General Principles of Statistical Arbitrage, Correlation-Based Trading, and Cointegration-Based Trading FINM 33170 and STAT 33910

Per Mykland

University of Chicago, Winter 2021

Outline

What is arbitrage?

To put it in a simple way, arbitrage is a riskless profit. Arbitrage strategy is a trading strategy that lock in a *riskless* profit by simulaneously buying and selling several securities, or by buying and selling the same securities on different occasions. The transactions could take place in the same market or in two or more markets.
In reality, absolute riskless opportunity does NOT exist. However, various anomaly and/or inefficient pricing exist, at least temporarily, and one can take advantage of these anomalies to creat “approximate” arbitrage opportunity.

Statistical arbitrage

Broadly speaking, statistical arbitrage strategy covers any trading strategies which use statistical tools to identify "approximate" arbitrage opportunity and to carry out the trading process,while evaluating the risk involved. We stress that *statistical strategy is not riskless*.
One should be aware that the term "statistical arbitrage" has been used in a carelessly manner. Often, practioners use statistical arbitrage strategy, pairs trading, and market neutral strategy interchangeably. In this class, we make a distinction for these strategies.

Market neutral strategy

It is a strategy that is neutral to market returns. Ideally, a market neutral portfolio combines long and short positions in such a way that market risk is reduced to zero. In other words, the return from a market neutral portfolio is uncorrelated to market returns. In practice, such a portfolio is designed to be indifferent to the market returns to the extent possible.
A market neutral portfolio can be profitable if one can exploit the mean-reverting behaviour of the residual portfolio series. Most of time, statistical arbitrage is market neutral. However, there are popular statistical arbitrage strategies which are not market neutral, this includes momentum strategy. Note that *statistical arbitrage strategy should NOT be a black box*, it relies on historical data and present data (both the price data and fundamentals).

Pairs trading

We view pairs trading as a *a market neutral strategy which involves only two stocks (or futures, currencies, etc.)*. Technically, pairs trading strategy could be entirely nonparametric, or semiparametric. Like any market neutral strategy, pairs trading is based on relative pricing.

Correlation-based trading

We here demonstrate nonparametric pairs trading.

- idea: If two stocks, A and B, have the exact same statistical properties and fundamental characteristics, they bound to have the same price. Hence, take opposite position of A and B would give a market neutral strategy.
- notations:
  Let $p_{t_i}^A$ and $p_{t_i}^B$ be the prices of stocks A and B at time t_i.
  Let $x_{t_i}^A$ and $x_{t_i}^B$ be the log prices of stocks A and B at time t_i.
  Let $r_{t_i}^A$ and $r_{t_i}^B$ be the returns of stocks A and B at time t_i.
  where $r_{t_i}^A = x_{t_i}^A - x_{t_{i-1}}^A$
- setup: suppose you ONLY have the price time series of several stocks in front of you. How to select pairs? how to choose the relative positions? what is the buy/sell signals?
- Pair selection: select the two which are close to each other.

selection I: pick the pair with the shortest euclidean distance between the return series.
selection II: pick the pair with the shortest euclidean distance between the NORMALIZED return series.
selection III: pick the pair with the maximum sample correlation between the return series.

To normalize the series $r^A_{t_i}$ and $r^B_{t_i}$, $i = 1, \cdots, n$,

$$\begin{aligned} Z^A_{t_i} &= \frac{r^A_{t_i} - \bar{r}^A}{s_A} \\ Z^B_{t_i} &= \frac{r^B_{t_i} - \bar{r}^B}{s_B}, \end{aligned}$$

where $\bar{r}^A$ and $\bar{r}^B$ are the sample means for the returns of A and B, and s^2_A and s^2_B are the sample variance.
The euclidean distance between the standardized series Z^A and Z^B is

$$\sqrt{\sum_{i=1}^{n} (Z^A_{t_i} - Z^B_{t_i})^2}.$$

Q_1: Are all three selection procedures plausible? Why?
Q_2: Any relation between selection II and III?
Ans : Let $\hat{\rho}_{A,B}$ be the sample correlation between r_A and r_B.
Recall that

$$\begin{aligned}\hat{\rho}_{A,B} &= \frac{\sum_{i=1}^{n}(r_{t_i}^A - \bar{r}^A)(r_{t_i}^B - \bar{r}^B)}{\left[\sum_{i=1}^{n}(r_{t_i}^A - \bar{r}^A)^2 \sum_{i=1}^{n}(r_{t_i}^B - \bar{r}^B)^2\right]^{1/2}} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(r_{t_i}^A - \bar{r}^A)}{s_A}\frac{(r_{t_i}^B - \bar{r}^B)}{s_B} = \frac{1}{n-1}\sum_{i=1}^{n} Z_{t_i}^A Z_{t_i}^B\end{aligned}$$

and the square of the euclidean distance between Z^A and Z^B is

$$\begin{aligned}&\sum_{i=1}^{n}(Z_{t_i}^A - Z_{t_i}^B)^2 \\ = &\sum_{i=1}^{n}(Z_{t_i}^A)^2 + \sum_{i=1}^{n}(Z_{t_i}^B)^2 - 2\sum_{i=1}^{n} Z_{t_i}^A Z_{t_i}^B\end{aligned}$$

Comment: because of selection III, the relevant pairs trading strategy is also called correlation-based strategy.

Test the trade.

First let us try in-sample test on the strategies.
Consider the following three simple strategies,
strategy I.
When $p^A - p^B > \Delta_1$, take a short position on one unit of A and simultaneouly long one unit of B.
When $p^A - p^B < \Delta_2$, liquidate the previous positions.
(or, how about consider the above with log prices??)

strategy II. first estimate β for the returns, the coefficient from the linear regression equation (1).

$$r^A = \beta r^B + \epsilon^A \tag{1}$$

When $\log(p^A) - \beta \log(p^B) < \Delta_1$, take a long position on $\frac{1}{p^A}$ shares of A and simultaneously short $\frac{\beta}{p^B}$ shares of B.
When $\log(p^A) - \beta \log(p^B) > \Delta_2$, rewind the positions.
Calculate the profit/loss in one transaction:
At time t, $\log(p_t^A) - \beta \log(p_t^B) < \Delta_1$: $-\frac{1}{p_t^A} p_t^A + \frac{\beta}{p_t^B} p_t^B$
At time $t+1$,$\log(p_{t+1}^A) - \beta \log(p_{t+1}^B) > \Delta_2$: $\frac{1}{p_t^A} p_{t+1}^A - \frac{\beta}{p_t^B} p_{t+1}^B$

$$\begin{aligned} profit/loss &= -\frac{1}{p_t^A} p_t^A + \frac{\beta}{p_t^B} p_t^B + \frac{1}{p_t^A} p_{t+1}^A - \frac{\beta}{p_t^B} p_{t+1}^B \\ &\approx r_{t+1}^A - \beta r_{t+1}^B \end{aligned}$$

strategy III. first estimate β', the coefficient from the linear regression (2).

$$r^B = \beta' r^A + \epsilon^B \tag{2}$$

Note: Δ_1 does NOT have to be the same as Δ_2.

◇ ◇ ◇ ◇ ◇ ◇ ◇

Q_3: Any relation between portfolio II and III?

Q_4: which is relatively better, portfolio II or III?

Q_5: What could go wrong in the above?

We emphasize that even if the strategy is quite profitable in in-sample data, it has to be "validated" out of sample. That is to say, if the pair selection is conducted over time $[t_1, t_2, \cdots, t_n]$, the trading strategy tested over the same time period is "in-sample" test. In-sample test is subject to the sample bias. The real implementation of the strategy is always "out-of-sample".

When to buy or sell? — determine Δ

Everything else being equal,

- the magnitude of correlation: higher correlation, lower threshhold
- risk control: higher threshhold, lower risk
- carry-out cost: higher threshhold, lower cost

Anything wrong in the above strategies?

- cost of implementing the strategy? (slippage? midpoint of bid-ask)
- implementable? how long/how short to hold the strategy?
- risk assessment: control test? worst senario? Sharpe ratio?

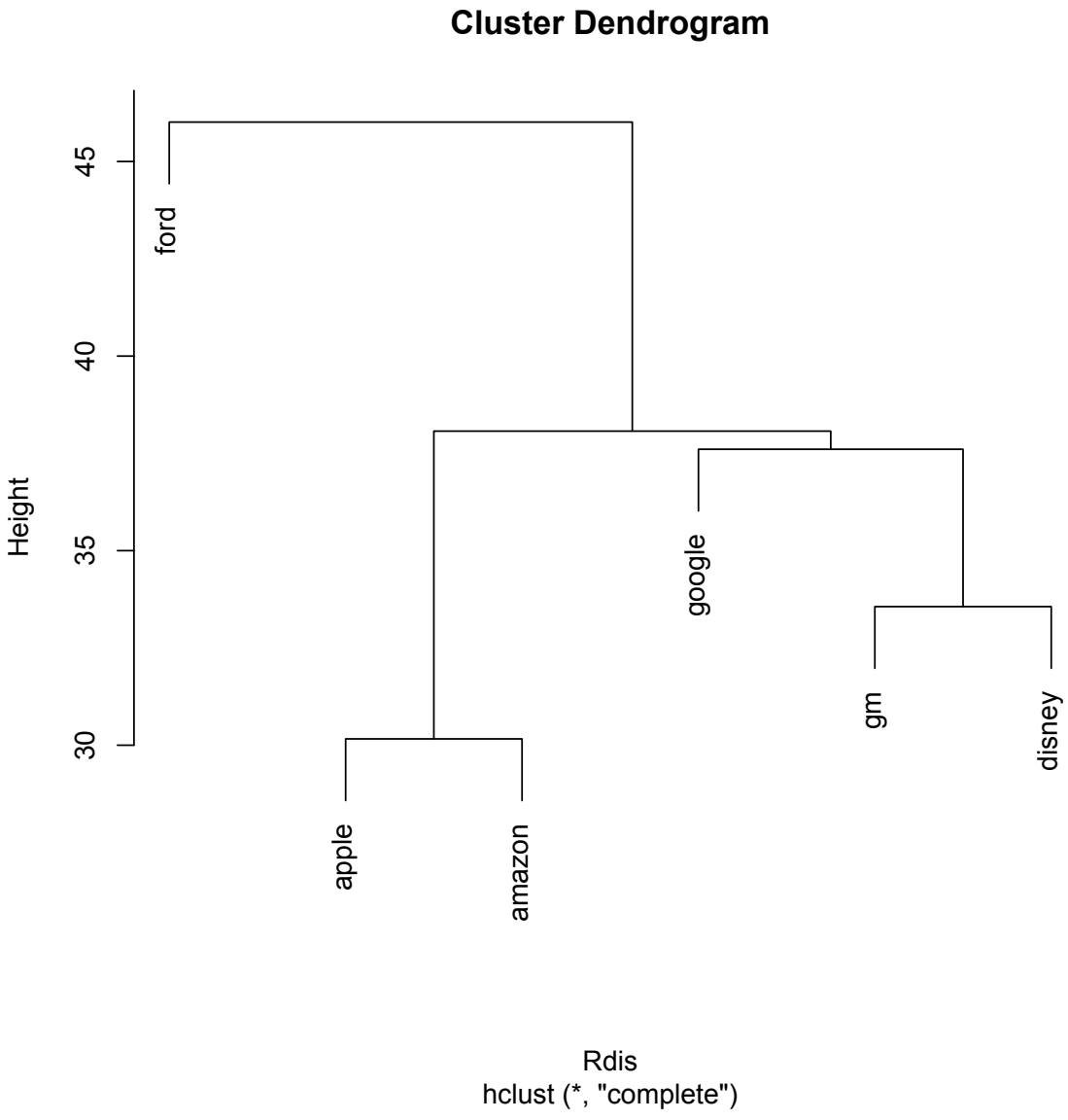
Specifically:

- On pairs selection, we have ignored the time order of the data, we have not considered incorporating the fundamental characteristics.
- On buy/sell signal,how to select Δ_1 and Δ_2 to maximize the profit?
- diversification: pairs $\Rightarrow$ portfolio involving more securities.

Data demonstration: first visit pairs trading Consider the following six stocks 13 Apr 2013 to 12 Apr 2018: Google, Apple, Amazon, GM, Ford, Disney Download AAPL.csv

```
aapl<-read.csv("AAPL.csv") #same with GOOG.csv, AMZN.csv, GM.csv, FORD.csv, DIS.csv
prices<-log(cbind(aapl[,6],goog[,6],amzn[,6],gm[,6],ford[,6],dis[,6]))
var.labels<-c("google","apple","amazon","gm","ford","disney")
colnames(prices)<-var.labels
returns<-apply(prices,2,diff)
Rpsel<-returns[1:1000,]
Rtest<-returns[1001:1259,]
cor(Rpsel)
             google     apple     amazon          gm         ford       disney
google  1.00000000 0.3255386 0.27456202  0.29220122 -0.059661449  0.323853937
apple   0.32553855 1.0000000 0.54465715  0.32577710  0.094163800  0.363153452
amazon  0.27456202 0.5446571 1.00000000  0.29014591  0.033510978  0.364523216
gm      0.29220122 0.3257771 0.29014591  1.00000000 -0.035836043  0.436238264
ford   -0.05966145 0.0941638 0.03351098 -0.03583604  1.000000000 -0.009105682
disney  0.32385394 0.3631535 0.36452322  0.43623826 -0.009105682  1.000000000
 #Q: returns vs. prices?
Rpsel.m<-apply(Rpsel,2,mean)
Rpsel.sd<-apply(Rpsel,2,sd)
Rpsel.std<-0*Rpsel
for (k in 1:6 ){ Rpsel.std[,k]<- (Rpsel[,k]-Rpsel.m[k])/Rpsel.sd[k]}
Rdis<-dist(t(Rpsel.std),method= "euclidean")   #calculate the distance among ROWS
plot(hclust(Rdis)) #see plot 1.
```

Cluster Analysis



Closest pair is (Apple, Amazon). The pair with the shortest Eucledean distance is the same pair with the greated correlation. Do this coincide with the pair which yields largest R-squared in linear regression??

```
m1<-lm(Rpsel.std[,2]~Rpsel.std[,3]-1)
> summary(m1)

Call:
lm(formula = Rpsel.std[, 2] ~ Rpsel.std[, 3] - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9757 -0.3968 -0.0277  0.3406  9.8765

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
Rpsel.std[, 3]  0.54466    0.02653   20.53   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.8387 on 999 degrees of freedom
Multiple R-squared:  0.2967,        Adjusted R-squared:  0.2959
F-statistic: 421.3 on 1 and 999 DF,  p-value: < 2.2e-16
```

What is "0.54466"?

What is the relative position?

```
m3<-lm(Rpsel[,2]~Rpsel[,3])

summary(m3)

Call:
lm(formula = Rpsel[, 2] ~ Rpsel[, 3])

Residuals:
      Min        1Q    Median        3Q       Max
-0.071595 -0.005710 -0.000398  0.004901  0.142114

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0002612  0.0003825   0.683    0.495
Rpsel[, 3]  0.4134798  0.0201535  20.516   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.01207 on 998 degrees of freedom
Multiple R-squared:  0.2967,        Adjusted R-squared:  0.2959
F-statistic: 420.9 on 1 and 998 DF,  p-value: < 2.2e-16

####Q: check the intercept
```

```
> plot(Rpsel[,2],Rpsel[,3],xlab="Apple",ylab="Amazon",main="scatterplot for returns")

>ts.plot(Rpsel[,2]-0.4134798*Rpsel[,3],xlab="time index",ylab="residuals
 from regression",main="residual time series")
           #note that ylab is same as r_{Apple} - 0.4134798*r_{Amazon}


>ts.plot(cumsum(Rpsel[,2]-0.4134798*Rpsel[,3]),xlab="time index",ylab="integrated
   residuals",main="time series of log(p_{Apple}) - 0.4134798*log(p_{Amazon})")

#IS this mean reverting?

>plot(cumsum(Rpsel[,2]-0.4134798*Rpsel[,3]),Rpsel[,2]-0.4134798*Rpsel[,3],xlab=
   "log(p_{Apple}) - 0.4134798*log(p_{Amazon})",ylab="r_{Apple} - 0.4134798*r_{Amazon}")

#note: one can consider the xlab is the log price of the portfolio
 consisting of (Apple,Amazon), ylab is then the diff(log price of the portfolio)
```

Cointegration-based trading: Engle-Granger approach

Integrated process of order d Most of financial time series are nonstationary. However, as you may have learned from time series class, a univariate nonstationary time series can often be made into a stationary time series by the technique of differencing.
Recall the differencing operator Δ and backshift operator B, where

$$\Delta X_t = X_t - X_{t-1} \quad \text{and} \quad BX_t = X_{t-1},$$

and the relations below, for some integer $d > 0$,

$$\begin{aligned}
\Delta X_t &= (1-B)X_t \\
\Delta^2 X_t &= \Delta(\Delta X_t) = \Delta(X_t - X_{t-1}) = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = (\\
&\vdots \\
\Delta^d X_t &= (1-B)^d X_t
\end{aligned}$$

A univariate time series $\{X_t\}$ is *integrated of order d*, if $\{\Delta^d X_t\}$ is stationary for some positive integer d, but $\{\Delta^{d-1} X_t\}$ is nonstationary. We denote this by $\{X_t\} \sim I(d)$. The concept of integrated process of order d extends to multivaiate times series, where $\{\Delta^d X_t\}$ means applying the operator Δ^d to each component of a multivariate time series $\{X_t\}$. For example, for a k-variate time series $\{X_t\}$, denote its i-th component series by $\{X_{t,i}\}$, then the i-th component series of $\{\Delta^d X_t\}$ is the differenced series $\{\Delta^d X_{t,i}\}$, where $i = 1, \cdots, k$.
A k-variate, $I(d)$ times series $\{X_t\}$ is said to be cointegrated if a linear combination of its component series is of order less than d. In other words, if there exists a $k \times 1$ vector γ such that $\{\gamma' X_t\}$ is of order less than d. γ is called cointegration coefficient. In practice, financial time series are frequently $I(1)$. In this context, one can say that a univariate time series $\{X_t\}$ and another univariate time series $\{Y_t\}$ are cointegrated, if there exists a γ, $\gamma \neq 0$, such that the time series $\{X_t + \gamma Y_t\}$ is stationary.

Example

Q 1: Suppose X is a random walk, and Y is a noisy random walk. Assume that the noise is independent of X. Is X or Y stationary? Are X and Y cointegrated?

$$\begin{aligned} \textit{A 1: Write } X_t &= \sum_{i=1}^{t} Z_i, \quad t = 1, 2, \cdots, \{Z_t\} \sim WN(0, \sigma^2), \\ Y_t &= X_t + \epsilon_t, \quad t = 1, 2, \cdots, \{\epsilon_t\} \sim WN(0, \nu), \end{aligned}$$

It's easy to see that X and Y are not stationary. For example, though $E(X_t) = 0$, the autovariance depends on t.

$$Cov(X_t, X_{t+h}) = E[X_t X_{t+h}] = \sum_{i=1}^{t} \sum_{j=1}^{t+h} E[Z_i Z_j] = \sum_{i=j=1}^{min(t,t+h)} \sigma^2$$

However, the linear combination of X and Y can be stationary. For example, choose $\gamma = (1, -1)'$. Here $\sum_{i=1}^{t} Z_i$ is the common trend, by linear combination, the common trend get canceled out!. Note that a trend is always nonstationary, but it can be deterministic or stochastic.

Example

Q2 : Let $X_{t,1} = \sum_{i=1}^{t} Z_{i,1} + Z_{t,2}$ $X_{t,2} = 3\sum_{i=1}^{t} Z_{i,1} + Z_{t,3}$ $X_{t,3} = 2Z_{t,2}$

where $Z_{t_1}, Z_{t_2}, Z_{t_3}$ are iid with mean 0 and variance ν. Is the 3-dimensional time series $\{X_t\}$ cointegrated?

A 2: Again the first two component series have (common) trend term, thus not stationary. The third component is stationary. However, ΔX_t is stationary, because,

$$\begin{aligned}\Delta X_t &= (I - B)X_t \\ &= (Z_{t,1} + Z_{t,2} - Z_{t-1,2}, 3Z_{t,1} + Z_{t,3} - Z_{t-1,3}, 2Z_{t,2} - 2Z_{t-1,2})'.\end{aligned}$$

It is true that $\{X_t\} \sim I(0)$. Moreover, the linear combination of the component series,

$$3X_{t,1} - X_{t,2} = 3Z_{t,2} - Z_{t,3}$$

is stationary. Here, the cointegrating vector $\gamma = (3, -1, 0)'$. Actually, the cointegrating vector is **not unique**.

Cointegration: the ideas

Cointegration is a term to describe the comovement of asset prices (typically in log scale). The cointegrated time series behave like being “tied” together (by a common trend). Even if each component series is nonstationary, the linear combination of cointegrated series displays stationarity, especially, mean reverting.

The idea of a cointegration test is to identify the linear combination of component series that is stationary and best defines the long-run equilibrium relation among the component series.

There are two well-established test for cointegration, the Engle-Granger method and the Johansen method. As we shall see in the following, the Engle-Granger method is LS-based, easy to implement. The Johansen method is one step procedure, delivers the equilibrium relation which is the most stationary.

We first focus on the Engle-Granger Methodology.

Engle-Granger Test for Cointegration

Engle-Granger method is a two-step tests. Step 1 involves regressing one integrated series on the other integrated series (OLS), step 2 conducts unit root test on the residual series. If the residual series has no unit root, it's stationary, and hence mean-reverting. To illustrate the idea, suppose the data $\{x_t, y_t\}$ follows from $I(1)$ process. Suppose regression of

$$x_t = \alpha + \beta y_t + \epsilon_t,$$

yields $\{\hat{\epsilon}_t\}$ being stationary. This is equivalent to say that $x_t = \hat{\alpha} + \hat{\beta} x_t$ is the long-run commom trend (or, equilibrium relation) between x and y. $(1, -\beta)'$ is called *cointegration vector*. Ordinary LS regression typically works for stationary data. In the current context, the data, the logrithmic price of the securities, is almost always nonstationary. The regression result is only meaningful when it yields stationary residual series, that is, when the component time series are cointegrated. In the latter case, the regression line gives the long-run equilibrium relation between the log prices.

validity of initiating the regression

Make sure that your input for the regression is $I(1)$.

- check the time series plot, at least inspecting the existence of mean-reverting behavior
- unit root test

Here, we consider the daily log price of SPY (ETF, proxy for S&P 500) and DAX (ETF, proxy for Germany's DAX index), from 23 October 2014 to 12 April 2018. The time series show pattern of nonstationarity. We here demonstrate the Dickey-Fuller (DF) test for unit root.

Unit Root Test

Consider the model

$$X_t = \alpha + \beta X_{t-1} + \epsilon_t. \tag{3}$$

For simplicity, assume $\alpha = 0$. A test against

$$H_0 : \beta = 1 \quad v.s. \quad H_1 : \beta < 1,$$

is called unit root test (or random walk test), since under H_0, $\{X_t\}$ follows random walk.
Dickey-Fuller test (DF) is among the earliest ideas in unit root test. DF test is based on the t-ratio from the OSL. The more powerful unit root tests (say, Phillips-Perron test, Durbin-Hausmann test) shall be discussed in later lectures, if time permits.

The idea behind the DF test

If $\{X_t\}$ has unit root, then in the regression

$$\Delta X_t = (\beta - 1)X_{t-1} + \epsilon_t,$$

the coefficient of X_{t-1} would be "close" to zero. Notice, however, X_t is not iid data, and the coefficient of X_{t-1} does NOT follow t-distribution. Instead,

Theorem

Let X_t follow model (3). Assume $\alpha = 0$. Then under $H_0 : \beta = 1$, the t-ratio from OLS converges in law,

$$T_{OSL} \xrightarrow{\mathcal{L}} \frac{\frac{1}{2}(W_1^2 - 1)}{\sqrt{\int_0^1 W_t^2 dt}} \tag{4}$$

$$\text{and } n(\hat{\beta} - 1) \xrightarrow{\mathcal{L}} \frac{\int_0^1 W_t dW_t}{\int_0^1 W_t^2 dt}. \tag{5}$$

The critical value from the distribution in (4) can be calculated from simulation. Assuming $\alpha = 0$:

```
>n<-100
>m<-9999
>A<-rnorm(m*n)
>Amat<-matrix(A,nrow=m)
>Amat<-Amat/sqrt(n)
>B<-apply(Amat,1,cumsum) #get row cumsum
>B<-t(B)
>Bsq<-B^2
>denom<-apply(Bsq,1,sum)/n
>tstatalt<-(1/2)*(B[,n]^2 -1)/sqrt(denom)
>sort(tstatalt)[500] ##9999*.05 = 499.95
[1] -1.934616
```

Suppose $\alpha \neq 0$, the asymptotic distribution of DF statistic is a modification of (4). For $n = 100$, 5-th percentile is then around -2.37.

Note that Dickey-Fuller test becomes less powerful at the presence of trend α, when the residuals $\{\epsilon_t\}$ have heteroscedastic variance, and/or are autocorrelated. In the latter case, the Augmented Dickey-Fuller (ADF) test can be used instead. The ADF test assumes the model

$$\Delta X_t = (\beta - 1)X_{t-1} + \theta_1 \Delta X_{t-1} + \cdots + \theta_p \Delta X_{t-p} + \epsilon_t.$$

The number of lags p is determined such that ϵ_t becomes "independent", hence $\hat{\beta}$ from OLS is an unbiased estimator. The test statistic of ADF(p) is still the OLS t-ratio for $\beta - 1$. The critical values for ADF are slightly different from DF test. For reference, the critical values of ADF(1) at .05 is

n	0.05
50	-2.93
100	-2.89
500+	-2.86

DF test in R

:
The hypothesis of $\beta = 1$ is the same as saying that $\{X_t\}$ is not stationary, hence, not mean-reverting.
In the following, we will test

$$H_0 : \{X_t\} \text{ is } I(1) \text{ or higher} \quad vs. \quad H_1 : \{X_t\} \sim I(0),$$

this is equivalent to test

$$H_0 : \beta = 1 \quad vs. \quad H_1 : \beta < 1.$$

To proceed with the above test, one can regress ΔX_t over X_{t-1} and then check the t-statistics.
Similary, consider,

$$H_0 : \{X_t\} \text{ is } I(2) \text{ or higher} \quad vs. \quad H_1 : \{X_t\} \sim I(1),$$

regress $\Delta^2 X_t$ over ΔX_{t-1}.

SPY and DAX

```
spy<-read.csv("SPY.csv")
dax<-read.csv("dax.csv")
dax[1,]
        Date  Open  High   Low Close Adj.Close Volume
1 2014-10-23 25.09 25.27 25.09 25.25  24.02423   7400
spy[388,]
         Date   Open  High    Low  Close Adj.Close    Volume
388 2014-10-23 194.62 196.2 194.26 194.93  181.6737 154944000
spy.longer<-spy
spy<-spy[388:1260,]
lspy<-log(spy[,6])
ldax<-log(dax[,6])
dlspy<-diff(lspy)
dldax<-diff(ldax)
```

```
m10<-lm(dlspy ~ lspy[2:873,6])
summary(m10) #test H_0: not I(0) vs. H_1: log(FTSE) is I(0)

Call:
lm(formula = dspy ~ spy[2:873, 6])

Residuals:
     Min       1Q   Median       3Q      Max
-11.7173  -0.6798   0.0055   0.8521   7.0503

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.579878   0.477873  -1.213    0.225
spy[2:873, 6]  0.003131   0.002195   1.427    0.154

Residual standard error: 1.75 on 870 degrees of freedom
Multiple R-squared:  0.002334,        Adjusted R-squared:  0.001187
F-statistic: 2.035 on 1 and 870 DF,  p-value: 0.154
```

Also visual inspection of the time series can be done by

```
ts.plot(lspy,main="time series of SPY (log scale)")
         #see plot for stationarity
ts.plot(dlspy,main="differenced time series of SPY (log scale)")
```

Clearly, the log price of FTSE 100 in the studies time period is $I(1)$.
Implement DF test: conclude that log price of FTSE 100 is not stationary.

OLS regression

Regress the log price of the SPY on the log price of the DAX:

```
m20<-lm(ldax~lspy)
summary(m20)
```

Check stationarity of the residual series

First by plot,

```
ts.plot(m20$res,main="time series of residuals from log(spy) ? log(dax)")
```

Then by DF test,

```
> dres.m20<-diff(m20$res)
> m21<-lm(dres.m20 ~ m20$res[1:872])
> summary(m21)

Call:
lm(formula = dres.m20 ~ m20$res[1:872])

Residuals:
      Min        1Q    Median        3Q       Max
-0.077691 -0.004680 -0.000012  0.004837  0.037945

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.668e-05  3.047e-04  -0.186    0.852
m20$res[1:872] -1.254e-02  5.243e-03  -2.392    0.017 *
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.008998 on 870 degrees of freedom
Multiple R-squared:  0.006534,        Adjusted R-squared:  0.005392
F-statistic: 5.722 on 1 and 870 DF,  p-value: 0.01696
```

Q: Can you draw conclusion on the DF test?

Also, how about reversing the regressor?

Comments

- the Engle-Granger method requires large sample size, and it inherits the drawbacks tied with LS regression (minimum variance). Why??
  To see why the OSL-based cointegration method may NOT be desirable, consider for $|\beta| < 1$,

$$\begin{aligned} Z_{t+1} &= \beta Z_t + \sigma \epsilon_{t+1}, \quad \text{where } \epsilon_t \sim IID(0,1), \\ W_{t+1} &= W_t + \gamma \eta_{t+1}, \quad \text{where } \eta_t \sim IID(0,1), \end{aligned}$$

  also assume ϵ_{t+1} and Z_t are independent, η_{t+1} and W_t are independent, ϵ and η are independent. Note that $Var(W_t) = \gamma^2 t$, and $Var(Z_t) = \frac{\sigma^2}{1-\beta^2}$.
  Now construct $\{X_t\}$ and $\{Y_t\}$ by setting

$$\begin{aligned} X_t &= aZ_t + bW_t \\ Y_t &= cZ_t + dW_t \end{aligned}$$

Consider the linear combination of the form $Y_t - \theta X_t$,

$$Y_t - \theta X_t = (c - a\theta)Z_t + (d - \theta b)W_t,$$

the variance would be

$$E(Y_t - \theta X_t)^2 = (c - \theta a)^2 \frac{\sigma^2}{1 - \beta^2} + (d - \theta b)^2 \gamma^2 t$$

In OLS: minimize the residual sum square $\sum_{t=1}^{T} E(Y_t - \theta X_t)^2$ will lead to

$$\hat{\theta} = \frac{2ac\frac{\sigma^2}{1-\beta^2} + (T+1)\gamma^2 bd}{2a^2\frac{\sigma^2}{1-\beta^2} + (T+1)\gamma^2 b^2}.$$

So,

- $\theta \neq \frac{d}{b}$, where the latter leads to $Y_t - \theta X_t$ being a nonstationary series, by canceling out the nonstationary common component W_t.
- if $T \to \infty, \theta \to \frac{d}{b}$.
- if $\sigma^2 \to \infty$, T is fixed, then $\theta \to \frac{c}{a}$. This is more unsettling if β is close to 0, or less than 0.

- If $\{X_t\}$ is k-variate time series, $k > 2$, one regression from E-G method will give only one linear combination of the component series. Which one component should be chosen as "regressor"??
  In principle, there could be at most $k - 1$ *independent* linear combinations.
- Johansen style co-integration instead?