

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI FISICA E ASTRONOMIA
"GALILEO GALILEI"

Laurea Triennale in Fisica

**Test di Kolmogorov-Smirnov nel caso di
Osservazioni Dipendenti**

Candidato:
Tommaso FURIERI

Relatore:
Prof. Attilio STELLA

Co-Relatore:
Dott. Michele CARAGLIO

Anno Accademico 2015-2016

INDICE

1	Test di Kolmogorov-Smirnov	3
1.1	Funzionamento del test	3
1.2	Determinazione dei valori critici	4
1.3	Caso di media e sigma stimati dal campione	5
2	Serie Temporal e Processi Autoregressivi	7
2.1	Serie temporali	7
2.2	Processo AR(p)	8
2.3	Processo ARCH(q)	10
3	Modifica della Statistica di Kolmogorov-Smirnov nel caso di Variabili Correlate	11
3.1	Modifica della statistica K-S nel caso di processi AR(2)	11
3.2	Modifica della statistica K-S nel caso di processi ARCH(1)	15
4	Conclusioni	19

INTRODUZIONE

Il test di Kolmogorov-Smirnov (test K-S) è un test non parametrico utilizzato per testare il grado di accettabilità che la distribuzione empirica $\mathcal{F}_n(x)$ di un campione di variabili casuali provenga da una determinata distribuzione di probabilità teorica $\mathcal{F}(x)$ (test K-S ad un campione). Il test può anche essere esteso al caso in cui si voglia verificare se due campioni indipendenti provengano dalla stessa distribuzione (test K-S a due campioni). In questa tesi ci occuperemo principalmente del primo caso.

Il test ha come requisiti fondamentali sia l'indipendenza tra il campione e la distribuzione di riferimento, sia l'indipendenza tra le variabili che costituiscono il campione. Sotto tali ipotesi il test risulta indipendente dalla forma di $\mathcal{F}(x)$ e dipende solamente dalla dimensione n del campione. Tuttavia è spesso utile stimare i parametri della distribuzione direttamente dal campione. Per esempio, se si suppone che la distribuzione abbia un andamento gaussiano, possiamo stimare media μ e varianza σ^2 dal campione ed eseguire successivamente un test K-S. In questo caso cade l'ipotesi di indipendenza tra campione e distribuzione di riferimento. Pertanto il test dovrà essere modificato per poter ancora produrre dei risultati attendibili (test di Lilliefors) [1]. Sarebbe poi estremamente utile poter applicare questo test nel caso di serie temporali, come possono essere: l'andamento di un titolo in borsa, il livello di polveri sottili in una città, la temperatura media nel corso degli anni di una zona geografica ecc. Tuttavia tutti gli esempi citati hanno in comune la correlazione tra i dati. Pertanto anche in questo caso il test di Kolmogorov-Smirnov non risulta più applicabile.

Lo scopo di questo lavoro è di studiare come il test K-S debba essere modificato per poterlo applicare anche a serie temporali correlate. La tesi sarà strutturata in tre parti:

- Nel Capitolo 1 vi sarà una presentazione della teoria e dei risultati che hanno portato allo sviluppo del test di Kolmogorov-Smirnov e delle modifiche da effettuare nel caso in cui i parametri della distribuzione siano stimati dal campione.
- Successivamente nel Capitolo 2 verranno esposti alcuni risultati riguardanti lo studio di serie temporali e processi autoregressivi di tipo AR(p) e ARCH(q).
- Infine nel Capitolo 3 verrà presentato un articolo [2] riguardante un primo tentativo di modificare il test K-S nel caso di serie temporali generate da un processo autoregressivo AR(2). Per concludere, con un lavoro originale, si è cercato di applicare le idee appena descritte nel caso di una serie temporale generata da un processo ARCH(1), ampiamente utilizzato in finanza.

1 | TEST DI KOLMOGOROV-SMIRNOV

Il test di Kolmogorov-Smirnov ad un campione è un test non parametrico, unidimensionale e valido nel continuo. Esso viene usato per valutare l'accettabilità o meno di un sample costituito da variabili indipendenti ed identicamente distribuite provenga da una distribuzione $\mathcal{F}(x)$.

FUNZIONAMENTO DEL TEST

Supponiamo di avere un insieme $X = \{x_1, x_2, \dots, x_n\}$ di n numeri e che si voglia verificare se tale insieme possa essere stato generato da una funzione densità di probabilità (PDF) $\mathcal{F}(x)$. Chiamiamo \mathcal{H}_0 l'ipotesi che il campione provenga da $\mathcal{F}(x)$. Sia $\mathcal{R}_n(x)$ la funzione di ripartizione empirica (cumulativa) del sample definita come:

$$\mathcal{R}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n J_{[-\infty, x]}(x_i) \quad (1.1)$$

dove $J_{[-\infty, x]}(X_i)$ è la funzione indicatrice che può assumere il valore 1 o 0 a seconda se $x_i < x$ o $x_i > x$; mentre chiamiamo $\mathcal{R}(x)$ la funzione di ripartizione di $\mathcal{F}(x)$.

Definiamo la statistica di Kolmogorov-Smirnov (statistica K-S) come:

$$D_n \equiv \sup_{x \in \mathbb{R}} |\mathcal{R}_n(x) - \mathcal{R}(x)|. \quad (1.2)$$

Se il sample proviene dalla distribuzione $\mathcal{F}(x)$, al tendere di $n \rightarrow \infty$, si avrà che $D_n \rightarrow 0$. Questo risultato, piuttosto intuitivo, è dovuto al teorema di Glivenko-Cantelli [8], il quale verrà enunciato in modo preciso nella prossima sezione.

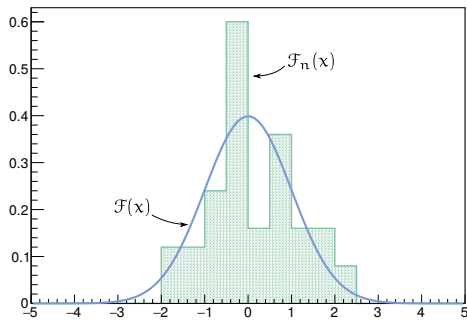


Figura 1.1: Sovrapposizione tra l'istogramma di un sample $\mathcal{F}_n(x)$ e la funzione $\mathcal{F}(x)$ da cui è stato generato.

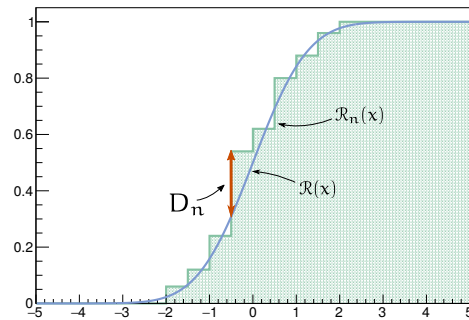


Figura 1.2: Sovrapposizione della funzione di ripartizione empirica $\mathcal{R}_n(x)$ del sample di sinistra e la funzione di ripartizione $\mathcal{R}(x)$.

Si deve quindi confrontare il risultato D_n ottenuto con un valore di riferimento $\mathcal{K}_{\alpha, n}$, chiamato anche valore critico. Questo valore dipende dai parametri α e n , dove α definito come

$$\alpha \equiv \Pr(\mathcal{K} > \mathcal{K}_{\alpha, n})$$

è il livello di confidenza, ovvero la probabilità che un campione di n elementi generato dalla distribuzione $\mathcal{F}(x)$ ottenga un valore $D_n = \mathcal{K} > \mathcal{K}_{\alpha, n}$. Pertanto se risulta $D_n > \mathcal{K}_{\alpha, n}$ rifiutiamo

l'ipotesi \mathcal{H}_0 con un livello di confidenza pari ad α ; o, nel caso opposto in cui $D_n < \mathcal{K}_{\alpha, n}$, accettiamo l'ipotesi \mathcal{H}_0 con un livello di confidenza pari ad $P = 1 - \alpha$.

DETERMINAZIONE DEI VALORI CRITICI

Dimostriamo innanzitutto un asserto che sarà utile in seguito.

Lemma. *Sia x una qualsiasi variabile casuale continua con funzione di ripartizione $\mathcal{R}(x)$, allora $\mathcal{U} \equiv \mathcal{R}(x)$ è una variabile casuale distribuita uniformemente in $[0, 1]$*

Dimostrazione. Sia $u \in [0, 1]$. Data la continuità di x , $\exists x_u \in \mathbb{R}$ tale che $\mathcal{R}(x_u) = u$. Per la definizione di funzione di ripartizione si ha:

$$\mathcal{R}(u) = \Pr(\mathcal{U} < u) = \Pr(\mathcal{R}(x) < \mathcal{R}(x_u)) = \mathcal{R}(x_u) = u$$

dove nel secondo passaggio si è usato il fatto che la funzione di ripartizione è monotona crescente. Pertanto la variabile \mathcal{U} è distribuita uniformemente in $[0, 1]$. \square

Da ciò si evince che una qualsiasi variabile casuale con funzione di ripartizione $\mathcal{R}(x)$ può essere mappata in $[0, 1]$ attraverso \mathcal{R} .

Corollario. *Sia $\mathcal{R}_n(x)$ definita come in 1.1. Se $\mathcal{R}(x)$ è continua allora la distribuzione di D_n non dipende da \mathcal{R}*

Dimostrazione. Siano $u = \mathcal{R}(x)$ e $x = \mathcal{R}^{-1}(u)$. Calcoliamoci la probabilità che D_n risulti minore di un certo valore η :

$$\Pr\left(\sup_{x \in \mathbb{R}} |\mathcal{R}_n(x) - \mathcal{R}(x)| \leq \eta\right) = \Pr\left(\sup_{u \in [0, 1]} \left| \mathcal{R}_n(\mathcal{R}^{-1}(u)) - \mathcal{R}(\mathcal{R}^{-1}(u)) \right| \leq \eta\right).$$

Usando la definizione 1.1 di funzione di ripartizione empirica si ha:

$$\mathcal{R}_n(\mathcal{R}^{-1}(u)) = \frac{1}{n} \sum_{i=1}^n \mathcal{J}_{[-\infty, \mathcal{R}^{-1}(u)]}(X_i) = \frac{1}{n} \sum_{i=1}^n \mathcal{J}_{[-\infty, u]}(\mathcal{R}(X_i)) = \frac{1}{n} \sum_{i=1}^n \mathcal{J}_{[-\infty, u]}(U_i)$$

Dove nell'ultimo passaggio si è introdotta la variabile casuale U_i , distribuita uniformemente in $[0, 1]$, come mostrato nel Lemma 1.2. Pertanto risulta:

$$\Pr\left(\sup_{x \in \mathbb{R}} |\mathcal{R}_n(x) - \mathcal{R}(x)| \leq \eta\right) = \Pr\left(\sup_{u \in [0, 1]} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{J}_{[-\infty, u]}(U_i) - u \right| \leq \eta\right)$$

che chiaramente non dipende da \mathcal{R} . \square

Partendo da questi risultati, Kolmogorov ha dimostrato che per $n \rightarrow \infty$ la distribuzione della variabile casuale $\sqrt{n}D_n$ tende a quella della variabile $\sup|B(\mathcal{R}(x))|$, dove $B(t)$ è il "Brownian Bridge", ovvero un moto browniano ad estremi fissati [3]. Questo risultato è noto anche con il nome di "Teorema di Kolmogorov".

A titolo di esempio in Fig.1.3 sono state graficate le distanze $|\mathcal{R}_n(x) - \mathcal{R}(x)|$, dove, in questo caso, $\mathcal{R}(x)$ è la funzione di ripartizione di una PDF uniforme in $[0; 1]$, mentre in Fig.1.4 è stato graficato un Brownian Bridge.

Se adesso per entrambi in grafici condiseriamo il Sup e ripetiamo n volte questa operazione otteniamo il risultato mostrato nelle Fig.1.5 e 1.6 per $n = 10^4$, ovvero che, in distribuzione:

$$\sqrt{n}D_n \rightarrow \sup|B(\mathcal{R}(x))|; \quad n \rightarrow \infty. \quad (1.3)$$

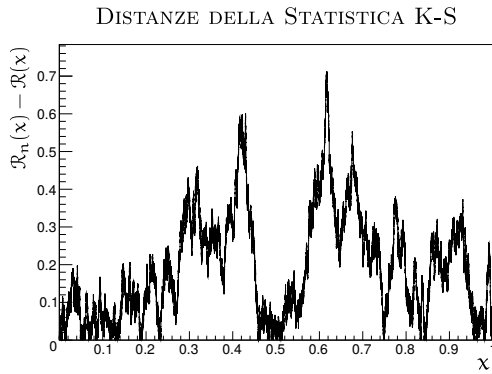


Figura 1.3: Grafico delle distanze calcolate attraverso la statistica di Kolmogorov per un campione uniforme

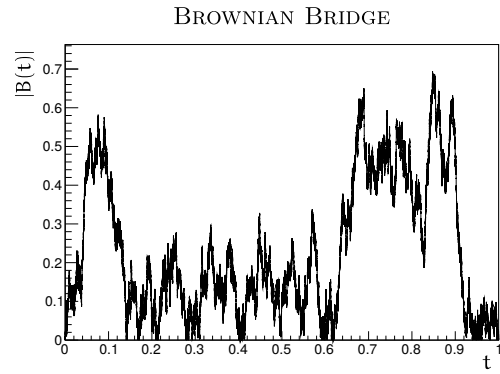


Figura 1.4: Moto browniano unidimensionale ad estremi fissati

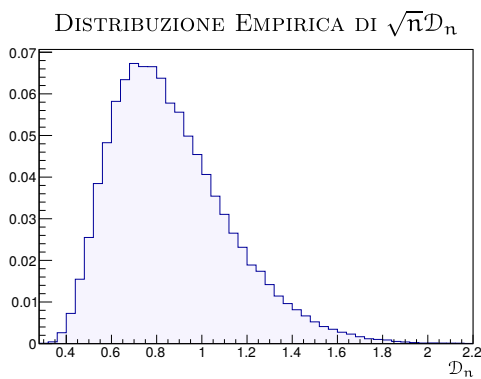


Figura 1.5

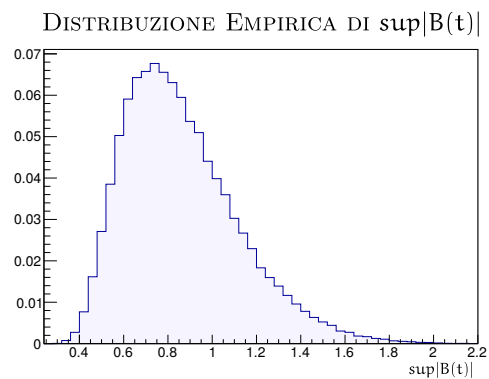


Figura 1.6

Questo è il risultato centrale mostrato da Kolmogorov, e ciò che più è importante è che la forma di questa distribuzione è indipendente da $\mathcal{R}(x)$. Questa distribuzione viene chiamata "Distribuzione di Kolmogorov" ed ha la forma:

$$\mathcal{D}_K(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2} \tag{1.4}$$

Nel caso in cui n non tenda ad infinito non è più vero che $D_n \rightarrow \mathcal{D}_K(x)$, pertanto la distribuzione di D_n deve essere ricavata per ogni n .

Nel corso degli anni sono stati diversi gli approcci per determinare i valori critici $\mathcal{K}_{\alpha,n}$, come l'uso di formule asintotiche ed attraverso risoluzioni approssimate di equazioni che legano $\mathcal{K}_{\alpha,n}$ ad α ed n [4]. È anche possibile utilizzare metodi Monte Carlo (MC), in modo simile a quanto fatto per la determinazione empirica della distribuzione $\mathcal{D}_K(x)$. L'idea di base è quella di generare un numero molto alto di campioni di lunghezza n e per ciascun campione calcolare la statistica di Kolmogorov D_n . In questo modo otterremo una distribuzione empirica dalla quale sarà possibile determinare i valori critici $\mathcal{K}_{\alpha,n}$ per ogni livello di confidenza α al variare della dimensione del campione. Con questo metodo attualmente è possibile ottenere dai valori attendibili alla terza cifra decimale in meno di un minuto di simulazione in single thread.

CASO DI MEDIA E SIGMA STIMATI DAL CAMPIONE

Nella sezione precedente ci siamo limitati ad esporre i principali risultati ottenuti da Kolmogorov e Smirnov, ma non abbiamo sottolineato sotto quali condizioni essi valgano.

Infatti la 1.3 poggia, come già detto, sul teorema di Glivenko-Cantelli.

Teorema (di Glivenko-Cantelli). *Sia ω una variabile casuale a cui è associata una funzione di ripartizione $\mathcal{R}(x)$ e sia $X = \{x_1, x_2, \dots, x_n\}$ una sequenza (vettore) di variabili casuali indipendenti ed identicamente distribuite generate da ω . Sia inoltre $\mathcal{R}_n(x)$ la funzione di ripartizione empirica di X definita come in 1.1.*

Allora si ha

$$\lim_{n \rightarrow \infty} \sup_x |\mathcal{R}_n(x) - \mathcal{R}(x)| = 0 \quad (1.5)$$

Da qui emerge che in presenza di variabili casuali correlate la 1.3 potrebbe non essere più valida. Inoltre, per essere valida, la statistica di Kolmogorov richiede indipendenza tra il campione e la funzione $\mathcal{F}(x)$. Per capire il motivo di questa affermazione, è utile ricordare che il test di K-S può essere utilizzato nella variante a due campioni, utilizzando la stessa procedura descritta finora. Alla luce di questo risulta più comprensibile il fatto che, per testare che i due campioni provengano dalla stessa distribuzione di probabilità, essi debbano essere indipendenti. Pertanto stimare alcuni parametri di $\mathcal{F}(x)$ dal campione in esame porta ad una procedura scorretta che ne influenza la distribuzione asintotica 1.4.

Si può mostrare [5] che se i parametri stimati dal campione sono parametri di scala o posizione, possono essere ancora ricavati nuovi valori critici in funzione di α e n . La procedura per la determinazione dei nuovi valori è chiamata "Test di Lilliefors", per la quale è necessario adottare metodi MC come descritto in [1]. È fondamentale notare che da questo momento in poi cade un'importante conseguenza del teorema di Kolmogorov, ovvero che il test è indipendente dalla scelta della PDF. Pertanto per ogni scelta di \mathcal{F} si avranno distribuzioni e di conseguenza valori critici diversi.

La procedura, simile a quella descritta per la determinazione dei valori critici per il test di Kolmogorov-Smirnov, è la seguente:

- Si simula un campione di n dati proveniente dalla PDF scelta.
- Si stimano i parametri della distribuzione dal campione.
- Si crea la funzione di ripartizione empirica del campione.
- Si calcola la statistica di Kolmogorov-Smirnov D_n tra la funzione di ripartizione stimata e quella empirica ottenuta dal sample.

Ripetendo un numero molto alto di volte questi passaggi, possiamo costruirci la distribuzione di D_n e da questa ricavarci i valori critici. Anche in questo caso è possibile eseguire questi passaggi in un moderno pc ottenendo una precisione alla terza cifra decimale rispetto ai valori critici corretti [6] in circa un minuto di simulazione.

2

SERIE TEMPORALI E PROCESSI AUTOREGRESSIVI

Il test di Kolmogorov-Smirnov può essere applicato anche nel caso in cui le variabili derivino da una serie temporale, se questa non presenta correlazione.

In questo capitolo esporremo brevemente alcuni concetti fondamentali riguardanti le serie temporali ed il loro studio.

Nel caso in cui la serie presenti correlazione, invece, questa può essere riprodotta attraverso particolari processi, chiamati "Autoregressivi", dove il valore della variabile casuale dipende dai valori che ha assunto precedentemente, secondo una certa legge dipendente dal modello utilizzato. Verranno quindi studiati due particolari modelli autoregressivi: modello AR(p) e ARCH(q).

I risultati che otterremo saranno utilizzati nel capitolo seguente per modificare il test di Kolmogorov-Smirnov nel caso in cui la serie temporale sia modellizzabile attraverso processi autoregressivi di tipo AR(2) o ARCH(1).

SERIE TEMPORALI

Definizione 1 (Processo Stocastico). Sia \mathcal{T} un sottoinsieme di $[0, +\infty)$. Una famiglia di variabili casuali $\{Y_t\}_{t \in \mathcal{T}}$ indicizzate da \mathcal{T} è detto processo stocastico.

Definizione 2 (Serie temporale). Una serie temporale è una collezione di dati ottenuti attraverso misure successive di una grandezza eseguite ad intervalli di tempo consecutivi ed equispaziati¹.

Quindi una serie temporale è una parte finita di un processo stocastico a tempo discreto.

Uno degli obiettivi fondamentali, quando si lavora con serie temporali, è quello di capirne la dinamica, ovvero come la serie evolve nel tempo. Si può dimostrare che la distribuzione di probabilità di un processo stocastico è la distribuzione di probabilità congiunta di tutte le distribuzioni di probabilità del vettore delle variabili $\{Y_{t_1}; \dots; Y_{t_n}\}$ del processo. Tuttavia nella pratica stimare dai dati tutte queste distribuzioni è impossibile, pertanto nello studio di serie temporali ci si limita a considerare particolari momenti del processo:

$$\text{media: } \mu_t = \mathbb{E}(Y_t);$$

$$\text{varianza: } \sigma_t^2 = \text{Var}(Y_t) = \mathbb{E}[Y_t^2] - (\mathbb{E}[Y_t])^2$$

$$\text{autocovarianza: } \gamma(t', t'') = \text{cov}(Y_{t'}, Y_{t''}) = \mathbb{E}[(Y_{t'} - \mathbb{E}(Y_{t'}))(Y_{t''} - \mathbb{E}(Y_{t''}))]$$

$$\text{autocorrelazione: } \rho(t', t'') = \frac{\gamma(t', t'')}{\sigma_{t'} \sigma_{t''}}$$

Inoltre è utile fare delle ipotesi sul processo in esame; una delle più comuni è l'ipotesi di stazionarietà.

Definizione 3 (Processo Stocastico Stazionario).

Un processo stocastico è detto **stazionario in senso stretto** se la distribuzione di probabilità congiunta del vettore $\{Y_{t_0+\delta}; \dots; Y_{t_n+\delta}\}$ è indipendente da δ .

Un processo stocastico è detto **stazionario in senso lato** se:

$$\begin{aligned} \mathbb{E}(Y_{t'}) &= \mathbb{E}(Y_{t''}) \\ \text{cov}(Y_{t'}, Y_{t''}) &= \text{cov}(Y_{t'+\delta}, Y_{t''+\delta}) \end{aligned}$$

¹ In realtà l'equispaziatura temporale dei dati non è necessaria per la definizione di serie temporale, ma ai fini di questa tesi assumiamo questa richiesta come valida.

$\forall \delta, t', t''$.

Descriviamo ora uno strumento utile nell'analisi di serie temporalì, ovvero lo spettro di potenza di una serie.

Definizione 4 (Spettro di Potenza).

Sia Y_t un processo stocastico stazionario in senso lato con funzione di autocovarianza $\gamma(0, t) \equiv \gamma(t)$. Lo spettro di potenza $\Gamma(\omega)$ di Y_t è la trasformata di Fourier della funzione di autocovarianza $\gamma(t)$.

$$\Gamma(\omega) \equiv \mathbf{F}[\gamma](\omega) = \sum_{t=-\infty}^{+\infty} \gamma(t) e^{-it\omega} \quad (2.1)$$

In molti casi quando si ha a che fare con serie temporalì è piú significativo studiare non tanto l'andamento di Y_t , quanto piuttosto la variazione della variabile al passare del tempo, rispetto al valore assunto precedentemente. Possiamo, per esempio, definire una nuova variabile $Z_{t_i} = Y_{t_i} - Y_{t_{i-1}}$ con $Z_{t_0} = 0$. Altre definizioni sono possibili, ma non siamo interessati a questo.

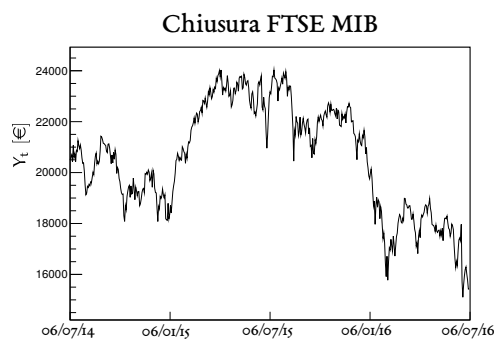


Figura 2.1: Andamento del prezzo di chiusura dell'indice FTSE MIB nel periodo dal 06/07/2014 al 06/07/2016

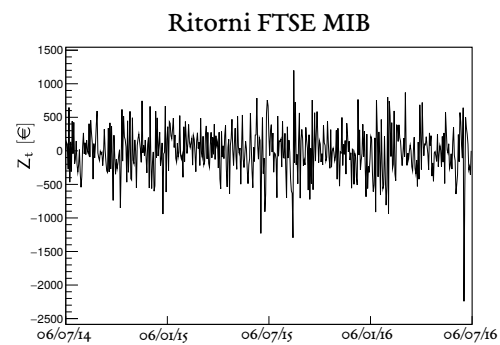


Figura 2.2: Grafico delle differenze del prezzo di chiusura Z_t dell'indice FTSE MIB nel periodo dal 06/07/2014 al 06/07/2016

A questo punto quello che si può fare è cercare di capire qual è il processo che ha generato tale serie temporale. Graficamente, per aiutarsi, si può costruire l'istogramma di Z_{t_i} . Per esempio, prendendo in analisi un Moto Browniano, questo sarà generato da una famiglia di variabili casuali gaussiane indipendenti, con media e sigma costanti. Pertanto l'istogramma di Z_{t_i} sarà ancora una gaussiana. Come già detto, però, nello studio di serie temporalì ci si trova spesso di fronte a forte correlazione fra i dati. Per cercare di riprodurre tali processi, si ricorre all'ausilio di modelli che riescano ad emulare le caratteristiche della serie temporale in esame. Esistono molti tipi di modelli, ma la nostra attenzione sarà principalmente focalizzata su due in particolare. Di questi ora verrà data una definizione e verranno espòste alcune proprietà che saranno utili in seguito:

- *modello lineare autoregressivo* (di ordine p), denominato $\text{AR}(p)$.
- *modello autoregressivo* (di ordine q) *ad eteroschedasticità condizionata*, denominato $\text{ARCH}(q)$.

PROCESSO $\text{AR}(p)$

Definizione 5. Un processo stocastico $\{Z_t\}$ si dice autoregressivo di ordine p se:

$$Z_{t_i} = \epsilon_t + c + \sum_{j=1}^p a_j Z_{t_{i-j}}; \quad i > p \quad (2.2)$$

dove ϵ_t è un processo di rumore bianco e $\{a_i, c\} \in \mathbb{R}$.

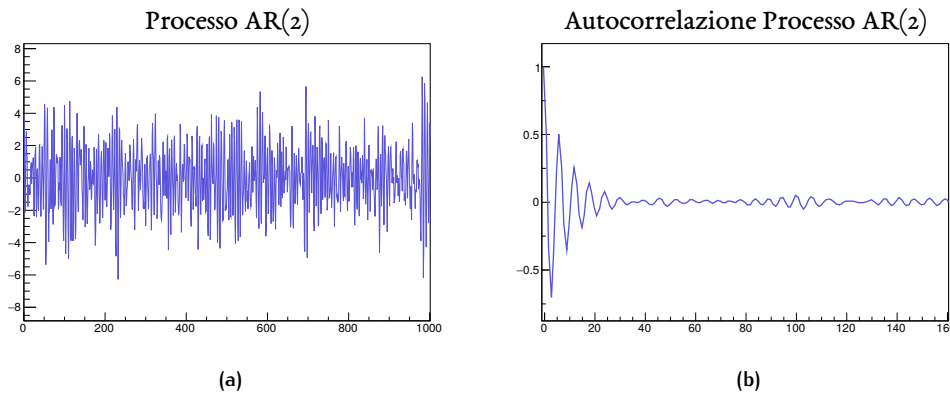


Figura 2.3: A sinistra (2.3a) viene mostrato un esempio di serie derivante da un processo AR(2) con $a_1 = 0.9$, $a_2 = -0.8$ ed $\epsilon_t = \mathcal{N}(0, 1)$. La varianza non condizionata del processo è $\text{var}_u \simeq 3.7$. A destra (2.3b) è graficata la sua funzione di autocorrelazione

Si ha [7] che la condizione di stazionarietà in senso lato per un processo AR(p) è soddisfatta se tutte le radici complesse z_i del polinomio $P(z) = 1 - \sum_{i=1}^p a_i z^i$ soddisfano $|z_i| > 1$, ovvero giacciono all'esterno della circonferenza unitaria. Inoltre si può mostrare [9] che per una serie AR(1), se il rumore bianco è generato da un processo gaussiano di media $\mu = 0$ e varianza σ^2 , la media e la varianza non condizionate del processo sono date da:

$$\mu_u = 0$$

$$\text{var}_u = \frac{\sigma}{1-a_1^2}$$

e la distribuzione non condizionata di $\{Z_t\}$ è gaussiana con media μ_u e varianza σ_u^2 . In generale si ha che un processo stabile di tipo AR(q) con rumore bianco gaussiano avrà una distribuzione non condizionata $\{Z_t\}$ a sua volta gaussiana di media e varianza non condizionate²:

$$\mu_u = 0$$

$$\text{var}_u = \sigma_u^2 = \frac{\sigma}{1 - \sum_{k=1}^p a_k \rho(k)}$$

dove $\rho(k)$ è la funzione di autocorrelazione del processo.

Studiando in dettaglio un processo AR(2) si ha che è stazionario per:

$$a_1 + a_2 < 1, \quad a_2 - a_1 < 1, \quad |a_2| < 1. \quad (2.3)$$

Infatti applicando la condizione di stazionarietà a $P(z) = 1 - a_1 z - a_2 z^2$ si ha:

$$z_{1,2} = \frac{a_1 \pm \sqrt{a_1^2 - 4a_2}}{2} > 0$$

che conduce proprio alle condizioni 2.3.

La funzione di autocorrelazione valutata agli istanti 1 e 2 è:

$$\rho(1) = \frac{a_1}{1-a_2}$$

$$\rho(2) = \frac{a_1^2 - a_2^2 + a_2}{1-a_2}$$

e la varianza non condizionata è data da :

$$\text{var}_u = \sigma_u^2 = \frac{1}{1 - a_1 \rho(1) - a_2 \rho(2)}.$$

² [econweb.rutgers.edu/ctamayo/teaching/AR\(1\)_process.pdf](http://econweb.rutgers.edu/ctamayo/teaching/AR(1)_process.pdf)

Si mostra inoltre che lo spettro di potenza normalizzato è dato da:

$$\Gamma(\omega) = \frac{1}{\pi\sigma_u^2} \frac{t_s\sigma^2}{1 + a_1^2 + a_2^2 + 2a_1(a_2 - 1)\cos(\omega t_s) - 2a_2\cos(2\omega t_s)}; \quad 0 \leq \omega t_s < \pi \quad (2.4)$$

dove t_s è l'intervallo di acquisizione.

PROCESSO ARCH(q)

Definizione 6. Un processo stocastico $\{Z_t\}$ si dice autoregressivo ad eteroschedasticità condizionata di ordine q se

$$Z_{t_i} = \sigma_{t_i} \epsilon_t \quad (2.5)$$

con

$$\sigma_{t_i}^2 = c + \sum_{j=1}^q a_j Z_{t_i-j}^2; \quad i > q, \quad c > 0, \quad a_j \geq 0$$

ed ϵ_t un processo di rumore bianco.

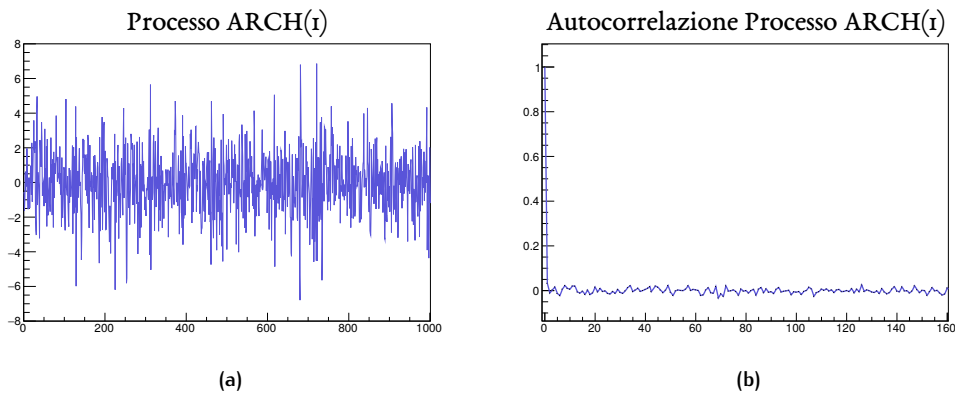


Figura 2.4: A sinistra (2.4a) viene mostrato un esempio di serie derivante da un processo ARCH(1) con $c = 20/9$, $a_1 = 0.5$ ed $\epsilon_t = N(0, 1)$. La varianza non condizionata del processo è $\text{var}_u \simeq 3.7$. A destra (2.4b) è graficata la sua funzione di autocorrelazione

Anche in questo caso il processo è stazionario in senso lato se le radici complesse z_i del polinomio $P(z) = 1 - \sum_{i=1}^q a_i z^i$ soddisfano $|z_i| > 1$, ovvero giacciono all'esterno della circonferenza unitaria e si hanno [9] le seguenti relazioni per la media e la varianza non condizionate nel caso in cui il rumore bianco sia generato da un processo gaussiano di media $\mu = 0$ e varianza $\sigma^2 = 1$:

$$\mu_u = 0$$

$$\text{var}_u = \frac{c}{1 - \sum_{i=1}^q a_i}$$

Affinchè quest'ultima sia positiva e finita si deve avere:

$$a_0 > 0 \quad \cap \quad \sum_{i=1}^q a_i < 1. \quad (2.6)$$

Inoltre, per un processo ARCH(1), si ha che la curtosi è funzione del parametro a_1 secondo la relazione:

$$\kappa = 3 \frac{1 - a_1^2}{1 - 3a_1^2}. \quad (2.7)$$

Non esiste una formula analitica per la distribuzione non condizionata di un processo ARCH(q) e per il suo spettro di potenza.

3

MODIFICA DELLA STATISTICA DI KOLMOGOROV-SMIRNOV NEL CASO DI VARIABILI CORRELATE

Come già detto, se l'insieme dei dati presenta correlazione, il test di Kolmogorov-Smirnov non produce più risultati attendibili. Tuttavia ci si trova spesso in situazioni dove è presente una forte correlazione tra i dati, come ad esempio con dati estratti da serie temporali finanziarie. In tali casi si può considerare la funzione di autocorrelazione per valutare la correlazione tra i dati a tempi diversi per poi opportunamente scartarne alcuni, in modo da annullarne la dipendenza (diminuzione del rate di acquisizione). Tuttavia questo non è un buon approccio al problema, poiché, nella pratica, ci si trova sempre con campioni di lunghezza finita ed in questo modo il numero di dati a disposizione viene ridotto notevolmente con una conseguente diminuzione di efficacia del test.

Un primo tentativo per determinare una procedura che possa modificare il test di Kolmogorov-Smirnov per adattarlo al caso di variabili correlate si trova in [2].

MODIFICA DELLA STATISTICA K-S NEL CASO DI PROCESSI AR(2)

In questo articolo è presente un metodo per modificare valori critici della statistica di K-S, nel caso in cui il campione di dati provenga da un processo AR(2), dato dalla Definizione 5, con $c = 0$ e rumore bianco ϵ_t gaussiano. Lo studio nasce dalla necessità di applicare il test K-S nel caso di un campione proveniente da dati encefalografici, caratterizzati da forte correlazione e da spettri di potenza con frequenze concentrate attorno ad un picco approssimabili attraverso processi AR(2). Verranno ora presentati direttamente il metodo utilizzato ed i risultati ottenuti.

IL METODO

In via preliminare sono stati ritrovati, attraverso simulazioni MC, i valori critici ottenuti da Miller [4], nel caso di variabili indipendenti, e da Lilliefors [1], nel caso di media e sigma stimati dal campione. Nel primo caso sono stati generati 1000 campioni di $n = 25$ elementi, mentre il secondo caso è stato analizzato più in dettaglio, generando 10000 campioni per diversi valori di n da 16 a 256. Dai dati ottenuti è stata estrapolata una legge per il calcolo dei valori critici, che è stata poi utilizzata per trovare i valori critici KS nella 3.2

$$KS(\alpha, n) = M(\alpha, n)(0.7 - 1.8n^{-3/2} + 3.15n^{-2} + 0.075\alpha + 0.01245 \ln \alpha) \quad (3.1)$$

con

$$M(1\alpha, n) = \sqrt{-\frac{n \ln \alpha}{2}} - 0.16693n^{-1} - 0.025864 \sqrt{-\frac{\ln \alpha^3}{n}} + 0.0028575 \frac{\ln^2 \alpha}{\sqrt{n^3}} - 0.08467 \frac{\alpha}{\sqrt{n^3}} - 0.11143 \frac{1}{\sqrt{n^3}}$$

Non ci soffermiamo ulteriormente su tale formula poiché attualmente è possibile ottenere valori estremamente più precisi.

Per la generazione di numeri pseudo-casuali è stato utilizzato un algoritmo con ciclo di 2^{22} . Per ogni simulazione vengono generati 1000 campioni provenienti da una serie AR(2) per n compreso tra 16 e 256 per vari valori dei parametri α_1 ed α_2 . Per ogni campione generato sono

stati stimati la media μ e la varianza σ^2 dal campione e si è eseguita una statistica di Kolmogorov. Giunti a questo punto è possibile ottenere i valori critici come nel caso non correlato.

Ogni simulazione è stata eseguita almeno tre volte mediando così i valori critici ottenuti per ogni simulazione.

Tornando ad analizzare la 2.3 possiamo definire delle regioni di piano che delineano comportamenti in frequenza differenti. Infatti si ha che per $a_2 > -a_1^2/4$ le radici $z_{1,2}$ sono reali, mentre per $a_2 < -a_1^2/4$ sono immaginarie. Questo si ripercuote sullo spettro della serie.

Per

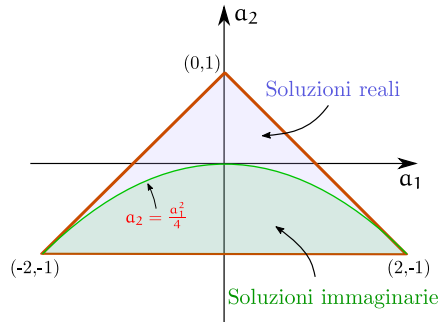
$$|a_1(1 - a_2)| < |4a_2| \cap |a_2| < 1$$

lo spettro ha un punto stazionario in

$$\omega_0 = \frac{\cos^{-1}\left(\frac{a_1(a_2-1)}{4a_2}\right)}{t_s}$$

In particolare nella regione $a_2 < -a_1^2/4 \cap a_2 < 0$ si ha un picco a bassa frequenza, e nella regione $a_2 > -a_1^2/4 \cap a_2 < 0$ si ha un picco ad alta frequenza. Vengono quindi studiate in dettaglio sia la regione

$a_1 > 0 \cup a_2 < 0$ sia la regione $a_1^2 + 4a_2 < 0$. Si è inoltre seguito lo stesso procedimento per un processo AR(1), ovvero ponendo $a_2 = 0$ ed indagando la regione $a_1 > 0$.



Spettri di Potenza Processi AR(2)

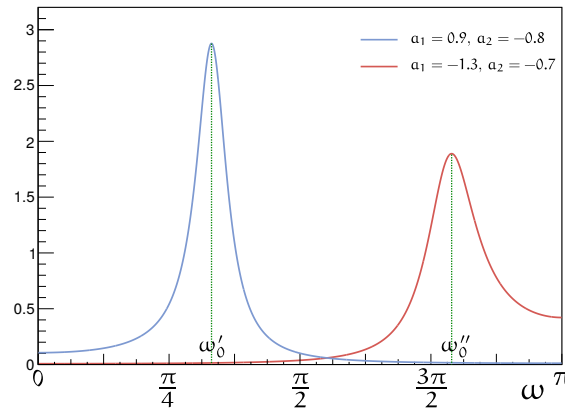


Figura 3.1: Esempio di spettri che presentano comportamenti in frequenza differenti.

I RISULTATI

Dai risultati non emerge alcuna relazione semplice tra i parametri a_1, a_2 e i valori critici, tuttavia è stato notato che, in generale, per serie con spettri aventi picchi stretti e a bassa frequenza si hanno valori critici più elevati rispetto a serie con spettri aventi picchi larghi e ad alta frequenza.

Si è cercato allora di trovare una relazione tra lo spettro della serie, ottenuto applicando la 2.1, ed i valori critici ottenuti. Analizzando i vari momenti dello spettro di potenza si è osservato che il parametro adimensionale

$$\tau = \sqrt{\frac{1/f_2 - 1/f_4}{t_s}}$$

costruito con f_2 , ovvero la radice quadrata del secondo momento dello spettro, ed f_4 , la radice quarta del quarto momento dello spettro, è linearmente correlato al rapporto

$$\kappa(n) \equiv \frac{KS_e(\alpha, n)}{KS(\alpha, n)}$$

tra i valori critici ottenuti dalla simulazione e di quelli ottenuti attraverso la 3.1, come si può notare in tabella 3.1. Il fatto che dipenda da soli due momenti è motivato dal fatto che la serie

Processo AR(2)					Processo AR(1)				
n	τ	κ	N	$\rho_{\tau,\kappa}$	n	τ	κ	N	$\rho_{\tau,\kappa}$
16	2.18	0.944	9	0.982	16	2.22	0.906	5	0.991
25	1.62	0.947	14	0.967	25	1.82	0.934	7	0.994
32	2.20	1.075	51	0.988	32	1.89	0.984	9	0.978
64	2.21	1.151	5	0.995	64	1.94	1.070	11	0.992
128	3.32	1.452	21	0.996	128	1.89	1.123	9	0.992
256	3.40	1.522	18	0.998					

Tabella 3.1: Sono riportati i risultati ottenuti per $\alpha = 0.05$, dove n è la dimensione del campione, τ e κ sono i valori medi ottenuti su N prove per diversi valori di a_1 e a_2 , mentre $\rho_{\tau,\kappa}$ è il coefficiente di correlazione lineare tra τ e κ .

dipende dai soli due parametri a_1 e a_2 . Per il caso $a_1 = 0$ $a_2 = 0$ si ottiene un valore di $\tau = 0.69$ il quale cresce all'aumentare di $a = a_1 + a_2$ assumendo un valore $\tau > 1$ se $a > 0.8$. Si è quindi cercata una funzione lineare del parametro τ che riproducesse l'andamento del rapporto $\kappa(n)$ ed empiricamente si è trovato che una funzione del tipo

$$\frac{KS_e(\alpha, n)}{KS(\alpha, n)} = \underbrace{\left(c_1 - \frac{c_2}{n} + \frac{c_3}{n^2} \right)}_{R(n)} (\tau - 1) + 1 \quad (3.2)$$

offre un fit in buon accordo con i dati ottenuti per valori di $\tau \geq 1$ ed $n \geq 16$, mentre per valori di $\tau < 1$ possono essere usati i valori critici non modificati, tenendo conto del fatto che il valore è sovrastimato. Dalla tabella 3.1 si può notare che per valori di τ tendenti ad 1, i valori critici KS_e risultano più piccoli dei valori non modificati KS , pertanto aumenta il rischio di errori di tipo I (falso positivo). Ovvero rifiutare l'ipotesi nulla se questa è vera. In ultima, fittando i dati ricavati dalla simulazione con la 3.2 si sono ottenuti i seguenti valori per c_1, c_2, c_3 :

	c_1	c_2	c_3
Processo AR(2)	0.45	14	120
Processo AR(1)	0.3	4.3	19

POTENZA DEL TEST

È stata infine valutata la potenza del test, ovvero la probabilità di rifiutare l'ipotesi nulla se questa è falsa. Per fare ciò sono state generate nuove serie AR(1) ed AR(2), nelle quali il rumore bianco ϵ_t non fosse più dato solo da una distribuzione normale. Vengono utilizzate una distribuzione uniforme \mathcal{U} , due distribuzioni t di Student t_1 e t_2 (rispettivamente con 1 e 2 gradi di libertà), due distribuzioni chi-quadro χ_1^2 e χ_2^2 (rispettivamente con 1 e 2 gradi di libertà) ed una distribuzione normale \mathcal{N} .

Per ogni rumore bianco ϵ_t sono state generate serie di lunghezza $n = 192$ per diversi valori dei parametri a_1 e a_2 . Da queste serie vengono poi estrapolate delle sottoserie di lunghezza $n < 192$ selezionando solo i termini $i * j$ -mi con $j = 4, 6, 12$ ed $i \leq 192/j$.

Per ognuna di queste viene quindi calcolato il parametro τ e, attraverso la 3.2, il valore critico KS_e per un livello di confidenza $\alpha = 0.05$. Ripetendo questo procedimento si esegue quindi una statistica di Kolmogorov-Smirnov, definita in 1.2, dove la distribuzione $\mathcal{F}(x)$ è data da una gaussiana con media e varianza stimati dal campione. Si annota quindi quante volte il valore ottenuto è superiore al valore critico KS_e .

a_1	a_2	n	τ	ϵ_t					
				u	t_1	t_2	χ_1^2	χ_2^2	N
1.81	-0.845	192	4.02	0.05	0.87	0.34	0.19	0.13	0.04
		48	1.49	0.06	0.84	0.27	0.16	0.11	0.04
		32	1.09	0.05	0.80	0.27	0.14	0.14	0.06
		16	0.65	0.05	0.69	0.19	0.10	0.08	0.04
1.65	-0.690	192	3.94	0.03	0.90	0.38	0.18	0.12	0.05
		48	1.62	0.03	0.86	0.30	0.17	0.11	0.03
		32	1.24	0.05	0.83	0.28	0.15	0.09	0.04
		16	0.82	0.04	0.71	0.20	0.10	0.07	0.05
1.60	-0.800	192	1.99	0.04	0.97	0.47	0.48	0.27	0.04
		48	0.58	0.04	0.90	0.37	0.41	0.16	0.04
		32	0.43	0.04	0.82	0.27	0.19	0.12	0.04
		16	0.80	0.04	0.70	0.21	0.11	0.09	0.05
1.30	-0.500	192	2.09	0.02	1.00	0.68	0.83	0.52	0.04
		48	0.82	0.04	0.99	0.56	0.48	0.34	0.04
		32	0.67	0.04	0.95	0.42	0.44	0.25	0.04
		16	0.68	0.04	0.79	0.26	0.21	0.13	0.05

Tabella 3.2: Percentuale dei campioni con una statistica di Kolmogorov-Smirnov maggiore del valore critico KS_e per un livello di confidenza $\alpha = 0.05$

Si noti che in generale la distribuzione risultante dal processo $AR(p)$ è gaussiana solo se il rumore di fondo ϵ_t è gaussiano, altrimenti questa sarà una convoluzione di PDF gaussiane, pesate dai parametri a_1 e a_2 . I risultati del test sono esposti in tabella 3.2.

Dalla tabella risulta evidente che il metodo permette di aumentare l'efficacia del test per serie $AR(2)$ per tutti i rumori di fondo testati. Fa eccezione il rumore con distribuzione uniforme. Questo riflette il fatto che, anche per serie non correlate, il test di Kolmogorov-Smirnov per una distribuzione normale non risulta molto efficace se il rumore di fondo è uniforme.

È da sottolineare inoltre che il procedimento di utilizzare serie di minor durata, estrapolate da un'unica serie di durata maggiore, non è corretto, poichè la serie derivata non rappresenta più un processo $AR(2)$. Tuttavia, dai risultati ottenuti, si nota che anche in questi casi il metodo permette di aumentare comunque l'efficacia del test.

Il procedimento è stato poi ripetuto per una serie $AR(1)$ ($a_2 = 0$) producendo qualitativamente gli stessi risultati.

DISCUSSIONE DEI RISULTATI

Dai risultati del test di potenza possiamo concludere che, per serie modellizzabili attraverso processi $AR(2)$ ed $AR(1)$ con spettri aventi picchi a bassa frequenza, l'efficacia del test di Kolmogorov-Smirnov è aumentata utilizzando i valori critici dati dalla 3.2. In tutto ciò non si è però considerato che la correzione dipende da τ , il quale non è un parametro di scala o posizione. Ciò significa che i risultati ottenuti potrebbero dipendere dal modo in cui è stato ricavato, tuttavia non sono stati indagati gli effetti derivanti da una diversa stima di τ .

MODIFICA DELLA STATISTICA K-S NEL CASO DI PROCESSI ARCH(1)

Capita spesso, specialmente in ambito finanziario, di trovarsi di fronte a serie temporali che presentano comportamenti non più modellizzabili attraverso processi AR. Come è già stato detto, la distribuzione non condizionata di un processo AR, dove il rumore bianco è dato da una distribuzione normale, è a sua volta normale. Ma se andiamo a studiare il grafico delle differenze del prezzo di chiusura dell'indice FTSE MIB dal 1988 ad oggi, notiamo che non presenta affatto un andamento gaussiano. È evidente dalla Figura 3.2 che le code della distribuzione sono molto più marcate rispetto ad un andamento gaussiano. Tale comportamento è ben modellizzato da processi ARCH.

Distribuzione Differenze prezzo FTSE MIB

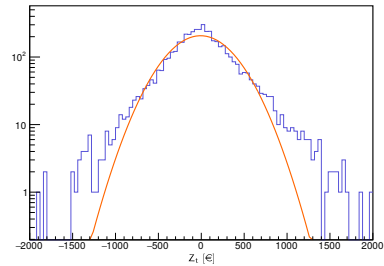


Figura 3.2: Distribuzione delle differenze di prezzo dell'indice FTSE MIB. In arancione è graficata la funzione gaussiana fittata sui dati

Traendo spunto dal procedimento descritto in precedenza, per serie derivanti da processi AR(2), si è tentato di modificare i valori critici KS del test di Kolmogorov-Smirnov per adattarli al caso di serie derivanti da processi ARCH(1) a varianza non condizionata unitaria.

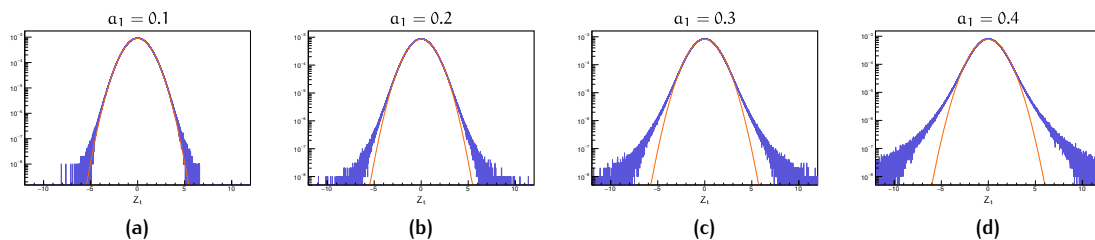


Figura 3.3: Da sinistra a destra vengono mostrate le distribuzioni della variabile Z_t per processi ARCH(1) con diversi valori del parametro α_1 . Ogni distribuzione è stata fittata con una funzione gaussiana graficata in arancione.

IL METODO

Riportiamo le formule per il calcolo della varianza non condizionata e della curtosi per un processo ARCH(1), già espone nella sezione 2.3:

$$\text{var}_u = \frac{c}{1 - \alpha_1}$$

$$\kappa = 3 \frac{1 - \alpha_1^2}{1 - 3\alpha_1^2}$$

Quest'ultima, per essere finita e positiva, impone:

$$0 < \alpha_1 < \frac{1}{\sqrt{3}}.$$

Possiamo quindi fare in modo che la serie abbia una varianza non condizionata unitaria ponendo $c = 1 - \alpha_1$.

Sono state eseguite due simulazioni indipendenti. In ciascuna di esse sono stati generati 4096 campioni di lunghezza n compresa tra 16 e 128 per i seguenti valori del parametro α_1 :

$$\alpha_1 = \{0.05; 0.1; 0.15; 0.2; 0.25; 0.3; 0.35; 0.4; 0.45; 0.5; 0.55\}$$

con $c = 1 - a_1$ ed $\epsilon_t = \mathcal{N}(0, 1)$.

Come detto in precedenza, non esiste una funzione analitica per la densità di probabilità non condizionata di un processo ARCH. Pertanto, per ogni campione generato, è stata calcolata la curtosi del campione, definita come:

$$\kappa = \frac{\mu_4}{\sigma^4}$$

con

$$\mu_4 = \mathbb{E}((x - \mu)^4)$$

e la σ definita con denominatore $n - 1$. Dalla curtosi si è poi ricavato il parametro a_1^{*1} e da questo si è generata una PDF empirica attraverso una processo ARCH(1). Va spesa una nota sulla creazione della PDF. Si è verificato infatti che all'aumentare di a_1^* , per ottenere un valore della curtosi stimato dalla distribuzione consistente con il valore atteso, la dimensione N della simulazione per generare la PDF segue approssimativamente la legge

$$N = 1000000(10 + 900(a_1^*)^3).$$

Successivamente si è potuta eseguire una statistica di Kolmogorov per ogni campione. Sono stati quindi ricavati i valori critici per $0.01 \leq \alpha \leq 0.2$ che sono stati poi mediati tra le due simulazioni.

I RISULTATI

Innanzitutto si è notato che il rapporto $\kappa(n) = \frac{KS_e(\alpha, n)}{KS(\alpha, n)}$ dipende fortemente da n e a_1 , mentre è quasi costante al variare di α . Come valori di $KS(\alpha, n)$ sono stati utilizzati i valori ottenuti da una simulazione effettuata con 10^7 campioni. Anche in questo caso non si è notata alcuna relazione semplice tra il parametro a_1 ed i valori critici KS_e . Si è quindi cercato di analizzare lo spettro della serie, in modo analogo a quanto fatto nello studio di serie derivanti da processi AR(2), per cercare un parametro correlato a $\kappa(n)$. Tuttavia, dallo studio empirico dello spettro, non emergono particolari picchi al variare del parametro a_1 . Infatti, già dal grafico in Figura 2.4b, si nota che l'autocorrelazione per un processo ARCH(1) diventa rumore bianco a partire da $t = 1$. Ciò trova conferma nello spettro empirico in Figura 3.4.

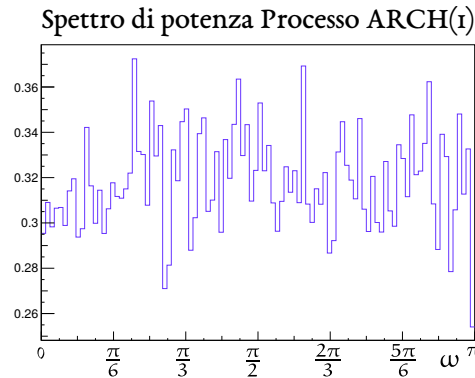


Figura 3.4: Spettro normalizzato di una serie temporale generata da un processo ARCH(1) con $a_1 = 0.4$

Si sono cercati allora nuovi candidati per la scelta di tale parametro e si è trovato che a_1^2 e a_1^3 sono fortemente correlati al valore di $\kappa(n)$, come si nota dalla Tabella 3.3. Pertanto si è cercata una funzione di tali parametri dipendente da n che potesse fittare bene l'andamento di $\kappa(n)$ al variare di n .

Si è trovato che la funzione

$$\kappa(n) = \frac{KS_e(\alpha, n)}{KS(\alpha, n)} = 1 + 1.25a_1^2 - 0.99^n a_1^3 \quad (3.3)$$

¹ Il parametro è identificato con un asterisco per indicare che è stato stimato dal campione e distinguerlo dal a_1 che è il parametro dal quale il campione è stato generato

riproduce correttamente i valori ottenuti dalla simulazione per $0.3 < \alpha_1 \leq 0.55$ entro un errore del 2%.

Valori correlazione lineare		
n	$\rho_{\alpha_1^2, \kappa}$	$\rho_{\alpha_1^3, \kappa}$
16	0.90511	0.89379
32	0.90769	0.90041
64	0.90580	0.90346
96	0.90429	0.90551
128	0.90461	0.90556

Tabella 3.3: Sono riportati i risultati ottenuti dove n è la dimensione del campione, mentre $\rho_{\alpha_1^2, \kappa}$ e $\rho_{\alpha_1^3, \kappa}$ sono i coefficienti di correlazione lineare tra α_1^2 e κ e tra α_1^3 e κ .

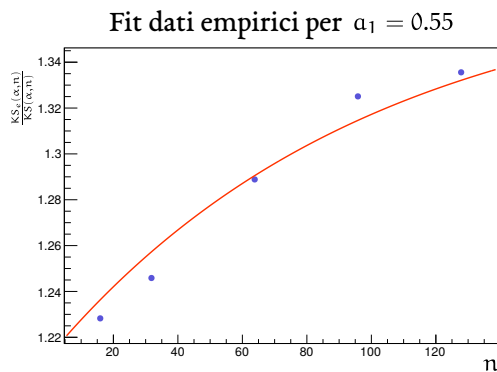


Figura 3.5: Sono mostrati i valori di $\kappa(n)$ (pallini blu) e la funzione 3.3 che li approssima (in rosso)

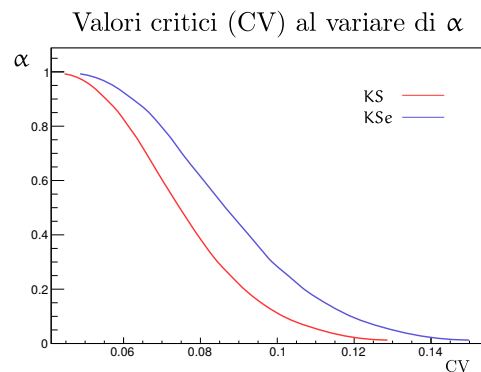


Figura 3.6: In questo grafico viene mostrato il legame tra i valori critici (CV) ed α . In rosso sono riportati i valori critici nel caso di variabili non correlate per $n = 64$, mentre in blu sono riportati i valori critici ottenuti dalla simulazione per $n = 64$ e $\alpha_1 = 0.4$.

Per le regioni $\alpha_1 \leq 0.3$ la 3.3 non risulta più applicabile. Tuttavia, per bassi valori del parametro α_1 , è emerso che i valori di $\kappa(n)$ dipendono sempre meno fortemente da n. In particolare, per $0.05 \leq \alpha_1 \leq 0.15$, risulta che i valori di $\kappa(n)$ sono approssimabili a costanti. Riportiamo in Tabella 3.4 i valori di κ ottenuti dalla simulazione per i tre valori di α_1 indagati, validi per $16 \leq n \leq 128$ e $0.01 \leq \alpha \leq 0.2$ entro un errore del 2%.

α_1	κ
0.05	1.06
0.05	1.07
0.05	1.08

Tabella 3.4

POTENZA DEL TEST

Per valutare l'efficacia di questa correzione è stato seguito lo stesso procedimento utilizzato per serie derivanti da processi AR(2). Sono state scelte tre distribuzioni per il rumore bianco ϵ_t : una distribuzione gaussiana $\mathcal{N}(0, 1)$, una distribuzione di Poisson $\mathcal{P}(10)$ di parametro $\lambda = 10$

(alla quale è stato sottratto il suo valore medio ed è stata divisa per la deviazione standard) e una distribuzione uniforme $\mathcal{U}[-\frac{\sqrt{12}}{2}, \frac{\sqrt{12}}{2}]$. Sono stati generati 4096 campioni di serie di lunghezza $n = 96$ per ogni distribuzione di ϵ_t e per ogni campione è stata eseguita una statistica di K-S. Si è quindi calcolato il valore critico per $\alpha = 0.05$ e si è verificato quante volte, in percentuale, il valore della statistica di K-S eccedeva il valore critico $KS_e(0.05, 96)$. I risultati sono riportati in Tabella 3.5. Come si può notare, il test produce risultati corretti per $\epsilon_t = \mathcal{N}(0, 1)$, mentre mostra che la potenza del test è aumentata se ϵ_t proviene da una distribuzione uniforme o di Poisson.

n	α_1	ϵ_t		
		\mathcal{U}	\mathcal{P}	\mathcal{N}
96	0.35	0.27	0.43	0.06
	0.45	0.15	0.32	0.05
	0.55	0.10	0.23	0.05

Tabella 3.5: Percentuale dei campioni con una statistica di Kolmogorov-Smirnov maggiore del valore critico KS_e per un livello di confidenza $\alpha = 0.05$

DISCUSSIONE DEI RISULTATI

Dai risultati del test di potenza possiamo concludere che, per serie modellizzabili attraverso processi ARCH(1) a varianza unitaria, l'efficacia del test di Kolmogorov-Smirnov è aumentata utilizzando i valori critici dati dalla 3.3 con $0.01 \leq \alpha \leq 0.2$, $0.3 < \alpha_1 \leq 0.55$ ed $16 \leq n \leq 128$. Per bassi valori di α_1 è possibile utilizzare i valori dati dalla Tabella 3.4. Non sono stati indagate le regioni di con $n > 128$, tuttavia l'andamento asintotico della 3.3 suggerisce che il rapporto $\kappa(n)$ tenda ad un valore finito all'aumentare di n . Questo può essere oggetto di studi futuri.

4 | CONCLUSIONI

Dopo aver compreso la teoria ed il funzionamento del test di Kolmogorov-Smirnov, si è studiato un metodo per modificare i valori critici del test per serie derivanti da processi AR(2). A partire da queste idee si è tentato, con successo, di modificare i valori critici del test per serie derivanti da processi ARCH(1) a varianza unitaria. Viene lasciata aperta la possibilità di indagare il comportamento di tali valori critici sia per $n > 128$, sia per serie provenienti da processi ARCH(1) a varianza non unitaria. Inoltre è possibile, cambiando il metodo di stima del parametro α_1 , indagare il comportamento di $\kappa(n)$ anche per valori di $\alpha_1 > \frac{1}{\sqrt{3}}$. Anche questo è lasciato ad eventuali studi futuri.

BIBLIOGRAFIA

- [1] H. W. Lilliefors, "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown", *Journal of the American Statistical Association* Vol. 62, (1967), pp. 399-402
- [2] M. S. Weiss, "Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data" *Journal of the American Statistical Association* Vol. 73,(1978), pp. 872-875
- [3] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione", *Giornale dell'Istituto Italiano degli Attuari* Vol. 4, (1933), pp. 83-91
- [4] L. H. Miller, "Table of Percentage Points of Kolmogorov Statistics", *Journal of the American Statistical Association* Vol. 51, (1956), pp. 111-121
- [5] F. N. David and N. L. Johnson, "The Probability Integral Transformation When Parameters are Estimated from the Sample", *Biometrika* Vol. 35,(1948), pp. 182-190
- [6] G. C. Blain, "Revisiting the critical values of the Lilliefors test: towards the correct agro-meteorological use of the Kolmogorov-Smirnov framework", *Bragantia* Vol. 73, (2014), pp. 192-202
- [7] P. J. Brockwell and R. A. Davis, "Introduction to Time Series and Forecasting", Springer Second Edition, (2002)
- [8] M. H. DeGroot and M. J. Schervish, "Probability and Statistics", Addison-Wesley Fourth Edition, (2012)
- [9] R. S. Tsai, "Analysis of Financial Time Series", Wiley-Interscience Second Edition, (2005)