Statistical arbitrage strategies : from static to time-varying approaches

Minichini Carmine

Abstract

1 Teoria

1.1 Co-integration

Per introdurre il concetto di co-integrazione tra due serie temporali è utile definire il concetto di integrazione.
Una serie temporale X_t è definita integrata di ordine 1, $I(1)$, se non è stazionaria, ma è stazionaria nelle prime differenze $X_t - X_{t-1}$.
Il concetto di co-integrazione prende invece in esame due serie temporali:
due serie temporali X_t e Y_t integrate di ordine $I(1)$, sono co-integrate se la loro combinazione lineare è un processo debolmente stazionario.

$$\hat{\epsilon} = Y_t - \hat{\beta} X_t - \hat{\alpha} \tag{1}$$

Seguendo l'approccio two-step Engle-Granger(1987), per verificare la stazionarietà del processo $\hat{\epsilon_t}$, dopo aver stimato α e β con una regressione lineare , si testa la presenza di una radice unitaria tramite il test aumentato di Dickey-Fuller (ADF).

1.2 Linear State Space Models

Gli State Space Models (Modelli nello Spazio degli Stati) forniscono una metodologia generale di riferimento, molto flessibile, per affrontare una vasta gamma di problemi nell'analisi delle serie storiche. L'idea metodologica alla base degli State Space Models è che lo sviluppo nel tempo del fenomeno in analisi, $y_1....y_n$, sia determinato da una serie di vettori non osservabili $\theta_1....\theta_n$. La relazione tra y_t e θ_t specifica il modello nella forma state space.

1.2.1 Time-varying coefficients

La relazione lineare tra due variabili co-integrate è descritta da:

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{2}$$

In cui α e β sono invarianti nel tempo. La formulazione dell'equazione 2 come un modello state space Gaussiano ci permette di considerare invece l'evoluzione degli stati α e β nel tempo. Assumendo che questi si evolvano secondo un processo random walk avremo che [1]:

$$y_t = \alpha_t + \beta_t x_t + \epsilon_t \tag{3}$$

$$\alpha_t = \alpha_{t-1} + \eta_{1,t} \tag{4}$$

$$\beta_t = \beta_{t-1} + \eta_{2,t} \tag{5}$$

Il modello state space gaussiano è quindi definito da due equazioni[2]:

$$\theta_{t+1} = T_t \theta_t + \eta_t \tag{6}$$

$$y_t = Z_t \theta_t + \epsilon_t \tag{7}$$

- $\theta_t \triangleq \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ è lo stato inosservato
- $T_t \triangleq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ è la matrice di transizione degli stati
- $\eta_t \sim \mathcal{N}(0, Q)$ è detto **state transition noise** ed è un errore gaussiano a media 0 e con matrice di covarianza $Q \triangleq \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$
- $Z_t \triangleq \begin{bmatrix} 1 & x_t \end{bmatrix}$ è la matrice di osservazione, dove x_t rappresenta,nel nostro caso,un vettore di prezzi.
- $\epsilon_t \sim \mathcal{N}(0, \sigma_e^2)$ è detto **observation noise**

L'equazione (6) è definita **equazione di transizione** ed esprime la transizione dello stato θ_t attraverso una relazione lineare.
L'equazione (7) è invece definita **equazione di misurazione** ed esprime la relazione

[1] J. Durbin and S. J. Koopman, Time Series Analysis by State Space Methods, 2nd Ed. Oxford University Press, 2012

[2] Daniel P. Palomar- Portfolio Optimization with R - The Hong Kong University of Science and Technology (HKUST)

lineare tra lo stato di osservazioni y_t e lo stato latente θ_t.
Una volta specificato il modello in forma state space, la stima degli stati inosservati θ può essere ottenuta in maniera ricorsiva attraverso l'utilizzo di un Kalman Filter.

1.2.2 Inizializzazione del Kalman Filter

Il modello definito dalle equazioni (6) e (7) viene inizializzato con l'ipotesi base che i parametri σ_1^2, σ_2^2 ,responsabili dell'incertezza attorno all'evoluzione stocastica dello stato θ, siano di ordine molto piccolo, prossimo allo zero. L'ipotesi base con cui viene inizializzato il modello è quindi di assumere $\sigma_1^2 = 0.0001, \sigma_2^2 = \sigma_1^2, \sigma_e^2 = 0.001$.

2 Dati e metodologie

L'analisi empirica è stata svolta considerando i prezzi di chiusura giornalieri, aggiustati per i dividendi, dei costituenti dell'indice FTSE MiB nel periodo dal 1° gennaio 2015 al 31 dicembre 2019.[3] Il campione ottenuto è stato poi diviso in:

- **Periodo di Training**: Dal 1° gennaio 2015 al 31 dicembre 2018 (4 anni)
- **Periodo di Testing**: Dal 1° gennaio 2019 al 31 dicembre 2019 (1 anno)

L'analisi procede come segue:

1. **Cointegration Test:** Nel periodo di training, per i 40 titoli dell'indice è stata testata la relazione di co-integrazione tramite il metodo two-step Engle-Granger.
2. **Pairs Selection:** Stimata la relazione di co-integrazione, escluse le coppie non co-integrate,sono state selezionate le coppie dello stesso settore di appartenenza e ordinate per l'indice di correlazione di Pearson.
3. **Statistical Arbitrage**: Nel periodo di testing, sono state implementate le 3 diverse strategie di trading.
4. **Performance**: Valutazione della performance

2.1 Risultati della selezione

Table 1: Risultati

[3] I prezzi di chiusura giornalieri, aggiustati per i dividendi, sono stati ottenuti tramite Yahoo Finance

Pair	ρ	ADF Statistic	p-value	Settore
DIA.MI - AMP.MI	0.97	-3.66	0.026	Sanitario
UCG.MI - BPE.MI	0.94	-3.73	0.022	Bancario
BMED.MI - BGN.MI	0.80	-4.43	0.01	Servizi Finanziari
BMED.MI - AZM.MI	0.61	-3.98	0.01	Servizi Finanziari

3 Statistical Arbitrage

Per il periodo di testing sono state considerate 3 strategie di trading, la prima basata sull'approccio classico di co-integrazione, in cui è stato calcolato lo spread con le stime dei parametri ottenuti tramite la regressione lineare nel periodo di training. La seconda metodologia utilizzata prevede la stima, a intervalli regolari di 2 mesi, dei coefficienti $\hat{\alpha}$ e $\hat{\beta}$ nel periodo di testing, i parametri stimati nella finestra precedente sono stati poi utilizzati per calcolare lo spread nella finestra successiva. Il terzo approccio prevede una stima dinamica dei parametri α e β attraverso l'utilizzo di un Kalman Filter. Calcoliamo lo spread tra le due serie considerate nel periodo di testing come:

$$z_t = log(stock_{1,t}) - \hat{\beta} log(stock_{2,t}) - \hat{\alpha} \tag{8}$$

Gatev et al.(2006)[4] nel loro lavoro, adottano come threshold per una strategia di Pairs Trading un valore pari a 2 volte la deviazione standard dello spread calcolato nel periodo di training. Tuttavia un problema dell'adottare una threshold con l'approccio di Gatev et al. risiede nel considerare una soglia eccessivamente alta per le oscillazioni nel periodo di testing, portandoci nella maggior parte dei casi a non avere segnali di entrata o uscita. Nella nostra casistica è stato scelto quindi di adottare una valore soglia costante per tutte le metodologie sviluppate, con due scopi: generare segnali molto aggressivi e riuscire ad equiparare le performance per tutti i modelli.
I segnali per la strategia (± 1) sono così generati:

- Se : $z_t > threshold$. significa che $stock_{1,t}$ è sopravvalutato rispetto a $stock_{2,t}$, per cui vendiamo 1 unità di $stock_{1,t}$ e compriamo $\hat{\beta}$ unità di $stock_{2,t}$ Chiudiamo la posizione quando $z_t \leq 0$
- Se: $z_t < -threshold$. significa che $stock_{1,t}$ è sottovalutato rispetto a $stock_{2,t}$, per cui compriamo 1 unità di $stock_{2,t}$ e vendiamo $\hat{\beta}$ unità di $stock_{1,t}$ Chiudiamo la posizione quando $z_t \geq 0$

[4]Gatev, E., Goetzmann,W. and Rouwenhorst, K. Pairs Trading: Performance of a Relative-Value Arbitrage Rule, Review of Financial Studies 19(3), 797–827. 2006

L'intera strategia di trading è equivalente ad avere un portafoglio con pesi[5]

$$w \triangleq \begin{bmatrix} 1 \\ -\hat{\beta} \end{bmatrix} \tag{9}$$

sotto il vincolo che $w = 1$ il portafoglio costruito per la strategia è :

$$w \triangleq \begin{bmatrix} 1/(1+\hat{\beta}) \\ -\hat{\beta}/(1+\hat{\beta}) \end{bmatrix} \tag{10}$$

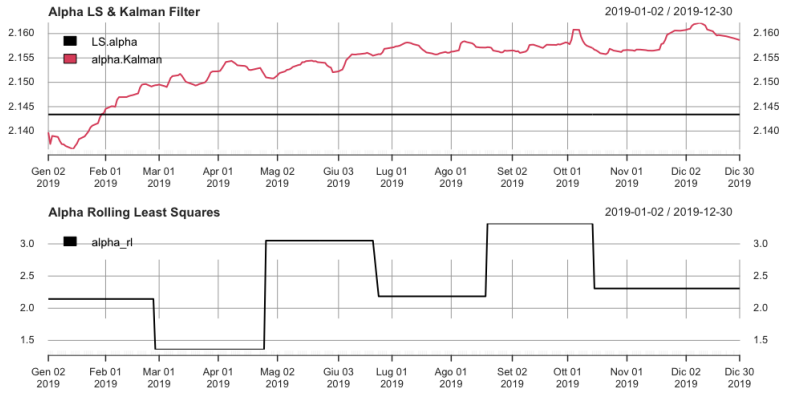
mentre i ritorni del portafoglio considerato sono stati poi calcolati come:

$$w^T \Delta log(stocks) \tag{11}$$

4 Performance

La prima coppia considerata è formata dalle multinazionali **DiaSorin S.p.A.** e **Amplifon S.p.A.**, entrambe operanti nel settore medico-sanitario.
La figure 1 e 2 mostrano i parametri $\hat{\alpha}$ e $\hat{\beta}$ stimati con le tre metodologie:

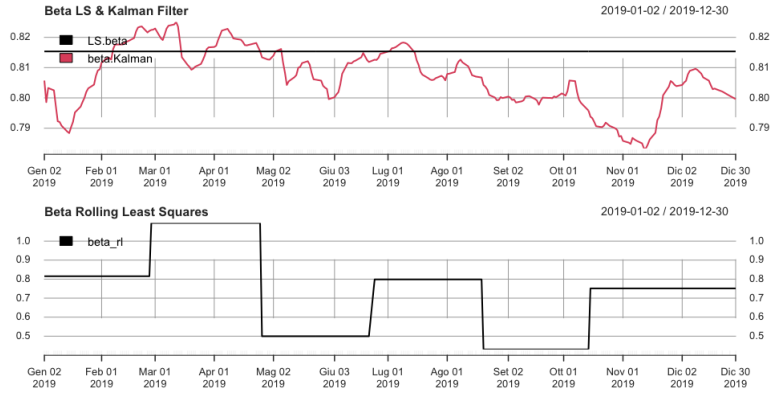
Figure 1: $\hat{\alpha}$



Con i parametri ottenuti con i tre diversi approcci è stato calcolato poi lo spread come definito in (8). Al fine di equiparare le 3 diverse metodologie, per la coppia **DIA.MI-AMP.MI** è stato scelto un valore soglia di 0.02.
I **P&L** delle tre diverse strategie sono visibili nella figura 4.

[5]Yiyong Feng and Daniel P. Palomar (2016), "A Signal Processing Perspective on Financial Engineering"

Figure 2: $\hat{\beta}$



Di seguito la valutazione delle tre performance:

Table 2: Performance evaluation

Metodologia	Sharpe Ratio	MDD	AR	RMSE
Kalman Filter	1.85	0.02	0.16	0.013
Co-Integration	3.45	0.05	0.42	0.037
Rolling Regression	3.56	0.05	0.41	0.032

Figure 3: $z_t = log(stock_{1,t}) - \hat{\beta} log(stock_{2,t}) - \hat{\alpha}$

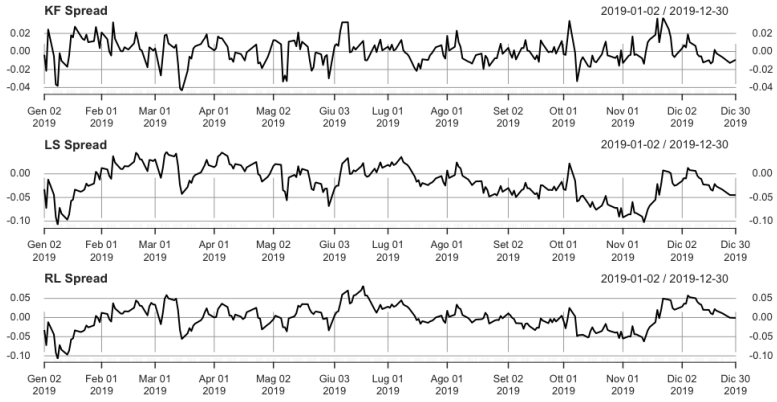
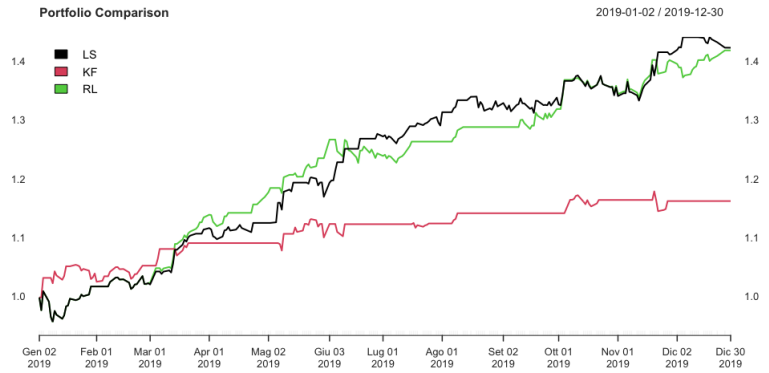


Figure 4: Portfolio P&L




```
#initial position
signal[1] <- 0
if (Z_score[1] <= threshold_long[1]) {
        signal[1] <- 1
} else if (Z_score[1] >= threshold_short[1])
signal[1] <- -1

# loop
for (t in 2:nrow(Z_score)) {
        if (signal[t-1] == 0) {  #if we were in no position
                if (Z_score[t] <= threshold_long[t]) {
                        signal[t] <- 1
                } else if(Z_score[t] >= threshold_short[t]) {
                        signal[t] <- -1
                } else signal[t] <- 0
        } else if (signal[t-1] == 1) {  #if we were in a long position
                if (Z_score[t] >= 0) signal[t] <- 0
                else signal[t] <- signal[t-1]
        } else {  #if we were in a short position
                if (Z_score[t] <= 0) signal[t] <- 0
                else signal[t] <- signal[t-1]
        }
}
```