Empirical Study of Statistical Arbitrage Chance in U.S Equity Market

——Take pairs trading as an example

- Name: Yang Xu (UNI: YX2378)
- Name: Lanting Jiao (UNI: LJ2403)
- Name: Jocelyn Ma (UNI: JM4407)
- Name: Chengcheng Yuan (UNI: CY2434)
- Name: Mengya Zhao (UNI: MZ2593)

Abstract

Statistical Arbitrage was firstly discovered by Nunzio Tartaglia led by Morgan Stanley's quantitative analysis team in the 1980s. It has been widely used in the financial market in foreign countries since it was put forward. Statistical Arbitrage is part of quantitative analysis, the essential principles can be concluded as theorems or lemma and easy to utilize to the real world. Therefore, the strategies has become a very well welcomed way that utilizing theory of statistics to make trading strategies and win great profits. In the more mature markets, this strategy is being used by hedge funds, mutual funds, investment companies and senior independent investors.

In this report, the main research methods are quantitative analysis and empirical verification. This paper analyzes the mathematical methods behind the statistical arbitrage, especially the matching trading strategy, and the reasons makes the quantitative analyzes effectively. Meanwhile, by learning the study of existing research results and conclusions, gain a better understanding about the grasp of U.S domestic stock market and its historical data trending. Looking forward to find potential pairing of the stock pair ultimately. The process in whole involves data selection, industry screening, time series process, correlation testing, and stability testing. While stability guarantees the arbitrage opportunity, we can use the nature of stability and make it applicable to the stock spread arbitrage.

1. Introduction

Talking about the trading principle more specifically in the stock market, it's more about the principle of profit acquisition which essentially based on market volatility and personal expectations. People buy undervalued stock and sell overvalued stock to earn the difference as a profit. And the trading principle of the Statistical Arbitrage is based on a pair of highly correlated stocks or assets. So when the stock price occur deviation, it can be concluded that in the future such a deviation will be corrected and the price will converge and then arbitrage opportunities appear. Compared to the deterministic arbitrage, Statistical arbitrage uses the quantitative analysis methods to make strategies rather than simply waiting for the opportunity to speculate. Among some mature markets in western countries, statistical arbitrage strategies have been successfully used and proved with rich profits. When used properly, this strategy can provide investors with a large amount of low risk benefits. However, in U.S markets this strategy is as popular or as known as that. Thus, along with the development and perfection of U.S equity market, researches of statistical arbitrage strategy and its application may provide participants in the market more chances.

2. Methodology

The basic idea of this paper is to apply cointegration theory and method to find the long-term properties of two stocks so that they can be used for pairs trading. The concept of cointegration is introduced as follows.

Definition Cointegration is a statistical property of a collection $(X_1, X_2, \cdots, X_k)$ of time series variables. If the series satisfy the following two conditions:

1. All of the series must be integrated of order one
2. A linear combination of this collection is integrated of order zero

Then this collection is said to be co-integrated.

There are three widely used methods for testing cointegration: Engle-Granger two-step method, Johansen test and Phillips-Ouliaris cointegration test. In this paper, we mainly use the Engle-Granger two-step method and the theory of this method is the following:

Engle-Granger two-step method If $X_t, Y_t \sim I(1)$ and X_t, Y_t are cointegrated, then there exists a constant β such that $Z_t = Y_t - \beta X_t$ is stationary, where $I(1)$ means integration of order 1.

From the idea of Engle-Granger two-step method, we could use ordinary least square regression to find the parameter β and run the stationarity test on Z_t series. Here we use Augmented Dickey

Fuller Test (ADF) to test stationarity. ADF test is a particular kind of unit root test and the hypotheses of the test is the following:

H_0: There is a unit root H_1: The time series is stationary

Using R function adf.test(), we are able to perform the ADF Test. And if the p-value given is less than 0.05, then we will reject the null hypothesis and conclude that the time series is stationary.

In conclusion, our methods of finding the stock pairs used for arbitrage can be summarized to four steps:

1. Test if series X_t and Y_t are stationary, where X_t and Y_t are the logarithm of two stock prices
2. Calculate the first order difference of X_t and Y_t, denoted by ΔX_t and ΔY_t, and test if ΔX_t and ΔY_t are stationary
3. Regress Y_t on X_t using ordinary least squares method and test if the residual is stationary
4. If ΔX_t and ΔY_t are non-stationary, ΔX_t , ΔY_t and residuals of OLS regression are stationary, then X_t and Y_t are cointegrated and can be used for pairs trading strategy

3. Data Selection

The scope of our empirical analysis are publicly traded companies in U.S stock market. We are aim to find the stock pairs which could meet the arbitrage premises within this range. The analyzing and testing processes are made according to the available historical data. These steps help us to gain understanding about the relationship between those pairs and develop strategies.

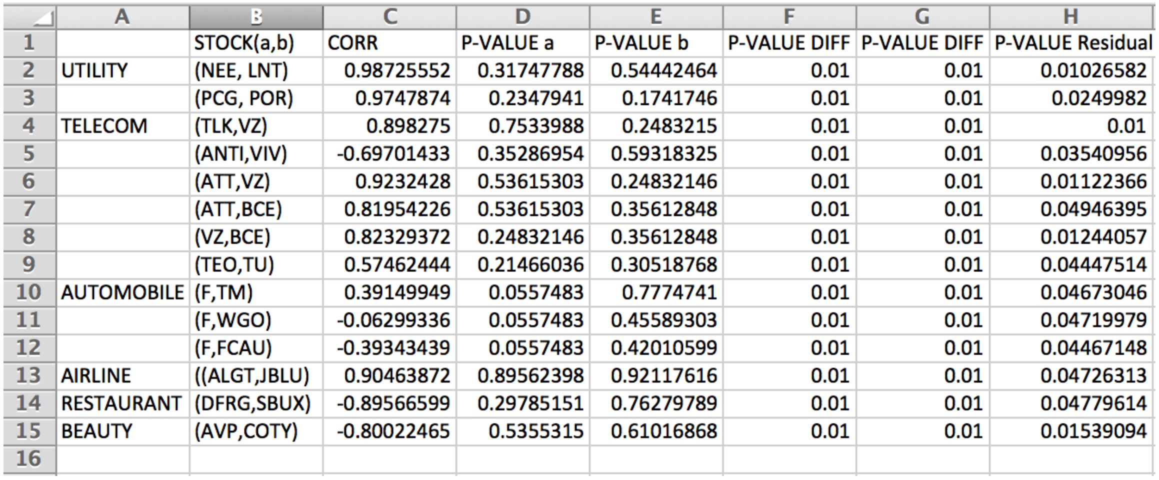
Since there are almost 4000 stocks are actively traded in NYSE or Nasdaq, and another 15,000 are traded over the counter, not in the major exchange. So, if using the pair trading strategy and testing all possible relevances about the stocks, we need to do the combination calculation. It's about 2 choose from 4000 such times pairing needed. This is too computationally expensive. In order to narrow down the searching range and saving energy, the strategy we used here is choosing the stocks within the same industries. From theoretical perspectives, due to the resource sharing, similar cost, the same market environment and some other factors, we believe the stocks from the same industry sector may likely to have cointegration properties. In order words, it means we will have a better chance and easier to take the pair trading strategy.

After decided the range of the raw data, the next step is the industry screening. We attempt to screening 11 industries in the first stage, including Real Estate, Airline, Beverage, Automobile, Public Utility, Beauty, Telecommunication etc, which including around 200 companies in total and all of the data are extracted from Yahoo Finance.

Within all these options, the stocks from Telecommunication field have an outstanding performance of our model. There are forty-five pair trails generated from ten companies in this industry. Besides, the industry is a burning topic in the business recently as the new products and services in the marketplace opens up some new vertical segments for the industry. Based on all these reasons, we decided to use this industry as a target sector and applied the related companies' stocks to our modeling. The data from this industry are also used for the back-testing part, at the very end of our project.

4. Application and results

After many trials, we found 14 pairs of stocks from 6 different industries successfully went through the test method. The 6 industries concluding Utility, Telecom, Automobile, Airline, Restaurant, and Beauty. We found most of our successful pairs are from Telecom Industry, including TLK, VZ, ANTI, VIV, ATT, BCE, TEO and TU stocks. Therefore, we choose Telecom Industry to make further test.

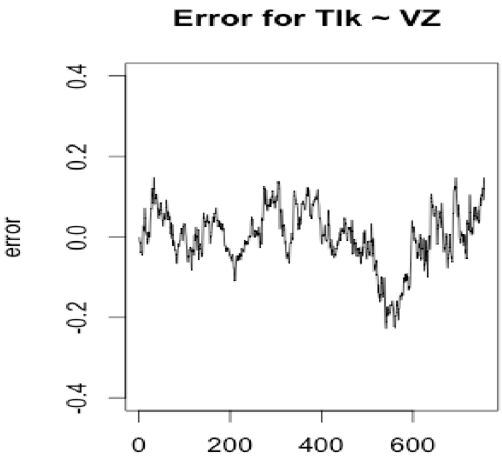


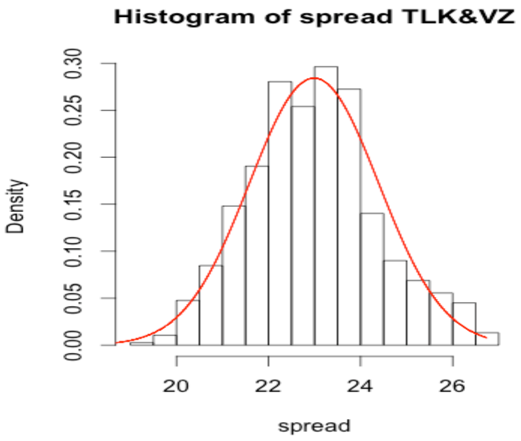
	A	B	C	D	E	F	G	H
1		STOCK(a,b)	CORR	P-VALUE a	P-VALUE b	P-VALUE DIFF	P-VALUE DIFF	P-VALUE Residual
2	UTILITY	(NEE, LNT)	0.98725552	0.31747788	0.54442464	0.01	0.01	0.01026582
3		(PCG, POR)	0.9747874	0.2347941	0.1741746	0.01	0.01	0.0249982
4	TELECOM	(TLK,VZ)	0.898275	0.7533988	0.2483215	0.01	0.01	0.01
5		(ANTI,VIV)	-0.69701433	0.35286954	0.59318325	0.01	0.01	0.03540956
6		(ATT,VZ)	0.9232428	0.53615303	0.24832146	0.01	0.01	0.01122366
7		(ATT,BCE)	0.81954226	0.53615303	0.35612848	0.01	0.01	0.04946395
8		(VZ,BCE)	0.82329372	0.24832146	0.35612848	0.01	0.01	0.01244057
9		(TEO,TU)	0.57462444	0.21466036	0.30518768	0.01	0.01	0.04447514
10	AUTOMOBILE	(F,TM)	0.39149949	0.0557483	0.7774741	0.01	0.01	0.04673046
11		(F,WGO)	-0.06299336	0.0557483	0.45589303	0.01	0.01	0.04719979
12		(F,FCAU)	-0.39343439	0.0557483	0.42010599	0.01	0.01	0.04467148
13	AIRLINE	((ALGT,JBLU)	0.90463872	0.89562398	0.92117616	0.01	0.01	0.04726313
14	RESTAURANT	(DFRG,SBUX)	-0.89566599	0.29785151	0.76279789	0.01	0.01	0.04779614
15	BEAUTY	(AVP,COTY)	-0.80022465	0.5355315	0.61016868	0.01	0.01	0.01539094
16								

The chart above indicates that the highest correlation occurs between ATT and VZ, which is 0.923, with ATT p-value 0.536 and VZ p-value 0.248. Both of the DIFF p-values for ATT and VA are 0.01, less than 0.05, which indicates their distribution is stationary. The residual p-value is 0.011, which is the smallest among all 6 pairs.

Literally speaking, there are two methods to compare stationary for different pairs of data, diagrammatic method and unit root test, while diagrammatic method gives us a more direct view of the data over the certain interval. By using diagrammatic method, we figure out that the pair (ATT, VZ), which is the pair between AT&T and Verizon Communications Inc., performs the best among all 6 pairs of stocks, since it has both the most stationary distribution and a high correlation.

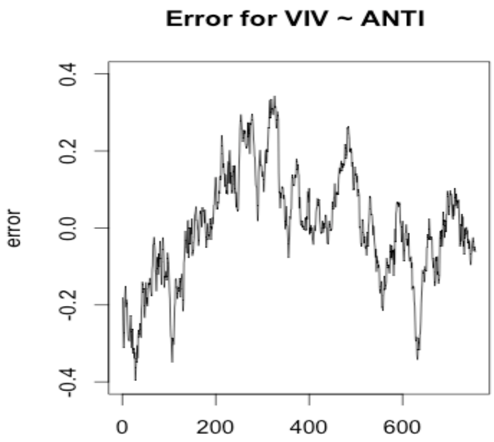
(TLK,VZ)

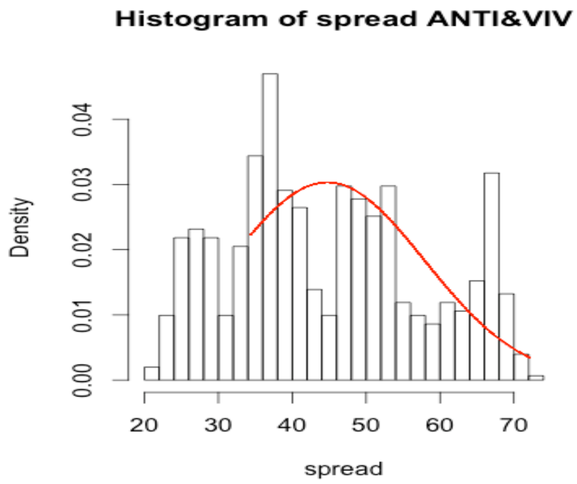




TLK~VZ error follows a stationary distribution. From the graph of TLK~VZ spread histogram, we have the relationship between TLK and VZ normal.

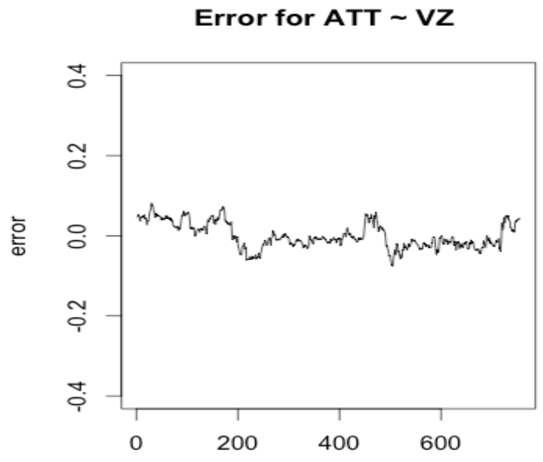
(ANTI, VIV)

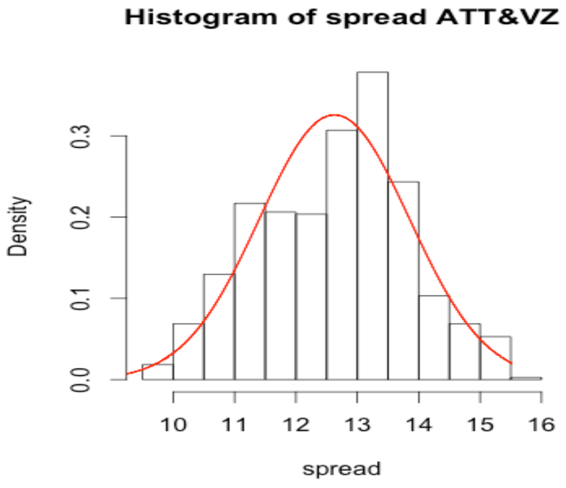




ANTI~VIV error does not follow a stationary distribution. From the graph of ANTI~VIV spread histogram, we have the relationship between TLK and VZ roughly normal.

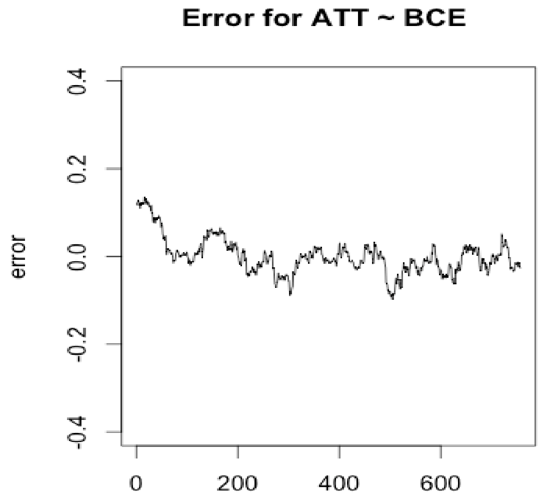
(ATT, VZ)

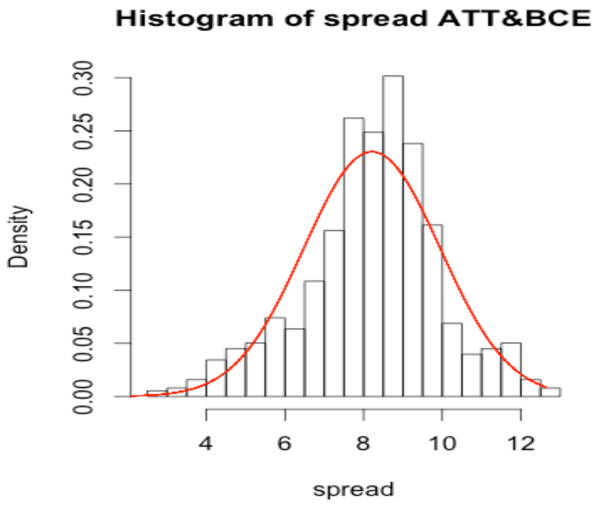




ATT~VZ error follows a stationary distribution. From the graph of ANTI~VZ spread histogram, we have the relationship between TLK and VZ nearly normal.

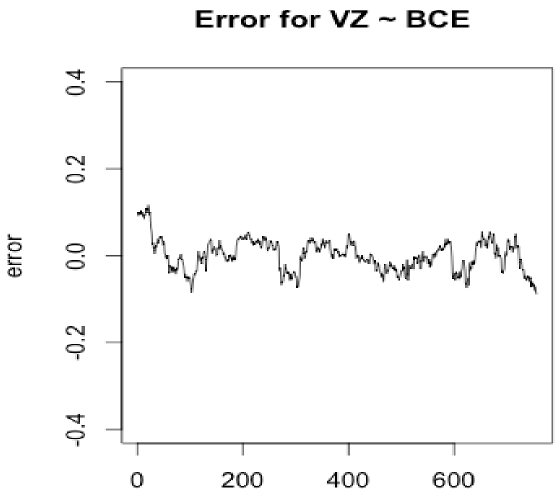
(ATT, BCE)

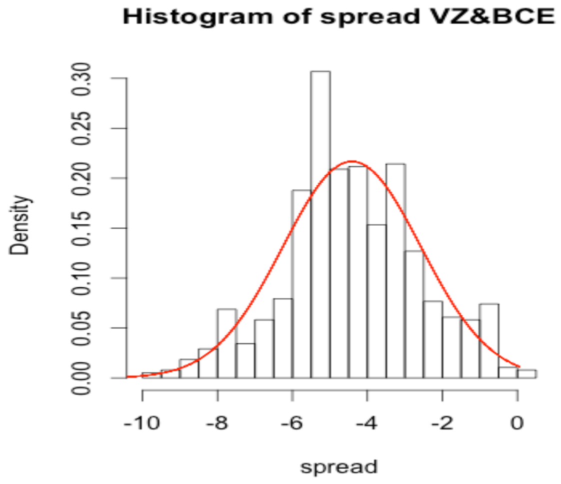




ATT~BCE error follows a roughly stationary distribution. From the graph of ATT~BCE spread histogram, we have the relationship between ATT and BCE nearly normal.

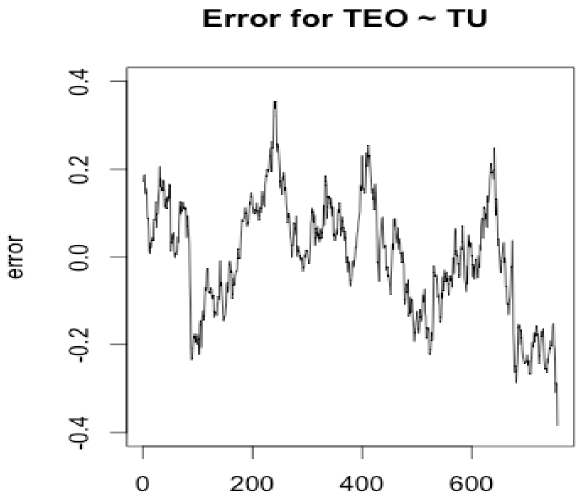
(VZ, BCE)

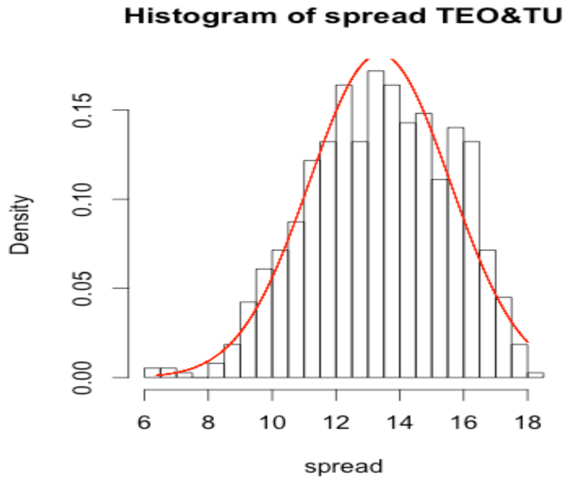




VZ~BCE error follows a stationary distribution. From the graph of VZ~BCE spread histogram, we have the relationship between VZ and BCE nearly normal.

(TEO, TU)





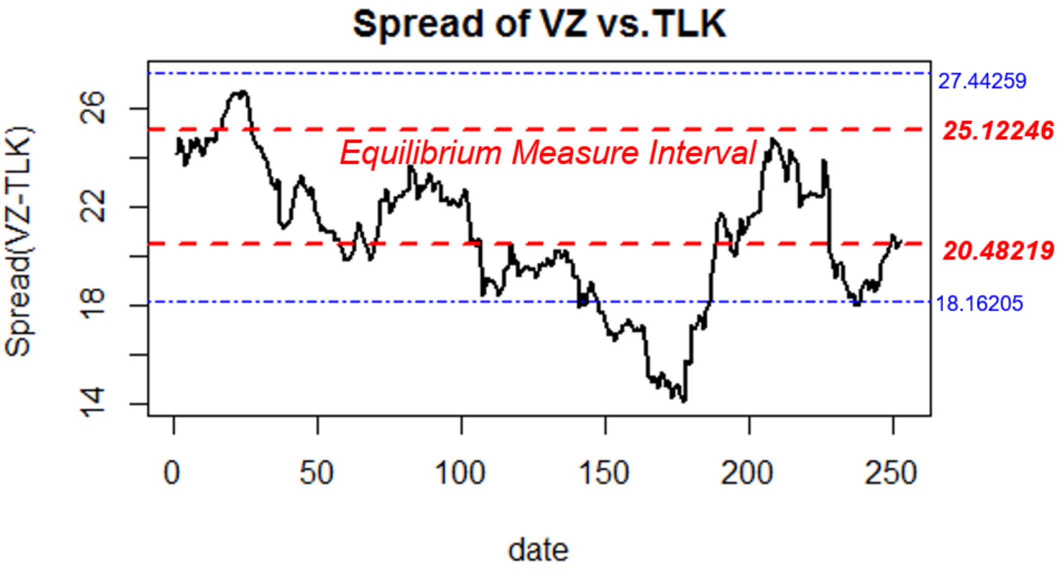
TEO~TU error does not follow a stationary distribution. From the graph of TEO~TU spread histogram, we have the relationship between TEO and TU nearly normal.

5. Back-testing

After a series of analysis on data of historical adjusted stock prices over the 3 year period (March 1, 2013 to March 1, 2016), we have successfully found 14 pairs of stocks that pass all the tests and satisfy cointegration, on which pairs trading strategy are very likely succeed. Among them, the distribution of spread of pair (VZ ,TLK) is very similar to shape of normal distribution. So we take it as the "best pair" and do back-testing on it using 1-year-period (March 1, 2016 to March 1, 2017) data. Thus, we can test whether the pair can be used for pairs trading, and how good the performance of our strategy can possibly be.

The strategy we used here is very simple, which is just giving an example to show a possibility and an adaption of our method. Using the historical data of stock VZ and TLK for the 3 year period, we calculated daily spread of this stock pair and get mean μ = 22.80232 and standard deviation σ = 1.160067. Then according to Gaussian distribution assumption, we calculated an Equilibrium Measure Interval μ±2σ, which is (20.48219, 25.12246) as shown in the plot below in red. This interval is where we expect the spread lies in a long term. Statistically, we are 95% confidence that the spread will lie in this interval in a long term. So once the spread of the chosen pair violate to the outside of this interval, we think it as an "abnormal" situation and take associated strategies to buy or sell.

The black line shows the spread trend of the stock pairs on 1 year period and we can see most of time the spread stays in the interval. Due to difference of risk tolerance level and expectation of return for different investors, arbitrage strategy for the same pair may differ. Here, the strategy is that one will enter market when the spread hit μ±4σ and expect the spread to return to the Equilibrium interval in the future. So we also draw these lines in blue, where spread is (18.16205, 27.44259). The points of intersection for the blue line and price trend is where we will enter the market.



Over the one year back-testing period, the first time we can enter market is 9/19/2016, where the spread is 17.88627 (<18.16205). However the spread didn't return to the Equilibrium Measure Interval as we expected until 11/23/2016, where spread is 20.48725 (> 20.48219). So using the strategy with μ±4σ as entering chance, we would enter the market on 9/19/2016 in a long position of equal share of both stocks and close the trade on 11/23/2016. Overall this strategy can help us get an arbitrage of 20.48725-17.88627 =2.60098 per share of stock.

References

[1] "Cointegration." Wikipedia. Wikimedia Foundation, 10 May 2017. Web. 11 May 2017., en.wikipedia.org/wiki/Cointegration

[2] "Augmented Dickey–Fuller Test." Wikipedia, Wikimedia Foundation, 1 May 2017., en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

Appendix A: R codes for Adftest and Plots

~~~
```{r}
library(tseries)
arbitrage_test <- function(y1,y2){
  y1 <- log(y1)
  y2 <- log(y2)
  cor <- cor(y1,y2)
  serie_y1 <- ts(y1)
  serie_y2<- ts(y2)
  adf.y1 <- adf.test(serie_y1)
  t1 <- adf.y1$p.value > 0.05
  adf.y2 <- adf.test(serie_y2)
  t2 <- adf.y2$p.value > 0.05
  d_y1 <- diff(serie_y1)
  d_y2 <- diff(serie_y2)
  adf.dy1 <- adf.test(d_y1)
  t3 <- adf.dy1$p.value <= 0.05
  adf.dy2 <- adf.test(d_y2)
  t4 <- adf.dy2$p.value <= 0.05
  reg<-lm(y1~y2)
  error <- residuals(reg)
  adf.error <- adf.test(error)
  t5 <- adf.error$p.value <= 0.05
  return(list(cor=cor,stock1=t1,stock2=t2,diff1=t3,diff2=t4,residual=t5))
}

rum <- read.csv("/Users/yangxu/Desktop/TELE_DATA_UPDATED.csv",header=T)
rum
n <- ncol(rum)-1
for (i in 2:n){
  for (j in (i+1):(n+1)){
    print(c(i,j))
    print(unlist(arbitrage_test(rum[,i],rum[,j])))
  }
}
adf.test(rum[,5])

ts.error <- function(y1, y2){
  logy1 <- log(y1)
  logy2 <- log(y2)
~~~

```
  fit <- lm(logy1~logy2)
  error <- ts(residuals(fit))
  return(error)
}
error <- ts.error(rum[,2],rum[,9])
plot(error, xlab = "", ylab ="error", main= "Error for Tlk ~ VZ ", ylim = c(-0.4,0.4))

error <- ts.error(rum[,7],rum[,5])
plot(error, xlab = "", ylab ="error", main= "Error for VIV ~ ANTI ", ylim = c(-0.4,0.4))

error <- ts.error(rum[,6],rum[,9])
plot(error, xlab = "", ylab ="error", main= "Error for ATT ~ VZ ", ylim = c(-0.4,0.4))

error <- ts.error(rum[,6],rum[,10])
plot(error, xlab = "", ylab ="error", main= "Error for ATT ~ BCE ", ylim = c(-0.4,0.4))

error <- ts.error(rum[,8],rum[,11])
plot(error, xlab = "", ylab ="error", main= "Error for TEO ~ TU ", ylim = c(-0.4,0.4))

error <- ts.error(rum[,9],rum[,10])
plot(error, xlab = "", ylab ="error", main= "Error for VZ ~ BCE ", ylim = c(-0.4,0.4))

y1 <- rum[,2]
y2 <- rum[,9]
spread <- y2-y1
range(spread)
x<-seq(14.06434,26.74792,by=0.001)
hist(spread, breaks = 20, probability = T, main = "Histogram of spread TLK&VZ")
lines(x,dnorm(x,mean = mean(spread), sd = sd(spread)), col = "red")

y1 <- rum[,7]
y2 <- rum[,5]
spread <- y2-y1
range(spread)
x<-seq(34.33930, 72.09493,by=0.001)
hist(spread, breaks = 20, probability = T, main = "Histogram of spread ANTI&VIV")
lines(x,dnorm(x,mean = mean(spread), sd = sd(spread)), col = "red")

y1 <- rum[,6]
y2 <- rum[,9]
```

~~~
spread <- y2-y1
range(spread)
x<-seq(6.25592,15.52728,by=0.001)
hist(spread, breaks = 20, probability = T, main = "Histogram of spread ATT&VZ")
lines(x,dnorm(x,mean = mean(spread), sd = sd(spread)), col = "red")

y1 <- rum[,6]
y2 <- rum[,10]
spread <- y2-y1
range(spread)
x<-seq(0.692634,12.635936,by=0.001)
hist(spread, breaks = 20, probability = T, main = "Histogram of spread ATT&BCE")
lines(x,dnorm(x,mean = mean(spread), sd = sd(spread)), col = "red")

y1 <- rum[,8]
y2 <- rum[,11]
spread <- y2-y1
range(spread)
x<-seq(6.370013,18,by=0.001)
hist(spread, breaks = 20, probability = T, main = "Histogram of spread TEO&TU")
lines(x,dnorm(x,mean = mean(spread), sd = sd(spread)), col = "red")
#6.370013, 17.569626

y1 <- rum[,9]
y2 <- rum[,10]
spread <- y2-y1
range(spread)
x<-seq(-10.428150,0.052518,by=0.001)
hist(spread, breaks = 20, probability = T, main = "Histogram of spread VZ&BCE")
lines(x,dnorm(x,mean = mean(spread), sd = sd(spread)), col = "red")
```
~~~

Appendix B: R codes for Backtesting

~~~
```{r}
setwd("C:/Users/Mengya/Desktop/Columbia Desk/GR5010/5010 Project")
tele10 <- read.csv("TELECOMM10.csv");dim(tele10) #getdata
tele_train  <- tele10[ 1:503, ]  #train dataset march 1,2013 to march 1,2016
tele_test   <- tele10[504:756, ] #test dataset march 1, 2016 to march 1 2017
```


#Best pair TLK 2 VZ 9;
```{r}
spread  <- tele_train$VZ_Adj.Close - tele_train$TLK_Adj.Close   #historical spread
mean(spread) + c(-1,1)*2*sd(spread)  #interval of u+-2*sigma
sp_test <- tele_test$VZ_Adj.Close - tele_test$TLK_Adj.Close
range(sp_test)
```
```{r}
mean(spread) + c(-1,1)*4*sd(spread)

plot(sp_test, lwd = 2, ylim = c(14.06434,27.44259),xlab = "date", ylab = "Spread(VZ-TLK) ", type="l", main ="Spread of VZ vs.TLK")
abline(h = 20.48219, untf = FALSE, lty=2, col ="red", lwd = 2)
abline(h = 25.12246, untf = FALSE, lty=2, col ="red", lwd = 2)

abline(h = 18.16205, untf = FALSE, lty=4, col ="blue", lwd = 1)
abline(h = 27.44259, untf = FALSE, lty=4, col ="blue", lwd = 1)
```
```{r}
which(sp_test<18.16205) #find the enter market chance
tele_test[141,]$Date
tele_test[185,]$Date

sp_test[141]  #spot spread value
sp_test[185]
~~~

~~~
which(sp_test>20.48219) #find the close trade chance near 200
sp_test[188]
tele_test[188,]$Date
```
~~~