Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Una misura di auto-dipendenza per serie storiche non lineari

Relatrice: Prof.ssa Luisa Bisaglia
Dipartimento di Scienze Statistiche

Laureando: Marco Ballarini
Matricola n. 1218942

Anno Accademico 2021/2022

Indice

Introduzione

L'utilizzo di appositi modelli per serie storiche che presentano caratteristiche di non linearità è diffuso in moltissimi campi applicativi, dalla finanza alle scienze sociale. Nonostante ciò, la trattazione di serie non lineari è molto complessa e molte tecniche, che risultano adatte alle serie lineari, possono portare a conclusioni errate se applicate al caso non lineare.
Un aspetto cruciale nell'analisi delle serie storiche è rappresentato dalla ricerca di strutture di dipendenza all'interno della serie stessa per capirne il comportamento e costruire un modello coerente anche a fini previsivi. L'indicatore più utilizzato in letteratura da questo punto di vista è il ρ di Pearson, la cui estensione al caso di serie storiche è rappresentata dalla funzione di autocorrelazione (ACF), introdotta da Box e Jenkins nel 1970. Questa funzione ha l'obiettivo di misurare la correlazione, ovvero la forza della dipendenza lineare tra le osservazioni della serie distanti k periodi temporali con $k > 0$. Nel caso di serie storiche non lineari l'ACF può risultare non significativa o nulla anche se tra k osservazioni è presente una struttura di dipendenza, qualora questa non sia lineare. In questo elaborato si studierà una misura alternativa all'ACF che ha come obiettivo la misura del grado di dipendenza, anche nel caso di serie storiche non lineari.
Nel capitolo 1 si farà una panoramica generale sull'utilizzo dell'indice ρ di Pearson, sottolineando le sue criticità e alcu-

ni metodi per evitare queste problematiche. Nel capitolo 2 si introdurrà la teoria alla base della correlazione basata sulla distanza *distance correlation* e nel capitolo 3 la si estenderà al caso di serie storiche, introducendo le funzioni di autocorrelazione basate sulla distanza (*auto distance correlation function.* Infine nel capitolo 4 si svolgeranno alcune simulazioni, utilizzando il software statistico R Core Team (2022), per evidenziare alcuni aspetti dell'utilizzo dell'*auto distance correlation function* su serie storiche non lineari.

Capitolo 1

Misurare la dipendenza

In questo capitolo si richiama l'indice ρ di Pearson, considerato come lo standard per la misura della correlazione tra due variabili aleatorie, e si analizzeranno alcuni suoi punti critici che possono minare, in alcuni casi, le analisi svolte. Si andrà poi a fare una panoramica generale su alcuni modi per misurare la dipendenza tra variabili che non risentono delle criticità del ρ di Pearson.

1.1 Indice ρ di Pearson

1.1.1 Definizione

La misurazione della dipendenza tra fenomeni risulta centrale in tutta la statistica, in particolare la correlazione misura la relazione lineare che intercorre tra due variabili. L'indice più diffuso per misurare tale tipo di dipendenza è il ρ di Pearson, introdotto da Francis Galton nella seconda meta del diciannovesimo secolo, fu sviluppato in maniera più rigorosa dal punto di vista matematico da Pearson (1895).
Date due variabili aleatorie X e Y, con il momento secondo finito, l'indice ρ di Pearson viene così definito:

$$\rho = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y} \tag{1.1}$$

dove il numeratore rappresenta la covarianza fra le due variabili x e y e il denominatore il prodotto tra le deviazioni standard delle singole variabili.
Il corrispettivo campionario, dato un insieme di osservazioni $(\mathbf{x},\mathbf{y}) = (x_1, y_1)$, ..., (x_n, y_n), risulta pari a:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1.2}$$

dove $\bar{x}$ e $\bar{y}$ sono le medie campionarie dei due vettori di osservazioni x e y.
Dalla disuguaglianza di Cauchy-Schwarz si può dimostrare che sia ρ che r variano nell'intervallo $[-1, 1]$, dove -1 indica la perfetta correlazione negativa, 1 indica la perfetta correlazione positiva e 0 indica assenza di correlazione.
La popolarità dell'indice ρ di Pearson si deve anche alla sua estrema versatilità, che ne permette l'uso in moltissimi campi applicativi, dalle scienze naturali fino alla finanza. In particolare, per studiare la struttura di dipendenza lineare presente all'interno di serie storiche, risulta importate adattare il coefficiente ρ di Pearson a questo caso. Dato un processo stazionario $\{Y_t\}_{t \in T}$ con media μ, varianza σ^2 e funzione di autocovarianza pari a $\gamma_k = Cov(Y_t, Y_{t+k})$, la funzione di autocorrelazione risulta pari a:

$$\rho_k = \frac{\gamma_k}{\sigma^2} \tag{1.3}$$

con $k = 0, \pm 1, \pm 2, ...$ La funzione di autocorrelazione, inoltre, presenta proprietà analoghe all'indice di Pearson, solamente declinate nel caso delle serie storiche:

1. $\rho_0 = 1$
2. $\rho_k = \rho_{-k}$
3. $\rho_k \in [-1, 1]$

A partire dalla funzione di autocorrelazione è possibile tracciare il correlogramma, un grafico dei valori di ρ_k per $k = 0, \pm 1, \pm 2, ...$ Il correlogramma risulta decisivo nella fase di identificazione e di modellazione di una serie storica e nella verifica della bontà del modello scelto.

1.2 Criticità del ρ di Pearson

Nonostante l'ampia popolarità di cui gode, il ρ di Pearson non è esente da criticità che, in situazioni particolari, possono minare i risultati delle analisi.

1.2.1 Robustezza

La prima criticità può essere riscontrata nella non robustezza dell'indice rispetto alla presenza di valori anomali o di *outliers*.
Una possibile soluzione è stata proposta in letteratura con l'introduzione del coefficiente di correlazione per ranghi da Spearman (1961) o con il coefficiente di correlazione τ da Kendall (1938).

1.2.2 Non-gaussianità

Vi è una forte connessione tra il concetto di correlazione e gaussianità; in particolare considerando una variabile aleatoria che si distribuisce come una normale multivariata $X \sim$

$\mathcal{N}_p(\mu, \Sigma_p)$, si può dimostrare che l'incorrelazione tra due o più componenti comporta l'indipendenza tra queste; dato che la distribuzione normale multivariata appartiene ad una classe più ampia di distribuzioni di probabilità, la classe delle distribuzioni ellittiche, ci si chiede se lo stretto legame tra indipendenza e incorrelazione può essere generalizzato anche per questa classe. La risposta a questa domanda è negativa: la normale rappresenta un *unicum* da questo punto di vista e la relazione tra incorrelazione e indipendenza non può essere generalizzata all'interno della classe delle distribuzioni ellittiche.
Si può definire la distribuzione ellittica in termini di funzione di densità, in particolare quest'ultima ha la forma:

$$f(x) = k \cdot g((x - \mu)'\Sigma(x - \mu)) \tag{1.4}$$

dove k è una costante di normalizzazione, x è un vettore n dimensionale, μ è il vettore di medie, Σ è una matrice definita positiva e g una funzione misurabile secondo Lesbegue, non negativa e definita su $[0, \infty)$. La distribuzione t-Student multivariata può essere fatta risalire alla classe di distribuzioni ellittiche, tuttavia, si può dimostrare, che per questa non vale l'equivalenza tra incorrelazione e indipendenza; in particolare, nel caso di una t-Student bivariata, se due componenti sono incorrelate, non è detto che siano anche indipendenti.
In generale, l'indice ρ di Pearson presenta un problema nella misurazione della dipendenza, non solo nel caso t-Student, ma per un più ampio numero di distribuzioni.
Si può, quindi, dire che l'incorrelazione non implica l'indipendenza, e che è necessario prestare molta attenzione nell'uso del ρ di Pearson per la misura della dipendenza, non solo

nel caso t-Student, ma in generale per ogni distribuzione non gaussiana.

1.2.3 Non linearità

L'indice ρ di Pearson è stato costruito per misurare la forza di relazioni lineari, fra due variabili x e y. In generale è possibile costruire numerosi esempi in cui è presente una forte dipendenza non lineare tra le variabili e questa non viene rintracciata dall'indice ρ di Pearson, dal momento che quest'ultimo misura solamente in grado di correlazione, ovvero di relazione lineare tra due variabili.
Nell'ambito delle serie storiche la modellazione di serie non lineari risulta decisiva e di estrema importanza, anche da un punto di vista delle applicazioni in finanza. Un esempio semplice di modello che presenta una struttura di dipendenza non lineare è:

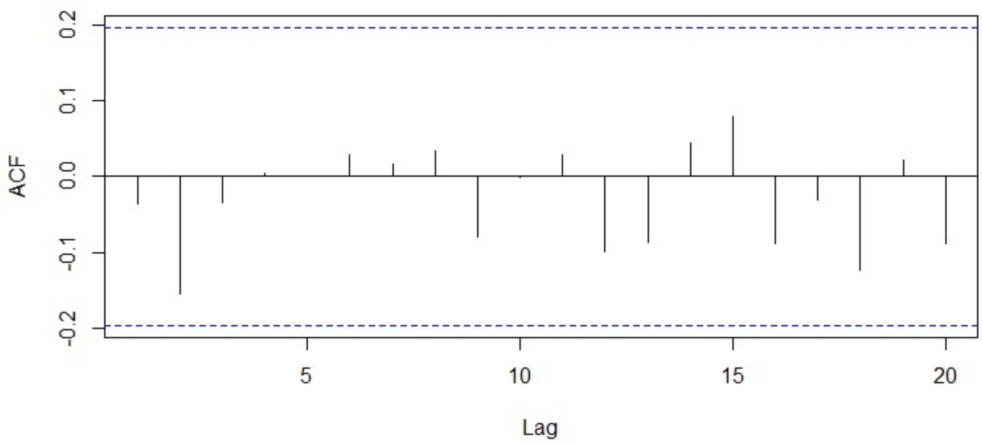
$$x_t = \sigma_t \epsilon_t \tag{1.5}$$

$$\sigma_t^2 = \alpha + \beta x_{t-1}^2$$

dove α e β sono parametri, ϵ_t è un *white noise* con varianza unitaria. Serie di questa tipologia mostrano una correlazione nulla o non significativa, anche se x_t è dipendente da x_{t-1}; dunque i metodi classici per la verifica di dipendenza lineare (come la verifica grafica tramite ACF e pACF) non porteranno ad alcun contributo.
Il modello (1.5) è un modello ARCH (autoregressive conditional heteroskedasticity), in particolare un ARCH(1). E' stata simulata una serie casuale di numerosità n = 100 con $\alpha = 1$ e $\beta = 0.5$, nonostante la forte dipendenza seriale implicata dal

processo generatore dei dati, il coefficiente di Pearson al ritardo 1 risulta pari a $\rho(1) = 0.064$ e il correlogramma empirico [Figura 1.1] non mostra per nessun ritardo una correlazione significativa. Questo sta ad indicare che non è presente una relazione lineare fra le osservazioni, tuttavia può essere presente un altro tipo di dipendenza; quindi, osservando il solo indice ρ di Pearson si potrebbe essere erroneamente portati a dire che le variabili non sono legate da alcuna relazione.
Si possono avanzare numerosi esempi, e non solo appartenenti all'ambito delle serie storiche, nei quali l'utilizzo dell'indice ρ di Pearson può portare a conclusioni ingannevoli: un esempio classico è rappresentato dal caso $Y = X^2$, dove Y risulta univocamente determinato dato X. E' possibile verificare la fallacia dell'indice di Pearson empiricamente generando 1000 osservazioni da normali standard indipendenti X e ϵ, dove $Y = X^2 + \epsilon$. Nonostante vi sia un evidente dipendenza tra X e Y, dato che Y è definito, a meno di un errore normale, da X, ρ risulta molto vicino a 0; la stessa problematica può essere rintracciata anche per l'indice di Spearmann e per il τ di Kendall.

Figura 1.1: Correlogramma di una serie di 100 osservazioni generate da un ARCH(1)



1.3 Oltre l'indice ρ di Pearson

In questa sezione si darà una panoramica sui metodi che permettono di misurare una dipendenza non-lineare e indagare strutture in cui non è presente la gaussianità.
Attraverso questa panoramica studiamo i rapporti di correlazione, bensì direttamente la possibile dipendenza tra variabili, rendendo possibile, quindi, l'individuazione di rapporti di dipendenza non lineari.
Un' ampia trattazione complessiva di come sia possibile aggirare i problemi esposti nella sezione sopra è stata fatta da Tjøstheim *et al.* (2022), in questa sezione verrà esposta una panoramica su due metodologie generali.

1.3.1 Copula

Sia $X = (X_1, ..., X_p)$ un vettore aleatorio tale che $F_i(x) = P(X_i \leq x)$ rappresenti la funzione di ripartizione marginale per l'i-esima componente del vettore, con quest'ultima continua. Secondo il teorema di Sklar (1959) è possibile decomporre la distribuzione congiunta come segue:

$$F(x_1, ..., x_p) = C(F_1(x_1), ..., F_p(x_p)) \tag{1.6}$$

con $C : [0, 1]^p \to [0, 1]$ chiamata copula. Grazie a tale decomposizione è possibile rintracciare nella copula le strutture di dipendenza tra le varie componenti, mentre le singole marginali F_i contengono le informazioni riguardanti X_i.
E' possibile, quindi, analizzare la distribuzione congiunta di una serie di variabili aleatorie in modo relativamente agevole, resta, tuttavia, il problema relativo ad una chiara interpre-

tazione dei parametri rispetto alla possibile dipendenza tra variabili.

1.3.2 Misure di dipendenza alternative

Un approccio alternativo consiste nella definizione di nuove misure di dipendenza e dei relativi test. In letteratura vi sono numerosi esempi di misure e test d'indipendenza che si basano sulle distribuzioni congiunte e marginali: in particolare considerando p variabili aleatorie $(X_1, ..., X_p)$, queste sono indipendenti se e solo se la loro distribuzione congiunta può essere fattorizzata come il prodotto delle marginali: $F_{X_1,..,X_p}(x_1, ..., x_p) = F_{X_1}(x_1)...F_{X_p}(x_p)$. Se le variabili sono continue è possibile esprimere quest'ultima identità in termini funzioni di densità.
La costruzione di una misura di dipendenza tipicamente si basa sulla distanza tra la distribuzione congiunta e il prodotto delle marginali, maggiore sarà questa distanza più ci si allontanerà dall'ipotesi che le variabili siano tra loro indipendenti.
Risulta interessante analizzare le seguenti caratteristiche, proposte da Rényi (1959), che una misura di dipendenza $\delta(X, Y)$, tra due variabili aleatorie X e Y, idealmente dovrebbe possedere:

1. $\delta(X, Y)$ definito per ogni X e Y

2. $\delta(X, Y) = \delta(Y, X)$

3. $\delta(X, Y) \in [0, 1]$

4. $\delta(X, Y) = 0$ se e solo se X e Y sono indipendenti

5. $\delta(X, Y) = 1$ se $X = g(Y)$ e $Y = f(X)$ dove f e g sono funzioni misurabili secondo Borel

6. se, date due funzioni f, g, definite iniettive e misurabili secondo Borel, è possbile scrivere $\delta(f(X), g(Y)) = \delta(X, Y)$

7. se la distribuzione congiunta di X e Y è normale allora $\delta(X, Y) = |\rho(X, Y)|$, dove ρ è l'indice di Pearson

Si può osservare come nel caso di misure di dipendenza con le proprietà 1-7, l'indice non riesce a distinguere tra dipendenza positiva o negativa, come accade per il ρ di Pearson, ma rintraccia solamente la forza della relazione che intercorre tra le due variabili.

Capitolo 2

Misure di auto-dipendenza

Come è riportato nel capitolo precedente, un approccio per la misura del grado di dipendenza tra due variabili è rappresentato dalla costruzione di un indice basato sulla distanza tra la distribuzione congiunta e il prodotto delle marginali. La funzione di ripartizione e la funzione di densità, utilizzate per la costruzione di questa misura, sono tuttavia un altro possibile approccio è rappresentato dalla costruzione di una nuova misura che si basa sulla funzione caratteristica.
Il lavoro di Székely *et al.* (2007) si focalizza sulla presentazione di una misura di dipendenza basata sulla distanza tra funzioni caratteristiche e, in particolare, sulla derivazione di una funzione di peso adatta a definire una misura di correlazione il più aderente possibile alle caratteristiche proposte da Rényi (1959), riportate nel capitolo 1, e con adeguate proprietà sia matematiche che statistiche. In questo capitolo si riportano i risultati teorici principali alla base della *distance correlation* e le sue proprietà.

2.1 Misure di dipendenza basate sulla distanza

Notazione: durante tutta la trattazione teorica si andranno a considerare due variabili aleatorie continue $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$; siano inoltre $f_X(t) = E[e^{it^TX}]$ e $f_Y(s) = E[e^{is^TY}]$ le funzioni caratteristiche relative, rispettivamente, di X e Y, e $f_{X,Y}(t,s) = E[e^{it^TX+is^TY}]$ la funzione caratteristica della variabile congiunta.
Vi sono molti tipi di distanza che possono essere utilizzati per misurare la lontananza tra due funzioni; la distanza più usata in letteratura è certamente la norma euclidea, Székely *et al.* (2007) utilizzano la norma indotta dallo spazio L_2 pesato.

Definizione 2.1.1. *Data una funzione $h : \mathbb{C} \to \mathbb{R}^q \times \mathbb{R}^p$, la norma indotta dallo spazio L_2 pesato è definita da:*

$$\|h(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |h(t,s)|^2 w(t,s)\, dt\, ds \qquad (2.1)$$

dove $w(t,s)$ è una generica funzione peso definita in modo tale che l'integrale esista.

La definizione della norma appena data risulta decisiva per l'esplicitazione della distanza tra due funzioni, essenziale per la formulazione della misura di dipendenza.
Per la costruzione della misura di dipendenza viene utilizzato nel metodo esposto in Székely *et al.* (2007) la funzione caratteristica; l'utilizzo della funzione caratteristica è appropriato viste alcune proprietà che la contraddistinguono. Innanzitutto, la funzione caratteristica esiste sempre per ogni variabile

aleatoria, la funzione caratteristica è inoltre unica, ovvero se due variabili aleatorie hanno la stessa funzione caratteristica, allora hanno anche medesima distribuzione, infine date due variabili X e Y indipendenti $f_{X,Y}(t,s) = f_X(t)f_Y(s)$.
In particolare, per la definizione della misura di dipendenza è decisiva la norma indotta dallo spazio L_2 pesato e la fattorizzazione della funzione caratteristica congiunta di due variabili aleatorie indipendenti; prima di definire la misura di dipendenza, dato che questa verrà costruita con la stessa struttura dell'indice ρ di Pearson, è opportuno definire nuove misure per la varianza e la covarianza di due variabili aleatorie X e Y

Definizione 2.1.2. *Data la norma indotta dallo spazio L_2 pesato, una misura di covarianza tra due variabili aleatorie X e Y è definita da:*

$$\begin{aligned}\mathcal{V}^2(X,Y;w) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)\,dt\,ds\end{aligned} \tag{2.2}$$

dove $\mathcal{V}^2(X,Y;w) = 0$ se e solo se X e Y sono indipendenti.

Analogamente alla misura di covarianza (2.2), per la definizione di una misura di correlazione è necessario definire anche una misura di varianza basata sulla distanza:

Definizione 2.1.3. *Data la norma indotta dallo spazio L_2 pesato, una misura di varianza per la variabile aleatoria X è definita da:*

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{2p}} |f_{X,X}(t,s) - f_X(t)f_X(s)|^2 w(t,s)\, dt\, ds \tag{2.3}$$

La misura di correlazione risultante basata sulla distanza sarà costruita similmente all'indice ρ di Pearson dividendo la $\mathcal{V}^2(X,Y)$ per $\sqrt{\mathcal{V}^2(X;w)\mathcal{V}^2(Y;w)}$, tuttavia prima è necessario trovare un'adeguata funzione di peso $w(\cdot)$.

2.2 Funzione peso, *distance covariance* e *variance*

Dal punto di vista strettamente matematico la scelta della funzione peso da utilizzare nelle formule esposte in precedenza è subordinata solamente all'esistenza dell'integrale che definisce la norma nello spazio L_2 pesato. Dal punto di vista statistico è necessario, invece, che la misura di dipendenza costruita sulle quantità descritte in precedenza aderisca a regole precise: per due variabili X e Y dipendenti la misura dev'essere diversa da zero, inoltre è conveniente che sia invariante rispetto a trasformazioni di scala e, nel caso della varianza, anche di traslazioni.
In Székely *et al.* (2007) viene proposta la seguente funzione di peso:

$$w(t,s) = (c_p c_q |t|^{1+p} |s|^{1+q})^{-1} \tag{2.4}$$

dove

$$c_d = C(d,1) = \frac{\pi^{(1+d)/1}}{\Gamma((1+d)/2)}$$

e $\Gamma(\cdot)$ rappresenta la funzione gamma completa. Questa non è l'unica scelta per una funzione peso, in Székely *et al.* (2007) viene anche proposta una generalizzazione della funzione (2.4) che fa uso di $C(d,\alpha) = \frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((\alpha+d)/2)}$, questa quantità tuttavia porta a risultati più complessi dal punto di vista computazionale, per questo viene preferita la formulazione (2.4) con $C(d,1)$.

Grazie alla funzione peso appena definita è possibile sostituire $w(t,s)$ all'interno delle formule (2.5) e (2.6) per ottenere, rispettivamente, le misure di *distance covariance* e *distance variance*.

Definizione 2.2.1. *Data la norma indotta dallo spazio L_2 pesato e la funzione peso (2.4), la distance covariance tra due variabili aleatorie X e Y è definita da:*

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 \\ &= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|^{1+p}|s|^{1+q}}\,dt\,ds\end{aligned} \tag{2.5}$$

dove $\mathcal{V}^2(X,Y;w) = 0$ se e solo se X e Y sono indipendenti.

In Székely e Rizzo (2012) viene anche mostrato come, data la funzione peso (2.4), la *distance covariance* risulta univocamente determinata dalla norma indotta dallo spazio L_2 pesato.

Definizione 2.2.2. *Data la norma indotta dallo spazio L_2 pesato e la funzione peso (2.4), la distance variance per la* *variabile aleatoria X è definita da:*

$$\mathcal{V}^2(X) = \frac{1}{c_p c_p} \int_{\mathbb{R}^{2p}} \frac{|f_{X,X}(t,s) - f_X(t) f_X(s)|^2}{|t|^{1+p}|s|^{1+p}} \, dt \, ds \qquad (2.6)$$

Analogamente per la variabile Y.

2.3 *Distance correlation*

Una volta definita la *distance covariance* e la *distance variance* per X e Y, la misura di correlazione basata su queste quantità può essere costruita basandosi sulla struttura del ρ di Pearson, ovvero il rapporto tra la covarianza e la radice del prodotto delle varianze.

Definizione 2.3.1. *Date due variabili aleatorie $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$, la distance correlation è definita dalla radice quadrata di*

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}} & se\ \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0 & se\ \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases} \qquad (2.7)$$

In Székely *et al.* (2007) vengono anche definite alcune proprietà che collegano il ρ di Pearson con la *distance correlation* nel caso in cui X e Y siano variabili aleatorie distribuite come normali standard, con correlazione pari a $\rho(X,Y) = \rho$.

1. $\mathcal{R}(X,Y) \leq |\rho|$
2. $\mathcal{R}^2(X,Y) = \frac{\rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin \rho/2 - \sqrt{4-\rho^2} + 1}{1+\pi/3-\sqrt{3}}$
3. $\inf_{\rho \neq 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \lim_{\rho \to 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \frac{1}{2(1+\pi/3-\sqrt{3})^{1/2}}$

2.4 *Distance correlation* campionaria

In questa sezione viene presentato uno stimatore finalizzato al calcolo delle quantità teoriche viste in precedenza.
Una prima formulazione, che ricalca direttamente le formule viste nelle sezioni precedenti, è basata sulla funzione caratteristica empirica; sia $(\mathbf{x}, \mathbf{y}) = \{(x_k, y_k) : k = 1, ..., n\}$ un campione casuale realizzazione della distribuzione congiunta di $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$, si definiscono le seguenti funzioni caratteristiche empiriche:

$$
\begin{aligned}
f^n_{x,y}(t,s) &= \frac{1}{n}\sum_{k=1}^{n} e^{i(t^T x_k + s^T y_k)} \\
f^n_{x}(t) &= \frac{1}{n}\sum_{k=1}^{n} e^{i(t^T x_k)} \\
f^n_{y}(s) &= \frac{1}{n}\sum_{k=1}^{n} e^{i(s^T y_k)}
\end{aligned}
$$

Grazie alla formulazione appena vista è possibile definire la *distance covariance*, *distance variance* e *distance correlation* campionarie.

Definizione 2.4.1. *Data una realizzazione* $(x_1, ..., x_n)$ *da una varaibile aleatoria* $X \in \mathbb{R}^p$*, la distance variance campionaria è definita dalla quantità*

$$\mathcal{V}^2_n(x) = \frac{1}{c_p c_p}\int_{\mathbb{R}^{2p}} \frac{|f^n_{x,x}(t,s) - f^n_x(t) f^n_x(s)|^2}{|t|^{1+p}|s|^{1+p}}\, dt\, ds \qquad (2.8)$$

Definizione 2.4.2. *Data una realizzazione* $\{(x_1, y_1), ..., (x_n, y_n)\}$ *da una coppia di variabili aleatorie* $X \in \mathbb{R}^p$ *e* $Y \in \mathbb{R}^q$*, la distance covariance campionaria viene definita come segue:*

$$\begin{aligned} \mathcal{V}_n^2(x, y) &= \|f_{x,y}^n(t, s) - f_x^n(t) f_y(s)\|_w^2 \\ &= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{x,y}^n(t, s) - f_x^n(t) f_y^n(s)|^2}{|t|^{1+p} |s|^{1+q}} \, dt \, ds \end{aligned} \tag{2.9}$$

É quindi possibile definire la *distance correlation* campionaria sostituendo nella formula (2.7) le quantità (2.9) e (2.8). Questa metodologia di stima delle quantità teoriche è vantaggiosa, poiché intuibile direttamente dalle formule teoriche, tuttavia comporta problemi computazionali non indifferenti, innanzitutto legati al calcolo dell'integrale.
In Székely *et al.* (2007) viene proposto un nuovo stimatore, più facile da calcolare; sia $(\mathbf{x}, \mathbf{y}) = \{(x_k, y_k) : k = 1, ..., n\}$ un campione casuale realizzazione della distribuzione congiunta di $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$, si definiscono le seguenti quantità campionarie:

$$S_1 = \frac{1}{n^2} \sum_{k,l=1}^{n} |x_k - x_l||y_k - y_l| \tag{2.10}$$

$$S_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} |x_k - x_l| \frac{1}{n^2} \sum_{k,l=1}^{n} |x_k - x_l| \tag{2.11}$$

$$S_3 = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l,m=1}^{n} |x_k - x_l||y_k - y_m| \tag{2.12}$$

Si può dimostrare che la formula (2.9), e quindi anche (2.8), può essere riscritta come:

$$\mathcal{V}_n^2(x, y) = \|f_{x,y}^n(t, s) - f_x^n(t)f_y^n(s)\| = S_1 + S_2 - 2S_3 \quad (2.13)$$

La notazione (2.13) può essere a sua volta scomposta algebricamente, introducendo i seguenti termini rispetto ad $a_{kl} = |x_k - x_l|$:

$$\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl} \qquad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{n}\sum_{k=1}^{n} a_{kl} \qquad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

dove $k, l = 1, ..., n$; allo stesso modo possono essere definite le medesime quantità per Y, dove $b_{kl} = |y_k - y_l|$ per ogni k, l, e analogamente per le quantità $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$, $\bar{b}_{\cdot\cdot}$. Con la nuova notazione la *distance covariance* e la *distance variance* campionaria possono essere riscritte come segue:

$$\mathcal{V}_n^2(x) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2$$

$$\mathcal{V}_n^2(x, y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}$$

Grazie alle varie metodologie di stima della *distance covariance*, viste in precedenza, è ora possibile definire lo stimatore per la *distance correlation*

Definizione 2.4.3. *La distance correlation campionaria dato un campione* $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_k, y_k) : k = 1, ..., n\}$ *è definita dalla radice quadrata di*

$$\mathcal{R}_n^2(x,y) = \begin{cases} \frac{\mathcal{V}_n^2(x,y)}{\sqrt{\mathcal{V}_n^2(x)\mathcal{V}_n^2(y)}} & se\ \mathcal{V}_n^2(x)\mathcal{V}_n^2(y) > 0 \\ 0 & se\ \mathcal{V}_n^2(x)\mathcal{V}_n^2(y) = 0 \end{cases} \tag{2.14}$$

2.5 Proprietà della *distance correlation* campionaria e teorica

Come prima cosa è possibile verificare alcune proprietà asintotiche degli stimatori, sia della *distance correlation* che della *distance covariance*.

Se $E[|X|] < \infty$ e $E[|Y|] < \infty$ allora quasi certamente:

$$\lim_{n\to\infty} \mathcal{V}_n^2 = \mathcal{V}^2$$

Se $E[|X| + |Y|] < \infty$ allora quasi certamente:

$$\lim_{n\to\infty} \mathcal{R}_n^2 = \mathcal{R}^2$$

In Székely *et al.* (2007) vengono anche citate e dimostrate altre tre proprietà fondamentali per la *distance correlation*:

1. se $E[|X| + |Y|] < \infty$, allora $0 \leq \mathcal{R}^2 \leq 1$ e $\mathcal{R}^2 = 1$ se e solo se X e Y sono prefettamente dipendenti.
2. $0 \leq \mathcal{R}_n^2 \leq 1$
3. se $\mathcal{R}_n^2 = 1$, allora esiste un vettore a, con $a \neq 0$, un valore reale b e un matrice C ortogonale, tale per cui $Y = a + bXC$

Oltre alle proprietà appena esposte risulta interessante studiare la *distance correlation* rispetto alle caratteristiche che una misura di dipendenza dovrebbe avere, secondo Rényi

(1959) e citate nel primo capitolo. É dimostrabile che solamente le prime quattro proprietà risultano soddisfatte, le ultime vengono, quindi, disattese.
Si può dimostrare, inoltre, che la *distance correlation* è invariante per trasformazioni di scala e pari a zero se e solo se X e Y sono indipendenti.
In Székely e Rizzo (2014) viene inoltre dimostrato che lo stimatore della *distance covariance* campionaria è uno stimatore distorto.

2.6 Test d'indipendenza

Una volta definite le quantità descritte nelle sezioni precedenti, in Székely *et al.* (2007) dimostrano la possibile costruzione di un test d'indipendenza basato sulla *distance covariance*. Il sistema di ipotesi di partenza comprende nell'ipotesi nulla il caso d'indipendenza tra le due variabili e nell'ipotesi alternativa il caso di dipendenza.

$$\begin{cases} H_0 : f_{XY}(t,s) = f_X(t)f_Y(s) \\ H_1 : f_{XY}(t,s) \neq f_X(t)f_Y(s) \end{cases}$$

Il test proposto si basa sulla statistica $n\mathcal{V}_n^2/S_2$ dove n è la numerosità campionaria, $\mathcal{V}^2$ è la *distance covariance* e S_2 è la formula (2.11). Si può dimostrare che, se le variabili X e Y presentano il primo momento finito, sotto ipotesi nulla la statistica test converge in distribuzione ad una forma quadratica:

$$n\mathcal{V}_n^2/S_2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j^2$$

dove Z_j sono variabili aleatorie distribuite come normali standard indipendenti tra loro, λ_j sono costanti non negative che dipendono dalla distribuzione di (X, Y) inoltre, $E[n\mathcal{V}_n^2/S_2] = 1$. L'ipotesi d'indipendenza viene, quindi, rigettata nel caso di valori alti della statistica test, in particolare se:

$$\frac{n\mathcal{V}_n^2(X,Y)}{S_2} > \left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2 \tag{2.15}$$

dove Φ indica la funzione di ripartizione di una normale standard e α il livello di significatività scelto per il test.

La *distance correlation* risulta quindi una misura di dipendenza che riesce ad ovviare ad alcuni dei problemi esposti nel capitolo 1. In particolare, basandosi sulle distanze tra distribuzioni marginali e congiunte di due variabili aleatorie, questa misura riesce a rintracciare strutture di dipendenza, anche non lineari, che indagando mediante la semplice correlazione non emergerebbero. Inoltre, dato che si indaga direttamente la dipendenza tra le variabili aleatorie in analisi, non emerge il problema legato alla non-gaussianità delle variabili di partenza.
Tuttavia, come già sottolineato in precedenza, la *distance correlation* misura solamente il grado di dipendenza, non la direzione; non è quindi possibile capire come le variabili in analisi siano legate tra loro, ma solamente quanto sono interconnesse.

Capitolo 3

Distance correlation e serie storiche

In questo capitolo si applica la teoria alla base della *distance correlation*, vista nel capitolo precedente alle serie storiche con l'obiettivo di rintracciare strutture di dipendenza anche all'interno di serie storiche non lineari.
Da questo punto di vista risulta decisiva l'analisi del lavoro svolto da Pitsillou e Fokianos (2016) e Zhou (2012).

3.1 ACDF: funzione di autocorrelazione basata sulla distanza

Sia $\{X_t \ , t \in \mathbb{Z}\}$ un processo stocastico strettamente stazionario, è possibile definire la *distance covariance* in termini di funzioni caratteristiche. Date, infatti, due generiche variabili del processo (X_t, X_{t+j}), la funzione caratteristica congiunta risulta pari a $f_{X_t,X_{t+j}}(u,v) = E[e^{iu^T X_t + iv^T X_{t+j}}]$.
É possibile generalizzare tutte le quantità viste nel capitolo precedente al caso delle serie storiche. In particolare si può definire l'*auto distance covariance* nel seguente modo:

Definizione 3.1.1. *Sia* $\{X_t, t \in \mathbb{Z}\}$ *un processo stocastico strettamente stazionario, univariato, per ogni ritardo* $j \geq 0$*, l'auto distance covariance è definita da:*

$$\begin{aligned}\mathcal{V}^2(X_t, X_{t+j}) &= \|f_{X_t,X_{t+j}}(u,v) - f_{X_t}(u)f_{X_{t+j}}(v)\|_w^2 \\ &= \frac{1}{\pi^2}\int_{\mathbb{R}^2} \frac{|f_{X_t,X_{t+j}}(u,v) - f_{X_t}(u)f_{X_{t+j}}(v)|^2}{|t|^2|s|^2}\,du\,dv\end{aligned} \tag{3.1}$$

La formula 3.1 risulta quindi una derivazione diretta della (2.5), con la notazione modificata per renderla coerente con la trattazione di serie storiche e la funzione di peso (2.4) adattata al contesto unidimensionale. Come già trattato nel capitolo precedente, lo stimatore più immediato per la *distance covariance* è quello basato sulla funzione caratteristica empirica, che nel caso di trattazione di serie storiche, viene scritta, nella sua forma marginale e congiunta, come segue:

$$f^n_{x_t,x_{t+j}}(u,v) = \frac{1}{n-j}\sum_{t=1}^{n-j} e^{i(u^T x_t + v^T x_{t+j})}$$

$$f^n_{x_t}(u) = \frac{1}{n-j}\sum_{t=1}^{n-j} e^{i(u^T x_t)}$$

$$f^n_{x_{t+j}}(v) = \frac{1}{n-j}\sum_{t=1}^{n-j} e^{i(v^T x_{t+j})}$$

la definizione della *distance covariance* campionaria deriva direttamente dalla formula (3.1):

Definizione 3.1.2. *Sia* $\{x_1, x_2, ..., x_n\}$ *una serie storica strettamente stazionaria univariata per ogni ritardo* $j \geq 0$*, l'auto* *distance covariance è definita da*

$$\begin{aligned}\mathcal{V}_n^2(x_t, x_{t+j}) &= \|f^n_{x_t,x_{t+j}}(u,v) - f^n_{X_t}(u) f^n_{x_{t+j}}(v)\|^2_w \\ &= \frac{1}{\pi^2}\int_{\mathbb{R}^2} \frac{|f^n_{x_t,x_{t+j}}(u,v) - f^n_{x_t}(u) f^n_{x_{t+j}}(v)|^2}{|t|^2|s|^2}\, du\, dv\end{aligned} \tag{3.2}$$

Tenendo presente il fatto che $\mathcal{V}_n^2(x_t, x_t)$ rappresenta la *distance variance* applicata al caso delle serie storiche, è possibile definire la *auto distance correlation function* e la *auto distance correlation function* campionaria, che sono rispettivamente esplicitabili come segue

Definizione 3.1.3. *Sia* $\{X_t\ , t \in \mathbb{Z}\}$ *un processo stocastico stazionario in senso stretto e univariato, per ogni ritardo* $j \geq 0$*, l'auto distance correlation function è definita dalla radice quadrata di*

$$\mathcal{R}^2(X_t, X_{t+j}) = \begin{cases} \frac{\mathcal{V}^2(X_t, X_{t+j})}{\mathcal{V}^2(X_t)} & se\ \mathcal{V}^2(X_t)\mathcal{V}^2(X_{t+j}) \neq 0 \\ 0 & se\ \mathcal{V}^2(X_t)\mathcal{V}^2(X_{t+j}) = 0 \end{cases} \tag{3.3}$$

l'auto distance correlation campionaria, data una realizzazione $\{x_1, x_2, ..., x_n\}$ *del processo stocastico, è definibile dalla radice quadrata di*

$$\mathcal{R}_n^2(x_t, x_{t+j}) = \begin{cases} \frac{\mathcal{V}_n^2(x_t, x_{t+j})}{\mathcal{V}_n^2(x_t)} & se\ \mathcal{V}_n^2(x_t)\mathcal{V}_n^2(x_{t+j}) \neq 0 \\ 0 & se\ \mathcal{V}_n^2(x_t)\mathcal{V}_n^2(x_{t+j}) = 0 \end{cases} \tag{3.4}$$

Come già scritto nel capitolo precedente, l'utilizzo della funzione caratteristica empirica risulta più immediato, ma poco conveniente dal punto di vista computazionale; per ovviare a questo problema si può utilizzare la stessa metodologia usata nel capitolo precedente. Definite le quantità $a_{rl} = |x_r - x_l|$ e $b_{rl} = |y_r - y_l|$ dove Y rappresenta un generico X_{t+j} con ritardo j:

$$\bar{a}_{k\cdot} = \frac{1}{n-j}\sum_{l=1}^{n-j} a_{kl} \qquad \bar{a}_{\cdot l} = \frac{1}{n-j}\sum_{k=1}^{n-j} a_{kl}$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{n-j}\sum_{k,l=1}^{n-j} a_{kl} \qquad A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

e allo stesso modo per b. Ricalcando la formula vista al capitolo precedente è possibile definire la *auto distance covariance* come:

$$\mathcal{V}_n^2(x_t, x_{t+j}) = \frac{1}{(n-j)^2}\sum_{k,l=1}^{n-j} A_{kl}B_{kl} \tag{3.5}$$

É quindi possibile sostituire la formulazione sopra descritta all'interno dell'espressione della *auto distance correlation*, per avere una forma più congeniale dal punto di vista computazionale.

3.2 Proprietà della ACDF

In generale le proprietà della *auto distance correlation*, sia nel caso empirico che nel caso teorico, rimangono le medesime esposte nel capitolo 2. É importante ricordarne due: $\mathcal{R}^2(X_t, X_{t+j}) = 0$ se e solo se X_t e X_{t+j} sono indipendenti e $\mathcal{R}^2(X_t, X_{t+j}) = 1$ se e solo se X_t e X_{t+j} sono perfettamente dipendenti.

Lo stimatore risulta inoltre invariante rispetto trasformazioni di scala e pari a zero se e solo se X_t e X_{t+j} sono indipendenti.
Come già scritto la *distance covariance* campionaria è uno stimatore distorto della *distance covariance*, anche nel caso di serie storiche è possibile formulare una sua versione non distorta.
E' inoltre possibile, come nel capitolo 2, definire la distribuzione della *distance covariance* per un ritardo fissato e da quella derivare una statistica test; emerge tuttavia il problema relativo a trovare una distribuzione della *distance covariance* nel caso delle serie storiche, ovvero considerando il progredire dei ritardi.

3.3 Test univariato per la dipendenza seriale

É possibile sfruttare il lavoro di Hong (1999) per verificare l'ipotesi di assenza di strutture di dipendenza fra le osservazioni di una serie storica. Sia $\{X_t, t \in \mathbb{Z}\}$ un processo stazionario in senso stretto, la trasformata di Fourier esiste ed è pari a

$$f(\omega, u, v) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j(u, v) e^{ij\omega} \tag{3.6}$$

dove $\sigma_j(u, v) = f_{X_t, X_{t+j}}(u, v) - f_{X_t}(u) f_{X_{t+j}}(v)$, come già definite, e $\omega \in [-\pi, \pi]$. Nel caso in cui $\sigma_j(u, v) = 0$ si può dire che X_t è indipendente da ogni valore X_{t+j} per ogni ritardo j; in questo caso è possibile scrivere la trasformata di Fourier come:

$$f_0(\omega, u, v) = \frac{1}{2\pi}\sigma_0(u, v) \tag{3.7}$$

Si può verificare un'eventuale presenza di struttura di dipendenza all'interno della serie storica tramite la valutazione dello scostamento di $f(\omega, u, v)$ da $f_0(\omega, u, v)$, ovvero il distanziarsi della trasformata di Fourier dal caso di indipendenza temporale.
La stima delle quantità sopra esposte avviene tramite la stima kernel della densità, in particolare, in Fokianos e Pitsillou (2017) è stata proposta la seguente statistica test:

$$T_n = \sum_{j=1}^{n-1}(n-j)^2 k^2(j/p)\mathcal{V}_n^2(X_t, X_{t+j}) \tag{3.8}$$

dove $k(\cdot)$ è una funzione kernel lipschitziana e p è una misura che definisce la larghezza di banda come $p = cn^{\lambda}$ per $\lambda \in (0, 1)$, $c > 0$ e n è la numerosità del campione; la scelta dell'ampiezza di banda è cruciale e influisce sul valore della statistica test. In generale si utilizza la *cross-validation* per la scelta del parametro P, tuttavia vi possono essere anche situazioni particolari in cui risulta maggiormente utile far variare p appositamente esaminando i risultati del test in analisi.
Sotto l'ipotesi nulla che i dati siano indipendenti, la versione di T_n standardizzata segue una distribuzione normale standard ed è asintoticamente consistente. Si può derivare allo stesso modo una versione della statistica test con $\mathcal{R}_n^2(X_t, X_{t+j})$ al posto di $\mathcal{V}_n^2(X_t, Y_{t+j})$.
Dunque grazie a queste quantità è possibile verificare la possibile esistenza di dipendenza temporale all'interno di una serie storica, l'ipotesi nulla comprende il caso in cui i dati

presenti in una serie sono indipendenti, l'alternativa rappresenta il caso in cui vi è dipendenza seriale all'interno dei dati.

É quindi possibile utilizzare la teoria alla base della *distance correlation* per determinare la presenza o meno di una struttura di dipendenza temporale all'interno della serie storica in esame; in particolare l'*auto distance correlation function*, a differenza dell'ACF che misura il grado di correlazione esistente all'interno della serie storica, quantifica la dipendenza presente all'interno della serie storica, variando tra 0 ed 1. Come già scritto nel capitolo precedente, dato che l'*auto distance correlation function* varia solamente tra 0 e 1, questa quantifica solamente la forza della dipendenza e non la sua direzione.
La possibilità di utilizzare l'*auto distance correlation function* permette di rintracciare strutture di dipendenza non lineari in serie storiche che, se analizzate con la sola ACF, risulterebbero, in molti casi, non significative e difficilmente riconducibili a modelli presenti in letteratura.

Capitolo 4

Esperimenti Monte Carlo

In questo capitolo verranno proposte alcune simulazioni svolte con R Core Team (2022), al fine di spiegare e comprendere al meglio l'utilizzo e le potenzialità della *distance correlation* applicata alle serie storiche. In particolare, ci si concentrerà sull'utilizzo della *distance correlation* per l'analisi delle strutture di dipendenza in serie storiche non lineari.
Durante l'analisi per l'applicazione su R degli strumenti analitici visti nel capitolo 3 si farà riferimento alla libreria R "dCovTS", sviluppato da Tsagris et al (2022); la trattazione dei modelli TAR viene svolta con il supporto del pacchetto "TSA" sviluppato da Chan et al (2022).

4.1 Modelli TAR

La classe dei modelli TAR (Threshold Autoregressive Model), o modelli a soglia, generalizza il concetto di modello autoregressivo al caso non lineare, rendendo possibile la modellazione di una serie con comportamento asimmetrico, tramite l'utilizzo di più sotto-modelli lineari e una variabile soglia che indica quando usare un sotto-modello piuttosto di un altro. Un caso particolare di questo tipo di modelli è rappresentato dal modello SETAR (Self-Exciting TAR); che utilizza come variabile soglia la stessa Y. In particolare la specificazione del modello seTAR di ordine uno e a due regimi è:

$$Y_t = \begin{cases} \phi_{1,0} + \phi_{1,1} \cdot Y_{t-1} + \sigma_1 \epsilon_t & Y_{t-d} \leq r \\ \phi_{2,0} + \phi_{2,1} \cdot Y_{t-1} + \sigma_2 \epsilon_t & Y_{t-d} > r \end{cases} \tag{4.1}$$

dove $\phi_{i,0}$ e $\phi_{i,1}$, per $i = 1, 2, ...$ sono i coefficienti della parte autoregressiva del modello, d rappresenta il ritardo per il quale viene valutata l'appartenenza ad un regime o all'altro, r è la soglia, σ_i sono le deviazioni standard caratteristiche per ogni sotto-modello e $\{\epsilon_t\}$ è una sequenza di variabili aleatorie indipendenti e identicamente distribuite con media nulla e varianza unitaria. Il modello (4.1) evidenzia come sia possibile trattare Y_t marginalmente come un modello AR(1) con parametri $(\phi_{1,0}, \phi_{1,1}, \sigma_1)$ quando Y_{t-d} è inferiore al valore soglia r, viceversa quando è superiore Y_t sarà modellato tramite un AR(1) con parametri $(\phi_{2,0}, \phi_{2,1}, \sigma_2)$.
Risulta inoltre utile per le simulazioni studiare la stazionarietà dei modelli TAR: in particolare non è possibile dire che in generale valgono le stesse regole che si potrebbero usare per la stazionarietà nei modelli AR, non si possono quindi valutare i singoli sotto-modelli per definire la stazionarietà del modello complessivo. Chan e Tong (1986) hanno dimostrato che un modello del primo ordine è ergodico e quindi stazionario quando i coefficienti dei due sotto-modelli soddisfano le seguenti condizioni:

$$\phi_{1,1} < 1 \qquad \phi_{2,1} < 1 \qquad \phi_{1,1}\phi_{2,1} < 1$$

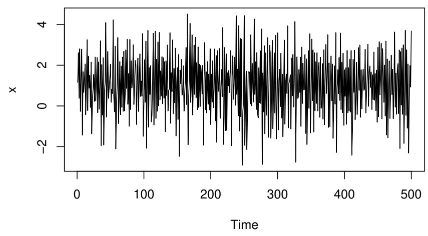
4.2 Simulazioni con modelli SETAR

In questa sezione si confronteranno i grafici dell'ACF e dell'ADCF per diverse serie storiche generate da modelli SETAR a due regimi. In particolare l'obiettivo è evidenziare le differenze che emergono nell'osservare i grafici dell'ACF e dell'ADCF, vedendo come sia possibile avere serie che presentano sia correlazione che dipendenza significative e altre serie con correlazione non significativa e dipendenza significativa.

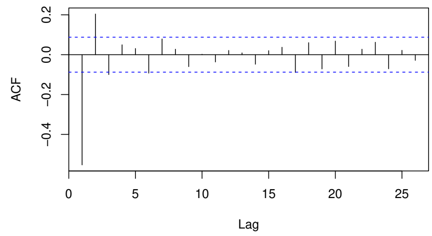
Il primo modello che prendo in esame è un modello SETAR a due regimi così specificato:

$$Y_t = \begin{cases} 2 + 0.5 \cdot Y_{t-1} + \epsilon_t & Y_{t-1} \leq 1.5 \\ 1 - 0.5 \cdot Y_{t-1} + \epsilon_t & Y_{t-1} > 1.5 \end{cases} \tag{4.2}$$

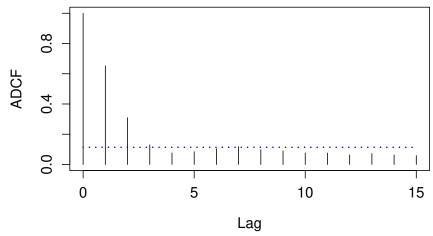
entrambi i sotto-modelli AR(1) risultano stazionari, sia in media che in varianza e l'errore ϵ_t è un *white noise* con distribuzione gaussiana. Le figure 4.1 riportano una realizzazione per n=500 osservazioni di tale processo generatore dei dati, assieme alle funzioni di autocorrelazione e autodistanza campionarie:



(a) grafico della serie



(b) ACF della serie



(c) ADCF della serie

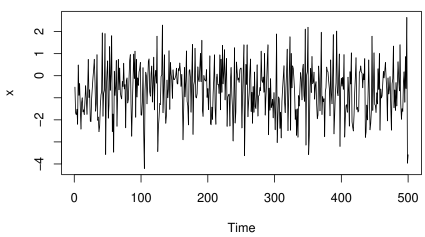
Figura 4.1: grafici della serie simulata da un modello SETAR

Come si può vedere osservando il grafico dell'ACF, la serie mostra valori significativi solo per i primi ritardi. Analogamente l'ADCF risulta significativo per i primi tre ritardi. Questo particolare risultato deriva dal fatto che, se vi è correlazione, i dati risultano dipendenti e quindi ci si aspetta un ADCF significativo.
L'utilizzo di questo particolare modello in cui è presente sia correlazione, rintracciata dall'ACF, sia dipendenza, rintracciata dall'ADCF, è stato proposto per mostrare come sia possibile utilizzare congiuntamente ACF e ADCF, per lo studio delle serie storiche, e di come le conclusioni a cui si può giungere guardando i singoli grafici sono in accordo.

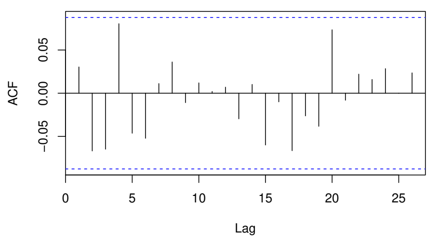
Il secondo modello in esame è un modello SETAR a due regimi così specificato:

$$Y_t = \begin{cases} 0.5 \cdot Y_{t-1} + \epsilon_t & Y_{t-1} \leq -1 \\ -0.8 \cdot Y_{t-1} + \epsilon_t & Y_{t-1} > -1 \end{cases} \tag{4.3}$$

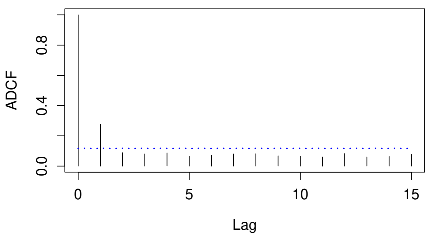
Entrambi i due regimi risultano singolarmente stazionari e anche l'intero modello è stazionario, l'errore ϵ_t è un *white noise* con distribuzione gaussiana.
Per studiare la struttura di dipendeza della serie generata dal modello si andranno a presentare i grafici della serie, dell'ACF e dell'ADCF, la figura 4.2 riporta una realizzazione per n=500 osservazioni da tale processo generatore dei dati, assieme alla funzione di autocorrelazione e autodistanza campionaria:



(a) grafico della serie



(b) ACF della serie



(c) ADCF della serie

Figura 4.2: grafici della serie simulata da un modello SETAR

Il grafico dell'ACF suggerisce la mancanza di una struttura di correlazione all'interno della serie storica dato che tutti i valori dei ritardi risultano all'interno delle bande di confidenza.

Nonostante la mancanza di correlazione l'ADCF mostra come, a ritardo uno, sia presente dipendenza nella serie storica. Questo è un esempio di come sia possibile utilizzare l'ADCF per misurare la dipendenza seriale in situazioni in cui l'ACF non riesce a rintracciare strutture di correlazione significative; così facendo è possibile analizzare una serie storica non solo da punto di vista della correlazione ma anche per la possibile presenza di dipendenza.

4.3 Test d'indipendenza per un ritardo singolo

In questa sezione si riporta un piccolo esperimento Monte Carlo con l'obiettivo di verificare le performance del test d'indipendenza nel caso di indipendenza a ritardo 1 ovvero con $\mathcal{R}_n^2(X_t, X_{t+1}) = 0$. Per fare questo vengono generate 250 replicazioni indipendenti, con numerosità crescente (pari a 100, 150 e 200) sia sotto H_0 che sotto H_1 e, quindi, si calcolerà le frequenze empiriche di rifiuto nei due casi (lavorando con $\alpha = 0.05$). Il processo generatore dei dati sotto H_0 è un AR(2) così specificato: $Y_t = \phi \cdot Y_{t-2} + \epsilon_t$ con ϕ pari a 0.2, 0.5 e 0.8. Il modello appena descritto genera dati che sono indipendenti a ritardo 1, quindi ci si aspetta una frequenza empirica di rigetto attorno al 0.05. Questo risultato viene in parte confermato dalle simulazioni i cui risultati sono riportati nella tabella 4.1.É interessante osservare come aumenti la frequenza empirica di rifiuti all'aumentare del valore del parametro del modello AR. Infatti, facendo avvicinare il coefficiente ad 1, il modello si avvicina ad una situazione di non stazionarietà e, oltre ad aumentare la dipendenza per il ritardo 2 e i suoi multipli, aumenta anche quella dei ritardi circostanti, facendo sembrare che vi sia dipendenza anche a ritardo 1, nonostante non sia presente guardando al processo generatore dei dati. Questo fenomeno va a spiegare il motivo per il quale all'aumentare del coefficiente si incrementa anche la frequenza empirica di rigetto portandola, nel caso di $\phi = 0.8$ a circa il 50%.

numerosità (n)	ϕ	frequenza empirica di rigetto
100	0.2	0.036
	0.5	0.156
	0.8	0.496
150	0.2	0.044
	0.5	0.152
	0.8	0.54
200	0.2	0.064
	0.5	0.192
	0.8	0.524

Tabella 4.1: Frequenza empirica di rigetto sotto H_0 per test d'indipendenza per un ritardo singolo

Lo stesso esperimento è stato effettuato sotto H_1: simulando i dati da un processo con struttura di dipendenza, in questo caso la scelta è ricaduta sul modello SETAR 4.1, descritto nella sezione precedente. Come si può vedere la potenza empirica del test aumenta all'aumentare della numerosità campionaria: già con 100 osservazioni la frequenza empirica di rigetto risulta molto alta, pari a 0.852; arrivando a n=200, questa converge ad 1, andando a testimoniare la bontà del test svolto.

numerosità (n)	frequenza empirica di rigetto
100	0.852
150	0.972
200	1

Tabella 4.2: Frequenza empirica di rigetto sotto H_1 per test d'indipendenza per un ritardo singolo

4.4 Test d'indipendenza per l'intera serie

In questa sezione si calcolerà la frequenza empirica di rigetto del test d'indipendenza sull'intera serie. Per tutti gli esperimenti svolti in questa sezione si utilizzeranno le medesime replicazioni e numerosità campionari presenti nella sezione precedente.
La verifica della frequenza empirica di rigetto verrà fatta sotto H_0, ovvero sotto ipotesi di indipendenza, e generando dati con le caratteristiche di un *white noise* con distribuzione gaussiana. I risultati confermano le aspettative, all'aumentare della numerosità campionaria la frequenza empirica di rigetto si attesta su 0.05. Con lo stesso procedimento può essere verificata la potenza empirica del test: gli esperimenti saranno svolti sotto H_1, ovvero sotto l'ipotesi di dipendenza, in particolare si utilizzerà lo stesso modello SETAR (4.1), descritto nelle simulazioni della sezione precedente. Anche in questo caso al crescere della numerosità campionaria la potenza empirica del test si avvicina ad 1; i risultati dell'esperimento sono riportati nella tabella 4.3

numerosità (n)	frequenza empirica di rigetto sotto H_0	frequenza empirica di rigetto sotto H_1
100	0.052	0.572
150	0.044	0.84
200	0.048	0.972

Tabella 4.3: Frequenze empiriche di rigetto per test d'indipendenza per l'intera serie

Conclusioni

In definitiva le simulazioni svolte nel capitolo 4 hanno evidenziato come l'*auto distance correlation function* risulti decisiva nella misurazione delle strutture di dipendenza all'interno di serie storiche non lineari e il suo utilizzo può, inoltre, risultare complementare all'ACF, nei casi in cui quest'ultimo risulta utilizzabile. L'*auto distance correlation function* infatti misura la forza della dipendenza sussistente all'interno della serie storica, senza dare informazioni sulla direzione di questa dipendenza; in particolare l'utilizzo di questa misura non è relegata al solo studio della presenza di dipendenza all'interno della serie, ma risulta anche propedeutica alla formulazione di test statistici che vanno a verificare l'esistenza o meno di dipendenza sia per l'intera serie che per ritardi fissati.
I risultati teorici esposti nei capitoli 2 e 3, oltre ad evidenziare le proprietà matematiche alla base dell'*auto distance correlation function* e della *distance correaltion*, mostrano anche la genesi della costruzione di queste formule a partire dalle criticità del ρ di Pearson.

Bibliografia

Chan K. S.; Tong H. (1986). On estimating thresholds in autoregressive models. *Journal of time series analysis*, **7**(3), 179–190.

Fokianos K.; Pitsillou M. (2017). Consistent testing for pairwise dependence in time series. *Technometrics*, **59**(2), 262–270.

Hong Y. (1999). Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *Journal of the American Statistical Association*, **94**(448), 1201–1220.

Kendall M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1/2), 81–93.

Pearson K. (1895). Notes on regression and inheritance in the case of two parents proceedings of the royal society of london, 58, 240-242. *K Pearson*.

Pitsillou M.; Fokianos K. (2016). dcovts: Distance covariance/correlation for time series. *R J.*, **8**(2), 324.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rényi A. (1959). On measures of dependence. *Acta mathematica hungarica*, **10**(3-4), 441–451.

Spearman C. (1961). The proof and measurement of association between two things.

Székely G. J.; Rizzo M. L. (2012). On the uniqueness of distance covariance. *Statistics & Probability Letters*, **82**(12), 2278–2282.

Székely G. J.; Rizzo M. L. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, **42**(6), 2382–2412.

Székely G. J.; Rizzo M. L.; Bakirov N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, **35**(6), 2769–2794.

Tjøstheim D.; Otneim H.; Støve B. (2022). Statistical dependence: Beyond pearson's ρ. *Statistical Science*, **37**(1), 90–109.

Zhou Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, **33**(3), 438–457.