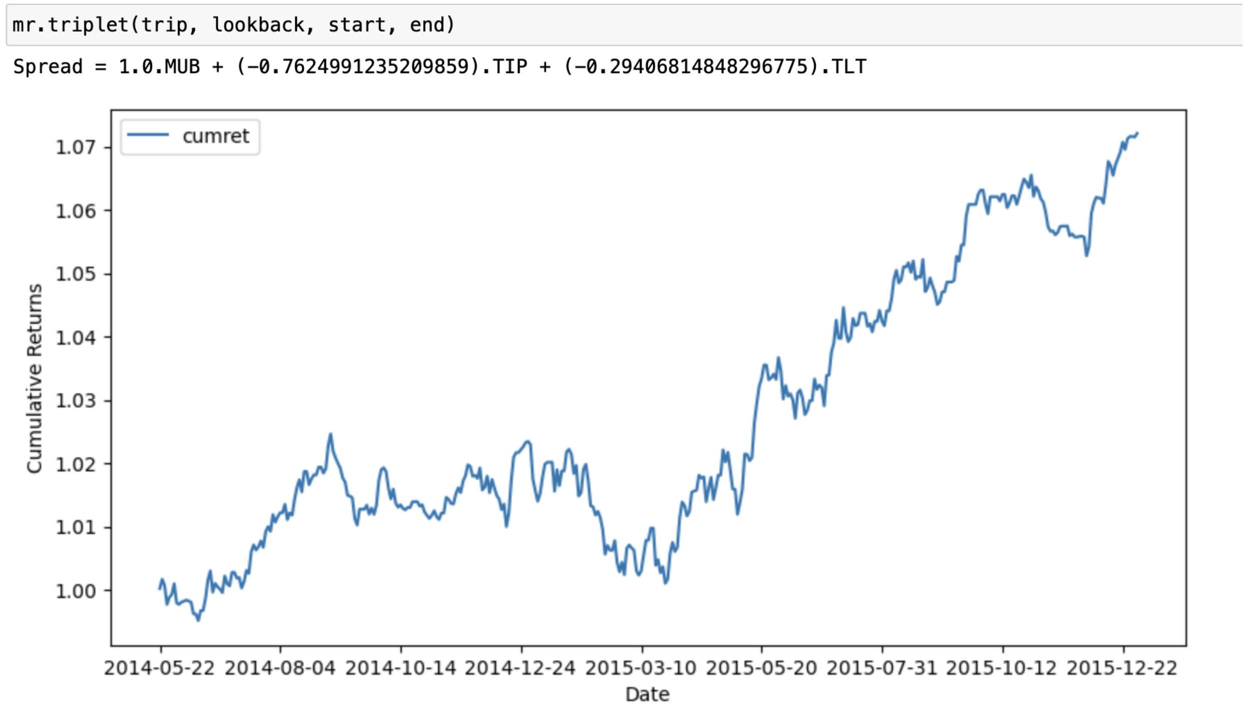
Johansen Test:

The CADF Test is order-dependent. CADF Test different test results when changing the order of the independent and dependent variables. Also, the CADF Test can only test two time series. Autoregressive model. The Vector Error Correction Model is the Johansen Test. Lambda eigenvalue, if all eigenvalues are zero, this linear combination is not stationary. Trace statistics of the Johansen Test allows us to know whether the sum of the eigenvalues is zero. The null hypothesis for this trace statistic can be rejected. The eigen statistics allows us to know how strongly cointegrated the time series are. The null hypothesis can be rejected if the eigen statistic is, in absolute terms, greater than the significance level chosen for critical values.

TRIPLETS STRATEGY EXPLAINED

The triplets trading strategy is the pairs trading strategy except with 3 - 12 assets instead of 2. The main difference here is just using the Johansen test instead of the ADF Test. We don't use the ADF test because it is less efficient to run on multiple assets because as mentioned earlier it is order dependent. Triplets strategy is probably best used on assets that share the same sector or have similar macroeconomic earnings drivers.

Here is an example of the triplets strategy with a 90 day formation period (ran Johansen test on preceding 90 days on the fixed income ETFs - TLT, TIP, and MUB). Results are below along with the hedge ratio is shown below.



```
mr.triplet(trip, lookback, start, end)
```

```
Spread = 1.0.MUB + (-0.7624991235209859).TIP + (-0.29406814848296775).TLT
```

Caveat: Transaction costs (commission costs and slippage) and the understanding of the securities you are trading like their drivers as well as the macro drivers that can affect their specific sector

Also, make sure cointegration doesn't break down while you are trading these pairs / groups. Check frequently and stop trading that group when it does break down. You can also employ layering which I won't go into since this is just meant to be an introductory article about cointegration pairs trading.

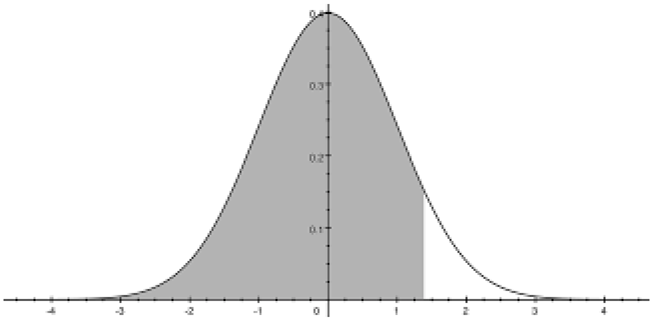
—---

Pairs Trading: Copulas Log Returns Trading Strategy

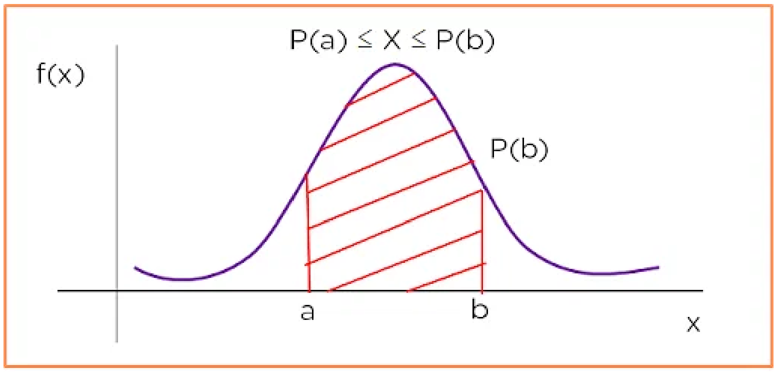
This is similar to the pairs trading cointegration strategy. This is particularly new math implemented in trading but it has been a popular tool in finance dating back to the Great Financial Crisis where quants were using Gaussian Copulas to price securities like CDOs.

Prior Math to Help Understand Copulas

Probability Density Functions - how likely something is to occur at different points in that distribution. [Normal Distribution with mean (mu = 0) and standard deviation (sigma) shown below]

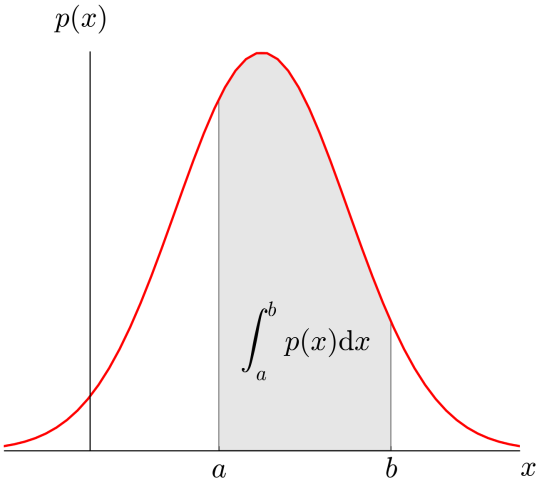


P(X<=1.4) = above picture (just an integral bounded by 1.4 and negative infinity.



Probability Density Function

This PDF above can be written as P(a<=X<=b).



You can find the probability of this happening by doing an integral with the formula above. You find the area under the curve between point B and point A. Probability Density Functions are just the derivative of cumulative density functions. I will come back to probability density functions.

Cumulative Density Functions - The cumulative density function is just the integral of the probability density function. These are important when talking about copulas because the inputs to the copula need to be standardized. To do this, you flatten or stretch your PDF while keeping the order the same while still making the distribution uniform. It is important to note that the copula joins the marginal distributions based on order and not value (Kendall’s Tau instead of Pearson’s Rho).

Probability Integral Transform - This is very important for copulas because you will need to get each marginal (univariate distribution) to be uniform [0,1] for it to be coupled into a joint distribution. Marginal Distributions => Uniform Distributions. This is also called the CDF Transform.

I won't go into the math too much however it is great to know. I will link three videos that are really great in explaining these topics.

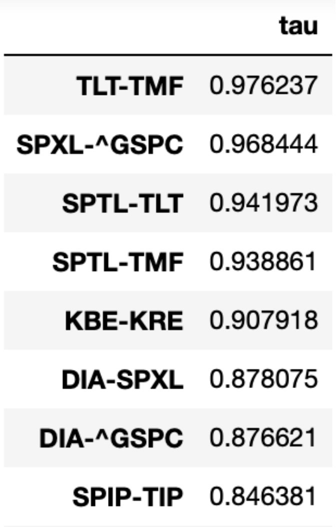
https://www.youtube.com/watch?v=b6aZJuwE3Cs (Intro By Paul Sweeting) [A]

https://www.youtube.com/watch?v=gzUxg0OUHU4 (Part 2 By Paul Sweeting) [B]

https://www.youtube.com/watch?v=WFEzkoK7tsE (Dirty Quant) [C]

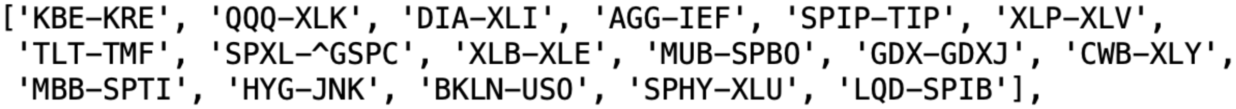
Trading Using Copulas:

1. **Select pairs using Kendall's Tau correlation coefficient. I used Kendall's Tau instead of Pearson's Rho because of the stress on how its calculated on order instead of value (Tau is more stable especially when we have to do the probability integral transform). Rank correlation is better in this case because it is calculated on order and not value. This doesn't change when the distribution is changed because order is not the same (Paul Sweeting's Part 2 Video (B).**
    a. Used 2015-01-01 to 2015-12-31 as my training or formation period to find the most correlated assets by their daily log returns.
    b. Screened and sorted the top 50 Tau correlated returns to be used on step two.
    c. My original universe consisted of GICS sector ETFs, popular fixed income ETFs and popular factor ETFs. I used yahoo finance and my SQL database to extract all the data. In future posts, I will use the BQNT Bloomberg API just because the data is better and easier to obtain.



	tau
TLT-TMF	0.976237
SPXL-^GSPC	0.968444
SPTL-TLT	0.941973
SPTL-TMF	0.938861
KBE-KRE	0.907918
DIA-SPXL	0.878075
DIA-^GSPC	0.876621
SPIP-TIP	0.846381

2. **Identify marginal distributions relative to the log returns of each stock pair selected in step 1.**
    a. I will use the 2015-01-01 to 2015-12-31 log returns of each stock found in the top 50 tau correlated pairs and fit them to the student t marginal distribution. I used the Student t because I believe security returns are usually not normal and usually have fatter tails. I used the Kolmogorov-Smirnov Test to see if these distributions fit the data well. I removed all the tickers who failed the KS Test (had p-values lower than 0.1 meaning that the selected parametric distribution (Student T) for the security wasn't a good fit. Only three securities, TYD (3x Bull 7-10 Year ETF), SPSB (SPDR's Short Term Corporate Bond ETF), and SPMB (SPDR's Mortgage Backed Bond), didn't fit its chosen distribution well, so they were removed. These are the new pairs that were highly correlated and fit student t distributions.



```
['KBE-KRE', 'QQQ-XLK', 'DIA-XLI', 'AGG-IEF', 'SPIP-TIP', 'XLP-XLV',
 'TLT-TMF', 'SPXL-^GSPC', 'XLB-XLE', 'MUB-SPBO', 'GDX-GDXJ', 'CWB-XLY',
 'MBB-SPTI', 'HYG-JNK', 'BKLN-USO', 'SPHY-XLU', 'LQD-SPIB'],
```

3. **Probability Integral Transform is used on the PDFs of the marginal returns to make the marginal returns uniform (0,1). You transform the Probability Density Function into the Cumulative Density Function to make the distribution uniform (or standardized).**
    a. You do this so copulas can be fit to the marginal distributions.
4. **Fit Copulas on Transformed Marginal Returns**
    a. I only used a few Archimedean Copulas to fit the marginal distributions
    b. I used Clayton, Gumbel, Joe, and Frank
    c. The optimal copula was selected to decouple the marginal distributions and it was selected using the lowest AIC according to the log likelihood.
5. **Use the fitted copula to calculate the conditional probabilities**
    a. The U and V are the random variables and the lowercase u and v's are the observed data. If this conditional probability is less than the 0.5 quantile (average), then the security is underpriced.

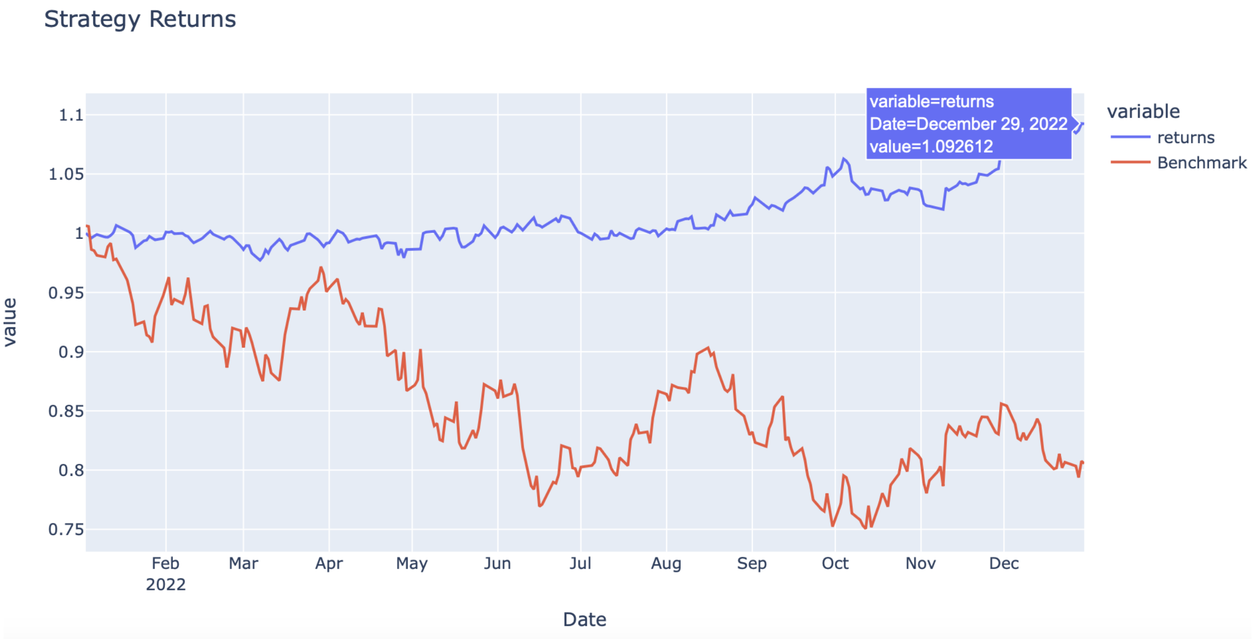
$$P(U \leq u|V = v) = \frac{\partial}{\partial v}C(u, v)$$

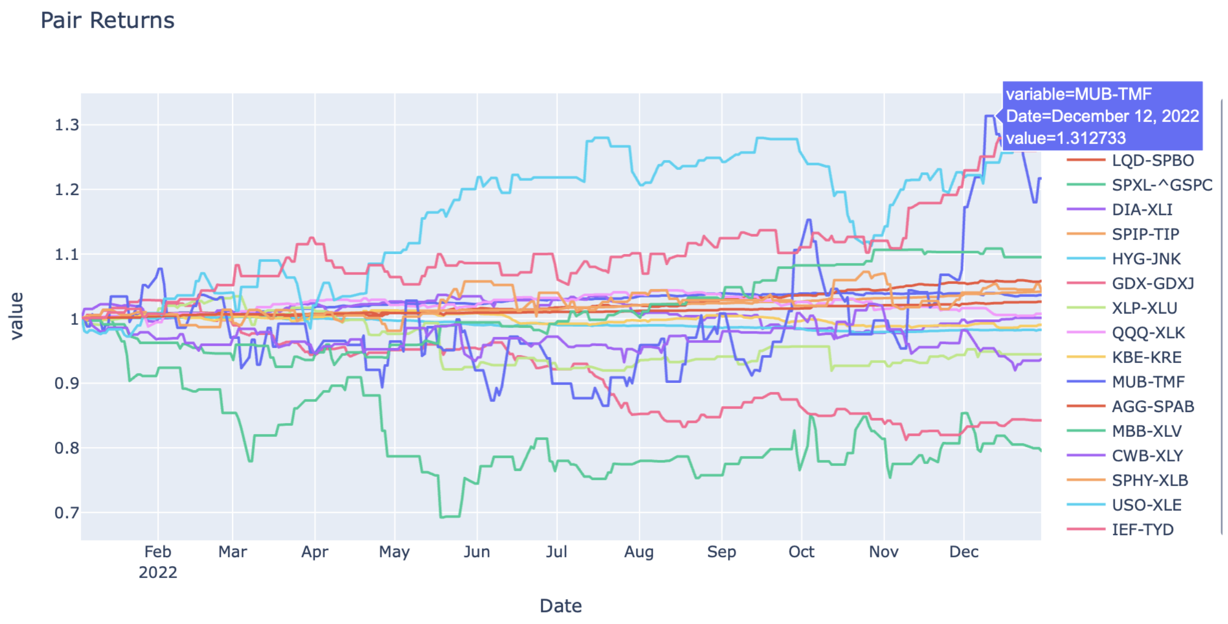
$$P(V \leq v|U = u) = \frac{\partial}{\partial u}C(u, v)$$

6. **Trading rules are based on these conditional probabilities**
    a. In most research papers I read, the logic or trading rules were quite similar and systematic. I used a different way of thinking and will try my best to describe. In

the past, most researchers set up a systematic rule where both marginal probabilities would simultaneously have to be met for a position to be made. For example, Security A would have to have a conditional probability over 95% meaning it was overpriced and Security B would have to be under 5% meaning it was underpriced. Both of these would have to be present at the end of the same trading day or hour / minute for a long and short position to be made. However, given I am less systematic, I found the conditional probabilities useful for seeing when a security was underpriced and overpriced without having a specific confidence band rule. It was also nice to create a historical excel sheet using Python's pandas to see the security's price movements day to day given the conditional probability at that same time step. The backtest that I will show using Plotly had the rules of 70% for overpriced and 30% for underpriced.

Results For Most Recent Year:





Caveat: Although I showed my results using a systematic strategy, I do not use any of these strategies as I am not a systematic investor / trader. Also, these strategies didn't account for slippage or transaction costs.

My Takeaway:

What is nice about these, especially the conditional probabilities presented by the copulas, is to see when assets are underpriced relative to securities they are historically dependent on. This pairs trading strategy that I learned by using links I supplied throughout this paper is a great resource for me as I use pandas to create a spreadsheet of tabular data with the conditional probabilities of the pairs of log returns to be used as analytics when I see an asset that is historically undervalued. I usually pair it with my top down strategy looking at assets with High Betas during good economic periods and Low Betas or defensive assets / sectors when the macro is not looking too positive.

Two great resources for macro investing are Keith McCoulough's Hedgeye and KeyYong Park's paper on a Sector Rotation Model where he talks about the Composite Macroeconomic Indicator. I will link those and hopefully they are easy to find as they were and still are great learning resources for me.

I hope you found this helpful. The code I showed is just meant to help with my process on getting the data when I don't have access to my work's Bloomberg data through BQNT. When I do look at stat arb and mean reversion pairs trading on my own time, I use a combination of OOP scripts in Python and R. I would be more than happy to discuss this topic more and show my code through my personal python scripts. Just reach out to me here:

Ryanfinegan5@gmail.com