Data assimilation in financial time series data

June 17, 2025

1 Teoria

1.1 Pairs trading

En el mercado de valores, pueden existir pares de activos que estan altamente relacionados entre si, por cuestiones de que comparten mercado y objetivos, de tal manera que los retornos de los precios de los activos suelen mostrar coherencia. Generalmente estos pares de activos correlacionados se encuentran en el mismo sector. Ejemplos paradigmaticos podrian ser Coca Cola y Pepsi o dos empresas de extraccion de petroleo.

Puede haber momentos en los cuales uno de los activos de estos pares puede estar sobrevaluado o subvaluado con respecto al valor del otro activo. Si somos capaces de detectar los momentos en lo que esto sucede, luego esperamos que el activo retorne a su valor esperado por lo cual si entramos en posicion cuando detectamos estos desajustes entre los valores de los activos podemos hacernos de una ganancia cuando estos activos retornen a los valores esperados.

Dicho de otra manera estamos usando los valores de un activo para estimar el valor del otro activo considerando que ambos activos tienen los mismos drivers. De ser asi, podemos detectar cuando un activo esta sobrevaluado con respecto a su predicción. Llamamos al proceso de ajuste de los activos como reversion a la media–mean reversion.

Entonces en el pairs trading buscamos detectar desbalances entre dos activos y cuando esto ocurre la estrategia consiste en comprar el activo que esperamos su precio aumente, el activo que esta barato, cuando haya reversion a la media y vender en corto el activo que su precio esta sobrevaluado, esperando una disminucion de valor cuando ocurra la reversion a la media. En general estas estrategias que se basan en retornos que acoplan la performance de posiciones en largo, con posiciones en corto se denominan market-neutral. El pair trading es un caso particular del pair trading.

La metodologia general entonces consiste en buscar pares de activos que esten correlacionados y esten gobernados por la reversion a la media, luego usar un método predictivo que nos permita identificar en esos activos los desbalances de precios, generalmente basados en el z-score, y luego evaluar si la estrategia podría ser robusta a traves de backtesting, es decir implementación en series de tiempo ya existentes.

Comenzando con la identificación de pares es necesario introducir el concepto de cointegración y ligado a este el de series de tiempo estacionarias. Pasemos a definir estos conceptos.

1.2 Series de tiempo estacionarias

DEFINICIÓN. *Una serie de tiempo x_t es estacionaria si su media y su varianza no cambian con el tiempo y su covarianza temporal solo depende del lag d.*

Las series de tiempo se dicen tienen un orden de integración $I(d)$ si d es el número mínimo de diferencias entre elementos consecutivos de la serie para que el proceso sea estacionario. Ejemplo una serie es integrada de orden 1, $I(1)$, significa que x_t no es estacionaria pero que $\Delta x_t = x_t - x_{t-1}$ es estacionaria.

1.3 Cointegración

DEFINICIÓN. *Sean dos activos representados a través de las series de tiempo correspondientes a los precios de los activos en función del tiempo, x_t, y_t, las cuales son integradas de orden 1, $I(1)$, si existe un valor de β tal que $y_t - \beta x_t$ es estacionaria entonces decimos que las dos series estan cointegradas.*

En esencia si la diferencia entre dos activos es estacionaria eso quiere decir que las diferencias van a revertir a la media. Intuitivamente si la media de la diferencia de los activos no esta cambiando con el tiempo eso significa que cualquier tendencia entre los activos se tiene que revertir a la media–que es constante temporalmente. Para demostrar la estacionaridad se introduce el concepto de representación del error.

El concepto de cointegración y de modelo de corrección del error fue introducido por Engle and Granger (1987), galardonados con el premio nobel de economía 2002. A través de los trabajos de Engle y Granger se introdujo una forma robusta de identificar pares de activos para pair trading, los cuales son denominados como pares de activos cointegrados.

1.4 Corrección del error de modelo

Consideremos nuevamente dos series de tiempo x_t, y_t, las cuales son integradas de orden 1, $I(1)$, las cuales son cointegradas con vector de cointegración bidimensinal $(1, -\beta)$. El modelo de corrección del error viene dado por

$$\Delta y_t = \phi_x(y_{t-1} - \beta x_{t-1}) + \sum_{i=1}^{p-1} \gamma_{xi} \Delta y_{t-i} + \sum_{i=1}^{q-1} \delta_{xi} \Delta x_{t-i} + \varepsilon_t \tag{1}$$

donde γ_x, γ_y y δ_x, δ_y son los que capturan la dinámica de corto plazo. ϕ_x, ϕ_y es la velocidad de ajuste al equilibrio (ver mas adelante). Este resultado es conocido como el teorema de representación de Granger.

Los órdenes de los lags deben ser determinados por criterios de información tales como AIC o BIC.

En la practica esto se realiza a traves de dos tests de cointegracion, el metodo de Engle-Grange de dos pasos que termina en los p-values o el metodo de Johansen basado en ECM vectorial (se puede trabajar con mas de una variable/asset dependiente).

Falta detallar/comentar estas metodologias!

1.5 z-score

Si consideramos la corrección del error de orden $p = 0$,

$$\Delta y_t = \phi_x[y_{t-1} - (\beta x_{t-1} + \alpha)] + \varepsilon_t \tag{2}$$

definimos al spread por $s_t = y_{t-1} - (\beta x_{t-1} + \alpha)$.

Supongamos ya demostramos que los dos activos son $I(1)$ y que el spread es $I(0)$ entonces sabes que el par esta cointegrado y podemos operar con el. Para realizar un análisis de reversión a la media le sacamos a la serie de tiempo del spread la media y de esta manera nos aseguramos que la media es 0, ademas si consideramos que el spread tiene una desviacion estandard σ entonces

$$z_t = \frac{s_t - \overline{s}}{\sigma_s} = [y_t - \overline{y} - \beta(x_t - \overline{x})]/\sigma_s \tag{3}$$

es el z-score el cual esperamos que sea una variable estocastica gaussiana de media 0 y varianza 1.

1.6 Unit root time series

Considerando un proceso AR(1) $x_t = \phi x_{t-1} + \epsilon_t$, si estimamos ϕ por ejemplo con regresión lineal entre la serie y la serie corrida en un tiempo se tiene que:

Si $\phi < 1$	proceso estacionario
	el proceso revierte a la media
Si $\phi = 1$	proceso unit root
	No es estacionario
	la varianza aumenta
	la media no es constante.

1.7 Augmented Dickey-Fuller test

Para comprobar que dos activos estan cointegrados necesito demostrar que el spread, $s_t = y_t - \beta x_t$, es una serie estacionaria. Para esto se utiliza un test de prueba, el Augmented Dickey-Fuller test cuya hipotesis nula es que el proceso es unit root. Modelamos la serie de tiempo del spread por

$$\Delta s_t = \phi s_{t-1} + \alpha + \beta t + \sum_{i=1}^{p} \gamma_i \Delta s_{t-i} + \epsilon_t \tag{4}$$

donde $\Delta s_t = s_t - s_{t-1}$ y ϕ es tiempo en que reversiona a la media.

La prueba estadistica es:

$$\tau = \frac{\hat{\phi} - 1}{\mathrm{SE}(\hat{\phi})} \tag{5}$$

donde SE es el error estandard de la regresión lineal.

1.8 Tiempo de vida medio de reversion a la media

Si en el modelo de corrección del error solo consideramos el orden p=1 resulta

$$\Delta y_t = \alpha(y_{t-1} - \beta x_{t-1}) + \epsilon_{yt}, \tag{6}$$

para que el proceso reversione a la media α debe ser negativo. La otra variable viene dado por

$$\Delta x_t = \alpha(x_{t-1} - \beta^{-1} y_{t-1}) + \epsilon_{xt}, \tag{7}$$

entonces se tiene que definiendo el spread o error como $z_t = y_t - \beta x_t$ y restando los modelos de error corrección para ambas variables se tiene que

$$z_{t+1} = (1+\alpha) z_t + \epsilon_t. \tag{8}$$

Si consideramos una condición inicial z_0 y por el momento nos concentramos en la tendencia sin tener en cuenta el ruido se tiene que

$$z_t = (1+\alpha)^t z_0. \tag{9}$$

Si queremos determinar cuanto tiempo le lleva al proceso a perder la mitad de su valor inicial, hacemos

$$(1+\alpha)^\tau z_0 = \frac{z_0}{2} \tag{10}$$

considerando que α es negativo.

$$\tau \log(1+\alpha) = \log 1/2 \tag{11}$$

Se tiene que $\tau = -\frac{\log 2}{\log(1+\alpha)}$.

1.9 Exponente de Hurst

El exponente de Hurst (H) se define por el comportamiento del rango de

$$\mathbb{E}\left[\frac{R(N)}{S(N)}\right] = C \cdot N^H \quad \text{as } N \to \infty, \tag{12}$$

donde N es el largo de la serie de tiempo $R(N)$ es el rango (max menos min) de las desviaciones acumuladas de la media, $S(N)$ es la desviacion estandar y C es una constante.

De acuerdo a si H nos da menor o mayor a $1/2$ nos dice si tenemos una serie con tendencias o una serie que revierte a la media.

$H = 0.5$	Random Walk	No memory (e.g., efficient markets).
$0.5 < H \leq 1$	Persistent Series	Long-term memory (e.g., trending prices).
$0 \leq H < 0.5$	Anti-Persistent Series	Mean-reverting (e.g., interest rates).

Table 1: Behavior of time series based on Hurst exponent

Una definicion alternativa presentada en Sarmento y Horta es:

1.10 Implementacion de pair trading

Por lo expuesto en las secciones anteriores lo primero que debemos determinar es de que orden de integracion son las series de tiempo.

Por lo que podemos comenzando probar si x_t e y_t son I(0). Esto lo podemos ver con el test ADF o Johansen como ya los hemos introducido. En este caso podriamos trabajar con cada una por separado, aunque no sea market neutral? Si ambas son $I(0)$ y además también $s_t = y_t - \beta x_t$ es

$I(0)$ entonces esperamos que s_t tenga reversion a la media, aun cuando el par no sea cointegrado. En este caso podemos operar con el z-score como si fueran cointegradas. Cual seria la diferencia con un par cointegrado???

Si probamos que x_t e y_t son $I(1)$, para esto trabajamos con el retorno r_t^x y r_t^y y hacemos el ADF test. Si lo cumplen estamos en condiciones de hacer el ADF test al spread s_t si lo cumple el par esta cointegrado y podemos trabajar con el z-score.

Finalmente, puede suceder que una es $I(0)$ y la otra es $I(1)$ en este caso por el momento no vamos a operar.

Tomemos a un conocido par Coca-Cola (KO) Pepsi (PEP.O), en este caso lsa series de tiempo entre el 2014 y 2024 tienen un p-value de 0.48, y 0.49 respectivamente. Si tomamos las diferencias los p-values son del orden de 10^{-20} es decir que podemos considerar ambas son $I(1)$.