Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

ARBITRAGGIO STATISTICO SU COMMODITY

Tesi di Laurea in Finanza Matematica

Relatore:
Chiar.mo Prof.
Andrea Pascucci

Presentata da:
Martina Tombari

II Sessione di Laurea
Anno Accademico 2012/13

...si campa anche senza sapere
che cos'è un logaritmo, senza sapere suonare
uno strumento musicale, senza conoscere
il nome del vento che ti soffia in faccia
o del fiore che regali a qualcuno.
Ma di sicuro ci si perde qualcosa.
Robert Ghattas

Indice

Introduzione

La Finanza Matematica è una disciplina che si occupa di risolvere problemi di carattere economico-finanziario attraverso teorie matematiche. Per risolvere matematicamente un problema finanziario, si realizza un *modello matematico probabilistico* che descriva il fenomeno considerato, cercando di semplificare la realtà. In seguito si applicano *principi economici* per passare razionalmente dal modello al problema che ci interessa. Questi principi sono alla base del funzionamento dell'economia. In questa tesi ci focalizzeremo in particolare sul *principio di non arbitraggio*, secondo il quale, se siamo certi che due titoli in una data futura avranno lo stesso valore, allora questi dovranno avere lo stesso valore anche in ogni data precedente. Se non valesse questo principio, si potrebbe facilmente fare **arbitraggio**, che consiste in un'operazione finanziaria illecita *a costo zero* e che permette di realizzare un *profitto privo di rischio.* Ad esempio, basterebbe vendere allo scoperto (cioè attestando presso la banca o un altro ente che in futuro tali titoli verranno riacquistati) il titolo più costoso e, con i soldi ricavati acquistare il meno costoso.
In ambito matematico non si può rappresentare il principio di non arbitraggio, ma in ogni *modello di mercato efficiente* si assume che debba valere, eliminando dunque la possibilità di fare soldi gratis.
Tuttavia, i mercati reali non sono sempre efficienti come si ipotizza, ma evidenze sperimentali mostrano la presenza di inefficienze temporanee, che generano opportunità di arbitraggio, dette *Intermarket spreads*, che vengono studiate nell'ambito dell'**arbitraggio statistico**, oggetto di questa tesi. In generale, si creano opportunità di arbitraggio statistico quando si riescono ad individuare delle componenti sistematiche nelle dinamiche dei prezzi di alcuni asset che si muovono con regolarità persistenti e prevalenti, e dunque per gli arbitraggisti è facile prevedere l'andamento di tali prezzi e sviluppare strategie di trading per speculare. Gli intermarket spreads riguardano simultanei acquisto e vendita di commodity differenti ma connesse tra loro, cioè che sono legate da una relazione stabile relativamente ai prezzi. Perturbazioni casuali della domanda e dell'offerta nei mercati possono causare divergenze nei prezzi, dando luogo a opportunità di intermarket spread.

In questa tesi cercheremo di fornire un quadro generale dell'arbitraggio statistico, rappresentando matematicamente concetti econometrici e statistici, che vengono ampiamente utilizzati.
Nel Capitolo 1 vengono presentati gli strumenti matematici e finanziari che sono indispensabili per affrontare la metodologia di arbitraggio statistico, come i processi stocastici e le strategie di arbitraggio puro e statistico. Vengono formalizzati molti concetti di carattere economico e statistico, come la rappresentazione

delle serie storiche mediante varie classi di processi stocastici, che vengono definite con le loro specifiche proprietà.
Nel Capitolo 2 viene descritto come i concetti matematici del capitolo precedente si applicano all'arbitraggio statistico, costruendo nuovi processi con particolari proprietà, che vengono verificate mediante appositi test statistici, di cui viene data una generica descrizione.
Infine nel Capitolo 3 viene presentata una particolare strategia di trading di tipo implicito (ISA), che viene poi applicata ad un esempio reale, sul quale viene implementata e verificata la metodologia di arbitraggio statistico proposta in questa tesi.

Capitolo 1

Background

In questo capitolo vengono analizzati gli strumenti matematici fondamentali per affrontare il tema dell'arbitraggio statistico.

1.1 Processi stocastici discreti

Definizione 1.1 (Processo stocastico discreto)**.** Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathrm{P})$, un processo stocastico discreto è una successione di variabili aleatorie definite sullo spazio di probabilità Ω:

$$(X_n)_{n\geq 0}$$

dove n è l'indice temporale, il termine "discreto" si riferisce al fatto che l'evoluzione temporale è discreta, cioè il numero di istanti è finito, X_n è la variabile aleatoria che in questa tesi rappresenta il fenomeno aleatorio che descrive il prezzo di un titolo all'istante n-esimo. Un processo stocastico discreto risulta dunque particolarmente utile nel nostro caso, in cui vogliamo modellizzare l'evoluzione sul mercato dei prezzi dei titoli. Lavoreremo dunque sempre su spazi di probabilità in cui $\Omega = \mathbb{R}_+$, poiché ogni numero reale positivo può essere visto come possibile prezzo.
La σ-algebra $\mathcal{F}$ è la σ-algebra di Borel $\mathcal{B}$, la più piccola σ-algebra che contiene come eventi gli intervalli aperti.
P è una misura di probabilità generica, in quanto ai fini della nostra trattazione non è necessario conoscerla.

Notazione 1.1. Ogni volta che si parla di processo stocastico discreto, è implicita l'assunzione di lavorare nello spazio di probabilità $(\mathbb{R}, \mathcal{B}, \mathrm{P})$

Osservazione 1.1. Un processo stocastico discreto $(X_n)_{n\geq 0}$ definisce in modo naturale una filtrazione $\mathcal{F}_n^x = \sigma(X_0 ... X_n) \quad \forall n$. $(\mathcal{F}_n^x)_{n\geq 0}$ si chiama *filtrazione associata al processo* $(X_n)_{n\geq 0}$, ed è la σ algebra generata da $X_0 .. X_n$.
Essa contiene tutte le informazioni sul processo istante per istante.

Definizione 1.2. Un processo stocastico X_n si dice adattato a $(\mathcal{F}_n)_{n\geq 0}$ se X_n è $\mathcal{F}_n$-misurabile $\forall n \geq 0$. Equivalentemente, $\mathcal{F}_n^x \subseteq \mathcal{F}_n \quad \forall n$

La filtrazione naturale è dunque la più piccola rispetto alla quale il processo stocastico è adattato.
Definiamo ora una classe importanti di precessi stocastici.

Definizione 1.3 (Martingala)**.** Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathrm{P}, (\mathcal{F}_n))$, un processo stocastico $(X_n)_{n\geq 0}$ si dice martingala se:

1. $E[|X_n|] < +\infty \quad \forall n$
2. $E[X_{n+1}|\mathcal{F}_n] = X_n \quad \forall n \geq 0$

Se X_n è il processo che rappresenta i prezzi di un titolo, e $\mathcal{F}_n$ rappresenta l'insieme delle informazioni al tempo n, la miglior stima del prezzo di domani in base alle informazioni che ho oggi è il prezzo di oggi.

Osservazione 1.2. $E[X_n]] = E[E[X_{n+1}|\mathcal{F}_n]] = E[X_{n+1}] \quad \forall n$, per una proprietà dell'attesa condizionata. cioè una martingala ha valore atteso costante per ogni n, dunque $E[X_n] = E[X_0] \quad \forall n$.

Una martingala è un processo stocastico che ha un'evoluzione nel tempo ma rimane costante in media.

Osservazione 1.3. $X_n = E[X_{n+1}|\mathcal{F}_n] = E[E[X_{n+2}|\mathcal{F}_{n+1}]|\mathcal{F}_n = E[X_{n+2}|\mathcal{F}_n]$, per una proprietà dell'attesa condizionata. In generale,
$X_n = E[X_{n+k}|\mathcal{F}_n]] \quad \forall n, k \geq 0$ Possiamo dunque estendere la seconda proprietà della martingala, che vale anche per tutti gli istanti futuri, non solo per quelli successivi.

Osservazione 1.4. $E[X_{n+1}|\mathcal{F}_n] = X_n \Rightarrow X_n$ è adattato a $\mathcal{F}_n$. Infatti, dalla definizione di attesa condizionata, X_n è $\mathcal{F}_n$-misurabile. Una martingala è quindi necessariamente adattata a $\mathcal{F}_n$.

Il seguente teorema fornisce una rappresentazione che useremo in seguito dei processi stocastici adattati.

Teorema 1.1 (di decomposizione di Doob)**.** *Sia $(X_n$ un processo stocastico adattato su $(\Omega, \mathcal{F}, P, (\mathcal{F}_n))$. Esistono e sono unici una martingala $(M_n)_{n\geq 0}$ e un processo predicibile $(A_n)_{n\geq 0}$, tale che $A_0 = 0$, t.c.:*

$$X_n = M_n + A_n \quad \forall n \geq 0 \tag{1.1}$$

cioè un processo adattato si decompone in modo unico nella somma di un processo costante in media e un processo predicibile.

1.2 Strategia

Definizione 1.4 (Portafoglio)**.** Un *portafoglio o strategia* è un processo stocastico in $\mathbb{R}^{d+1}$

$$(\alpha, \beta) = (\alpha_n^1, ..., \alpha_n^d, \beta_n)_{n=1...N}$$

dove α_n^i rappresenta il numero dei titoli S^i e β_n il numero dei bond, presenti nel portafoglio nel periodo n-esimo $[t_{n-1}, t_n]$.

Definizione 1.5. Si definisce *valore della strategia* (α, β) *nel periodo n-esimo* $[t_{n-1}, t_n]$

$$V_n^{(\alpha,\beta)} = \alpha_n S_n + \beta_n B_n = \sum_{i=1}^{d} \alpha_n^i S_n^i + \beta_n B_n \quad n = 1...N \tag{1.2}$$

Definiamo inoltre il *valore iniziale della strategia*

$$V_0^{(\alpha,\beta)} = \alpha_1 S_0 + \beta_1 B_0 = \sum_{i=1}^{d} \alpha_1^i S_0^i + \beta_1 B_0 \quad n = 1...N \tag{1.3}$$

Osserviamo che α_n^i e β_n possono assumere anche valori negativi. Questo accade nel caso della *vendita allo scoperto*, che consiste nella vendita di titoli non posseduti direttamente dal venditore.

Definizione 1.6. Una strategia (α, β) si dice *autofinanziante* se vale la seguente relazione

$$V_{n-1} = \alpha_n S_{n-1} + \beta_n B_{n-1} \quad \forall n = 1, ...N \tag{1.4}$$

Dall'Eq.1.2, segue che vale la seguente uguaglianza per un portafoglio autofinanziante

$$\alpha_{n-1} S_{n-1} + \beta_{n-1} B_{n-1} = \alpha_n S_{n-1} + \beta_n B_{n-1} \tag{1.5}$$

In pratica, al tempo t_{n-1}, si costruisce la strategia (α, β) per il periodo $[t_{n-1}, t_n]$, avendo a disposizione il capitale V_{n-1}, in modo da non mutare il valore complessivo del portafoglio.

Definizione 1.7. Una *strategia* (α, β) *è predicibile se* (α_n, β_n) è $\mathcal{F}_{n-1}$-misurabile $\forall n = 1...N$ Un portafoglio è predicibile dunque se siamo a conoscenza delle informazioni sul titolo fino all'istante t_{n-1}, quando costruiamo la nostra strategia.

Osservazione 1.5. Dalle Eq.1.2 e 1.4 segue

$$V_n - V_{n-1} = \alpha_n (S_n - S_{n-1}) + \beta_n (B_n - B_{n-1})$$

La variazione dell'investimento dipende solo dalla variazione dei prezzi e da come si è investito, non dal fatto che si sono aggiunti o tolti soldi al capitale iniziale. Sommando su n si ottiene

$$V_n - V_0 = \sum_{k=1}^{n} \alpha_k (S_k - S_{k-1} + \beta_k (B_k - B_{k-1})$$

Il rendimento della strategia adottata è uguale alla somma dei rendimenti dei singoli periodi.

Uno dei principi economici più importanti, sul quale si basano quasi tutti i modelli di valutazione di Finanza Matematica, è il *Principio di non arbitraggio'*.

1.3 Principio di non arbitraggio

Se X_t e Y_t rappresentano il prezzo al tempo t di due titoli rischiosi, secondo il principio di non arbitraggio

$$X_T = Y_T \quad \Longrightarrow \quad X_t = Y_t \quad \forall t \leq T \tag{1.6}$$

Ciò significa che se attualmente due titoli hanno lo stesso valore, allora anche in ogni istante precedente ad oggi avevano lo stesso valore. Se $X_t < Y_t$, l'arbitraggio consiste nel vendere allo scoperto Y_t e, con i soldi guadagnati nella vendita, comprare X_t. Se poi, al tempo T, $X_T = Y_T$ vendiamo X_T e compriamo Y_T. In questo modo abbiamo guadagnato soldi a costo 0, cioè senza investire nulla, e senza accollarci dei rischi. Questa operazione è chiamata arbitraggio ed è un'operazione illecita.

1.4 Arbitraggio puro e arbitraggio statistico

Introduciamo, di seguito, la nozione formale di strategia d'arbitraggio puro:

Definizione 1.8 (Arbitraggio)**.** Un *arbitraggio* è una strategia (α, β) autofinanziante e predicibile che verifica le seguenti condizioni:

1. $V_0^{(\alpha,\beta)} = 0$
2. $V_N^{(\alpha,\beta)} \geq 0 \quad \forall \omega \in \Omega$
3. $P(V_N^{(\alpha,\beta)} > 0) > 0 \quad \forall \omega \in \Omega$

Un arbitraggio è dunque una strategia che non richiede un investimento iniziale poiché ha costo zero (1), non espone ad alcun rischio poiché il suo valore finale è positivo o nullo (2) e ha probabilità positiva di assumere un valore positivo (3).

Osservazione 1.6. Dalla Definizione 1.1 segue che

$$E[V_N^{(\alpha,\beta)}] > 0 \quad \forall \omega \in \Omega$$

cioè il valore atteso della strategia è positivo

Nei modelli di mercato si cerca di fare in modo di non avere arbitraggi.

Definizione 1.9. Un *modello di mercato* (S,B) è *libero da arbitraggi* se la famiglia delle strategie autofinanzianti e predicibili non contiene arbitraggi.

Realizzare modelli di mercato con questa proprietà è abbastanza complicato. Tali modelli dovrebbero rispecchiare infatti i mercati reali, che si assume siano efficienti, cioè tali che non sia possibile costruire strategie di trading che generano profitti. Tuttavia, alcuni risultati empirici sull'andamento dei prezzi di alcuni beni, mostrano temporanee inefficienze di mercato. Si genera così l'opportunità di intraprendere un tipo di strategia nota come *arbitraggio statistico.*

Definizione 1.10 (Arbitraggio statistico)**.** Un *arbitraggio statistico* è una strategia (α, β) autofinanziante e predicibile che verifica le seguenti condizioni:

1. $V_0^{(\alpha,\beta)} = 0$
2. $V_N^{(\alpha,\beta)} \in \mathbb{R} \quad \forall \omega \in \Omega$
3. $E[V_N^{(\alpha,\beta)}] \geqslant 0 \quad \forall \omega \in \Omega$

Al contrario dell'arbitraggio puro, l'arbitraggio statistico può dare valori della strategia negativi in alcuni istanti temporali. In questo modo il valore medio della stategia può non essere sempre positivo in ogni tempo.

In generale, si creano opportunità di arbitraggio statistico quando si riescono ad individuare delle componenti sistematiche nelle dinamiche dei prezzi di alcuni asset che si muovono con regolarità persistenti e prevalenti. Di conseguenza per gli arbitraggisti è facile prevedere landamento di tali prezzi e sviluppare strategie di trading per speculare.

1.5 Rappresentazione di serie storiche

Definizione 1.11. Una serie storica è una successione di numeri reali $S_t = \{S_0, ., S_N\}$ che rappresenta un insieme di osservazioni effettuate nell'intervallo di tempo $[0, T]$, partizionato in N+1 istanti $\{t_i | i \in \mathbb{N}_0 \quad i = 0..N\}$.

In questa tesi le osservazioni delle serie storiche sono costituite dai prezzi dei titoli presenti sul mercato azionario.
Per poter lavorare finanziariamente su una serie storica ed utilizzare le informazioni in essa contenute, è necessario realizzare un modello matematico che la rappresenti. Nel caso di una serie storica, tale modello è rappresentato da un processo stocastico discreto. Le due complicazioni chiave che si incontrano in questo procedimento sono le proprietà, insite nelle serie, di "**rumore**" e "**non stazionarietà**", che determinano ampiamente la scelta della rappresentazione da adottare.

La scelta di un processo target o di una variabile "dipendente" y_t è forse l'unica decisione più importante nell'intero processo di modellizzazione. All'interno di questa tesi, considereremo un particolare tipo di processo stocastici discreti y_t, la cui rappresentazione è

$$y_t = z_t + \varepsilon_t \quad t \in \mathbb{N} \quad \varepsilon_t \sim \mathcal{N}_{(0,\sigma^2)} \tag{1.7}$$

dove z_t è un processo stocastico discreto e il processo stocastico ε_t è costituito da una successione di variabili, che hanno distribuzione normale con media zero e varianza costante σ^2. Se y_t è un processo adattato, per il teorema di Doob è possibile rappresentarlo in modo unico come in Eq.1.7, dove z_t è una martingala, ottenendo un caso particolre della Eq.1.1 del teorema di Doob.

Definizione 1.12. Un processo stocastico discreto y_t si dice autoregressivo se è della forma

$$y_t = f(y_{t-1}) + \varepsilon_t \quad t \in \mathbb{N} \quad \varepsilon_t \sim \mathcal{N}_{(0,\sigma^2)} \tag{1.8}$$

dove $f(y_{t-1})$ è un processo stocastico discreto e il processo stocastico ε_t è costituito da una successione di variabili, che hanno distribuzione normale con media zero e varianza costante σ^2.

In molte applicazioni nell'ambito della previsione, il tema della rappresentazione è trascurato, in gran parte perché la serie target è spesso data. Tuttavia in un'applicazione come l'arbitraggio statistico ci sono molte possibili combinazioni di beni che potrebbero essere usate come serie target, ed identificare combinazioni di processi stocastici che abbiano la componente stocastica con varianza più piccola possibile e dunque una forte componente potenzialmente predicibile, è un obiettivo chiave del primo passo della nostra metodologia di arbitraggio statistico.

La seconda proprietà chiave che è determinata dalla rappresentazione adottata è quella di "non stazionarietà", o variazione temporale delle proprietà statistiche. La stazionarietà si riferisce alle caratteristiche del processo stocastico che modellizza la serie storica.

Definizione 1.13. Si consideri un processo stocastico discreto autoregressivo y_t, $t \in \mathbb{N}_0$.
Tale processo si dice stazionario se verifica le seguenti condizioni:

$$\exists \mu \in \mathbb{R} \quad \text{t.c.} \quad E[y_t] = \mu \quad \forall t \tag{1}$$

$$\exists \gamma(\tau), \tau \in \mathbb{N} \quad \text{t.c.:} \quad \begin{cases} E[(y_t - \mu)^2] = \gamma(0) & \forall t \qquad (2) \\ E[(y_t - \mu)(y_{t-\tau} - \mu)] = \gamma(\tau) & \tau = 1, 2, ... \quad \forall t \qquad (3) \end{cases}$$

$$p(y_t, y_{t+1}, y_{t+2}, ...) = p(y_{t+\tau}, y_{t+\tau+1}, y_{t+\tau+2}, ...) \qquad \forall t \quad \forall \tau \tag{4}$$

Le prime tre condizioni si riferiscono rispettivamente a valore atteso, varianza e covarianze del processo, mentre la quarta riguarda la **stazionarietà assoluta**:

1. Un processo stocastico stazionario deve avere valore atteso costante, indipendente dall' istante temporale considerato.
2. Un processo stocastico stazionario deve avere varianza costante, mentre una processo non stazionario ha una varianza che tende all'infinito.
3. Le covarianze tra le variabili aleatorie t e $t - \tau$ dipendono solo da quanto esse sono lontane, cioè sono funzioni di τ, ma non di t. Non tendono a crescere o decrescere con t.
4. La quarta condizione è una condizione per la "stazionarietà forte" riguarda la funzione di densità di probabilità e richiede che la distribuzione congiunta di una successione di variabili aleatorie sia indipendente dal tempo di riferimento in cui sono state ricavate.

Quando le caratteristiche del processo stocastico cambiano nel tempo si può parlare di processo **non stazionario**. All'interno della classe dei processi non stazionari è presente un tipo di processo autoregressivo di grande utilità nella nostra trattazione, il processo **random walk**.

Definizione 1.14. Si consideri un processo stocastico discreto non stazionario e autoregressivo y_t, $t \in \mathbb{N}_0$.
Tale processo si dice **random walk** se vale la seguente condizione autoregressiva:

$$y_t = y_{t-1} + \varepsilon_t \quad \forall t \in \mathbb{N} \tag{1.9}$$

$$\varepsilon_t \quad i.i.d. \quad \varepsilon_t \sim \mathcal{N}_{0,\sigma^2} \quad \forall t \in \mathbb{N}$$

Questo processo è un processo autoregressivo, dunque può essere descritto come in Eq.1.8, in cui $f(y_{t-1}) = y_{t-1}$. Dalla definizione segue che in un processo random walk le variabili aleatorie degli incrementi $y_t - y_{t-1}$, oltre ad essere tra loro indipendenti, sono indipendenti anche dalle variabili $y_{t-1}..y_0$.
Ponendo $y_0 = \eta$, si ha che:

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i \quad \forall t \tag{1.10}$$

Il valore atteso e la varianza delle variabili aleatorie del processo hanno la seguente forma:

$$\begin{aligned}
E(y_t) &= E(y_0 + \sum_{i=1}^{t} \varepsilon_i) \\
&= E(\eta) + \sum_{i=1}^{t} E(\varepsilon_i) \quad \text{per l'indipendenza} \\
&= \eta \quad \forall t \\
V(y_t) &= V(y_0 + \sum_{i=1}^{t} \varepsilon_i) \\
&= V(\eta) + \sum_{i=1}^{t} V(\varepsilon_i) \quad \text{per l'indipendenza} \\
&= t\sigma^2
\end{aligned}$$

In definitiva:

$$E(y_t) = \eta \quad \forall t \tag{1.11}$$

$$V(y_t) = t\sigma^2 \quad \forall t \tag{1.12}$$

Tale processo, come già precisato, è non stazionario. Infatti, la varianza di un processo random walk cresce linearmente in funzione del tempo, incrementando all'infinito.

Osservazione 1.7. Un processo random walk è una martingala, cioè è tale che il valore assunto domani da una variabile aletoria sia pari al valore di oggi più un valore imprevedibile.

$$E(\varepsilon_t | y_{t-1}, y_{t-2}, .., y_0) = 0 \quad \text{per indipendenza}$$

$$\begin{aligned}
E(y_t | y_{t-1}, y_{t-2}, .., y_0) &= E(y_{t-1} | y_{t-1}, y_{t-2}, .., y_0) \quad \text{da 1.9 e per indipendenza} \\
&= y_{t-1} \quad \text{per proprietà della media condizionata}
\end{aligned}$$

Dunque per un processo random walk vale il Teorema di Doob, e si può rappresentare dunque come da Eq.1.1.

Questo modello implica che l'evoluzione del prezzo di uno strumento finanziario è essenzialmente casuale e la migliore previsione del prezzo dell'istante

successivo non può che essere fatta sulla base del prezzo precedente e delle informazioni disponibili sul mercato. Pertanto i movimenti dei prezzi e dei rendimenti non seguono alcun trend o regolarità, quindi i movimenti passati non possono essere usati per previsioni future.
Infine, un ultimo tipo di processo stocastico discreto (che può essere stazionario o meno) descriviamo è il processo di **mean reversion**, che significa propriamente "ritorno al livello medio".

Definizione 1.15. Si consideri un processo stocastico discreto autoregressivo y_t, $t \in \mathbb{N}_0$.
Tale processo si dice **mean reverting** se vale la seguente condizione ricorsiva:

$$\begin{cases} X_0 > 0 \\ X_t = a(b - X_{t-1}) + \varepsilon_t \quad a \in \mathbb{R}, a > 0 \end{cases}$$

$$\varepsilon_t \quad i.i.d. \quad \varepsilon_t \sim \mathcal{N}_{0,\sigma^2} \quad \forall t \in \mathbb{N}$$

Osservazione 1.8. Per un processo mean reverting valgono le seguenti condizioni sulla media $\begin{cases} E[X_t] > X_{t-1} & se \quad X_{t-1} < b \\ E[X_t] < X_{t-1} & se \quad X_{t-1} > b \end{cases}$

Il fenomeno della **mean reversion** riguarda dunque la tendenza delle osservazioni di una serie storica, ad esempio di prezzi di titoli, ad essere attratte verso un determinato valore, che rappresenta il loro valore medio nel lungo periodo. Tale comportamento si può osservare bene nella figura seguente:

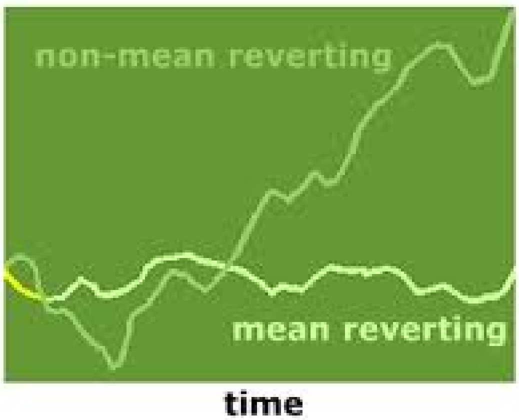

Figura 1.1: Confronto tra un comportamento mean-reverting e non. Nel caso di mean-reversion si osservano continue crescite e decrescite dei prezzi attorno ad un determinato valore, il valore medio dei prezzi calcolato nel perodo di osservazione. Al contrario, nel caso generale, l'andamento dei prezzi è puramente casuale, che non è indice di alcun comportamento particolare.

Tornando ai processi stazionari, se il processo che modellizza una serie storica è tale, possiamo utilizzare le informazioni passate della serie storica e prevedere il comportamento futuro del prezzo del titolo osservato. Tale tecnica di modellizzazione è chiamata *autoregressione* e si basa proprio su quest'assunzione di **stazionarietà**, o stabilità delle proprietà statistiche del sistema nel tempo. Si può mostrare facilmente che un'applicazione soddisfacente di strumenti di modellizzazione statistica dipende spesso dall'assunzione di stazionarietà; si consideri un processo stocastico non stazionario in media, cioè che viola la prima condizione, come mostrato in Figura 1.2.

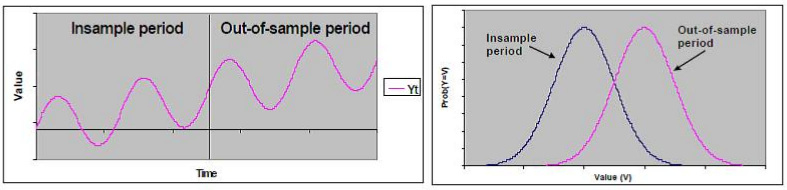

Figura 1.2: Un esempio di processo stocastico che mostra non stazionarietà in media; il grafico di sinistra mostra il processo stocastico y_t; il grafico di destra illustra come la distribuzione di probabilità della normale sia differenti durante un periodo campione e il periodo futuro, in cui si riporta una previsione dell'andamento del modello

Vediamo ora alcune proprietà dei processi stocastici autoregressivi.

Definizione 1.16 (Integrazione)**.** Si consideri un processo stocastico discreto autoregressivo y_t, $t \in \mathbb{N}$.
Le variabili aleatorie che formano il processo stocastico $\Delta y_t = y_{t+1} - y_t, \quad t \in \mathbb{N}_0$ sono dette *incrementi di primo ordine*.
Analogamente, le variabili aleatorie che formano il processo stocastico $\Delta y_{d,t} = y_{t+1} - y_t, \quad t \in \mathbb{N}_0, d = 1, 2, ...$ sono dette *incrementi di ordine d*.
Il processo y_t si dice integrato di ordine 1, e si indica con $y \sim I(1)$ se il processo degli *incrementi di primo ordine* $\Delta y_{1,t}$ è stazionario.
Il processo si dice integrato di ordine d, e si indica con $y \sim I(d)$, se il processo degli *incrementi di ordine d* $\Delta y_{d,t}$ è stazionario.
Da questa prospettiva, un processo stocastico stazionario è integrato di ordine 0, $y \sim I(0)$.

Alcuni processi stocastici possono non essere né completamente stazionari né completamente non stazionari, ma invece possono essere *debolmente integrate*, cioè $I(d)$ dove $0 < d < 1, d \in \mathbb{R}$.

Un'altra nozione molto importante nel contesto dei processi stocastici è quello di *cointegrazione*. Tale concetto è strettamente legato al concetto di *integrazione* ma riguarda una relazione tra un certo numero di processi stocastici piuttosto che una proprietà individuale del processo. Diamo ora due definizioni equivalenti di cointegrazione.

Definizione 1.17 (Cointegrazione 1.1)**.** Si consideri il processo stocastico vettoriale discreto $y_{i,t} = (y_{0,t}, y_{1,t}, ..y_{n,t})$, $y_{i,t} \in \mathbb{R}^{n+1}$, $t \in \mathbb{N}_0$, le cui componenti sono anch'esse processi stocastici discreti e autoregressivi $y_{i,t}$, $t \in \mathbb{N}_0 \quad i = 0, 1, .., n$ integrati di ordine d, cioè $\exists d \in \mathbb{N}$ t.c. $\forall i \quad y_{i,t} \sim I(d)$. Tale processo stocastico si dice cointegrato di ordine $b<d$, e si indica come $y_{i,t} \sim CI(d, b)$, se esiste

un processo stocastico discreto z_t integrato di ordine b, $z_t \sim I(b), \quad b \in \mathbb{N}$, che è una combinazione lineare delle n+1 componenti, cioè se esistono n coefficienti $\beta_i \in \mathbb{R} \quad i = 2, .., n$ tali che $z_t = \sum_{i=1,2..n} \beta_i y_{i,t}$, $t \in \mathbb{N}$.

Nel caso bidimensionale è possibile dare la definizione seguente:

Definizione 1.18 (Cointegrazione 1.2)**.** Si consideri un processo stocastico vettoriale discreto (x_t, y_t), $y_{i,t} \in \mathbb{R}^2$, $t \in \mathbb{N}_0$, , le cui componenti sono entrambe processi stocastici discreti e autoregressivi x_t, y_t, $t \in \mathbb{N}_0$ integrati di ordine uno I(d), $x_t \sim I(d), y_t \sim I(d)$. Tale processo stocastico si dice cointegrato $CI(d, 0)$ esiste un processo stocastico discreto z_t stazionario, $z_t \sim I(0)$, che è una combinazione lineare dei processi dati, cioè se $\exists \beta \in \mathbb{R}$ tale che $z_t = y_t - \beta x_t \quad t \in \mathbb{N}_0$. In particolare, se x_t, y_t, $t \in \mathbb{N}_0$ sono integrati di ordine uno I(1), $x_t \sim I(1), y_1 \sim I(1)$, tale processo stocastico si dice cointegrato.

In generale, il procedimento di regressione fra due variabili integrate dello stesso ordine, ad esempio I(1), porta risultati positivi secondo i canoni convenzionali anche quando non esiste nessuna relazione fra le variabili.

Definizione 1.19 (Cointegrazione 2)**.** Si consideri il processo stocastico vettoriale discreto $y_{i,t} = (y_{0,t}, y_{1,t}, ..y_{n,t})$,$y_{i,t} \in \mathbb{R}^{n+1}$, $t \in \mathbb{N}_0$, le cui componenti sono anch'esse processi stocastici discreti e autoregressivi $y_{i,t} \quad t \in \mathbb{N}_0 \quad i = 0, 1, .., n$ integrati di ordine d, $\forall i \quad y_{i,t} \sim I(d)$. Tale processo si dice cointegrato di ordine *d,b*, $y_{i,t} \sim CI(d, b)$, se $\exists \alpha \in \mathbb{R}^{n+1}, \alpha \neq 0$, detto *vettore di cointegrazione*, tale che $\alpha^T y \sim I(d - b)$, con b>0.

Un ampio range di metodi sono stati proposti per stimare il "vettore di cointegrazione" α. Il metodo più popolare di testing per la cointegrazione si basa sul concetto di **regressione di cointegrazione**. In questo approccio viene applicato un procedimento di regressione su una particolare serie storica (**serie target**) modellizzata dal processo target $y_{0,t}$, utilizzando come variabili indipendenti le serie storiche rimanenti (**serie cointegranti**) rappresentate dai processi cointegranti $y_{1,t}, ...y_{n,t}$:

$$y_{0,t} = \alpha + \beta_1 y_{1,t} + \beta_2 y_{2,t} ... + \beta_n y_{n,t} + e_t \tag{1.13}$$

Se il processo vettoriale $(y_{0,t}, 1, y_{1,t}, ..y_{n,t})$ è cointegrato, allora il termine di deviazione d_t sarà un processo stazionario. La motivazione usuale per cui si utilizza questo approccio è che i processi stazionari tendono ad avere una varianza più piccola di quelli non stazionari; in questo modo la procedura di regressione, che consiste nella minimizzazione della varianza dei residui e_t, cioè della somma degli scarti quadratici delle osservazioni portate dalle serie storiche $X_t = \{X_0, .., X_N\}$ e $Y_t = \{Y_1, ..Y_N\}$ nell'intervallo [0,T], partizionato negli istanti $\{t_i | i \in \mathbb{N}_0 i = 0..N\}$, dalla retta stimata, tenderà a massimizzare la probabilità di identificare segnali di eventuale cointegrazione del processo stocastico vettoriale. Nel caso bidimensionale, effettuiamo una semplice regressione lineare per stimare i coefficienti, che consiste nel minimizzare la somma degli scarti quadratici dalla retta stimata:

$$RSS := \sum_{t=0}^{T} e_t^2 = \sum_{t=0}^{T} (Y_t - \hat{Y}_t)^2 = \sum_{t=0}^{T} (Y_t - \hat{a} - \hat{b} X_t)^2 \tag{1.14}$$

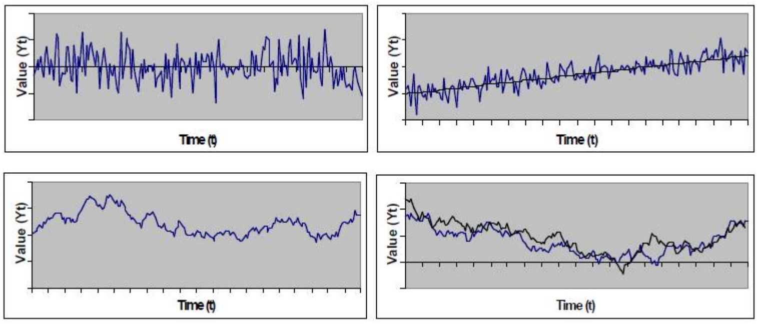

Figura 1.3: Processi stocastici con caratteristiche diverse, con particolare riguardo alla stazionarietà: in alto a sinistra un processo stocastico stazionario, un candidato adatto per essere inserito come variabile dipendente o indipendente in un modello senza creare un rischio di estrapolazione; in alto a destra un processo stocastico "ad andamento stazionario", che è stazionario intorno ad un andamento noto che è una funzione del tempo; in basso a sinistra un processo stocastico integrato; in basso a destra un processo vettoriale cointegrato o "set cointegrato".

La Figura 1.3 mostra diverse classi di processi stocastici.

La classe di processi stocastici con cui viene modellizzata una serie storica o un insieme di serie, che sia stazionario, integrato, o cointegrato, è fondamentale per individuare una potenziale componente predicibile che la serie storica può contenere.

Capitolo 2

Il contesto della cointegrazione nell'arbitraggio statistico

Questo capitolo descrive la metodologia che si può usare per costruire combinazioni di prezzi di asset chiamati *mispricings statistici*, che verranno definiti in seguito. Viene descritto il contesto base della cointegrazione che viene usato per generare serie storiche di mispricing.

2.1 Pairs trading: trading di una coppia di asset

Nella sezione precedente abbiamo visto il concetto di cointegrazione. La cointegrazione può essere vista come una relazione matematica di lungo periodo, *statistical fair-price relationship* tra due processi stocastici x_t e y_t. Tale relazione si realizza sul mercato attraverso il fenomeno del *relative pricing*: i due asset modellizzati da x_t e y_t sono prezzati uno in relazione all'altro. La cointegrazione dunque consente di individuare due asset che presentano caratteristiche simili e i cui prezzi si muovono insieme.

Talvolta però si verifica nel breve periodo un disallineamento dei prezzi, detto *mispricing*, dei due asset, che poi ritornano al precedente equilibrio di lungo periodo. Tale comportamento del fenomeno aleatorio considerato, che è possibile osservare solo statisticamente in quanto costituisce una deviazione dal modello matematico, fornisce un'opportunità interessante per fare arbitraggio.

Nel periodo di deviazione è possibile costruire un portafoglio, detto *portafoglio di mispricing*, sul quale si applicano strategie di trading, comprando l'asset sottoprezzato, cioè assumendo una "posizione lunga", e vendendo allo scoperto l'asset sovraprezzato, assumendo una "posizione corta", con l'aspettativa che si torni ad una posizione di equilibrio nota, che corrisponde all'annullamento del mispricing. In tale istante, uscendo dalle posizioni prese, avremo realizzato un profitto a costo zero.

2.2 Costruzione di processi stocastici di mispricing

Definiamo ora un processo stocastico chiamato *mispricing* e descriviamo nel dettaglio la sua metodologia di costruzione.

L'obiettivo della nostra metodologia è identificare combinazioni di serie storiche di titoli che rappresentano relazioni statistiche su cui basare le strategie di **arbitraggio statistico**.

In modo più specifico, dato un insieme di titoli (o asset) che modellizziamo con l'insieme di processi stocastici $U_A = \{T_i | i \in \mathbb{N}\}$ e un particolare **asset target**, che corrisponde a $T := T_i \in U_A$ per un certo indice i, il nostro obiettivo è creare un **asset sintetico** il cui prezzo possa essere considerato un **prezzo equo** per il titolo target, cioè coincida con il suo valore medio. Il processo stocastico associato $SA(T)$ coincide dunque con il processo generato dal valore atteso del processo target:

$$E[T_t] = SA(T)_t \quad t \in \mathbb{N}_0 \tag{2.1}$$

Definizione 2.1. Si consideri un insieme di processi stocastici $U_A = \{T_i | i \in \mathbb{N}\}$ e un particolare processo target, $T := T_i \in U_A$ per un certo indice i. Sia $SA(T)$ un processo sintetico, che soddisfa la relazione 2.1.
Si definisce **mispricing** il processo stocastico costruito nel modo seguente:

$$M_t = T_t - SA(T)_t \quad t \in \mathbb{N}_0 \tag{2.2}$$

Definizione 2.2. Si definisce **mispricing statistico** la serie storica $SM_t = SM_0, ., SM_N$ di osservazioni effettuate nell'intervallo [0,T], partizionato negli istanti temporali $\{t_i | i \in \mathbb{N}_0 \quad i = 0..N\}$, modellizzata dal processo M_t.

Il mispricing statistico contiene una componente predicibile che può essere utilizzata come base per costruire una strategia di trading di arbitraggio statistico.

Il nostro metodo per costruire mispricings statistici è basato sull'uso delle tecniche di cointegrazione. Una **regressione di cointegrazione** è usata per stimare una combinazione lineare di processi relativi a titoli che presenta la correlazione di lungo periodo massima possibile con il processo target T. I coefficienti della combinazione lineare sono stimati effettuando una regressione sulla serie storica di T rispetto alle serie storiche di un insieme di titoli detti **costituenti** $C \subset U_A - T$. In questo modo si ottiene un processo sintetico della forma:

$$SA(T)_t = \sum_{C_i \in C} \beta_i C_{i,t} \quad \{\beta_i\} = argmin \sum_{t=1..n} (T_t - \sum_{C_i \in C} \beta_i C_{i,t})^2 \tag{2.3}$$

Il **vettore di cointegrazione** è dato da $\beta = [\beta_1 ... \beta n_c]^T$.
L'asset sintetico può essere considerato una copertura statistica ottimale, condizionata dall'insieme degli asset costituenti C, nel senso che la proprietà standard della procedura usata nella regressione assicura sia che l'asset sintetico sarà uno stimatore oggettivo per l'asset target, cioè $E[T_t] = SA(T)_t$, sia che anche la deviazione tra le due serie di prezzi sarà minima in termini di errore

quadratico medio, e dunque la loro correlazione sarà massima.

Possiamo ora dare una definizione formale di processo sintetico.

Definizione 2.3. Si definisce un **processo stocastico sintetico** o modello di asset sintetico, la tripla $SA = \{T \in U_A; C \subset U_A - \{T\}; \beta \in R^{|C|}\}$, dove U_A è l'insieme dei processi dei titoli, $T \in U_A$ è il **processo target**, $C \subset U_A - \{T\}$ è l'insieme dei "processi costituenti" e $\beta \in R^{|C|}$ è il vettore dei pesi costituenti.

Dato un tale modello, è possibile derivare le serie storiche che rappresentano il mispricing statistico associato al processo definito come in Eq.2.2.

Definizione 2.4. Si definisce *strategia o portafoglio di mispricing* la strategia data dal vettore di cointegrazione $(1, -\beta_0, -\beta_1 ... - \beta_{n_c})$, da applicare rispettivamente ai titoli $\{T, C_0, C_1 ... C_{n_c}\}$.
Il valore di questa strategia rappresenta il valore in eccesso dell'asset target T, rispetto alla combinazione lineare di titoli $SA(T) = \sum_{C_i \in C} \beta_i C_i$ e si può pensare come una versione "stocasticamente detrendizzata" del prezzo originale del titolo T, cioè detrendizzata rispetto alle serie storiche osservate che sono generate almeno parzialmente da processi stocastici.

In questo contesto l'insieme degli asset costituenti C permette di tenere conto nel nostro modello quei fattori di rischio inosservati che agiscono comunemente sui prezzi del mercato. Nel massimizzare la correlazione tra asset target e asset sintetico la procedura di costruzione ha l'obiettivo di massimizzare indirettamente le sensibilità alle fonti comuni di rischio economico. Nel contesto dell'Eq.2.2 l'effetto della procedura di costruzione è quello di creare artificialmente una coppia di asset (l'asset target T_t e l'asset sintetico $SA(T)_t$) che hanno esposizioni simili ai fattori di rischio a cui sottostanno (ma che non sono necessariamente osservabili), che guidano le dinamiche dei prezzi dei titoli.

2.3 Test di predicibilità

Descriviamo ora i test che sono stati sviluppati con lo scopo di identificare componenti potenzialmente predicibili nelle dinamiche delle serie storiche, che applicheremo soprattutto alla serie di mispricing statistico. Questi test permettono di identificare un modello matematico che rappresenti una determinata serie storica, cioè una classe di processi stocastici a cui si può associare una serie storica. Tali test sono basati sul tipo di comportamento che intendono identificare, e sono i test autoregressivi, test unit root (radice unitaria), test variance ratio e test di cointegrazione.

2.3.1 Test per autoregressione

I test autoregressivi sono stati creati per identificare serie storiche che possano essere associate ad un modello di processo stocastico autoregressivo della forma descritta in Eq.1.8, distinguendo dunque serie storiche che contengono una componente predicibile da serie storiche che producono puramente rumore.

Data una serie storica $Y_t = \{Y_1, ..Y_N\}$ di osservazioni effettuate nell'intervallo [0,T], partizionato negli istanti $\{t_i | i \in \mathbb{N}_0 i = 0..N\}$, si definisce la Funzione di Autocorrelazione (Auto-Correlation Function, ACF), nota anche come *correlogramma*, come una funzione della variabile $k \in \mathbb{N}$:

$$r(k) = \frac{\sum_{t=k+1}^{N}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{N}(Y_t - \overline{Y})^2} \quad \overline{Y} = \frac{\sum_{t=0}^{N} Y_t}{N+1} \tag{2.4}$$

I valori r_k sono detti *coefficienti di autocorrelazione di ordine k* e sono tali che $0 < r_k < 1$. Tali coefficienti vengono utilizzati come statistica test. Essi hanno una distribuzione

Fissato un livello di significatività $\alpha \in \mathbb{R}_+, 0 < \alpha < 1$, si confrontano i valori assunti dalla ACF, dette *autocorrelazioni*, con tale valore. Autocorrelazioni **significativamente** diverse da zero, $r(k) > \alpha$ indicano che i valori futuri delle serie storiche sono **significativamente** legati ai valori passati, e quindi la presenza di una componente predicibile nelle dinamiche delle serie storiche.

2.3.2 Test unit root

I test unit root sono una particolare classe di test, usati per identificare modelli di serie storiche matematicamente complessi. Quelli che utilizzeremo sono utili per verificare l'ipotesi nulla H_0 che una serie storica possa essere modellizzata con un processo stocastico random walk, sia in un contesto statistico generale sia nell'ambito dell'analisi di cointegrazione. Se l'ipotesi nulla viene rigettata, è possibile concludere che il modello da adottare è quello di processo stocastico stazionario. Nelle seguenti sottosezioni analizziamo tre test standard chiamati test Dickey-Fuller (DF), Augmented Dickey-Fuller (ADF) e test di cointegrazione.

Test Dickey-Fuller

In primo luogo si sceglie un modello di processo stocastico su cui effettuare il test Dickey Fuller. Il modello da noi utilizzato è quello autoregressivo, descritto dall'Eq.1.8 in cui $f(y_{t-1}) = \beta y_{t-1}$

$$y_t = \beta y_{t-1} + \varepsilon_t \quad \forall t \tag{2.5}$$

$\beta \in \mathbb{R}$ Per testare l'andamento delle singole variabili che formano il processo si procede applicando un procedimento di autoregressione, che consiste nel fare una regressione sulla serie $Y_t \quad t = 1...N$ rispetto alla serie $Y_{t-1} \quad t = 1..N$. Poi si sottrae ad ambo i membri y_{t-1}, ottenendo:

$$\Delta y_t = (\beta - 1) y_{t-1} + \varepsilon_t$$

Conduciamo un test su β ponendo:

$$H_0 : \beta = 1 \tag{2.6}$$

La statistica test che si utilizza è applicata al coefficiente β $T = n(\beta - 1)$, dove n è la dimensione del campione. Tale statistica segue un'apposita distribuzione che è stata tabulata da Fuller, ma che non ci interessa conoscere ai nostri fini, in quanto utilizzeremo software che implementano tale test autonomamente, restituendo soltanto il "p-value". Il "p-value" è un valore che rappresenta la probabilità che corrisponde alla statistica test osservata nel campione, che in questo caso è pari a $P(T \leq t)$, che si può calcolare conoscendo la distribuzione di T e rappresenta l'area che sta sotto il grafico a sinistra della t osservata. Dopo aver fissato un livello di significatività al di sotto del quale l'ipotesi nulla è rigettata, che dipende dalla forma della distribuzione e in questo caso, poiché la forma è asimmetrica a destra, è pari a $\alpha = 0.05$, si confronta questo valore con il p-value. Si possono dunque verificare le seguenti situazioni:

1. Se $pvalue > \alpha$ l'ipotesi di non stazionarietà non è rigettata, in quanto abbiamo una confidenza del 95% che la serie possa essere non stazionaria.

2. Se $pvalue < \alpha$ l'ipotesi è rigettata e la serie può essere stazionaria

Se l'ipotesi di non stazionarietà non è rigettata, si può modellizzare la serie con il modello testato ponendo $\beta = 1$.
Spesso il test Dickey Fuller, come nel nostro caso, è utilizzato per individuare se è possibile associare ad una serie un modello di processo integrato di un certo ordine d. Si procede quindi a creare una nuova serie storica, le cui osservazioni sono date dagli *incrementi di primo ordine* $\Delta Y_{d,t} = Y_{t+d} - Y_t, \quad t \in [0, N], t \in \mathbb{N}_0$. Applicando il test su tale serie, se l'ipotesi è rigettata, siamo in presenza di un modello di tipo $I(d)$, in quanto la serie degli incrementi di ordine d può essere modellizzata tramite un processo stazionario. Il test DF si può applicare dunque per individuare modelli integrati di ordine d. Il fatto che l'ipotesi di non stazionarietà implichi $\beta = 1$ è la ragione per cui questo test viene spesso chiamato test **unit root** (radice unitaria).

Test Augmented Dickey-Fuller

La forma "incrementata" del test DF, detto Dickey-Fuller Aumentato (Augmented-Dickey-Fuller test) o test ADF, si effettua su un'ipotesi diversa sulla regressione degli incrementi di primo ordine ΔY_t rispetto alla serie $Y_0 ... Y_{t-1}$:

$$y_t - y_{t-1} = \Delta y_t = \beta y_{t-1} + \sum_{j=1}^{t-2} \beta_j \Delta y_{t-j} + \varepsilon_t \quad \forall t \tag{2.7}$$

$\beta, \beta_j \in \mathbb{R} \quad j \in [1, t-2]$ A questo punto si procede come per il test DF, cioè l'ipotesi di non stazionarietà viene rigettata solo se il pvalue trovato è minore del livello di significatività α. Se invece si vuole verificare se la serie è modellizzabile con un processo integrato di ordine d, si generano i processi stocastici degli *incrementi di ordine d.*

I test DF e ADF si possono considerare come test per individuare un tipo particolare di componente predicibile nelle dinamiche di una serie storica. Ciò è reso chiaro dalla forma dell'eq.2.7: se il coefficiente β è significativamente negativo questo implica una relazione significativamente negativa tra i cambiamenti *futuri* e i livelli *correnti*, indicativa di una componente **mean-reverting** nelle serie storiche.

Test per la cointegrazione

I test di cointegrazione servono ad identificare se un insieme di serie storiche modellizzate tramite processi stocastici integrati dello stesso ordine contengono una o più combinazioni lineari (ad esempio mispricing) che sono esse stesse modellizzabili, ma con processi cointegrati.

Come osservato, la più popolare forma di testing per la cointegrazione è quella che si basa sul concetto di **regressione di cointegrazione**. In questo approccio si effettua una regressione su una serie storica **target** modellizzata dal processo $y_{0,t}$ rispetto alle **serie cointegranti** modellizzate dall'insieme di processi stocastici $\{y_{1,t}, ..., y_{n,t}\}$.
Una condizione necessaria per applicare il test è che tutte le serie storiche di partenza siano integrate di ordine d. Si determina così un nuovo processo stocatico vettoriale y_t. La presenza (o assenza) di cointegrazione è individuata testando il termine di deviazione d_t che è rappresentato dagli errori residuali della regressione di cointegrazione, che applichiamo cercando di modellizzare la serie target con una combinazione lineare delle altre serie:

$$d_t = y_{0,t} - [\gamma + \beta_1 y_{1,t} + \beta_2 y_{2,t} ... + \beta_n y_{n,t}] = \alpha^t y_t \tag{2.8}$$

$$\gamma, \beta_i \in \mathbb{R} \quad \forall i = 1..n$$

Se il processo stocastico vettoriale $y_t = (y_{0,t}, 1, y_{1,t}, ...y_{n,t})$ è cointegrato CI(d,b), i test statistici unit root, indicheranno che d_t è un processo integrato di ordine d, I(d). Il vettore dei coefficienti $\alpha = (1, -\gamma, -\beta_1, .. - \beta_n)$ è detto *vettore cointegrante.*
Nell'esempio che vedremo non applicheremo la regressione in più variabili, ma useremo una serie target e una serie cointegrante, realizzando così una semplice regressione lineare.

2.3.3 Test Variance Ratio

I test variance ratio sono stati ideati per identificare la situazione in cui una serie storica devia dal comportamento **random walk**, definito come in Eq.1.9, e può essere modellizzata mediante un determinato processo stocastico mean-reverting.
I cosiddetti test *variance ratio* si basano sulla proprietà descritta dall'Eq.1.5, secondo la quale la varianza degli *incrementi di primo ordine*, che sono le variabili aleatorie che formano un processo random walk, è una funzione lineare del tempo a partire del quale gli incrementi sono misurati. Una semplice intuizione di questa proprietà è presentata nella figura sottostante.

La Funzione di Variance Ratio è definita come il rapporto normalizzato della varianza campionaria degli incrementi di ordine τ (detta varianza di lungo periodo) rispetto alla varianza campionaria degli incrementi di primo ordine (varianza di breve periodo)ed è dunque:

$$VR(\tau) = \frac{\sum_t (\Delta^\tau Y_t - \overline{\Delta^\tau Y})^2}{\tau \sum_t (\Delta Y_t - \overline{\Delta Y})^2} \tag{2.9}$$

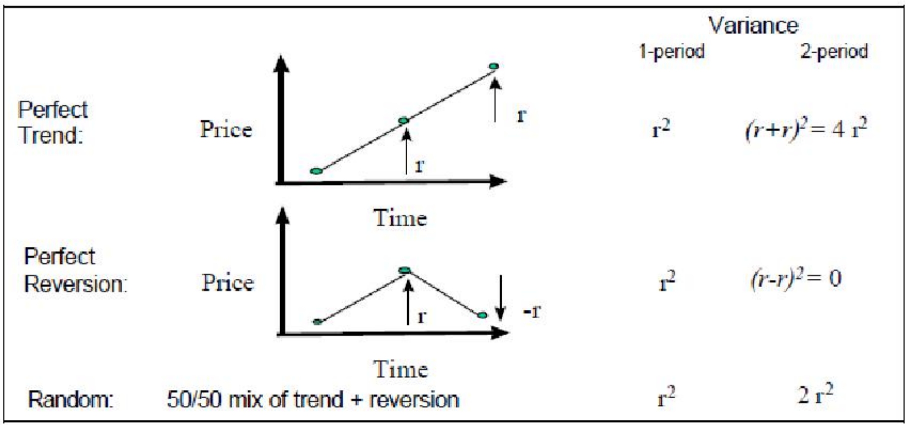

Figura 2.1: La relazione tra la varianza campionaria e il tempo per una serie storica di prezzi di titoli: nel caso limite in cui tutti i prezzi seguono un andamento lineare, la varianza campionaria della serie crescerà come una funzione quadratica del tempo, mentre nel grafico sottostante si osserva un esempio reversione pura, in cui la varianza della serie è dipendente dal tempo (e vicina a zero). Una serie storica random walk sarà una combinazione ponderata di entrambi i comportamenti e mostrerà una varianza che cresce linearmente con il tempo.

L'uso comune del test VR è quello di testare la presenza di una componente mean-reverting significativa nelle dinamiche delle serie storiche. Si osservano un numero di valori significativo della funzione VR e, fissato un valore critico o livello di significatività basato sul livello di importanza richiesto, si effettuano dei confronti, che rendono possibile il verificarsi di tre situazioni:

1. $1 - d < VR(\tau) < 1 + d$ indica che gli incrementi sulle serie sono scorrelati, cioè siamo in presenza di un *random walk behaviour*

2. $VR(\tau) > 1 + d$ indica una varianza di lungo periodo più alta di quanto si sarebbe aspettato data la varianza di breve periodo, e dunque che l'effetto **netto** sulla scala temporale τ è di autocorrelazione positiva o *trending behaviour*

3. $VR(\tau) < 1 - d$ indica una varianza di lungo periodo più bassa di quella aspettata, data la varianza di breve periodo, e dunque l'effetto **netto** sul periodo è di autocorrelazione negativa o *mean-reverting behaviour*

La figura seguente mostra un esempio di serie storiche con diverse caratteristiche, insieme alle loro VRF associate.

Osservazione 2.1. Una particolare forma di VRF è equivalente ad una combinazione lineare dei coefficienti di autocorrelazione, cioè esiste un legame tra le due funzioni ACF e VRF:

$$VR(\tau) = \sum_{j=1}^{q-1} \frac{2(q-j)}{q} r(j) \qquad (2.10)$$

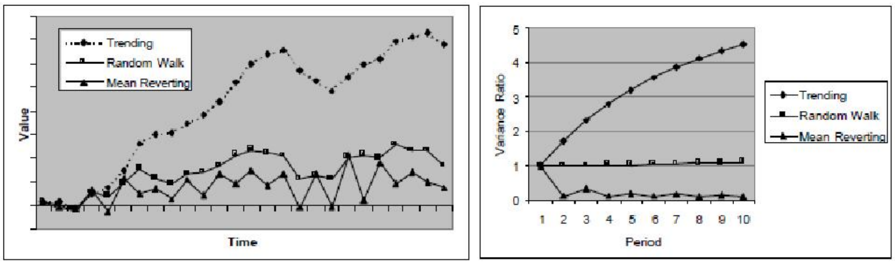

Figura 2.2: Esempio di serie storiche con diverse caratteristiche (a sinistra) e le loro funzioni di variance ratio o VRF (a destra). Per il random walk, la varianza cresce linearmente con τ e dunque la VRF rimane vicina a uno; per una serie seguente un trend deterministico la varianza cresce ad un passo più che lineare e così la VRF sale sopra uno alla crescita dell'ordine degli incrementi; per serie mean-reverting vale il contrario: la varianza cresce in modo sottolineare e dunque la VRF scende sotto l'unità.

dove r_j sono i coefficienti di autocorrelazione calcolati secondo l'eq.2.4.

La forza e la debolezza dei vari test si può definire in termini della loro dimensione relativa e potere nel distinguere tra l'ipotesi nulla di dinamiche puramente stocastiche (comportamento random walk) e varie ipotesi alternative che corrispondono a diversi tipi di dinamiche parzialmente deterministiche.

Capitolo 3

Implementazione di strategie di trading

3.1 Strategie implicite di StratArb

Descriviamo ora una strategia che rientra in un insieme di strategie di trading di **arbitraggio statistico implicito** (ISA), chiamate così perché le regole di trading su cui si basano contano *implicitamente* sul comportamento mean-reverting delle serie di mispricing.
Dato un modello di mispricing della forma $M_t = T_t - \sum_{C_i \in C} \beta_i C_{i,t}, \quad t \in \mathbb{N}_0$, l'assunzione che sottosta alle strategie ISA è che i cambiamenti dei prezzi futuri saranno tali da tendere in media a ridurre il mispricing tra il titolo target e la combinazione di titoli costituenti.

Le strategie ISA agiscono servendosi di ciò che potrebbe essere considerato un **effetto di correzione del mispricing**, e realizzeranno profitti nei casi in cui la componente di mean-reversion nelle dinamiche di mispricing sia sufficientemente ampia da superare i costi di transazione che sono legati alla scelta e alla realizzazione della strategia. Ciò si manifesta visivamente con la presenza di crolli o picchi significativi del mispricing, in corrispondenza dei quali è possibile intraprendere strategie di arbitraggio. In pratica, il grafico del mispricing dovrebbe mostrare un comportamento mean-reverting attorno all'asse delle ascisse.

Le strategie ISA sono implementate attraverso regole di trading parametrizzate che definiscono il segno (cioè indicano se acquistare o vendere il titolo) e l'ammontare dei titoli attualmente oggetto di interesse nel **portafoglio di mispricing**, che consiste nella strategia $(1, -\beta_0, -\beta_1, ..., -\beta_{n_c})$, applicata rispettivamente all'insieme di titoli $\{T_t, C_0, C_1, ..., C_{n_c}\}$.
Se M_t è la serie storica del titolo mispricing, la strategia adottata si basa sulla quantità di denaro che si vuole guadagnare, anche se non propriamente, in quanto venderemo titoli allo scoperto. Il guadagno G_t che si ottiene in ogni istante t in cui esercito si può rappresentare come:

$$G_t = k_t |M_t| \tag{3.1}$$

k_t è un processo stocastico che funge da parametro di sensibilità, che indica in che misura il valore attuale del mispricing influisce sullammontare di patrimonio da investire. Esso consente al numero dei titoli su cui investire di variare come una funzione del valore del mispricing attuale.
Tale parametro è soggettivo ed è scelto dallo stesso arbitraggista, nel momento in cui viene stabilita la strategia da adottare. In pratica, quando il titolo mispricing è positivo lo vendiamo, mentre quando è negativo lo compriamo, secondo la seguente strategia:

$$(\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t}) = -k_t(1, -\alpha, -\beta) \quad \forall t \tag{3.2}$$

dove $(1, -\alpha, -\beta)$ rappresenta il *portafoglio di mispricing* associato a M_t. Tale strategia va applicata per un certo orizzonte temporale, al termine del quale chiudiamo le posizioni, e stabiliamo se abbiamo ottenuto un profitto o una perdita, utilizzando la seguente formula:

$$PL = GT_T + \frac{\lambda_{1,T}}{k_T} M_T \tag{3.3}$$

dove GT_T è il guadagno totale fino all'istante T e M_T è il mispricing osservato il giorno della chiusura, moltiplicato per il numero di titoli Y_t acquistati o venduti, Se vi sono costi di transazione la formula sarà:

$$PLC = PL - \sum_{i=0}^{T} TC_i \tag{3.4}$$

dove TC_i è il costo di transazione all'istante i-esimo.

Nella seguente sezione applicheremo una strategia implicita di arbitraggio, lavorando sulle serie storiche dei prezzi dell'oro e dell'argento.

3.2 Esempio

L'arbitraggio statistico si effettua su particolari tipi beni, detti *commodity*. Una *commodity* è un bene per cui c'è domanda ma che è offerto senza differenze qualitative sul mercato ed è fungibile, cioè il prodotto è lo stesso indipendentemente da chi lo produce, come per esempio il petrolio o i metalli. Si tratta di bene indifferenziato, cioè ottenibile comodamente, pratico. Una commodity deve essere facilmente stoccabile e conservabile nel tempo, cioè non perdere le caratteristiche originarie. L'elevata standardizzazione che caratterizza una commodity ne consente l'agevole negoziazione sui mercati internazionali. Una delle caratteristiche di una commodity è che il suo prezzo viene determinato dal mercato. Generalmente le commodity sono prodotti agricoli o prodotti di base non lavorati come l'oro, il sale, lo zucchero e il caffè.
In particolare, prendiamo come esempio due commodity, l'oro e l'argento, e applichiamo la metodologia di arbitraggio statistico, che riassumiamo brevemente:

1. Costruzione di asset sintetici che replichino uno o più asset target e individuazione di una relazione statistica di lungo periodo, il *mispricing statistico*, che soddisfi una condizione di prevedibilità.

2. Individuazione di opportunità di arbitraggio statistico attraverso stime o previsioni statistiche sul mispricing.

3. Implementazione di strategie di trading adeguate

Analizziamo ora nel dettaglio questi passi, applicandoli agli asset che ci interessano.

3.2.1 Costruzione del mispricing

La scelta primaria da fare è quella dell'asset target da replicare, che nel nostro caso è l'**oro**, il cui prezzo nel tempo è dato dalla serie storica $Y_t = \{Y_0...Y_N\}$.
I prezzi dei beni che osserveremo sono dati in dollari per oncia. Un'oncia corrisponde a 28.34 grammi.
Selezioniamo dunque gli asset con cui costruire il portafoglio di replica, che è formato da un solo asset sintetico costruito con l'**argento**, il cui prezzo nel tempo è rappresentato dalla serie storica $X_t = \{X_0...X_N\}$.
Ricaviamo dal sito *Investing.com* le serie storiche dell'oro e dell'argento relative ai prezzi di chiusura giornalieri dell' anno 2012, osservando il seguente andamento.

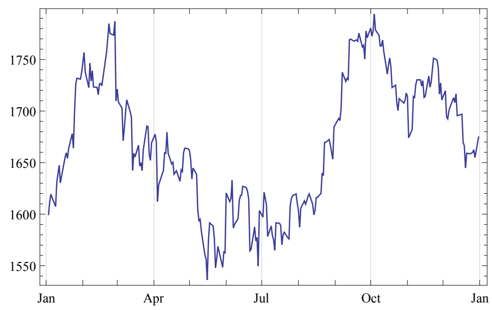

Figura 3.1: Grafico dell'andamento del prezzo dell'oro nel 2012

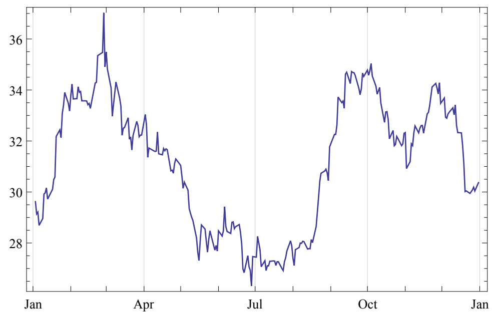

Figura 3.2: Grafico dell'andamento del prezzo dell'argento nel 2012

Per poter effettuare un arbitraggio statistico, le due serie storiche in esame devono poter essere modellizzare tramite processi stocastici che possono essere legate da una relazione di lungo periodo. Prima di cercare tale relazione, è necessario realizzare diversi test, che implementeremo utilizzando il software

Mathematica.
In primo luogo dobbiamo verificare che le due serie storiche siano modellizzabili con processi stocastici integrati dello stesso ordine, che ipotizziamo in questo caso essere uno. Applichiamo ad entrambe le serie storiche prima il test per autoregressione, per verificare se possono essere modellizzate con un processo autoregressivo, fissando un livello soglia $\alpha = 0.75$, al di sopra del quale accetteremo ottenendo i seguenti valori per i primi 5 coefficienti di autocorrelazione: Poiché tutti i coefficienti sono maggiori di 0.75, è possibile concludere

r(k)	r(1)	r(2)	r(3)	r(4)	r(5)
Oro	0.96264	0.934726	0.907038	0.886181	0.859535,
Argento	0.979517	0.960868	0.94222 9	0.923489	0.901805

che le serie storiche dei prezzi dell'oro e dell'argento sono modellizzabili con processi autoregressivi. Per verificare l'ipotesi che tale modello autoregressivo sia un modello integrato di ordine 1, I(1), applichiamo applichiamo il test DF ad entrambe le serie storiche, fissando un livello di significatività $\alpha = 0.05$:

1. *pvalue* oro=0.689293 > 0.05
2. *pvalue* argento=0.679458 > 0.05

Dunque l'ipotesi di non stazionarietà è accettata. Ora dobbiamo verificare l'ipotesi che i processi autoregressivi dell'oro e dell'argento siano integrati , costruendo le serie storiche degli incrementi di ordine uno $\Delta Y_{1,t}$ e verificando se sono stazionarie, applicando nuovamente il test DF:

1. *pvalue* $\Delta Y_{1,t}$=0 < 0.05
2. *pvalue* $\Delta X_{1,t}$=0 < 0.05

Il test restituisce il valore 0 e, poiché la sua precisione è la precisione di macchina, ciò significa che il pvalue è molto prossimo allo 0, dal momento che ha tante cifre decimali nulle. L'ipotesi di non stazionarietà è dunque rigettata per entrambe le serie storiche, che sono modellizzabili con processi stazionari.
Si può concludere così che le due serie storiche dell'oro e dell'argento Y_t e X_t sono modellizzabili con processi integrati di ordine 1, cioè $y_t \sim I(1)$ e $x_t \sim i(1)$.

Ora possiamo costruire il portafoglio di replica effettuando una regressione lineare di cointegrazione. I coefficienti della combinazione lineare sono stimati effettuando una regressione sulla serie storica dell'oro rispetto alle serie storiche dell'argento ottenendo il seguente modello:

$$y_t = \alpha + \beta x_t + \varepsilon_t \quad \{\alpha, \beta\} = argmin \sum_{t=1..n} (Y_t - \alpha - \beta X t)^2 \tag{3.5}$$

Utilizzando Mathematica, implementiamo il procedimento di regressione lineare, ottenendo $\alpha = 936.336$ e $\beta = 23.5$.
Otteniamo la serie storica di mispricing come deviazione dalla relazione di lungo periodo trovata:

$$M_t = Y_t - \alpha - \beta X_t = Y_t - 936.336 - 23.5 X_t \tag{3.6}$$

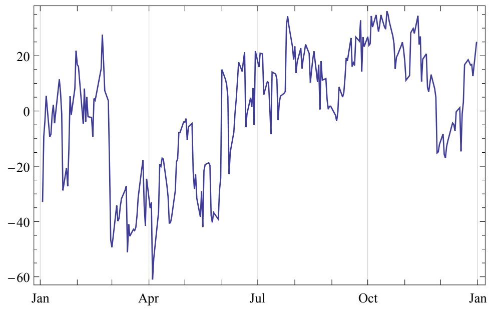

Figura 3.3: Grafico dell'andamento del mispricing nel 2012

3.2.2 Verifica di prevedibilità del mispricing

Ora che abbiamo trovato la serie storica di mispricing, verifichiamo se può essere modellizzata mediante un processo stocastico stazionario. Applichiamo al mispricing prima il test per autoregressione, per verificare se può essere modellizzata con un processo autoregressivo, fissando un livello soglia $\alpha = 0.55$, al di sopra del quale accetteremo ottenendo i seguenti valori per i primi 5 coefficienti di autocorrelazione: Poiché tutti i coefficienti sono maggiori di 0.55, è possibile

r(k)	r(1)	r(2)	r(3)	r(4)	r(5)
Mispricing	0.904122	0.837255	0.760223	0.710026	0.666359

concludere che il mispricing è modellizzabile con un processo autoregressivo.
Per verificare l'ipotesi che tale modello autoregressivo sia un modello stazionario, I(0), applichiamo applichiamo il test DF, fissando un livello di significatività $\alpha = 0.05$: otteniamo un *pvalue*=0.000675729 < 0.05 Dunque l'ipotesi di non stazionarietà è rigettata e possiamo rappresentare il mispricing con un modello stazionario.
Per confermare ulteriormente la stazionarietà del mispricing, applichiamo il Variance Ratio Test, che individua una deviazione dal comportamento random walk. Fissiamo come previsto dal test un livello di significatività $\alpha = 0.05$, che ci permette di vedere quale delle tre situazioni possibili si è verificata, e osserviamo il valore della funzione di Variance Ratio che corrisponde a $tau = 2$:

$$VR(2) = 0.8804$$

Sicuramente la deviazione dal comportamento random walk è confermata, in quanto $VR(2) \notin [-0.95, 0.05]$. Siamo dunque nel terzo caso, poiché $0.8804 < 1 - 0.05 = 0.95$, che significa che siamo in presenza di un comportamento *mean reverting* del mispricing.
Concludiamo dunque che la serie storica di mispricing è modellizzabile con un processo stazionario e mean-reverting.

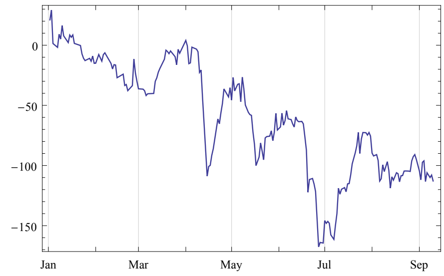

Figura 3.4: Grafico del'andamento del mispricing nell'anno 2013

3.2.3 Implementazione di una strategia di trading

Dopo aver fatto tutte le verifiche necessarie, implementiamo la strategia implicita di trading descritta in Sezione 3.1. Se M_t è la serie storica del titolo mispricing, la strategia adottata si basa sulla quantità di denaro che si vuole guadagnare, anche se non propriamente in quanto venderemo titoli allo scoperto. Il guadagno G_t che si ottiene in ogni istante t in cui esercito si può rappresentare come:

$$G_t = k_t|M_t| \tag{3.7}$$

dove k_t è un parametro che dipende dal valore del titolo mispricing osservato ed indica in che misura il valore attuale del mispricing influisce sull'ammontare di patrimonio da investire.

Definiamo in modo soggettivo il valore che deve assumere k_t, a seconda che il mispricing cada in uno degli intervalli fissati, che hanno ampiezza decrescente con l'allontanarsi dall'origine. Questi dipendono anche dai valori osservati nel 2012, in cui si sono ottenuti un mispricing massimo e minimo rispettivamente di 36.139\$ e -60.996\$.
$k_t \in \mathbb{N}_0, k_t \in [0, 10]$:

$$k_t = \begin{cases} 0 & se \quad -11.1 \leq M_t \leq 6.6 \\ 1 & se \quad -21.1 < M_t \leq -11.1 \quad o \quad 6.6 \leq M_t \leq 12.5 \\ 2 & se \quad -29.9 < M_t \leq -21.1 \quad o \quad 12.5 \leq M_t \leq 17.7 \\ 3 & se \quad -37.7 < M_t \leq -29.9 \quad o \quad 17.7 \leq M_t \leq 22.3 \\ 4 & se \quad -44.4 < M_t \leq 37.7 \quad o \quad 22.3 \leq M_t \leq 26.3 \\ 5 & se \quad -49.9 < M_t \leq -44.4 \quad o \quad 26.3 \leq M_t \leq 29.6 \\ 6 & se \quad -54.3 < M_t \leq -49.9 \quad o \quad 29.6 \leq M_t \leq 32.2 \\ 7 & se \quad -57.7 < M_t \leq -54.3 \quad o \quad 32.2 \leq M_t \leq 34.2 \\ 8 & se \quad -59.9 < M_t \leq -57.7 \quad o \quad 34.2 \leq M_t \leq 35.5 \\ 9 & se \quad -61 < M_t \leq -59.9 \quad o \quad 35.5 \leq M_t \leq 36.14 \\ 10 & se \quad M_t \leq -61 \quad o \quad M_t \geq 36.14 \end{cases}$$

In pratica, quando il titolo mispricing è positivo lo vendiamo, mentre quando è negativo lo compriamo, secondo la seguente strategia:

$$(\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t}) = -k_t(1, -\alpha, -\beta) \quad \forall t \tag{3.8}$$

dove $(1, -\alpha, -\beta)$ rappresenta il *portafoglio di mispricing* associato a M_t. Stabiliamo un periodo di applicazione della strategia, che corrisponde in questo

caso ai primi 4 mesi del 2013. Assumiamo che la componente α del vettore di cointegrazione, che dovrebbe essere associato ad un asset artificiale di prezzo costantemente unitario, rappresenti un ammontare di denaro che ogni giorno viene preso in prestito o restituito alla banca.
Infine, quando terminiamo l'applicazione della strategia, stabiliremo, chiudendo tutte le posizioni prese, se abbiamo guadagnato o meno.
Lavoriamo dunque sul mispricing seguente: $M_t = Y_t - 936.336 - 23.5X_t$ Le tabelle seguenti mostrano i risultati ottenuti da Gennaio ad Aprile 2013, analizzando i diversi Profit&Loss che si ottengono in diversi momenti di chiusura della posizione, senza e con un costo di transizione del 1. Indichiamo con GT_t le somme parziali dei guadagni dei singoli istanti, mentre TC_t rappresentano i costi di transazione ad ogni istante t.

Data	M_t	k_t	G_t	$(\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t})$	GT_t	TC_t
02.01.2013	21.0665	3	63.2	(-3,2809,70.5)	63.2	10.072
03.01.2013	29.1915	5	145.96	(-5,4681.68,117.5)	209.16	16.6
04.01.2013	1.406	0	0	(0,0,0)	209.16	0
07.01.2013	-1.7815	0	0	(0,0,0)	209.16	0
08.01.2013	9.119	1	9.119	(-1,933.336,23.5)	218.279	3.31
09.01.2013	5.1215	0	0	(0,0,0)	218.279	0
10.01.2013	16.3415	2	32.683	(-2,1872.672,47)	250.962	6.65
11.01.2013	7.807	1	7.807	(-1,936.336,23.5)	258.769	3.3
14.01.2013	2.184	0	0	(0,0,0)	258.769	0
15.01.2013	8.506	1	8.506	(-1,+936.336,+23.5)	267.275	3.36
16.01.2013	6.631	1	6.631	(-1,936.336,23.5)	273.9	3.36
17.01.2013	8.5265	1	6.631	(-1,936.336,23.5)	282.43	3.37
18.01.2013	1.507	0	0	(0,0,0)	282.43	0
22.01.2013	-0.236	0	0	(0,0,0)	282.43	0
23.01.2013	-7.253	0	0	(0,0,0)	282.43	0
24.01.2013	-10.799	0	0	(0,0,0)	282.43	0
25.01.2013	-12.7835	1	12.7835	(1,-936.336,-23.5)	295.21	3.3
28.01.2013	-10.9085	0	0	(0,0,0)	295.21	0
29.01.2013	-13.301	1	13.301	(1,-936.336,-23.5)	308.5	3.3
30.01.2013	-9.0235	0	0	(0,0,0)	308.5	0
31.01.2013	-15.046	1	15.046	(1,-936.336,-23.5)	323.55	3.3

Tabella 3.1: Se dovessi chiudere la mia posizione ora (-12 oro,+282 argento), Il mio guadagno totale sarà dunque di come da Eq.3.3 $PL = 323.55 + 180.552 = 504.102$\$.Tenendo conto dei costi di transazione, invece, come da Eq.3.4, $PLC = 504.102 - 59.9 = 444.2$\$. Proseguiamo ora ad esercitare nel mese di Febbraio.

Data	M_{t-1}	k_t	G_t	$(\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t})$	GT_t	TC_t
01.02.2013	-14.941	1	14.941	(1,-936.336,-23.5)	338.491	3.35
04.02.2013	-7.7485	0	0	(0,0,0)	338.491	0
05.02.2013	-13.2335	1	13.2335	(1,-936.336,-23.5)	351.73	3.36
06.02.2013	-7.816	0	0	(0,0,0)	351.73	0
07.02.2013	-6.3035	0	0	(0,0,0)	351.73	0
08.02.2013	-8.5415	0	0	(0,0,0)	351.73	0
11.02.2013	-15.1085	1	15.1085	(1,-936.336,-23.5)	366.84	3.3
12.02.2013	-19.7315	1	19.7315	(1,-936.336,-23.5)	386.57	3.3
13.02.2013	-16.171	1	16.171	(1,-936.336,-23.5)	402.74	3.3
14.02.2013	-16.459	1	16.459	(1,-936.336,-23.5)	419.2	3.29
15.02.2013	-27.0135	2	54.03	(2,-1872.672,-47)	473.33	6.49
19.02.2013	-24.059	2	48.12	(2,-1872.672,-47)	521.45	6.46
20.02.2013	-33.468	3	100.4	(3,-2809,-70.5)	621.86	9.57
21.02.2013	-32.6565	3	145.96	(3,-2809,-70.5)	767.82	9.58
22.02.2013	-37.8	4	151.2	(4,-3745.344,-94)	919.02	12.73
25.02.2013	-33.6335	3	100.9	(3,-2809,-70.5)	1019.9	9.6
26.02.2013	-11.566	1	11.566	(1,-936.336,-23.5)	1031.5	3.2
27.02.2013	-23.106	2	46.212	(2,-1872.672,-47)	1077.712	6.4
28.02.2013	-30.031	3	90.1	(3,-2809,-70.5)	1167.8	9.56

Tabella 3.2: Chiudendo la posizione ora, in cui ci troviamo allo scoperto di 399.5 titoli dell'argento ed abbiamo acquistato 17 titoli oro, il profit&loss sarà $PL = 1167.8 - 17 * 30.031 = 657.27\$$. Tenendo conto dei costi di transazione, invece, $PLC = 657.27 - 59.9 - 93.49 = 503.88\$$. Vediamo i risultati per il mese di Marzo

Data	M_t	k_t	G_t	$(\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t})$	GT_t	TC_t
01.03.2013	-36.2305	3	108.7	(3,-2809,-70.5)	1276.5	9.3
04.03.2013	-36.524	3	109.6	(3,-2809,-70.5)	1386.1	9.5
05.03.2013	-37.831	4	151.3	(4,-3745.344,-94)	1537.4	12.75
06.03.2013	-41.826	4	167.3	(4,-3745.344,-94)	1704.7	12.76
07.03.2013	-40.404	4	121.2	(4,-3745.344,-94)	1826	12.76
08.03.2013	-40.296	4	161.2	(4,-3745.344,-94)	1987.2	12.77
11.03.2013	-40.083	4	160.3	(4,-3745.344,-94)	2147.5	12.78
12.03.2013	-30.049	3	90.15	(3,-2809,-70.5)	2237.65	9.64
13.03.2013	-27.139	2	54.3	(2,-1872.672,-47)	2292	6.4
14.03.2013	-22.4185	2	44.8	(2,-1872.672,-47)	2336.8	6.4
15.03.2013	-19.414	1	19.414	(1,-936.336,-23.5)	2356.2	3.2
18.03.2013	-11.638	1	11.638	(1,-936.336,-23.5)	2367.8	3.2
19.03.2013	-4.139	0	0	(0,0,0)	2367.8	0
20.03.2013	-5.2365	0	0	(0,0,0)	2367.8	0
21.03.2013	-7.0205	0	0	(0,0,0)	2367.8	0
22.03.2013	-4.774	0	0	(0,0,0)	2367.8	0
25.03.2013	-9.5465	0	0	(0,0,0)	2367.8	0
26.03.2013	-16.2785	1	16.2785	(1,-936.336,-23.5)	2384	3.2
27.03.2013	-3.411	0	0	(0,0,0)	2384	0
28.03.2013	-6.7035	0	0	(0,0,0)	2384	0

Tabella 3.3: Chiudendo la posizione ora, in cui ci troviamo allo scoperto di 1245.5 titoli dell'argento e possediamo 53 titoli oro, il profit&loss sarà $PL = 2384 - 53 * 6.7035 = 2028.7$\$. Tenendo conto dei costi di transazione, invece,$PLC = 2028.7 - 59.9 - 93.49 - 114.66 = 1760.65$\$. Vediamo i risultati per il mese di Aprile

Data	M_t	k_t	G_t	$(\lambda_{1,t}, \lambda_{2,t}, \lambda_{3,t})$	GT_t	TC_t
01.04.2013	3.972	0	0	(0,0,0)	2384	0
02.04.2013	-0.5065	0	0	(0,0,0)	2384	0
03.04.2013	-15.3335	1	15.3335	(1,-936.336,-23.5)	2399.3	3.1
04.04.2013	-14.6885	1	14.6885	(1,-936.336,-23.5)	2414	3.1
05.04.2013	-1.6165	0	0	(0,0,0)	2414	0
08.04.2013	-3.254	0	0	(0,0,0)	2414	0
09.04.2013	-5.4335	0	0	(0,0,0)	2414	0
10.04.2013	-22.641	2	45.3	(2,-1872.672,-47)	2459.3	6.3
11.04.2013	-20.636	1	20.636	(1,-936.336,-23.5)	2480	3.15
12.04.2013	-44.3385	4	177.35	(4,-3745.344,-94)	2657.35	12.2
15.04.2013	-108.834	10	1088.34	(10,-9363.36,-235)	3745.7	28.3
16.04.2013	-100.799	10	1007.99	(10,-9363.36,-235)	4753.7	28.7
17.04.2013	-99.8765	10	998.765	(10,-9363.36,-235)	5752.5	28.6
18.04.2013	-91.651	10	916.51	(10,-9363.36,-235)	6669	28.8
19.04.2013	-86.0715	10	860.715	(10,-9363.36,-235)	7529.7	28.8
22.04.2013	-61.3585	10	613.585	(10,-9363.36,-235)	8143.3	29
23.04.2013	-65.2985	10	652.985	(10,-9363.36,-235)	8796.3	28.8
24.04.2013	-56.209	8	449.7	(8,-7490.7,-188)	9246	23.2
25.04.2013	-48.4765	5	242.4	(5,-4681.68,-117.5)	9488.4	14.86
26.04.2013	-36.3435	3	109	(3,-2809,-70.5)	9597.4	8.9
29.04.2013	-42.994	4	172	(4,-3745.344,-94)	9769.4	11.9
30.04.2013	-35.3035	3	106	(3,-2809,-70.5)	9875.4	8.9

Tabella 3.4: Chiudendo la posizione ora, in cui ci troviamo allo scoperto di 3642.5 titoli dell'argento e possediamo 155 titoli oro, il profit&loss sarà $PL = 9875.4 - 155 * 35.3035 = 4403.3575\$$. Tenendo conto dei costi di transazione, $PLC = 4403.3575 - 59.9 - 93.49 - 114.66 - 296.61 = 3838.7\$$

Data inizio	Data fine	PL (Profit&Loss in$)	PLC (Profit&Loss in $ con transazione)
01/01/2013	31/01/2013	504.102	444.2
01/01/2013	28/02/2013	657.27	503.88
01/01/2013	28/03/2013	2028.7	1760.65
01/01/2013	30/04/2013	4403.36	3838.7

Tabella 3.5: La tabella riassume i Profit&Loss ottenuti, a seconda dell'orizzonte di tempo in cui esercito.

Appendice A

Notazioni generali

- $\emptyset$ è l'insieme vuoto
- $\mathbb{N} = \{1, 2, 3...\}$ è l'insieme dei numeri naturali
- $\mathbb{N}_0 = \{0, 1, 2, 3...\}$ è l'insieme dei numeri interi non negativi
- $\mathbb{R}$ è l'insieme dei numeri reali
- $\mathbb{R}_+$ è l'insieme dei numeri reali positivi
- se A è un insieme contenuto in uno certo spazio, $A \in \Omega$, A^c è l'insieme ad esso complementare in quello spazio, $A_n, n \in \mathbb{N}$, è una successione di insiemi in Ω
- $(\Omega, \mathcal{F}, \mathrm{P})$ è uno spazio di probabilità, dove Ω è lo spazio campione, $\mathcal{F}$ è una σ-algebra e P è una misura di probabilità, discreta o continua.
- $\sigma(X)$ è la σ-algebra generata dalla variabile aleatoria X
- $p(x)$ è la funzione densità di probabilità della variabile aleatoria x
- $p(x_1, ..., x_n)$ è la funzione densità di probabilità congiunta delle variabili aleatorie $x_1, ..., x_n$
- $\mathcal{N}_{\mu,\sigma^2}$ è la distribuzione normale con media μ e varianza σ^2
- χ^2_ν è la distribuzione chi-quadro con ν gradi di libertà
- χ è la funzione indicatrice
- i.i.d indica la proprietà di variabili aleatorie stocasticamente indipendenti ed identicamente distribuite

Appendice B

Elementi di finanza matematica

In questo capitolo richiamiamo le nozioni fondamentali che ci servono per trattare il tema dell'arbitraggio statistico.

Definizione B.1 (Variabile aleatoria)**.** Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathrm{P})$, si definisce variabile aleatoria la funzione:

$$\begin{aligned} X: &\quad \Omega \to \mathbb{R} \\ : &\quad \omega \to X(\omega) \end{aligned}$$

tale che $X^{-1}(H) \in \mathcal{F} \quad \forall H \in \mathcal{F}$, dove $X^{-1}(H) = \{\omega \in \Omega | X(\omega) \in H\}$ Questo concetto è equivalente all'espressione *X è misurabile rispetto a $\mathcal{F}$*.

Rivediamo ora alcuni concetti di base, per semplicità solo nel caso in cui una variabile assuma valori in un insieme discreto.

Definizione B.2 (Valore atteso)**.** Sia X una variabile aleatoria discreta definita su $(\Omega, \mathcal{F}, \mathrm{P})$, $X = \{x_1, ...x_n\}$. si definisce *valore atteso* di X la media dei possibili valori di X pesati con le rispettive probabilità:

$$E[X] := \sum_{k=1}^{n} x_k P(X = x_k)$$

Nell'ambito della finanza matematica, il valore atteso è utilizzato nel contesto di *valutazione neutrale al rischio.* Si definisce il *prezzo equo* di un titolo secondo le probabilità degli eventi come il valore della variabile aleatoria che rapppresenta tale titolo.

Definizione B.3 (Media campionaria)**.** Sia X una variabile aleatoria definita su $(\Omega, \mathcal{F}, \mathrm{P})$. Si definisce *media campionaria* di X, calcolata su un campione di N elementi, la media degli N valori osservati di X $x_1, ..x_N$:

$$\overline{X} = \frac{\sum_{1=1}^{N} x_i}{N}$$

Definizione B.4 (Varianza)**.** Sia X una variabile aleatoria definita su $(\Omega, \mathcal{F}$, P), $X = \{x_1, ...x_n\}$. si definisce *varianza* di X il valore atteso della variabile aleatoria scarto quadratico di X dalla propria media:

$$V[X] := E[(X - E[X])^2] = \sum_{k=1}^{n} (x_k - E[X])^2 P(X = x_k)$$

La varianza stima mediamente quanto la variabile aleatoria X è distante dal suo valore atteso. In finanza stima la rischiosità di un'azione: quanto più è grande, tanto più l'azione considerata è rischiosa.

Definizione B.5 (Varianza campionaria)**.** Sia X una variabile aleatoria definita su $(\Omega, \mathcal{F}$, P). Si definisce *varianza campionaria* di X, calcolata su un campione di N elementi, $x_1, ..x_N$, la quantità:

$$S^2 = \frac{\sum_{k=1}^{N} (x_k - \overline{X})^2}{N}$$

Definizione B.6 (Covarianza)**.** Siano X e Y due variabili aleatorie definite su $(\Omega, \mathcal{F}$, P). Si definisce *covarianza* di X e Y:

$$COV(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Definizione B.7 (Indipendenza)**.** Siano X e Y due variabili aleatorie definite su $(\Omega, \mathcal{F}$, P). X e Y si dicono indipendenti in P se lo sono le σ-algebre $\sigma(X)$ e $\sigma(Y)$

Vale il seguente teorema, che non dimostriamo:

Teorema B.1. *Se X e Y sono due variabili aleatorie indipendenti in P allora* $E[XY] = EXEY$*.*

Osservazione B.1. Se X e Y sono due variabili aleatorie indipendenti, vale:

$$V(X + Y) = V(X) + V(Y)$$

Infatti, $V(X + Y) = E[(X - E[X] + Y - E[Y])^2] = V(X) + V(Y) + 2E[(X - E[X])(Y - E[Y])] =$
$= V(X) + V(Y) + E[(X - E[X])]E[(Y - E[Y])] =$ perché indipendenti
$V(X) + V(Y)$ $(E[X - E[X]] = E[X] - E[X] = 0)$

Definizione B.8 (Attesa condizionata)**.** Sia X una variabile aleatoria su su $(\Omega, \mathcal{F}$, P) e sia $\mathcal{G}$ una σ-algebra contenuta in $\mathcal{F}$, $\mathcal{G} \subseteq \mathcal{F}$. Una variabile aleatoria Z tale che:

1. Z è $\mathcal{G}$-misurabile
2. $E[Z\chi_A] = E[X\chi_A] \quad \forall A \in \mathcal{G}$

è detta attesa condizionata di X a $\mathcal{G}$ e si denota con $Z = E[X|\mathcal{G}]$.

Definizione B.9. Date due variabili aleatorie X e Y, si definisce attesa di X condizionata a Y

$$E[X|Y] := E[X|\sigma(Y)] \tag{B.1}$$

Essa può rappresentare finanziariamente la stima del prezzo di un titolo X, date le informazioni sul titolo Y.

Ricordiamo anche alcune proprietà dell'attesa condizionata che vengono utilizzate spesso:

1. Se X è $\mathcal{G}$-misurabile, $E[X|G] = X$. Ad esempio $E[X|G] = X$.
2. Se X e $\mathcal{G}$ sono indipendenti, $E[X|\mathcal{G}] = E[X]$. Dunque se X e Y sono indipendenti, $E[X|Y] = E[X]$
3. $E[cX + Y|\mathcal{G}] = cE[X|\mathcal{G}] + E[Y|\mathcal{G}]$
4. Se Y è $\mathcal{G}$-misurabile, $E[XY|\mathcal{G}] = YE[X|\mathcal{G}]$
5. $E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{G}] = E[E[X|\mathcal{G}]|\mathcal{H}] \quad \mathcal{G} \subseteq \mathcal{H}$

Introduciamo infine il concetto di filtrazione, un oggetto che farà parte della definizione di spazio di probabilità.

Definizione B.10. Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \text{P})$, una filtrazione $(\mathcal{F}_n)_{n \geqslant 0}$ è una famiglia crescente di sotto-sigma algebre di $\mathcal{F}$, cioè $\mathcal{F}_n \subseteq \mathcal{F}_{n+1} ... \subseteq \mathcal{F} \quad \forall n \geqslant 0$

cioè ad ogni istante temporale è associata una σ-algebra $\mathcal{F}_n$ che rappresenta le informazioni possedute fino al tempo n. In finanza, tali informazioni possono essere rappresentate ad esempio dai prezzi dei titoli sul mercato in fino al tempo n.

Bibliografia

[1] A. Pascucci, 'Calcolo stocastico per la finanza', Springer- Verlag, Milano, 2007

[2] Burgess A.N.(1999) A Computational Methodology for Modelling the Dynamics of Statistical Arbitrage. University of London, London Business School

[3] Avellaneda, M., Lee, J.H. (2008). Statistical arbitrage in the U.S. equity market

[4] Bondarenko, O. (2003). Statistical arbitrage and securities prices. Review of Financial Studies.

[5] Burgess, A. N. (1999b). Statistical arbitrage models of the FTSE 100. Computational finance, edited by Abu-Mostafa.

[6] Granger, C. W. J. (1983). Cointegrated variables and error-correcting models. UCSD Discussion Paper.

[7] Pole, A. (2007). Statistical Arbitrage. WILEY FINANCE.