

Pair Trading

The main question

what is pair trading ?

- Pick two assets whose historical price move together
- monitor the “ Spread ”
- If diverge :
 - short the winner
 - long the loser

Spread construction

- 1 v 1 : univariate
 - 例如，假設有兩支股票 AAA 和 BBB，你可以定義其價差 $\text{Spread} = A - B$ 。
 - 舉例：常見的例子是兩隻同一行業或同一公司股權結構下的股票，如可口可樂與百事可樂的股票價格比較。配對交易者會根據它們之間的價差開多/空頭頭寸。
- 1 v multi : quasi multivariate
 - 假設你要交易公司 AAA 的股票，但你認為它與行業內其他公司如 BBB、CCC、DDD 的股價有一定關聯性。你可以將 AAA 的股票與 BBB、CCC、DDD 股票的加權組合進行價差分析，並基於此做交易決策。
 - extension to HFT application
 - alt :
 - 使用波動率、成交量、成交量加權價格 (VWAP) 、RSI (相對強弱指數) 等非價格變量來構建價差。
 - 這些替代指標可以提供更複雜的分析方法，幫助識別傳統價格價差無法捕捉到的機會。
 - 舉例：你可以構建一個策略，基於兩支股票的波動率變化來進行配對交易。例如，如果兩支股票的「波動率之間的差異」擴大到某個閾

- multi v multivariate 值，則做多波動率較低的股票，做空波動率較高的股票，假設波動率最終會收斂。

Approaches

- Distance Approach – using nonparametric distance metrics.
- Cointegration Approach – using formal cointegration testing.
- Time Series Approach – finding optimal trading rules for mean-reverting spreads.
- Stochastic Control Approach – identifying optimal portfolio holdings relative to other available securities.
- Other Approaches – relevant pairs trading frameworks with only a limited set of supporting literature.
 - Machine Learning Approach;
 - Copula Approach;
 - PCA & Other Approaches.

Distance :

不假設特定dist

Basic Distance Strategy and Adjustments

Pairs trading: Performance of a relative-value arbitrage rule. (2006)
by Gataev, E., Goetzmann, W.N., and Rouwenhorst, K.G.

Does simple pairs trading still work? (2010)
by Do, B. and Faff, R.

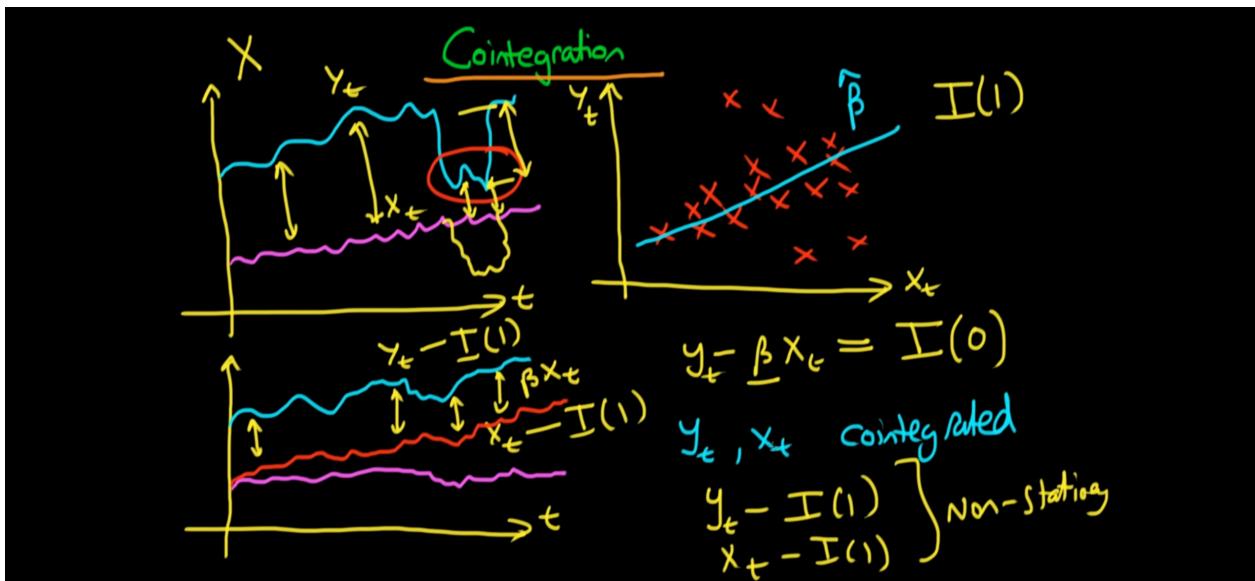
- Most cited work in Pairs Trading Domain
- Simple algorithm, robust to data snooping bias.
- Yielding annualized excess returns (prior to 2002), later profitability degraded.
- Adjustments show improved results.

1. 概述：距離法基於兩個資產的價格序列之間的歐幾里得距離或其他距離度量來構建價差。

- a. 例如，如果兩個資產的價格走勢非常相似（距離很短），當它們出現明顯偏離時，配對交易策略會進行套利。
2. 適用情況：這種方法適合用於對兩個資產之間具有明顯的歷史相似性的場景，且不需要考慮資產之間的複雜共整合關係。
3. 通常用於較簡單的均值回歸策略。

Cointegration approach : long term relationship

degree of comovement between pairs is assessed by cointegration testing



1. 概述：協整法利用「資產價格」之間的長期均衡關係，通過協整檢驗來確認資產之間的共整合關係，然後利用這種關係進行交易。
當協整關係中的資產價格出現短期偏差時，假設它們最終會回歸均衡。
2. 適用情況：這種方法適合用於兩個或多個資產存在長期均衡關係的情況，如同一產業內的公司股票或同類商品。
3. 通常用於較長期的策略，並假設資產之間的共整合關係穩定存在。

Idea

$$\Delta y_t = y_t - y_{t-1}$$

差分的目的是將非平穩序列轉換為平穩序列。如果一個序列經過一次差分（即 $y_t - y_{t-1}$ ）後變得平穩，我們稱這個序列是 $I(1)$ 的。這意味著這個序列的差分（變化量）是平穩的。

差一次：I(1)

I(1) 的例子：

假設有一個隨機遊走的過程 y_t :

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

其中， ϵ_t 是白噪聲（均值為 0，方差為常數的隨機過程）。這樣的過程是 I(1) 的，因為它的值會隨著時間隨機變動，而且均值和方差都不是固定的。

但如果對這個序列進行一次差分：

$$\Delta y_t = y_t - y_{t-1} = \epsilon_t$$

差分後的序列是白噪聲過程，它是平穩的，因為它的均值是 0，方差是固定的，且不隨時間變化。因此，這個序列 y_t 是 I(0) 的，因為它經過一次差分後變成 I(0) 的過程。

- 本身：I(1) : non-stationary, I(0) : stationary
 - 差一次, 差0次
- Both x, y non stationary
- Spread can be stationary (except outliers) → 求出beta → 認為具 cointegration
- $y - \text{beta} * x = I(0) \rightarrow \text{stationary}$

Three steps framework

- preselection of cointegrated pairs
- testing for tradibility
 - test for 共整合殘差
 - ADF (augmented dickey fuller : 檢驗殘差的單跟性質
- trading rule design with nonparametric method

ADF : auto corr 檢驗

```
import numpy as np
import pandas as pd
from statsmodels.tsa.stattools import adfuller
import matplotlib.pyplot as plt
```

```

# 定義一個函數來進行ADF檢驗並顯示結果
def adf_test(series, stock_name):
    print(f'ADF Test for {stock_name}:')
    result = adfuller(series.dropna()) # 去掉缺失值
    labels = ['ADF Test Statistic', 'p-value', '#Lags Used', 'Number of Observations Used']
    for value, label in zip(result, labels):
        print(f'{label}: {value}')

    if result[1] <= 0.05:
        print(f"=> {stock_name} 的數據沒有單根（也就是定態），拒絕單根假設")
    else:
        print(f"=> {stock_name} 的數據具有單根（也就是非定態），無法拒絕單根假設")

# 對兩支股票進行ADF檢測
adf_test(tw_common['Adj Close'], '2465')
adf_test(us_common['Adj Close'], 'SMCI')

```

```

ADF Test for 2465:
ADF Test Statistic: -1.6690679574417104
p-value: 0.44711589822947684
#Lags Used: 2
Number of Observations Used: 229
=> 2465 的數據具有單根（非定態），無法拒絕單根假設 (H0)

ADF Test for SMCI:
ADF Test Statistic: -1.3462387233799578
p-value: 0.6077617164412027
#Lags Used: 4
Number of Observations Used: 227
=> SMCI 的數據具有單根（非定態），無法拒絕單根假設 (H0)

```

- 在 ADF 檢驗結果中，"lag used" 表示檢驗過程中所使用的滯後期數。
 - 滯後期數決定了我們在檢驗中考慮了多長時間範圍內的過去數據，用來解釋當前數據點的變動
 - 例如，當我們使用 2 個滯後項 (lag=2)，這意味著我們用當前時間 t 的前兩個時間點 $t-1$ 和 $t-2$ 的數據來調整當前時間點的數據，這樣我們考慮了這些滯後項的影響，來減少當前數據的自相關性。
 - 這些滯後項是基於時間序列的歷史值構建的，目的是移除數據中的自相關，讓檢驗結果更加準確。

- 如果滯後期數選擇過高，可能會導致檢驗結果過度擬合數據，增加模型複雜性；如果滯後期數選擇過低，可能會忽略數據中的自相關性，導致檢驗結果不準確。
 - tradeoff:
 - 如果滯後項選擇得太少，可能無法完全消除自相關性，這樣檢驗的結果可能不準確。
 - 如果滯後項選擇得太多，可能會引入過多的變數，增加模型的複雜性和過擬合的風險。
-

可以拿來玩的東西

- volatility in
 - close
 - pct_change
 - spread btw 2 asset
 - return A - return B
 - order book : how to start ?
 - more about the details mechanism
 - how to get the data
 - volume + price
 - liquidity
-

Stochastic control approach : 市場中立

1. 概述：這種方法使用隨機控制理論來識別**最優投資組合**，即根據其他可用資產來動態地調整投資組合的持倉。該方法可以識別在不同市場條件下，如何動態調整投資，以達到最佳的風險回報平衡。
2. 適用情況：適合於面對不確定市場條件，並且需要動態調整投資組合的場景。特別是在風險管理或需要應對市場波動的情況下，這種方法能夠幫助找到最佳的資產配對交易策略。

e.g. Minimum profit optimization

- loss protection
- optimization of entry and exit rules to max and min the total profit

Time series approach

1. 概述：時間序列法基於配對資產之間的時間序列模型來構建交易策略。
 - a. 尋找能夠捕捉均值回歸的最佳交易規則。這通常包括使用自回歸模型（如 ARIMA）來分析配對的價差動態，並預測價差的回歸行為。
2. 適用情況：適合用於需要分析資產之間的動態時間行為的場景，尤其是在捕捉短期波動和均值回歸時。這種方法特別適合於金融市場中的價格序列預測，例如基於歷史數據的套利交易。

pairs trading with the Kalman filter

$$S_t = \alpha + \beta P_t + \epsilon_t$$

- **Kalman Filter** 在配對交易中被用來動態估計和更新資產之間的關聯性。該方法基於狀態空間模型進行估計，並能夠在短期內捕捉資產之間的均值回歸行為。
- Method:
 - 通常我們都會想要讓資料變成平穩，然後找到一個「固定」的underlying relationship
 - 但這邊引用狀態空間模型的概念：State space model
 - 狀態方程：描述系統的隱藏狀態hidden state 如何隨時間演變

$$x_t = F x_{t-1} + G u_t + w_t$$

- x_t ：當前時刻 t 的隱藏狀態。在配對交易中，這可能是資產價格之間的隱藏價差（即我們想捕捉但不能直接觀測到的價差）。
- F ：狀態轉移矩陣，描述隱藏狀態如何從上一步 x_{t-1} 過渡到當前狀態 x_t 。
- u_t ：可能的外生變量，在配對交易中可以包含市場的其他變量或指標，用來捕捉對價格變動的影響。
- w_t ：狀態噪音，代表隨機誤差或其他未被捕捉的因素。

- 觀測方程：

$$y_t = Hx_t + v_t$$

- y_t : 觀測數據，即我們可以觀察到的資產價格或價差。例如，兩個資產之間的價差。
- H : 觀測矩陣，描述隱藏狀態如何映射到我們觀測到的數據中。在配對交易中，這可能是兩個資產價格之間的線性關係 (例如： $y_t = \beta S_t - P_t$)。
- v_t : 觀測噪音，代表觀測過程中的隨機誤差或噪音。

- 用來描述我們實際觀測到的數據與隱藏狀態之間的關聯性。
- 也就是說，觀測到的數據是隱藏狀態的某種變現，但通常會受到噪音的影響。

ML approach

- finding profitable pairs
- framework
 - demensionality reduction
 - unsupervised learning
 - pairs selection criteria

General Method

- Spread [t] = Return A [t] - return B[t]
 - 注意是哪個資產 – 哪個資產
- Normalization
 - z score : spread [t] - spread.mean() / spread .std()
- signal :
 - enter : z score >< threshod
 - exit : z score >< threshod

Uni root

1. what is a uni root ?
- 在 time series 中，一 AR model 存在一個根 = 1
 - 若在自迴歸模型(AR) 中，序列的單根表示該序列是non stationary

- 也就是說其平均值、變異數或自相關結構隨時間變化。

2. why do we do uni root test to see if a series is stationary or non-stationary ?

most of the statistical model assume the time series to be stationary. Hence, if nonstationary (which is most of the case) , we need to make it stationary.

- 差分: to cancel trend and stabilize variance → in order to build the model
- **定態序列**：其均值、變異數和自相關結構 → 不隨時間變化
 - 數據波動較小，並且會圍繞一個固定的長期均值波動。
- **非定態序列**：其統計特性隨時間變化，例如均值和變異數隨著時間增長而發散
 - 這樣的序列往往會有趨勢或強烈的隨機漂移特徵。
- 若一個序列是非定態的：

我們需要對其進行差分或其他轉換以將其轉換 → 定態序列
這樣才能在模型中應用

Hedge ratio

1. Hedging problem is posed as the following equation

$$S_t = P_{1,t} + \sum_{n=2}^N \omega_n P_{n,t}$$

- a. where P_{-1} represents the market value at observation t of a portfolio we wish to hedge
 - b. and P_n represents a set of variables(instruments or portfolios) available for building a hedge
2. The hedging problem is in computing the vector w_n (holdings of each variable)

Code:

Ordinary least Squares regression (OLS)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

其中：

- y 是因變量（被解釋的變量）。
- x_1, x_2, \dots, x_n 是自變量（解釋變量）。
- β_0 是截距項， $\beta_1, \beta_2, \dots, \beta_n$ 是自變量的係數。
- ϵ 是誤差項。

OLS Regression Setup:

- The code sets up a regression where `price_asset1` is regressed on `price_asset2` to find the linear relationship between the two.
- The line of best fit is defined as:

$$\text{Asset 1 Price} = \alpha + \beta \times \text{Asset 2 Price} + \epsilon$$

Where:

- α is the intercept (constant).
- β is the hedge ratio (slope).
- ϵ is the residual (error term).

```
import numpy as np
import statsmodels.api as sm

# Example price data for two assets
price_asset1 = np.array([100, 102, 104, 103, 105]) # Asset 1 prices
price_asset2 = np.array([50, 51, 53, 52, 54]) # Asset 2 prices

# Add constant to the independent variable (Asset 2) for OLS regression
X = sm.add_constant(price_asset2)
model = sm.OLS(price_asset1, X).fit()

# Hedge ratio is the slope coefficient of the regression
hedge_ratio = model.params[1]
print(f"Hedge Ratio: {hedge_ratio}")
```

Note: 添加常數項是 OLS 回歸的常見操作，用來讓模型能夠估計截距，而不是強制擬合穿過原點的直線。

Hedging Error. (epsilon)

3. The hedging error is as follow

$$e(h) = S_{T+h} - S_T$$

- a. which is the error after h observations
4. Whether $e(h)$ is stationary or non-stationary in variance is a crucial problem to hedge ratio estimations (Lopez de Pedro, 2012)

- 指在進行避險交易時，[實際的避險效果] 與 [理論上] 應有的效果之間的偏差或差異。
- 避險的目的是透過建立一個相對應的對沖頭寸，來減少或消除市場風險。
然而，因為市場價格波動或模型的不準確性，實際避險並不能完美地對沖風險，這就產生了避險誤差。
- 造成避險誤差的原因：
 1. **市場波動性**：價格波動過大或過於劇烈，可能導致避險頭寸不能完全覆蓋原始頭寸的風險。
 2. **模型不準確性**：在計算避險比率時，使用的模型（如Black-Scholes模型）可能基於一些假設，但實際市場條件可能偏離這些假設，從而導致錯誤的避險比率。
 3. **時間間隔問題**：市場價格是連續波動的，而交易是間斷進行的，因此在兩次調整避險頭寸的時間之間，市場價格的變動可能會導致避險誤差。
 4. **基差風險 (Basis Risk)**：在期貨避險中，標的資產與期貨合約的價格可能並非完全同步變動，這會導致避險不完全，從而產生誤差。

Code

1. Use Price

```
import numpy as np

# 假設我們有資產 1 和資產 2 的價格數據
price_asset1 = np.array([100, 102, 105, 104, 107]) # 資產 1 的價格
price_asset2 = np.array([50, 51, 53, 52, 54]) # 資產 2 的價格

# 對沖比例（這個假設事先已經計算出來）
hedge_ratio = 1.5

# 實際價格變化: t - [t-1]
actual_price_change_asset1 = np.diff(price_asset1) # 資產 1 的價格變化
actual_price_change_asset2 = np.diff(price_asset2) # 資產 2 的價格變化

# 預期價格變化（基於資產 2 並乘以對沖比例）
expected_price_change_asset1 = hedge_ratio * actual_price_change_asset2

# 計算對沖誤差
hedging_error = actual_price_change_asset1 - expected_price_change_asset1

# 打印結果
print(f"Hedge Ratio: {hedge_ratio}")
print(f"Actual Price Changes (Asset 1): {actual_price_change_asset1}")
print(f"Expected Price Changes (Hedged): {expected_price_change_asset1}")
print(f"Hedging Error: {hedging_error}")
```

2. Use Return

```
# Actual Return
return_asset1 = np.diff(price_asset1) / price_asset1[:-1]
return_asset2 = np.diff(price_asset2) / price_asset2[:-1]

# Expected return
expected_return_asset1 = hedge_ratio * return_asset2

# Calculate the hedging error
hedging_error = return_asset1 - expected_return_asset1

# Print results
print(f"Hedge Ratio: {hedge_ratio}")
print(f"Actual Returns (Asset 1): {return_asset1}")
```

```
print(f"Expected Returns (Hedged): {expected_return_asset1}")  
print(f"Hedging Error: {hedging_error}")
```

Two main method

single period method : assume return to be iid

OLSD : OLS in difference 最小平方法回歸

ordinary least squares : min MSE

- 不考慮時間序列的自相關性或異方差性

1. Because of its simplicity, this is one of the most widely used methods (Moulton and Seydoux, 1998)

$$\Delta P_{1,t} = \alpha + \sum_{n=2}^N \beta_n \Delta P_{n,t} + \varepsilon_t$$

- a. where ΔP represents the change in market value between observations
2. Necessary condition
 - a. alpha(α) is statistically insignificant
3. Solution : $\omega_n = -\beta_n$

- Limitations

- alpha has to be zero
- epsilon(ε) is IID Normal
- assuming any change in the target portfolio(P_1) must be offset by the hedging portfolio(weighted sum of 2 to n portfolios)

MVP : min variance portfolio

Assumption :

- Efficient market
- rational investors
- all players are risk averse

1. Introduced by Markowitz (1952)
2. Settings:
 - a. ΔP observations are IID Normal
 - b. V is the covariance matrix of ΔP , where its first column represents the covariances against the portfolio we wish to hedge (P_{-1})

$$\begin{aligned} \text{Min}_{\beta} \quad & \beta' V \beta \\ \text{s.t.} \quad & \beta' a = 1 \end{aligned}$$

Target: 固定回報，讓投資組合總波動度最小

PCA :

1. The target is to compute the vector of weightings β such that $\Delta P^* \beta$ is hedged against moves of the m largest principal components (typically, $m=N-1$)
2. Our goal:
 - a. Find β which follows $W^* \beta = 0_m$
 - b. where $(W^*)`$ is the transposed eigenvector matrix after having removed the columns associated with the unhedged eigenvectors
3. This approach presents the advantage of searching for a solution which hedges against the principal sources of risk

multi period method : dynamic model

OLSL : OLS in level

1. The goal looks similar to OLSD method, but the condition that the hedge is effective when S is stationary in mean and variance gives a different approach to OLSD

$$P_{1,t} = \sum_{n=2}^N \beta_n P_{n,t} + S_t$$

2. However, as the error correction component is not separated from the observed levels in the equation, the calculated weight, beta, may not be the optimal

condition :

- **定態數據**：當數據是定態的（例如均值、變異數不隨時間變動），就可以直接使用OLS in Level進行回歸分析。這種數據的統計性質穩定，因此直接回歸可以給出可靠的結果。
- **非定態數據的風險**：如果數據是非定態的，如具有趨勢或隨機漫步特性，則使用OLS in Level可能會導致「虛假回歸（spurious regression）」，即回歸結果看似有顯著關係，但實際上這是由於數據的趨勢或單根問題導致的。

ECM :error correction model

1. If an error correction representation is verified, the series of it are cointegrated (Engle and Granger, 1987)
2. Although this model only shows a hedge ratio between two portfolios, the extension of this method will be more discussed in the advanced methods

$$\Delta p_{1,t} = \beta_0 + \beta_1 \Delta p_{2,t} + \gamma(p_{2,t-1} - p_{1,t-1}) + \varepsilon_t$$

- a. where p is the natural log of market value P
 - b. γ has to be tested positive(>0) in order to be effectively hedged
3. The optimal holdings will be $(\omega_1, \omega_2) = (1, -K)$

Advanced Method

extended version of ECM

DICKEY-FULLER OPTIMAL (DFO)

1. In the previous chapter, ECM is a dynamic model limited to two dimensions
2. This limitation could be solved through a canonical transformation of a multivariate, multi-equation specification as we did for BTCD
 - a. This approach will give stronger structure through a system of equations, each imposing an individual autoregressive equilibrium condition
3. The target is to find an optimal w where the probability of having a unit root in the spread(S) is minimized

$$S_t = P_{1,t} + \sum_{n=2}^N \omega_n P_{n,t}$$

Course and references

Book

- statistical arbitrage pairs trading strategies : review and outlook
 - christopher krauss
 - 5 distinct method