Politecnico di Torino

Politecnico di Torino

Corso di Laurea

A.a. 2025/2026

Sessione di laurea Marzo 2026

Test di bontà di adattamento per copule Archimedee: simulazioni e applicazione a rendimenti azionari

Relatore:

Gianluca Mastrantonio

Candidata:

Gaia Galizia

Sommario

Questa tesi replica il test di bontà di adattamento per famiglie parametriche di copule Archimedee proposto da Savu e Trede (2008). Dopo la definizione delle famiglie Archimedee e dopo aver richiamato la metodologia basata sulla Kendall distribution function, si implementa il test con stima CML del parametro sotto H_0 e approssimazione della distribuzione della statistica di test T ottenuta attraverso bootstrap parametrico. L'accuratezza del test viene analizzata con simulazioni Monte Carlo in termini di dimensione e potenza, variando campione e dimensione. L'applicazione empirica riguarda i rendimenti azionari di sei titoli dell'indice S&P500 Chemicals, con stima dei parametri delle famiglie candidate e confronto dei p-value. Il lavoro fornisce una replica della procedura e un quadro operativo per applicazioni su dati reali.

Indice

Elenco delle figure

Capitolo 1

Introduzione

La modellazione della dipendenza tra variabili aleatorie è un problema centrale in molte applicazioni quantitative, e in finanza in particolare, dove la struttura congiunta dei rendimenti incide direttamente su misure di rischio, costruzione e stress test di portafoglio, diversificazione e aggregazione delle perdite. In tali contesti, descrivere la dipendenza tramite la sola correlazione lineare è spesso insufficiente, perché non cattura aspetti non lineari e, soprattutto, il comportamento congiunto nelle code (ad esempio co-movimenti estremi) che è cruciale nell'analisi del rischio. Un punto critico inoltre è che l'assunzione Gaussiana può risultare troppo restrittiva in questi contesti in cui ciò che interessa descrivere è la tendenza delle variabili a muoversi insieme in condizioni di mercato avverse.
La teoria delle copule offre un quadro matematico naturale per separare la modellazione delle marginali dalla modellazione della dipendenza, rendendo possibile combinare marginali anche molto diverse con strutture di dipendenza flessibili e interpretabili [3].
Il punto di partenza è il teorema di Sklar, secondo cui ogni distribuzione congiunta multivariata può essere rappresentata come composizione di una copula e delle distribuzioni marginali; viceversa, assegnata una copula e assegnate le marginali, si ottiene una distribuzione congiunta valida [1]. Questa rappresentazione chiarisce perché le copule siano strumenti particolarmente adatti in finanza: consentono di lavorare con trasformazioni monotone (ad esempio passando dai prezzi ai rendimenti, o dai rendimenti a pseudo-osservazioni) senza alterare l'informazione essenziale sulla dipendenza catturata dalla copula stessa.
Tra le famiglie di copule più utilizzate in applicazioni multivariate si collocano le copule Archimedee, che sono definite a partire da una funzione generatrice e che, nella forma a un parametro, inducono una struttura exchangeable (ossia simmetrica e omogenea rispetto alle coppie) [6]. Questa semplicità le rende attraenti per applicazioni con dimensione moderata o elevata, dove modelli più ricchi possono diventare difficili da stimare o da interpretare. D'altra parte, proprio la natura

parametrica e la specifica struttura di dipendenza implicita rendono inevitabile la domanda metodologica: "quale copula è quella giusta per i dati?". In ambito finanziario questa domanda è stata esplicitata in modo emblematico da Durrleman et al. (2000) e, più in generale, dalla letteratura sulla selezione e validazione dei modelli di dipendenza [15].

La verifica di bontà di adattamento (goodness-of-fit, GOF) per copule è tuttavia un tema complesso: esistono diversi approcci (basati su trasformazioni tipo Rosenblatt, sul processo della copula empirica, su statistiche di tipo χ^2, su processi legati alla Kendall distribution, ecc.), e non esiste un metodo universalmente raccomandato che domini gli altri in ogni contesto applicativo [10]. Inoltre, l'estensione multivariata è particolarmente rilevante in finanza (portafogli con più asset) ma storicamente più difficile della controparte bivariata.

In questo scenario si inserisce il contributo di Savu e Trede (2008), i quali propongono un test GOF specifico per famiglie Archimedee multivariate, costruito su una proiezione univariata della copula stimata e sulla Kendall distribution associata, e che ricorre a bootstrap parametrico per approssimare la distribuzione della statistica sotto ipotesi nulla composta [10]. Il test, pur mantenendo una struttura concettualmente vicina a statistiche di tipo χ^2, adotta una procedura simulativa coerente con l'intera pipeline di stima (costruzione di pseudo-osservazioni, stima CML, ricampionamento parametrico) per i valori critici e p-value [10].

Questa tesi si colloca nel filone delle repliche metodologiche e applicative. L'obiettivo principale è implementare in modo riproducibile il test GOF di Savu e Trede (2008) e verificare, tramite simulazioni Monte Carlo, che l'implementazione riproduca i risultati chiave riportati dagli autori in termini di dimensione empirica (probabilità di rifiuto sotto H_0) e potenza empirica (probabilità di rifiuto sotto alternative) [10]. Una seconda componente, altrettanto centrale, è l'applicazione empirica del test a rendimenti azionari: si replica l'illustrazione empirica del paper, che studia la dipendenza tra rendimenti mensili di titoli del settore chimico dell'indice S&P 500 nel periodo 1996–2006, originariamente estratti da Datastream [16]. Poiché l'accesso a Datastream è vincolato da licenza e perché il tempo trascorso dalla pubblicazione rende più probabili discontinuità nei dati (fusioni, delisting, cambi ticker, rettifiche per split), la replica richiede anche un lavoro di ricostruzione e validazione del dataset tramite fonti alternative, con controlli sulle statistiche descrittive e sulla coerenza della dipendenza stimata.

Dal punto di vista organizzativo, la tesi introduce dapprima la teoria essenziale delle copule e delle famiglie Archimedee, insieme agli strumenti statistici necessari come pseudo-osservazioni, Kendall's τ, stima via CML e Kendall distrbution function. Successivamente viene descritto in dettaglio il test GOF adottato, con particolare attenzione alla costruzione della Kendall distribution e alla motivazione del bootstrap parametrico nella presenza di ipotesi composte e parametri stimati [14]. Infine, vengono presentati: (i) lo studio Monte Carlo di replica delle tabelle di

dimensione e potenza del paper, e (ii) l'applicazione empirica ai rendimenti azionari, includendo verifiche di robustezza (ad esempio rispetto a scelte operative come binning e numero di repliche bootstrap), una discussione delle differenze numeriche osservate rispetto allo studio originale e un'estensione dell'analisi a un periodo più recente, applicando lo stesso test su dati di mercato moderni per valutare la stabilità qualitativa delle conclusioni al variare della finestra temporale.
In sintesi, il contributo di questo lavoro è duplice: da un lato fornisce una replica documentata e riproducibile di un test GOF multivariato per copule Archimedee, con verifica simulativa della sua performance; dall'altro mostra, attraverso un caso empirico realistico, come la validazione di modelli di dipendenza richieda sia scelte statistiche coerenti sia un'attenzione concreta alla qualità e alla tracciabilità dei dati.

Capitolo 2

Copule e dipendenza: fondamenti teorici e modelli Archimedei

Lo scopo di questo capitolo è introdurre la teoria delle copule necessaria per replicare i risultati del paper di riferimento e, in particolare, per comprendere:

- la separazione tra margini e struttura di dipendenza tramite il teorema di Sklar;
- le principali misure di dipendenza usate (Kendall's tau, Spearman's rho, tail dependence);
- la classe delle copule Archimedee e la loro caratterizzazione tramite il generatore;
- le tre famiglie parametriche considerate nello studio: Gumbel–Hougaard, Cook–Johnson e Frank.

2.1 Definizioni fondamentali

2.1.1 Copule e proprietà di base

Sia $\mathbf{X} = (X_1, \ldots, X_d)$ un vettore aleatorio con funzione di ripartizione congiunta $F(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d)$ e marginali $F_i(x) = \mathbb{P}(X_i \leq x)$. L'idea chiave della teoria delle copule è che è possibile rappresentare la dipendenza tra le componenti di $\mathbf{X}$ separatamente rispetto alla forma delle marginali.
Formalmente, una copula d-dimensionale è una funzione di distribuzione multivariata su $[0,1]^d$ con marginali uniformi standard [10].

Definizione 2.1 (Copula d-dimensionale)**.** Una funzione $C : [0,1]^d \to [0,1]$ è una copula se: [1]

1. è grounded: per ogni $\mathbf{u} \in [0,1]^d$, $C(\mathbf{u}) = 0$ se almeno una coordinata di $\mathbf{u}$ è 0;
2. è d-increasing su $[0,1]^d$: per ogni $\mathbf{a}, \mathbf{b} \in [0,1]^d$ con $\mathbf{a} \leq \mathbf{b}$, si ha $V_C([\mathbf{a}, \mathbf{b}]) \geq 0$, dove
$$V_C([\mathbf{a}, \mathbf{b}]) = \sum_{\mathbf{c}} \operatorname{sgn}(\mathbf{c})\, C(\mathbf{c}),$$
la somma è presa su tutti i vertici $\mathbf{c}$ di $[\mathbf{a}, \mathbf{b}]$ e
$$\operatorname{sgn}(\mathbf{c}) = \begin{cases} 1, & \text{se } c_k = a_k \text{ per un numero pari di indici } k, \\ -1, & \text{se } c_k = a_k \text{ per un numero dispari di indici } k. \end{cases}$$
3. ha margini uniformi: per ogni $\mathbf{u} \in [0,1]^d$, se tutte le coordinate di $\mathbf{u}$ sono 1 eccetto u_k, allora $C(\mathbf{u}) = u_k$.

Proprietà:

- poiché una copula è una funzione di distribuzione congiunta su $[0,1]^d$, induce una misura di probabilità su $[0,1]^d$; in particolare vale [2]:
$$V_C([0, u_1] \times \cdots \times [0, u_d]) \;=\; C(u_1, \ldots, u_d).$$
Equivalentemente, se $\mathbf{U}$ è un vettore aleatorio con copula C, allora per ogni $\mathbf{u} \in [0,1]^d$, $V_C([0, \mathbf{u}])$ rappresenta la probabilità che $\mathbf{U}$ assuma valori in $[0, \mathbf{u}]$, dove $[0, \mathbf{u}] = [0, u_1] \times \cdots \times [0, u_d]$. [4]
- continuità e disuguaglianza di tipo Lipschitz: per ogni $\mathbf{u}, \mathbf{v} \in [0,1]^d$,
$$\big|C(\mathbf{u}) - C(\mathbf{v})\big| \leq \sum_{n=1}^{d} |u_n - v_n|.$$
Di conseguenza, C è uniformemente continua sul suo dominio.

[15]

2.1.2 Teorema di Sklar: enunciato e conseguenze

Il risultato centrale della teoria delle copule è il teorema di Sklar, che formalizza la separazione tra marginali e dipendenza ed è alla base di tante, se non quasi tutte, le applicazioni di questa teoria alla statistica.

Teorema 2.1 (Sklar)**.** *Sia H una funzione di distribuzione d-dimensionale con marginali $F_1, \dots, F_d$. Allora esiste una copula C tale che, per ogni $(x_1, \dots, x_d) \in \overline{\mathbb{R}}^d$,*

$$H(x_1, \dots, x_d) = C\Big(F_1(x_1), \dots, F_d(x_d)\Big). \tag{2.1}$$

Se le marginali sono continue, la copula C è unica; altrimenti, C è unica su $\mathrm{Ran}(F_1) \times \cdots \times \mathrm{Ran}(F_d)$.
Viceversa, se C è una copula d-dimensionale e $F_1, \dots, F_d$ sono funzioni di ripartizione, allora la funzione H definita da (2.1) *è una distribuzione d-dimensionale con marginali $F_1, \dots, F_d$ [9].*

L'equazione (2.1) esprime una congiunta in funzione di una copula e delle marginali; può però essere "invertita" per esprimere la copula in funzione della congiunta e delle inverse delle marginali.
Se una marginale non è strettamente crescente, però, l'inversa usuale può non esistere: si introduce quindi una nozione di inversa generalizzata: [1]

Definizione 2.2. Sia F una funzione di distribuzione univariata. La funzione F^{-1} tale che, per $t \in [0,1]$,

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\} = \sup\{x \in \mathbb{R} : F(x) \leq t\}.$$

è detta inversa generalizzata di F.
Se F è strettamente crescente, questa coincide con l'inversa ordinaria; inoltre, per $t \in \mathrm{Ran}(F)$, vale $F(F^{-1}(t)) = t$ [2].

Corollario 2.1. *Se H è una funzione di distribuzione d-dimensionale con marginali continue $F_1, \dots, F_d$ e C è una copula che soddisfa le ipotesi del Teorema 2.1, allora per ogni* $\mathbf{u} \in [0,1]^d$

$$C(u_1, \dots, u_d) = H\Big(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)\Big).$$

[2]

Questo risultato fornisce un metodo per costruire le copule da funzioni di distribuzione congiunte.
Nel linguaggio delle variabili aleatorie, se $X_1, \dots, X_d$ hanno funzioni di distribuzione $F_1, \dots, F_d$ e congiunta H, allora esiste una copula C tale che (2.1) è soddisfatta. Se le marginali sono continue, C è unica; altrimenti, è unicamente determinata su $\mathrm{Ran}(F_1) \times \cdots \times \mathrm{Ran}(F_d)$. [1]

Conseguenze pratiche. Il teorema di Sklar implica non solo che ogni copula con distribuzioni marginali come argomento è una distribuzione multivariata valida,

ma anche che ogni distribuzione multivariata può essere rappresentata come copula delle sue marginali; quindi, assegnate le marginali, la costruzione di un modello per distribuzioni multivariate si riduce alla scelta di una copula appropriata (spesso unica), anche se questa scelta non è banale in applicazioni reali [8].
In particolare, vale:

- invarianza della copula rispetto a trasformazioni monotone crescenti delle marginali: se $\mathbf{X} = (X_1, \dots, X_d)$ ha copula $C_{\mathbf{X}}$ e $h_1, \dots, h_d$ sono strettamente crescenti su $\mathrm{Ran}(X_i)$, allora $(h_1(X_1), \dots, h_d(X_d))$ ha ancora copula $C_{\mathbf{X}}$ [1].

- nel caso di distribuzioni continue, è possibile modellare le marginali F_i e la copula C separatamente;

- molte misure di dipendenza invarianti per trasformazioni monotone dipendono solo dalla copula;

- a partire dai dati osservati, è naturale lavorare con pseudo-osservazioni (ranks) per stimare la dipendenza senza imporre un modello parametrico sulle marginali. [6]

2.1.3 Limiti di Fréchet–Hoeffding

Per ogni copula d-dimensionale C e per ogni $\mathbf{u} \in [0,1]^d$ valgono i vincoli universali

$$W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}), \tag{2.2}$$

dove

$$W(\mathbf{u}) = \max\Big\{\sum_{i=1}^{d} u_i - d + 1,\, 0\Big\}, \qquad M(\mathbf{u}) = \min\{u_1, \dots, u_d\}.$$

Interpretando le copule come superfici sul cubo unitario $[0,1]^d$, i limiti (2.2) rappresentano, rispettivamente, la superficie superiore e quella inferiore, tra cui si collocano tutte le altre copule ammissibili [5].

Copula di comonotonicità (upper bound): La copula

$$M(u_1, \dots, u_d) = \min\{u_1, \dots, u_d\}$$

è il limite superiore di Fréchet e rappresenta la dipendenza perfetta positiva: è la funzione di ripartizione congiunta di $(U, \dots, U)$, con $U \sim \mathcal{U}(0,1)$. In particolare, descrive il caso in cui $X_i = h_i(X_1)$ con trasformazioni strettamente crescenti h_i, $i = 2, \dots, d$. $M(\mathbf{u})$ è una copula per ogni d.

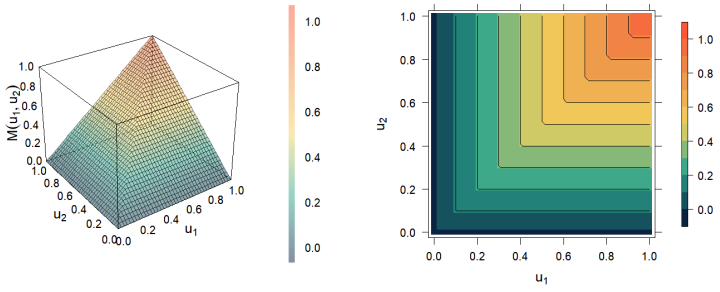

Figura 2.1: Copula di comonotonicità M: wireframe plot (sinistra) e contour plot (destra) nel caso bivariato.

Copula di contromonotonicità (lower bound): Nel caso bivariato, il limite inferiore è la copula di contromonotonicità

$$W(u_1, u_2) = \max\{u_1 + u_2 - 1,\, 0\},$$

che è la funzione di ripartizione congiunta di $(U, 1-U)$ e descrive la dipendenza perfetta negativa. Per $d > 2$, W non è una copula ma come bound rappresenta il "meglio possibile", in quanto per ogni $d \geq 3$ e ogni $\mathbf{u}$ in $[0,1]^d$ esiste una copula C d-dimensionale tale che $C(\mathbf{u}) = W(\mathbf{u})$.

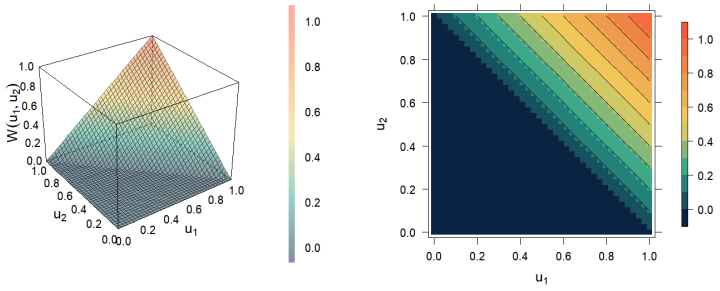

Figura 2.2: Copula di contromonotonicità W: wireframe plot (sinistra) e contour plot (destra) nel caso bivariato.

Copula di indipendenza: La copula di indipendenza è

$$\Pi(u_1, \ldots, u_d) = \prod_{i=1}^{d} u_i,$$

e corrisponde al caso in cui le componenti di $\mathbf{X}$ sono indipendenti. Nel caso bivariato, Π si colloca tra W e M.

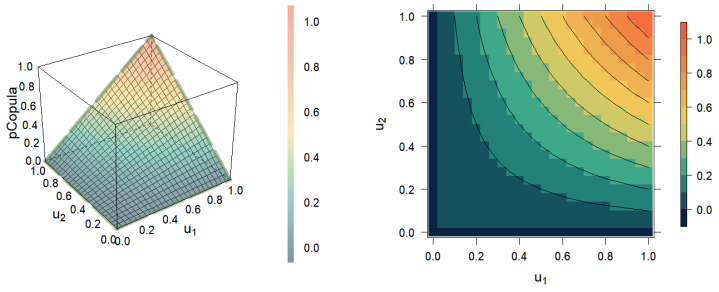

Figura 2.3: Copula di indipendenza Π: wireframe plot (sinistra) e contour plot (destra) nel caso bivariato.

2.2 Misure di dipendenza

2.2.1 Perché non basta la correlazione lineare

Spesso da un punto di vista applicativo è desiderabile misurare la dipendenza tra variabili aleatorie e, a tale scopo, viene frequentemente utilizzato il coefficiente di correlazione lineare:

Definizione 2.3. Sia (X, Y) un vettore di variabili aleatorie con varianze finite non nulle, il coefficiente di Pearson è

$$Corr(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

La correlazione di Pearson $Corr(X, Y)$ è semplice e utile per misurare la dipendenza lineare, ma può essere fuorviante in contesti non ellittici: non è invariante rispetto a trasformazioni non lineari strettamente crescenti; non è definita per tutti i vettori aleatori (X, Y) bensì solo per quelli con momenti secondi finiti; $Corr(X, Y) = 0$ non implica indipendenza e dipende dalle marginali di (X, Y) anche nel caso continuo [6]. Queste criticità chiariscono perché la correlazione lineare, pur diffusissima in pratica, non è una misura basata sulla copula e motivano l'uso di misure di associazione rank-based e tail-based.

2.2.2 Concordanza

Siano (x_i, y_i) e (x_j, y_j) due osservazioni da un vettore (X, Y) di variabili aleatorie continue. (x_i, y_i) e (x_j, y_j) sono detti concordanti se $(x_i - x_j)(y_i - y_j) > 0$ e discordanti se $(x_i - x_j)(y_i - y_j) < 0$. Informalmente, due variabili aleatorie sono concordanti se valori grandi di una tendono ad essere associati a valori grandi dell'altra e valori piccoli di una a valori piccoli dell'altra [1].

2.2.3 Kendall's τ

Siano (X_1, Y_1) e (X_2, Y_2) vettori aleatori indipendenti e identicamente distribuiti, ciascuno con distribuzione congiunta H, allora la τ di Kendall è definita come

$$\begin{aligned}\tau_{X,Y} &= \mathbb{P}(\text{concordanza}) - \mathbb{P}(\text{discordanza}) \\ &= \mathbb{P}\Big((X_1 - X_2)(Y_1 - Y_2) > 0\Big) - \mathbb{P}\Big((X_1 - X_2)(Y_1 - Y_2) < 0\Big),\end{aligned}$$

Alternativamente, sia (X, Y) un vettore di variabili aleatorie continue con copula C, si ha

$$\tau_{X,Y} = \tau_C = 4 \iint_{[0,1]^2} C(u, v)\, dC(u, v) - 1. \tag{2.3}$$

ed è invariante per trasformazioni monotone strettamente crescenti (quindi dipende solo dalla copula associata) [2].
In generale, la valutazione della τ di Kendall richiede il calcolo del doppio integrale in (2.3). Per una copula Archimedea, però, la situazione è più semplice, poiché la Kendall's τ può essere ricavata direttamente dal generatore della copula (vedi Sez. 2.3) [1].
Si ha che $\tau \in [-1,1]$: valori positivi indicano prevalenza di concordanza (tendenza a muoversi insieme), valori negativi prevalenza di discordanza.
Per molte famiglie parametriche di copule (in particolare quelle a un parametro), la relazione tra la τ di Kendall e il parametro θ della copula è strettamente monotona e, pertanto, invertibile. Di conseguenza, la τ campionaria può essere utilizzata per interpretare o inizializzare la stima parametrica [6].

2.2.4 Spearman's ρ

Siano (X_1, Y_1), (X_2, Y_2) e (X_3, Y_3) vettori aleatori indipendenti con funzione di distribuzione congiunta comune H (e marginali F e G) e copula C. La ρ di Spearman è definita come

$$\rho_{X,Y} = 3\Big(\mathbb{P}\Big((X_1 - X_2)(Y_1 - Y_3) > 0\Big) - \mathbb{P}\Big((X_1 - X_2)(Y_1 - Y_3) < 0\Big)\Big),$$

Alternativamente, sia (X, Y) un vettore di variabili aleatorie continue con copula C, si ha

$$\rho_{X,Y} = \rho_C = 12 \iint_{[0,1]^2} C(u, v)\, du\, dv - 3 = 12 \iint_{[0,1]^2} \Big(C(u, v) - uv\Big)\, du\, dv,$$

quindi ρ_C misura la "distanza media" tra la distribuzione di X e Y (rappresentata dalla copula C) e l'indipendenza (rappresentata dalla copula Π) [1].
Se $X \sim F$ e $Y \sim G$, ponendo $U = F(X)$ e $V = G(Y)$, allora

$$\rho_{X,Y} = \frac{\mathrm{Cov}(U, V)}{\sqrt{\mathrm{Var}(U)}\sqrt{\mathrm{Var}(V)}} = \mathrm{Corr}(U, V).$$

Proprietà: τ e ρ dipendono solo dalla copula sottostante (quindi sono invarianti rispetto a trasformazioni strettamente crescenti delle variabili aleatorie); $\tau, \rho \in [-1,1]$; $\tau(C) = \rho(C) = -1 \iff C = W$ e $\tau(C) = \rho(C) = 1 \iff C = M$; in caso di indipendenza $\tau = \rho = 0$ [2].

2.2.5 Dipendenza nelle code (tail dependence)

Per applicazioni finanziarie o di rischio è cruciale descrivere la co-occorrenza di eventi estremi. I coefficienti di tail dependence misurano la probabilità limite di osservare un estremo in una variabile condizionatamente a un estremo nell'altra.

Definizione. Siano X e Y variabili aleatorie continue con funzione di distribuzione F e G rispettivamente, i coefficienti di dipendenza in coda superiore e inferiore sono

$$\lambda_U = \lim_{t\to 1^-} \mathbb{P}\Big(Y > G^{-1}(t) \mid X > F^{-1}(t)\Big),$$
$$\lambda_L = \lim_{t\to 0^+} \mathbb{P}\Big(Y \leq G^{-1}(t) \mid X \leq F^{-1}(t)\Big).$$

quando i limiti $\lambda_U, \lambda_L \in [0,1]$ esistono [1]. In termini di copula, quando i limiti definiti sopra esistono, si ottiene

$$\lambda_U = 2 - \lim_{t\to 1^-} \frac{1 - C(t,t)}{1-t} \qquad \lambda_L = \lim_{t\to 0^+} \frac{C(t,t)}{t}.$$

In particolare,

- $\lambda_U > 0$ indica tendenza a muoversi insieme in "grandi valori" (co-rialzi o grandi estremi positivi).
- $\lambda_L > 0$ indica tendenza a muoversi insieme in "piccoli valori" (co-crolli o estremi negativi).
- Per alcune famiglie $\lambda_L = \lambda_U = 0$: la dipendenza non si manifesta negli eventi estremi; può però restare significativa per valori moderati.

2.3 Copule Archimedee

2.3.1 Definizione tramite generatore

Le copule Archimedee costituiscono una classe di copule molto usata per varie ragioni: sono facili da costruire, molte famiglie parametriche rilevanti sono Archimedee, la classe permette una grande varietà di strutture di dipendenza e, a differenza delle copule ellittiche, le copule Archimedee comunemente usate ammettono espressioni

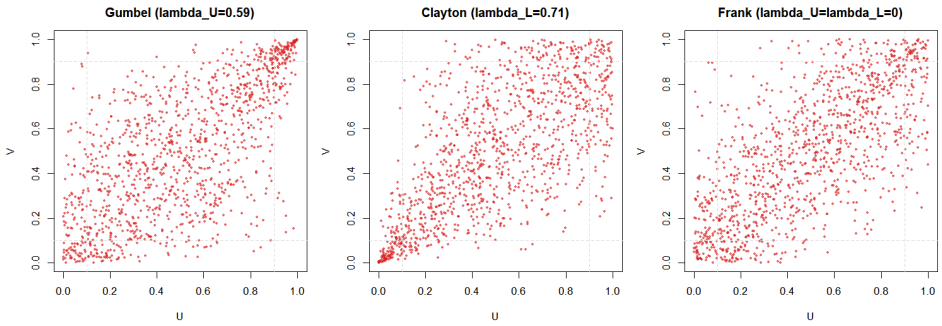

Figura 2.4: Scatter plot di campioni simulati da copule Gumbel, Clayton e Frank a parità di Kendall's τ ($\tau = 0.5$); le linee tratteggiate delimitano le regioni di coda. Si osserva concentrazione nella coda superiore per Gumbel, nella coda inferiore per Clayton e assenza di tail dependence per Frank.

in forma chiusa. [2]
Una copula Archimedea d-dimensionale $C : [0,1]^d \to [0,1]$ è costruita a partire da un generatore φ:

$$C(u_1, \dots, u_d) = \varphi^{-1}\Big(\varphi(u_1) + \cdots + \varphi(u_d)\Big), \tag{2.4}$$

dove $\varphi : [0,1] \to [0,\infty]$ è continua, strettamente decrescente e convessa, con $\varphi(1) = 0$ e $\varphi(0) = \infty$; la funzione φ ammette un'inversa $\varphi^{-1} : [0,\infty] \to [0,1]$ con le stesse proprietà di φ, eccetto che $\varphi^{-1}(0) = 1$ e $\varphi^{-1}(\infty) = 0$.
Inoltre, C è una copula archimedea per ogni $d \geq 2$ se e solo se φ^{-1} è completamente monotona su $[0,\infty)$, cioè

$$(-1)^k \frac{d^k}{du^k} \varphi^{-1}(u) \geq 0, \qquad k = 1,2,\dots$$

[10].

2.3.2 Proprietà chiave e limitazioni

Sia C una copula Archimedea con generatore φ. Allora:

1. C è simmetrica per permutazioni dei suoi d argomenti; ad esempio, nel caso bivariato $C(u_1, u_2) = C(u_2, u_1)$ per ogni $u_1, u_2 \in [0,1]$. Questo implica che C è la funzione di distribuzione di d variabili uniformi exchangeable;

2. C è associativa; ad esempio, nel caso trivariato

$$C\Big(C(u_1, u_2), u_3\Big) = C\Big(u_1, C(u_2, u_3)\Big), \qquad u_1, u_2, u_3 \in [0,1];$$

3. se $a > 0$ è una costante, allora $a\varphi$ è ancora un generatore di C. Di conseguenza il generatore determina la copula archimedea solo a meno di un fattore scalare.

[10]
Un ulteriore vantaggio è che, nel caso di copule Archimedee, la dipendenza è relativamente semplice da quantificare perché la τ di Kendall si esprime direttamente in funzione del generatore [2]:

$$\tau = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)}\, dt.$$

Nonostante la semplicità costruttiva, le copule Archimedee presentano alcuni limiti. In particolare la struttura di dipendenza è piuttosto limitata in quanto tutti i k-margini di una copula Archimedea d-dimensionale sono identici; in altre parole, C è la funzione di distribuzione di d variabili uniformi $\mathcal{U}(0,1)$ scambiabili. Per superare l'impossibilità di modellare la dipendenza tra le varie coppie in modo separato, vengono adottate estensioni multivariate parzialmente scambiabili [7].

2.3.3 Kendall distribution function $K(t)$

Un oggetto fondamentale (anche per i test di bontà di adattamento del paper) è la Kendall distribution function. Se $\mathbf{U} = (U_1, \ldots, U_d)$ ha copula C, si definisce la variabile

$$V = C(U_1, \ldots, U_d) \in [0,1],$$

e la sua funzione di ripartizione

$$K(t) = \mathbb{P}(V \leq t), \qquad t \in [0,1]. \tag{2.5}$$

Per le copule Archimedee, $K(t)$ ammette una forma esplicita (in termini del generatore e di funzioni ausiliarie) che consente di confrontare la struttura di dipendenza stimata con quella osservata. Nel caso bivariato, tale espressione si semplifica in modo notevole [11]:

$$K(t) = t - \frac{\varphi(t)}{\varphi'(t)}.$$

In particolare, una copula Archimedea è determinata dalla funzione $K(t)$ sull'intervallo [0,1]: questo risultato è molto utile quando si vuole stabilire quale famiglia parametrica si adatti meglio ai dati.

La teoria asintotica del processo empirico associato alla Kendall distribution function $K(t)$ (Kendall's process) fornisce la base per costruire test di bontà di adattamento e statistiche di scarto tra modello e dati. Questo quadro è sviluppato in modo formale nella letteratura dedicata [14].

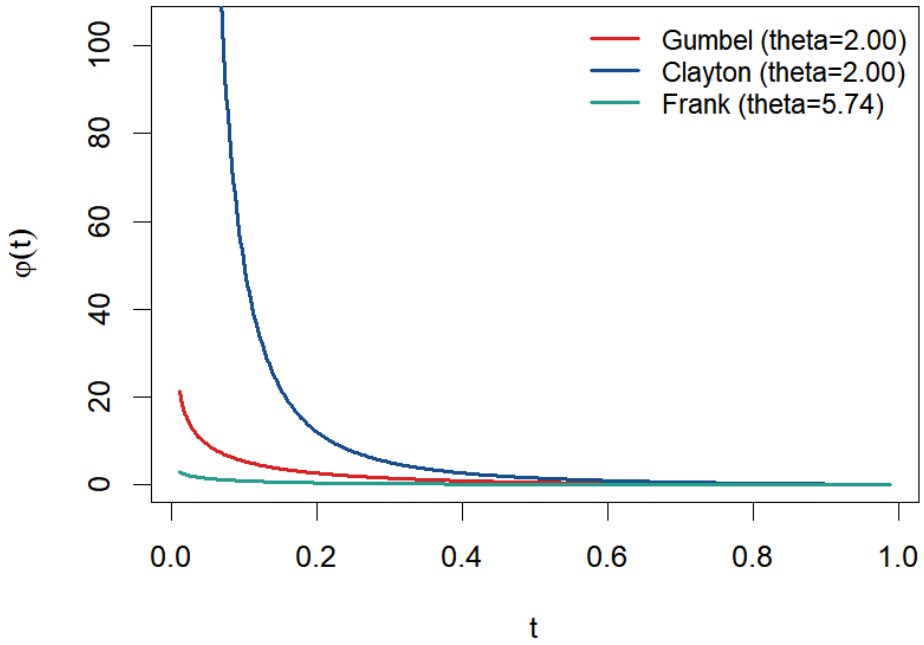

Figura 2.5: Generatori Archimedei $\varphi(t)$ per le famiglie Gumbel, Clayton e Frank (parametri scelti a parità di τ di Kendall, con $\tau = 0.5$), tracciati per $t \in (0.01,0.99)$.

2.4 Famiglie parametriche studiate

In questa sezione vengono presentate le tre famiglie Archimedee a singolo parametro considerate nell'analisi: Gumbel–Hougaard, Cook–Johnson (Clayton) e Frank. Per ciascuna famiglia si riportano: generatore, forma della copula, range del parametro, legame con τ di Kendall e proprietà di tail dependence. Le espressioni qui riportate sono standard in letteratura e sono quelle effettivamente utilizzate per interpretare e replicare i risultati empirici e i test di bontà di adattamento.

2.4.1 Gumbel–Hougaard ($\theta \geq 1$)

La famiglia di Gumbel–Hougaard è adatta a descrivere dipendenza più marcata nella coda superiore.

Generatore e copula. Il generatore è

$$\varphi(t) = (-\ln t)^{\theta}, \qquad \theta \geq 1,$$

che produce la copula $d-$dimensionale ($d \geq 2$)

$$C_\theta(u_1, \ldots, u_d) = \exp\left(-\left((-\ln u_1)^\theta + \cdots + (-\ln u_d)^\theta \right)^{1/\theta} \right).$$

Relazione con Kendall. Per questa famiglia, la τ di Kendall è

$$\tau = 1 - \frac{1}{\theta},$$

quindi θ cresce con la forza di dipendenza: in particolare $C_\theta = \Pi$ per $\theta = 1$, mentre $C_\theta \to M$ quando $\theta \to \infty$. [2].

Tail dependence. La Gumbel presenta dipendenza in coda superiore:

$$\lambda_U = 2 - 2^{1/\theta}, \qquad \lambda_L = 0.$$

Interpretazione: maggiore probabilià di co-movimenti in eventi "al rialzo" (o grandi estremi positivi), ma non nella coda inferiore.

2.4.2 Cook–Johnson / Clayton ($\theta > 0$)

La famiglia di Clayton è spesso usata per catturare dipendenza in coda inferiore (co-crolli).

Generatore e copula. Un generatore standard è

$$\varphi(t) = \frac{t^{-\theta} - 1}{\theta}, \qquad \theta > 0,$$

che induce per ogni $d \geq 2$ la copula

$$C_\theta(u_1, \ldots, u_d) = \left(u_1^{-\theta} + \cdots + u_d^{-\theta} - d + 1 \right)^{-1/\theta}.$$

Relazione con Kendall. Per Clayton vale

$$\tau = \frac{\theta}{\theta + 2},$$

ancora monotona e utile per interpretare il parametro: in particolare $C_\theta \to \Pi$ quando $\theta \to 0$, mentre $C_\theta \to M$ quando $\theta \to \infty$. [2].

Tail dependence. Questa famiglia presenta dipendenza in coda inferiore, ma non in coda superiore:

$$\lambda_L = 2^{-1/\theta}, \qquad \lambda_U = 0.$$

Interpretazione: maggiore probabilità di co-occorrenza di eventi estremi negativi (o "piccoli valori") rispetto a famiglie senza tail dependence [1].

2.4.3 Frank ($\theta > 0$)

La copula di Frank è simmetrica e non presenta tail dependence.

Generatore e copula. Generatore dato da

$$\varphi(t) = -\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right), \qquad \theta > 0,$$

e la copula associata in dimensione d ammette la forma chiusa

$$C_\theta(u_1, \ldots, u_d) = -\frac{1}{\theta}\ln\left(1 + \frac{\prod_{i=1}^{d}(e^{-\theta u_i} - 1)}{(e^{-\theta} - 1)^{\,d-1}}\right).$$

Per $\theta < 0$ l'inversa del generatore non è completamente monotona, quindi in quel caso la Frank è una copula solo per $d = 2$ [1].

Relazione con Kendall. La τ di Kendall per questa famiglia si esprime tramite la funzione di Debye $D_1(\theta)$:

$$\tau = 1 - \frac{4}{\theta}\Big(1 - D_1(\theta)\Big),$$

dove, per $k \in \mathbb{N}$,

$$D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t - 1}\,dt.$$

Tail dependence. Per la Frank vale

$$\lambda_L = \lambda_U = 0,$$

quindi c'è dipendenza ma non è concentrata agli estremi. [2].

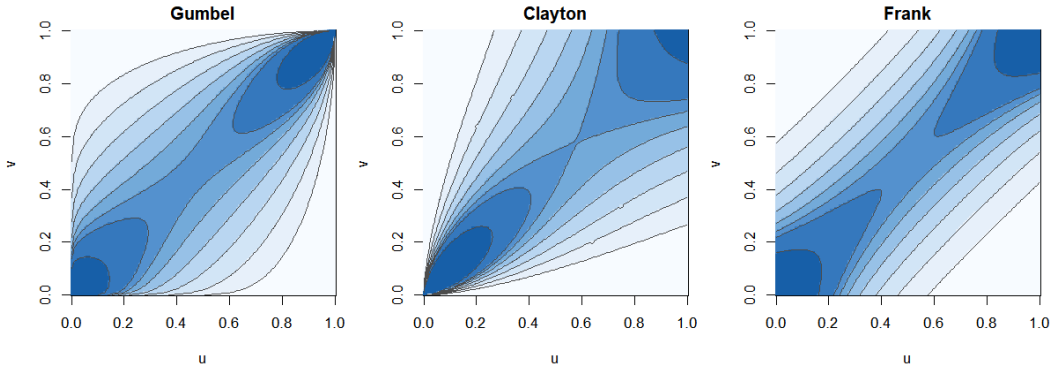

Figura 2.6: Confronto qualitativo tra copule Archimedee: livelli di densità basati su quantili comuni (Gumbel, Clayton, Frank) a parità di τ di Kendall ($\tau = 0.5$).

2.5 Scelta del modello e bontà di adattamento

Nelle applicazioni reali la scelta della famiglia di copula non è banale: famiglie diverse possono avere la stessa τ ma comportamenti molto differenti, specialmente nelle code. Questo aspetto motiva i test di bontà di adattamento costruiti confrontando la $K(t)$ teorica e quella empirica per stabilire quale copula si adatti meglio ai dati [10].
In generale la scelta può dipendere anche da cosa si vuole catturare e quindi se il focus sono co-crolli o se invece interessano i co-rialzi.

Capitolo 3

Metodologia di replica e implementazione del test di bontà di adattamento

3.1 Panoramica della replica e struttura della procedura

L'obiettivo della tesi è replicare il test di bontà di adattamento (goodness-of-fit test) proposto da Savu e Trede (2008) [10] per famiglie parametriche di copule Archimedee, con attenzione alla situazione di ipotesi nulla composita in cui, quindi, il parametro non è noto neanche sotto l'ipotesi nulla e deve, perciò, essere stimato. In tale contesto, il paper si concentra su un test che riguarda solo la specificazione della copula, evitando di imporre un modello parametrico per le distribuzioni marginali [10].

3.1.1 Struttura della procedura

Sia $\{(X_{11}, \ldots, X_{d1}), \ldots, (X_{1n}, \ldots, X_{dn})\}$ un campione casuale di dimensione n dal vettore aleatorio continuo $X = (X_1, \ldots, X_d)$ con copula incognita C. Lo scopo è testare se C appartenga o meno ad una famiglia parametrica $\{C_\theta : \theta \in \Theta\}$.
Quindi, la formulazione è data da

$$H_0 : \ C \in \{C_\theta : \theta \in \Theta\} \qquad \text{vs} \qquad H_1 : \ C \notin \{C_\theta : \theta \in \Theta\}, \tag{3.1}$$

Con questa procedura ciò che si fa è stimare i margini non parametricamente tramite la funzione di ripartizione empirica, mentre l'informazione di dipendenza resta concentrata nella copula [10].

Input. Osservazioni reali (es. rendimenti) in dimensione d e numerosità n.
Output. (i) stima $\hat{\theta}$ per ciascuna famiglia candidata massimizzando la pseudo-log-verosimiglianza, (ii) statistica test T calcolata sui dati, (iii) p-value ottenuto via bootstrap parametrico, (iv) decisione di rifiuto o meno dell'ipotesi nulla H_0 al livello prefissato.

3.1.2 Idea chiave: stima dei margini via ranghi e riduzione univariata

Il test di Savu e Trede sfrutta due ingredienti fondamentali:

1. **Inferenza semiparametrica.** Le distribuzioni marginali vengono stimate non-parametricamente e vengono sostituite da pseudo-osservazioni basate sui ranghi. La stima del parametro θ della copula avviene massimizzando la pseudo-log-verosimiglianza [10], nota come canonical maximum likelihood (CML) [15].

2. **Proiezione tramite Kendall distribution function.** La dipendenza multivariata viene "proiettata" sulla variabile unidimensionale

$$V = C(U_1, \ldots, U_d),$$

dove $U_1, \ldots, U_d$ sono d variabili aleatorie con funzione di distribuzione C e generatore φ. In particolare, la distribuzione univariata $K(t) = \mathbb{P}(V \leq t)$ (Kendall distribution function) diventa il centro del GOF per determinare quale famiglia parametrica si adatta meglio ai dati.

Poichè $K(t)$ definisce la funzione di distribuzione della copula Archimedea con generatore φ sfruttando le proprietà di queste copule, il test è informativo solo all'interno della classe Archimedea: copule non Archimedee con la stessa $K(t)$ non vengono distinte. Quindi viene mantenuta l'ipotesi dell'Archimedeità delle copule in considerazione senza un test formale [10].
In Savu e Trede (2008) [10], l'uso della quantità $\widehat{V}_j = \varphi_{\hat{\theta}}^{-1}\big(\varphi_{\hat{\theta}}(U_{1j}) + \cdots + \varphi_{\hat{\theta}}(U_{dj})\big)$ dipende dalla stima $\hat{\theta}$ (motivo per cui la proiezione sull'intervallo unitario è randomica), e questo rende non immediata l'applicazione dell'asintotica χ^2 standard. Per questo motivo, la distribuzione della statistica test sotto H_0 è approssimata tramite bootstrap parametrico [10].

3.1.3 Schema operativo

Questa tesi ricalca i passi centrali di Savu e Trede (2008):

1. trasformazione dei dati in pseudo-osservazioni;
2. stima del parametro $\hat{\theta}$ della famiglia candidata via CML;
3. calcolo della variabile proiettata $\widehat{V}_j$ e della statistica di test T;
4. bootstrap parametrico: generazione di J campioni artificiali dalla copula stimata, ristima di $\hat{\theta}$ e ricalcolo di T^*;
5. determinazione di valore critico e/o p-value.

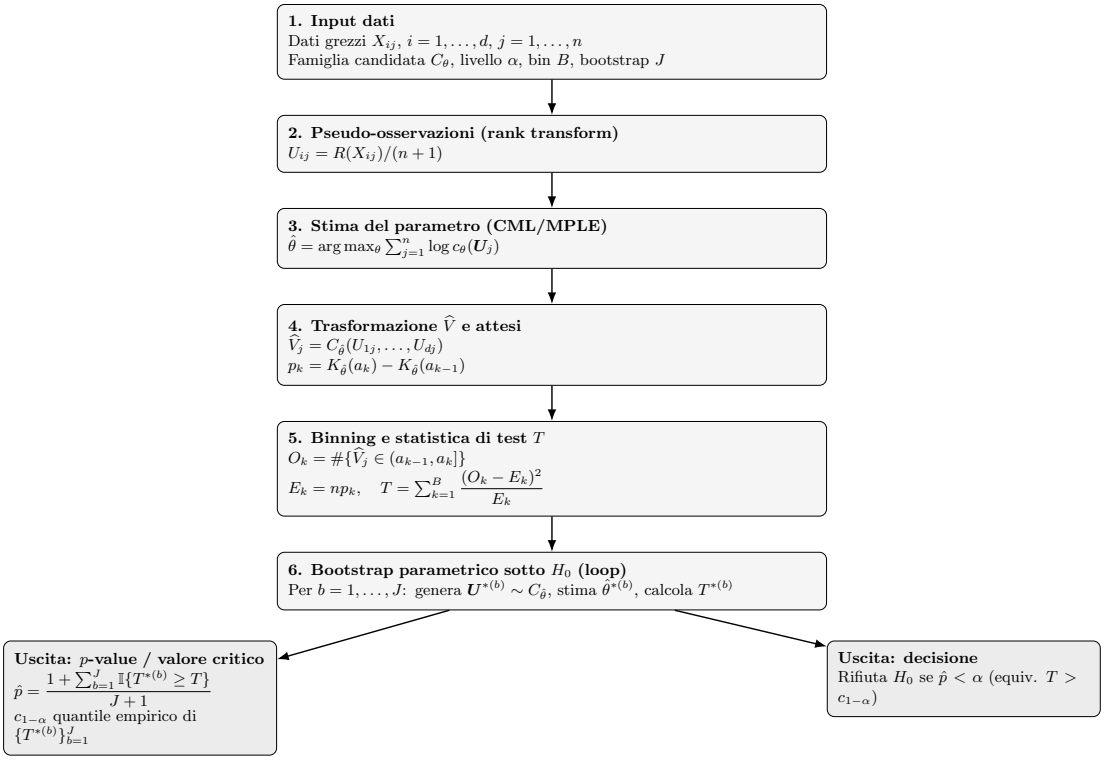

Figura 3.1: Schema della pipeline di replica del test GOF (flowchart): costruzione delle pseudo-osservazioni, stima CML del parametro, calcolo della statistica test e bootstrap parametrico per valore critico/p-value.

Algorithm 1 Schema generale del goodness-of-fit test

Require: Dati grezzi $\{x_{ij}\}$, $i = 1, \dots, d$, $j = 1, \dots, n$; famiglia candidata $\{C_\theta\}$; livello α; numero bootstrap J

1: Costruisci pseudo-osservazioni u_{ij} tramite ranghi
2: Stima $\hat{\theta}$ massimizzando la pseudo-log-verosimiglianza su $\{u_{ij}\}$ (CML)
3: Calcola la statistica T sui dati (proiezione $\widehat{V}_j = C_{\hat{\theta}}(u_{1j}, \dots, u_{dj})$, binning, conteggi O_k, E_k)
4: **for** $b = 1$ **to** J **do**
5: Genera campione bootstrap $\tilde{\mathbf{U}}_1^{(b)}, \dots, \tilde{\mathbf{U}}_n^{(b)} \sim C_{\hat{\theta}}$
6: Costruisci pseudo-osservazioni $u_{ij}^{*(b)}$ tramite ranghi
7: Ristima $\hat{\theta}^{*(b)}$ via CML su $\{u_{ij}^{*(b)}\}$
8: Calcola $T^{*(b)}$ replicando il passo 3 sul campione bootstrap
9: **end**
10: Calcola il p-value bootstrap $\hat{p} = \frac{1+\sum_{b=1}^{J} \mathbb{I}\{T^{*(b)} \geq T\}}{J+1}$ oppure valore critico $c_{1-\alpha}$ come quantile empirico $(1-\alpha)$ di $\{T^{*(b)}\}_{b=1}^{J}$
11: **Rifiuta** H_0 se $T > c_{1-\alpha}$ (equivalentemente se $\hat{p} < \alpha$)

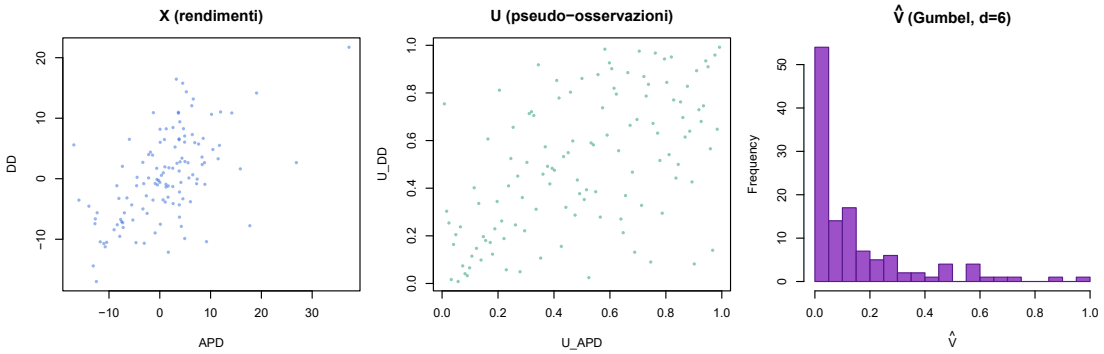

Figura 3.2: Trasformazione $X \to U \to \widehat{V}$: i primi due pannelli sono bivariati (qui APD–DD) e mostrano dati grezzi (sinistra) e pseudo-osservazioni (centro); il pannello di destra mostra $\widehat{V}$ costruita su tutte le 6 variabili dell'applicazione empirica ($d = 6$) e riassume la dipendenza complessiva tramite la copula stimata.

3.2 Costruzione delle pseudo-osservazioni

In questa sezione viene descritto l'input statistico della replica e la trasformazione che consente di lavorare sulla copula senza specificare le distribuzioni per i margini.

Si osservi un campione di ampiezza n da una variabile aleatoria multivariata continua $X = (X_1, \ldots, X_d)$. Indichiamo con

$$(X_{11}, \ldots, X_{d1}),\ \ldots,\ (X_{1n}, \ldots, X_{dn})$$

le realizzazioni campionarie. In particolare, X_{ij} con $i \in \{1, \ldots, d\}$ e $j \in \{1, \ldots, n\}$ rappresenta la i-esima componente (margine) della j-esima osservazione [10].

3.2.1 Pseudo-osservazioni tramite ranghi (margini non parametrici)

Seguendo Savu e Trede [10], per ogni marginale i si considerano i ranghi (con ordinamento crescente) di X_{ij} all'interno della serie $\{X_{i1}, \ldots, X_{in}\}$. Definendo $R(X_{ij})$ come rango di X_{ij}, le pseudo-osservazioni sono

$$U_{ij} \;=\; \frac{1}{n+1} \sum_{k=1}^{n} \mathbb{I}\{X_{ik} \leq X_{ij}\} \;=\; \frac{R(X_{ij})}{n+1}, \qquad i = 1, \ldots, d,\ \ j = 1, \ldots, n. \quad (3.2)$$

In questo modo, i margini vengono stimati non parametricamente tramite la funzione di ripartizione empirica, mentre l'informazione di dipendenza resta concentrata nella copula [10].

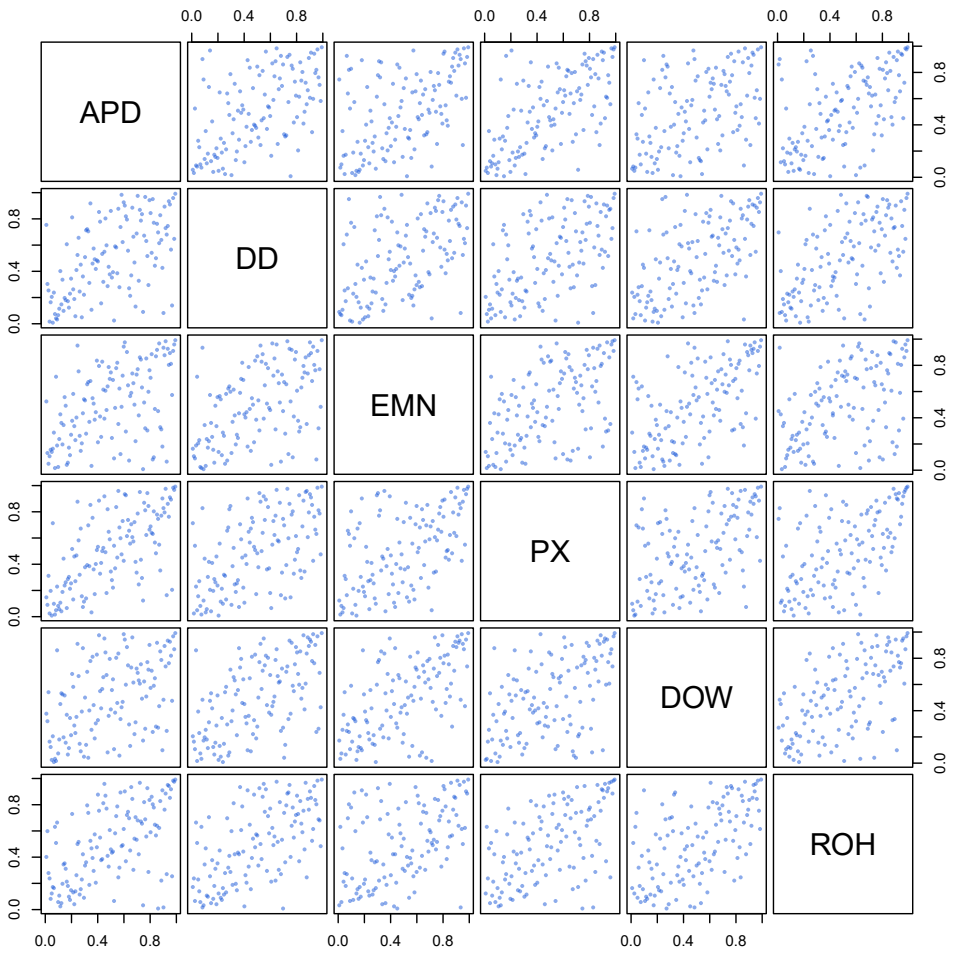

Figura 3.3: Pairplot delle pseudo-osservazioni U_{ij} ($d = 6$). La trasformazione $X \mapsto U$ rimuove l'effetto delle marginali; la dipendenza emerge nella struttura congiunta.

3.3 Stima del parametro della copula

Come discusso precedentemente, l'ipotesi nulla del test di bontà di adattamento proposto è composita: il parametro θ della famiglia candidata non è noto e deve essere stimato a partire dalle pseudo-osservazioni U_{ij} costruite tramite ranghi. [10] Sia C_θ una famiglia parametrica di copule Archimedee e sia c_θ la sua densità. Indicando con $\mathbf{U}_j = (U_{1j}, \ldots, U_{dj})^\top \in (0,1)^d$ la j-esima pseudo-osservazione, lo stimatore canonical maximum likelihood (CML), spesso indicato anche come maximum pseudo-likelihood estimator (MPLE), è definito come

$$\hat{\theta} \;=\; \arg\max_{\theta\in\Theta} \sum_{j=1}^{n} \ln c_\theta(\mathbf{U}_j). \tag{3.3}$$

Nella replica, l'ottimizzazione di (3.3) viene svolta separatamente per ciascuna famiglia candidata (Gumbel–Hougaard, Cook–Johnson, Frank), rispettando i vincoli sul parametro (ad esempio $\theta \geq 1$ per Gumbel–Hougaard e $\theta > 0$ per Cook–Johnson). [10]

Algorithm 2 Stima CML del parametro θ

Require: Pseudo-osservazioni $\{\mathbf{U}_j\}_{j=1}^{n}$, famiglia candidata C_θ, $\theta \in \Theta$

1: Scegli un valore iniziale θ_0 e bounds $[\theta_{\min}, \theta_{\max}] \subseteq \Theta$

2: $\hat{\theta} \leftarrow \arg\max_{\theta\in[\theta_{\min},\theta_{\max}]} \sum_{j=1}^{n} \log c_\theta(\mathbf{U}_j)$

3: **return** $\hat{\theta}$

3.4 Proiezione univariata e Kendall distribution

3.4.1 Variabile proiettata $\widehat{V}_j$

Il test di Savu e Trede (2008) si basa su una riduzione dimensionale della dipendenza: partendo dalle pseudo-osservazioni multivariate $\mathbf{U}_j \in (0,1)^d$, e quindi dallo studio di una copula multivariata, si costruisce una variabile univariata $\widehat{V}_j \in (0,1)$ applicando la copula stimata:

$$\widehat{V}_j \;=\; C_{\hat{\theta}}(\mathbf{U}_j) \;=\; \varphi_{\hat{\theta}}^{-1}\Big(\varphi_{\hat{\theta}}(U_{1j}) + \cdots + \varphi_{\hat{\theta}}(U_{dj})\Big), \qquad j = 1, \ldots, n. \tag{3.4}$$

Questa proiezione rende il GOF trattabile tramite strumenti univariati. Inoltre, nelle simulazioni del paper in questione, la potenza del test non peggiora all'aumentare di d, e questo è dovuto proprio al passaggio dall'ipercubo unitario $d-$dimensionale all'intervallo unitario [10].

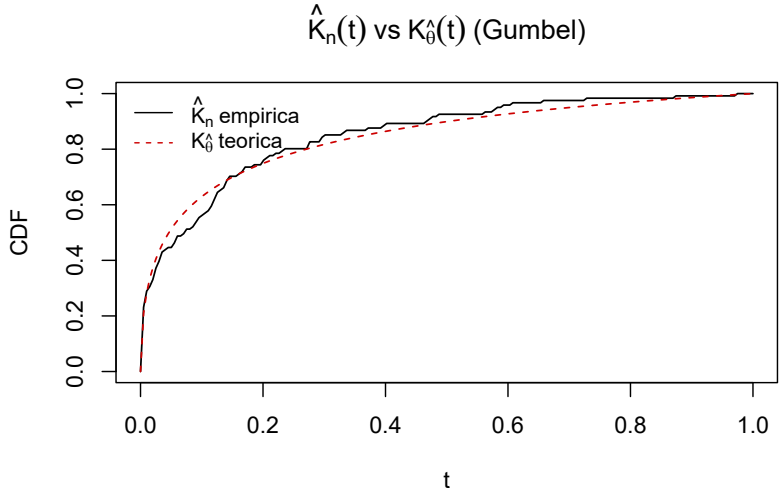

Figura 3.4: Confronto tra $\hat{K}_n(t)$ empirica e $K_{\hat{\theta}}(t)$ teorica (Gumbel). La linea nera rappresenta $\hat{K}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{\widehat{V}_j \leq t\}$, con $\widehat{V}_j = C_{\hat{\theta}}(U_{1j}, \ldots, U_{dj})$. La linea rossa tratteggiata è $K_{\hat{\theta}}(t) = \mathbb{P}(C_{\hat{\theta}}(\mathbf{U}) \leq t)$, ossia la Kendall distribution function indotta dalla copula Gumbel stimata.

3.4.2 Kendall distribution function

Siano $U_1, \ldots, U_d$ d variabili aleatorie con funzione di distribuzione C, si definisce la Kendall distribution function come la funzione di ripartizione della variabile casuale $V = C(U_1, \ldots, U_d)$ [9]:

$$K(t) = \mathbb{P}\Big[C(U_1, \ldots, U_d) \leq t\Big]. \tag{3.5}$$

In particolare, nel caso di copule Archimedee, $K(t)$ ammette una forma esplicita in termini del generatore φ:

$$K(t) \;=\; t + \sum_{i=1}^{d-1} (-1)^i \, \frac{\varphi^i(t)}{i!} \, f_{i-1}(t), \tag{3.6}$$

dove le funzioni ausiliarie sono definite da $f_0(t) = 1/\varphi'(t)$ e, per $i \geq 1$, $f_i(t) = f'_{i-1}(t)/\varphi'(t)$. [10] Nel caso bivariato, ad esempio, la formula sopra si semplifica a

$$K(t) = \mathbb{P}[\, C(U,V) \leq t \,], \qquad t \in [0,1],$$

e, per copule Archimedee bivariate, si ottiene la forma chiusa

$$K(t) = t - \frac{\varphi(t)}{\varphi'(t)}.$$

[12]
In particolare, nel test di Savu e Trede (2008) [10], sotto H_0 e per una famiglia candidata C_θ, la distribuzione di $V = C_\theta(U_1, \ldots, U_d)$ dipende dal parametro θ ed è descritta da K_θ. Non essendo noto, però, il parametro della copula neanche sotto l'ipotesi nulla, si impiega la versione stimata $K_{\hat{\theta}}$ coerente con (3.4) [10]. Di conseguenza, se il modello è corretto, le frequenze osservate nei bin di $\hat{V}$ dovrebbero essere compatibili con le probabilità teoriche

$$p_k = K_{\hat{\theta}}(a_k) - K_{\hat{\theta}}(a_{k-1}),$$

dove a_k sono i punti di taglio del binning.

3.5 Binning su [0,1] e statistica χ^2

3.5.1 Scelta dei bin e conteggi osservati/attesi

Fissata una partizione $0 = a_0 < a_1 < \cdots < a_B = 1$ dell'intervallo [0,1], si definiscono i conteggi osservati

$$O_k \;=\; \#\{\, j :\; \widehat{V}_j \in (a_{k-1}, a_k] \,\}, \qquad k = 1, \ldots, B$$

e i conteggi attesi, che, sotto H_0, si ottengono tramite la Kendall distribution stimata:

$$p_k \;=\; K_{\hat{\theta}}(a_k) - K_{\hat{\theta}}(a_{k-1}), \qquad E_k \;=\; n\, p_k.$$

Nel caso del paper in questione l'intervallo unitario viene suddiviso in $B = 20$ intervalli $[0,0.05], \ldots, (0.95,1]$.

3.5.2 Statistica di test

La procedura di test utilizza una classica statistica di tipo χ^2 basata sul confronto tra conteggi osservati e attesi,

$$T \;=\; \sum_{k=1}^{B} \frac{(O_k - E_k)^2}{E_k}. \tag{3.7}$$

Nella letteratura sui GOF compaiono frequentemente statistiche χ^2-based. Però in questo caso specifico, la distribuzione di T sotto H_0 non segue una distribuzione χ^2, in quanto lo stimatore $\hat{\theta}$ non rientra solo negli attesi ma anche negli osservati, infatti la trasformazione $\widehat{V}_j = C_{\hat{\theta}}(\mathbf{U}_j)$ dipende da $\hat{\theta}$ e E_k dipende da $K_{\hat{\theta}}$. Per questo motivo l'approssimazione della distribuzione di T è ottenuta via bootstrap parametrico [10].

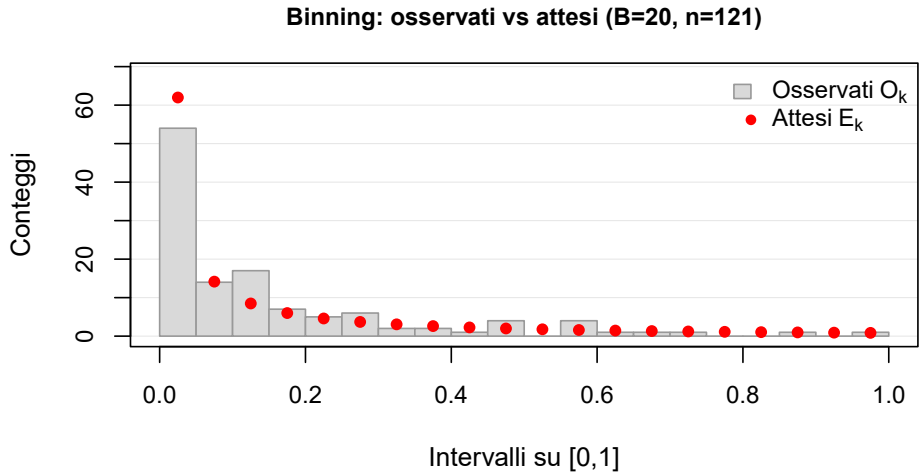

Figura 3.5: Binning su [0,1] con $B = 20$ e $n = 121$: conteggi osservati O_k (barre) e attesi E_k (punti) per ciascun intervallo uniforme $(a_{k-1}, a_k]$. I punti E_k sono tracciati ai midpoint dei bin. L'esempio è calcolato usando una specifica famiglia candidata (Gumbel), ma la costruzione è identica per qualunque famiglia candidata.

Algorithm 3 Calcolo della statistica T

Require: Pseudo-osservazioni $\{\mathbf{U}_j\}_{j=1}^n$, copula stimata $C_{\hat{\theta}}$, breakpoints $0 = a_0 < \cdots < a_B = 1$

for $j = 1$ **to** n **do**
  $\widehat{V}_j \leftarrow C_{\hat{\theta}}(\mathbf{U}_j)$
end
for $k = 1$ **to** B **do**
  $O_k \leftarrow \#\{j : \widehat{V}_j \in (a_{k-1}, a_k]\}$
  $p_k \leftarrow K_{\hat{\theta}}(a_k) - K_{\hat{\theta}}(a_{k-1})$
  $E_k \leftarrow n\, p_k$
end
$T \leftarrow \sum_{k=1}^{B} (O_k - E_k)^2 / E_k$
return T, $\{O_k\}$, $\{E_k\}$

3.6 Bootstrap parametrico sotto H_0

3.6.1 Motivazione

Poichè, come già detto, nel test proposto da Savu e Trede (2008) [10] il parametro θ è stimato (anche sotto l'ipotesi nulla non si conosce la copula esatta) e la statistica

test dipende dalla proiezione sull'intervallo unitario attraverso $\widehat{V}_j = C_{\hat{\theta}}(\mathbf{U}_j)$, la distribuzione di T sotto H_0 viene ottenuta con un bootstrap parametrico e i valori critici sono calcolati empiricamente (in particolare, nel paper, con 5000 repliche) [10].

3.6.2 Algoritmo bootstrap e p-value

Si generano un numero J di repliche $\mathbf{U}_1^{*(b)}, \ldots, \mathbf{U}_n^{*(b)}$ dalla copula stimata $C_{\hat{\theta}}$, per ogni campione si ristima il parametro $\hat{\theta}^{*(b)}$ attraverso CML e si ricalcola $T^{*(b)}$ replicando la pipeline completa (stima $\rightarrow$ proiezione $\rightarrow$ binning $\rightarrow$ confronto osservati/attesi). Il p-value è ottenuto come proporzione di repliche che eccedono la statistica osservata, oppure si calcola direttamente il quantile empirico $c_{1-\alpha}$ della distribuzione bootstrap [10].

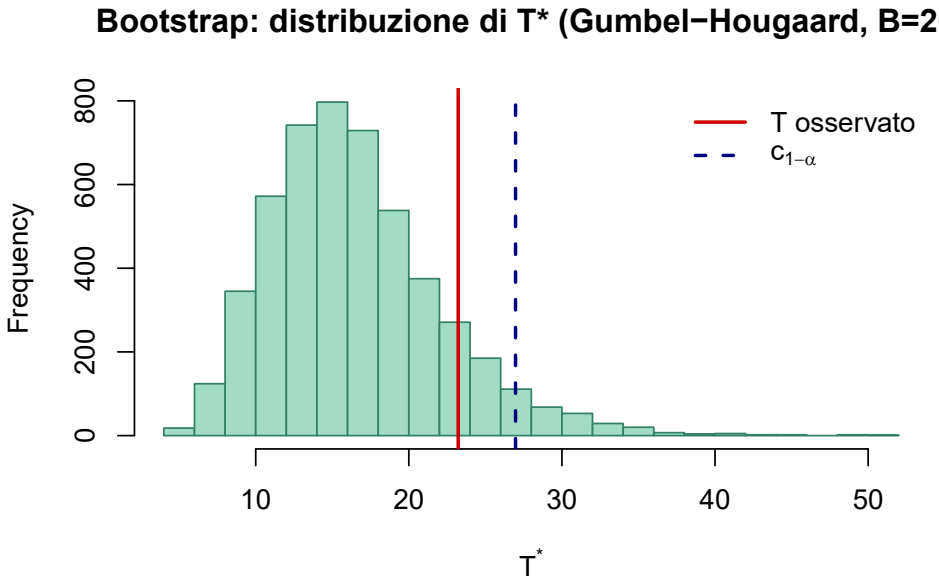

Figura 3.6: Bootstrap parametrico: distribuzione di $\{T^{*(b)}\}_{b=1}^{J}$. La linea continua indica T osservato, la linea tratteggiata il quantile critico $c_{1-\alpha}$.

Algorithm 4 Bootstrap parametrico per valori critici e p-value

Require: Pseudo-osservazioni $\{\mathbf{U}_j\}_{j=1}^n$, famiglia C_θ, livello α, breakpoints $\{a_k\}$, numero repliche J

1: Stima $\hat{\theta}$ su $\{\mathbf{U}_j\}$ (Alg. 2)
2: Calcola T sui dati (Alg. 3)
3: **for** $b = 1$ **to** J **do**
4: Genera $\{\mathbf{U}_j^{*(b)}\}_{j=1}^n \sim C_{\hat{\theta}}$
5: Stima $\hat{\theta}^{*(b)}$ a partire da $\{\mathbf{U}_j^{*(b)}\}$
6: Calcola $T^{*(b)}$ replicando Alg. 3 a partire da $\{\mathbf{U}_j^{*(b)}\}$ e $C_{\hat{\theta}^{*(b)}}$
7: **end**
8: p-value $\leftarrow \frac{1+\sum_{b=1}^J \mathbb{I}\{T^{*(b)} \geq T\}}{J+1}$
9: $c_{1-\alpha} \leftarrow$ quantile empirico $(1-\alpha)$ di $\{T^{*(b)}\}_{b=1}^J$
10: **return** p-value, $c_{1-\alpha}$, decisione (rifiuta se $p < \alpha$ oppure $T > c_{1-\alpha}$)

3.7 Aspetti implementativi e riproducibilità della replica

La procedura di replica è organizzata in tre script principali per l'applicazione empirica (pipeline sui dati reali) e in ulteriori script dedicati alla replica dei risultati Monte Carlo del paper (dimensione empirica e potenza). In particolare, i tre script dell'applicazione empirica coprono:

- import e preprocessing dei dati, costruzione delle pseudo-osservazioni;
- stima CML del parametro per le famiglie candidate;
- calcolo di $\widehat{V}$, binning su [0,1], statistica T e bootstrap parametrico.

Questa separazione ricalca i passaggi metodologici del test e facilita la riproducibilità su dataset alternativi.

Le scelte minime adottate sono:

- fissare i semi random per simulazioni e bootstrap;
- salvare gli output principali come pseudo-osservazioni, $\hat{\theta}$ per ciascuna famiglia candidata, statistica T, p-value e valore critico;
- riportare esplicitamente n, d, B, J (con $n = 121$, $d = 6$, $B = 20$ e $J = 5000$ come nel paper).

Gli script Monte Carlo replicano i risultati del paper su dimensione e potenza mantenendo B, d e n uguali al paper mentre i parametri di simulazione R e J sono ridotti per motivi computazionali.

Capitolo 4

Studio Monte Carlo: dimensione empirica e potenza empirica del test

Questo capitolo ha come scopo quello di verificare, tramite simulazione Monte Carlo, che l'implementazione del test di bontà di adattamento (GOF) descritta nel Capitolo 3 riproduca le principali evidenze riportate da Savu e Trede (2008) [10]. In particolare:

- la Sez. 4.2 studia la **dimensione empirica** del test, cioè la probabilità di rifiuto sotto H_0, che idealmente dovrebbe essere vicina al livello nominale;
- la Sez. 4.3 studia la **potenza empirica**, cioè la probabilità di rifiuto quando i dati sono generati da una copula alternativa.

4.1 Impostazione dello studio di simulazione

L'idea è replicare in ambiente controllato la stessa pipeline utilizzata nell'applicazione empirica: per ogni replica si genera un campione i.i.d. da una copula nota, ma il test viene applicato come in un problema realistico, cioè con famiglia specificata e parametro ignoto. Di conseguenza, in ogni replica si (ri)stima $\hat{\theta}$ con CML e si ottiene il valore critico tramite bootstrap parametrico basato su $\hat{\theta}$ [10].

4.1.1 Fattori dello studio e scelta della dipendenza

Seguendo Savu e Trede [10], le famiglie Archimedee considerate come ipotesi nulla sono: Cook–Johnson (Clayton), Frank e Gumbel–Hougaard e viene valutata la

dimensione del test al livello $\alpha = 0.05$. Per ciascuna famiglia si valutano diverse combinazioni di dimensione d del vettore e numerosità campionaria n:

$$d \in \{2,3,5\}, \qquad n \in \{100,200,500\}.$$

Per rendere confrontabile il grado di dipendenza tra famiglie diverse, il parametro θ è scelto in modo tale che Kendall's τ sia fissata a $\tau = 0.3$ in tutti gli scenari, come nel paper [10].[1]

4.1.2 Binning e bootstrap: scelta di B, J e R

Coerentemente con il disegno sperimentale del paper, l'intervallo [0,1] viene suddiviso in $B = 20$ bin uniformi, $[0,0.05], (0.05,0.10], \ldots, (0.95,1]$.

Compromesso computazionale. Il costo computazionale dello studio è dominato dalla struttura annidata della procedura: per ogni replica Monte Carlo (R) è necessario eseguire un bootstrap interno (J) per stimare la distribuzione della statistica del test sotto H_0. Nel paper, per ogni coppia (d, n), le configurazioni adottate sono:

- **Dimensione empirica:** $R = 5000$ repliche Monte Carlo e $J = 5000$ repliche bootstrap [10];
- **Potenza empirica:** $R = 1000$ repliche Monte Carlo e $J = 5000$ repliche bootstrap [10].

In questa tesi, per rendere praticabile l'esperimento su tutte le combinazioni (d, n) e sulle tre famiglie considerate, si è adottata la seguente configurazione:

- **Dimensione empirica (Sez. 4.2):** $R = 1000$, $J = 500$;
- **Potenza empirica (Sez. 4.3):** $R = 500$, $J = 350$.

Queste scelte aumentano la variabilità simulativa rispetto al paper, ma permettono comunque di verificare l'obiettivo principale della replica: (1) che la dimensione resti prossima al livello nominale e (2) che emergano i medesimi pattern qualitativi (ad esempio comportamenti conservativi in scenari specifici).

[1] Operativamente, θ viene ricavato invertendo numericamente la relazione $\tau(\theta)$ della famiglia considerata.

4.1.3 Variabilità simulativa con R e J ridotti

Le quantità riportate in Savu e Trede (2008) [10] sono frequenze di rifiuto, quindi stime di probabilità ottenute come proporzioni su R repliche. Con R più piccolo, è naturale osservare oscillazioni attorno al valore teorico (per la dimensione, attorno a α) anche quando l'implementazione è corretta. Inoltre, nel GOF considerato il valore critico è stimato via bootstrap a ogni replica: ridurre J introduce ulteriore rumore perché la coda della distribuzione bootstrap (vista attraverso il quantile critico) è stimata con meno repliche.

In sintesi, rispetto al paper ci si aspetta:

- minore precisione nelle stime di dimensione/potenza (variabilità Monte Carlo dovuta a R);
- maggiore variabilità dei quantili critici (variabilità bootstrap dovuta a J);
- una replica comunque informativa a livello qualitativo (pattern e confronti tra scenari).

4.2 Dimensione empirica

In questa sezione vengono riportati i risultati principali relativi alla dimensione empirica del test al livello $\alpha = 0.05$. In ciascuno scenario, la famiglia testata (quindi l'ipotesi nulla) coincide con la famiglia usata per generare i dati. Di conseguenza, la probabilità di rifiuto dovrebbe essere prossima al livello nominale $\alpha = 0.05$.

4.2.1 Procedura di simulazione per la size

Per ogni combinazione (d, n) e per ciascuna famiglia nulla, lo schema è il seguente: si genera un campione i.i.d. dalla copula C_{θ_0} (con $\tau(\theta_0) = 0.3$), si applica il test GOF completo (stima CML e bootstrap), e si registra se H_0 viene rifiutata al livello α (Algoritmo 5). La dimensione empirica è quindi la proporzione di rifiuti su R repliche valide.

4.2.2 Risultati complessivi e confronto con Savu–Trede

La Tabella 4.1 riporta le frequenze di rifiuto osservate nella replica per tutte le combinazioni (d, n) e con $R = 1000$, $J = 500$. Per facilitare il confronto, tra parentesi è riportato il valore corrispondente riportato nel paper [10].
Nel complesso, le frequenze di rifiuto ottenute nella replica sono dell'ordine del 5% e quindi coerenti con il comportamento atteso sotto H_0. Le differenze rispetto ai valori del paper sono contenute e plausibili considerando che:

Algorithm 5 Monte Carlo per la stima della dimensione empirica

Require: Famiglia nulla C_{θ_0}, livello α, dimensione d, numerosità n, numero repliche Monte Carlo R, bin B, bootstrap J

Scegli θ_0 in modo che $\tau(\theta_0) = 0.3$
for $r = 1$ **to** R **do**
  Genera $U_1^{(r)}, \dots, U_n^{(r)} \overset{iid}{\sim} C_{\theta_0}$ in $(0,1)^d$
  Stima $\hat{\theta}$ (CML) sul campione simulato
  Calcola la statistica osservata $T_{\text{obs}}^{(r)}$
  for $j = 1$ **to** J **do**
    Genera $U_1^{*(r,j)}, \dots, U_n^{*(r,j)} \sim C_{\hat{\theta}}$
    Stima $\hat{\theta}^{*(r,j)}$ e calcola $T^{*(r,j)}$
  end
  Determina $c_{1-\alpha}^{(r)}$ come quantile $1-\alpha$ della $\{T^{*(r,j)}\}_{j=1}^{J}$
  Registra $I^{(r)} = \mathbb{I}\{T_{\text{obs}}^{(r)} > c_{1-\alpha}^{(r)}\}$
end
return $\widehat{\text{size}} = \frac{1}{R}\sum_{r=1}^{R} I^{(r)}$

Tabella 4.1: Dimensione empirica del test GOF al livello $\alpha = 0.05$ con binning $B = 20$ e $\tau = 0.3$. Valori principali: replica con $R = 1000$, $J = 500$; tra parentesi: risultati del paper [10].

Copula sotto H_0	d	$n = 100$	$n = 200$	$n = 500$
Cook–Johnson	2	0.051 (0.052)	0.055 (0.046)	0.058 (0.049)
	3	0.060 (0.048)	0.053 (0.048)	0.046 (0.052)
	5	0.047 (0.053)	0.057 (0.052)	0.058 (0.052)
Frank	2	0.045 (0.054)	0.045 (0.047)	0.056 (0.052)
	3	0.057 (0.052)	0.053 (0.046)	0.057 (0.048)
	5	0.054 (0.051)	0.044 (0.051)	0.059 (0.048)
Gumbel–Hougaard	2	0.053 (0.058)	0.051 (0.048)	0.055 (0.056)
	3	0.057 (0.046)	0.058 (0.049)	0.058 (0.051)
	5	0.053 (0.048)	0.030 (0.029)	0.027 (0.024)

- la stima della dimensione è una proporzione su un numero finito di repliche Monte Carlo (qui $R = 1000$, nel paper $R = 5000$);
- i valori critici sono stimati via bootstrap e qui viene introdotta ulteriore variabilità (qui $J = 500$, nel paper $J = 5000$).

Il pattern più caratteristico evidenziato nel paper, ossia una dimensione conservativa per Gumbel–Hougaard nel caso $d = 5$, soprattutto per $n = 200$ e $n = 500$, emerge in modo netto anche nella replica.

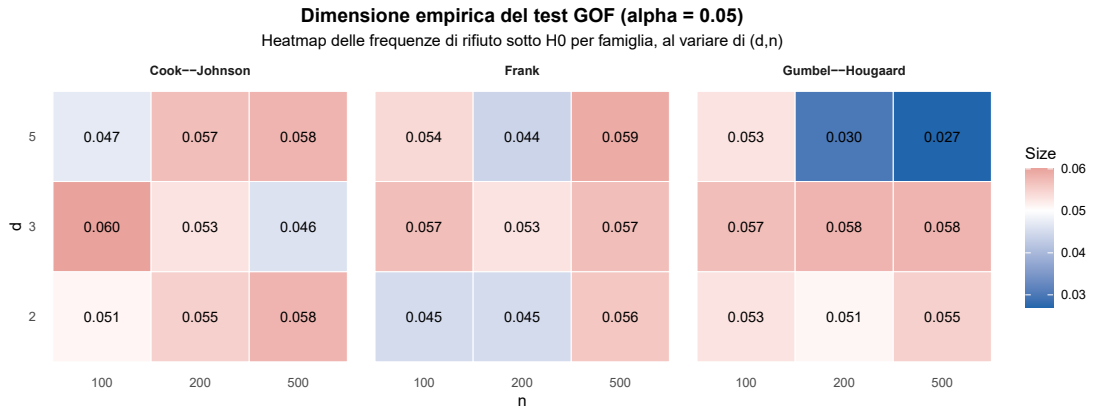

Figura 4.1: Heatmap della dimensione empirica stimata: per ciascuna famiglia sotto H_0, i valori numerici sono riportati nelle celle e il colore (centrato su $\alpha = 0.05$) evidenzia deviazioni dalla dimensione nominale. La figura riassume come la frequenza di rifiuto varia al variare di d e n.

4.3 Potenza empirica

In questa sezione si studia la potenza empirica del test, ossia la probabilità di rifiutare H_0 quando l'ipotesi nulla è falsa. Lo schema segue Savu e Trede (2008) [10]: i dati sono generati da una copula alternativa (DGP: Processo di generazione dei dati), ma il test viene applicato assumendo come ipotesi nulla una delle famiglie Archimedee a un parametro (Cook–Johnson, Frank, Gumbel–Hougaard), stimandone il parametro via CML e ottenendo i quantili critici tramite bootstrap parametrico.

4.3.1 Scenari alternativi (DGP) e impostazione

Come nel paper, oltre alle alternative date da una famiglia nulla Archimedea testata su dati generati da un'altra famiglia Archimedea, si considerano due alternative ellittiche: una copula Gaussiana e una copula t con $\nu = 2$ gradi di libertà (indicata come t_2). In tutti gli scenari il livello nominale è fissato a $\alpha = 0.05$, il binning è uniforme con $B = 20$ e l'intensità della dipendenza è mantenuta comparabile ponendo $\tau = 0.3$ (tramite scelta del parametro della copula o, per le copule ellittiche, tramite calibrazione della correlazione coerente con τ).[2]

[2]Per la copula Gaussiana e la copula t il coefficiente di correlazione è $\rho = sin(\tau/2) = 0.4540$ [10]

Algorithm 6 Monte Carlo per la stima della potenza empirica

Require: Copula alternativa C_{alt} (DGP), famiglia nulla C_θ, livello α, dimensione d, numerosità n, numero repliche Monte Carlo R, bin B, bootstrap J

for $r = 1$ **to** R **do**
  Genera $U_1^{(r)}, \dots, U_n^{(r)} \overset{iid}{\sim} C_{\text{alt}}$ in $(0,1)^d$
  Stima $\hat{\theta}^{(r)}$ (CML) assumendo la famiglia nulla C_θ
  Calcola la statistica osservata $T_{\text{obs}}^{(r)}$
  for $j = 1$ **to** J **do**
    Genera $U_1^{*(r,j)}, \dots, U_n^{*(r,j)} \sim C_{\hat{\theta}^{(r)}}$ (sotto H_0)
    Stima $\hat{\theta}^{*(r,j)}$ e calcola $T^{*(r,j)}$
  end
  Determina $c_{1-\alpha}^{(r)}$ come quantile $1-\alpha$ della $\{T^{*(r,j)}\}_{j=1}^{J}$
  Registra $I^{(r)} = \mathbb{I}\{T_{\text{obs}}^{(r)} > c_{1-\alpha}^{(r)}\}$
end
return $\widehat{\text{power}} = \frac{1}{R}\sum_{r=1}^{R} I^{(r)}$

4.3.2 Procedura Monte Carlo per la potenza

Per ogni combinazione (d, n) e per ogni coppia (DGP alternativa, famiglia nulla), si ripete la seguente procedura: si genera un campione i.i.d. dalla copula alternativa, si applica il GOF assumendo la famiglia nulla, e si registra se l'ipotesi nulla viene rifiutata. La potenza empirica è la proporzione di rifiuti su R repliche valide.
Coerentemente con quanto discusso in Sez. 4.1.2, per rendere computazionalmente praticabile lo studio su tutte le combinazioni, in questa tesi si utilizza $R = 500$ e un bootstrap interno di ampiezza $J = 350$ (mentre nel paper $R = 1000$ e $J = 5000$). Questa riduzione aumenta la variabilità delle stime (sia per via Monte Carlo sia per la stima bootstrap del quantile critico), ma consente comunque di verificare i pattern qualitativi registrati nel paper [10].

4.3.3 Risultati e confronto con Savu–Trede (2008)

Le Tabelle 4.2–4.4 riportano le stime della potenza empirica per le tre ipotesi nulle. In ciascuna cella è riportato il valore della replica (con $R = 500$, $J = 350$) e, tra parentesi, il corrispondente valore del paper [10]. Nel complesso, i risultati replicano bene l'andamento atteso e cioè che la potenza cresce con n. I risultati mostrano però anche che il test può avere una potenza relativamente elevata già con campioni piccoli per alcune combinazioni. Inoltre, in accordo con quanto osservato dagli autori, la potenza non peggiora al crescere della dimensione d; al contrario, tende spesso ad aumentare con d e questo è dovuto allo step di proiezione dall'ipercubo unitario d-dimensionale all'intervallo [0,1], che permette di individuare meglio i casi

Tabella 4.2: Potenza empirica del test GOF (livello $\alpha = 0.05$) quando l'ipotesi nulla è Cook–Johnson (Clayton). Valori: replica ($R = 500$, $J = 350$) e, tra parentesi, valori del paper.

DGP	$d = 2$			$d = 3$			$d = 5$		
	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
Frank	0.102 (0.14)	0.182 (0.21)	0.454 (0.52)	0.400 (0.43)	0.728 (0.76)	0.984 (0.99)	0.810 (0.81)	0.988 (0.99)	1.000 (1.00)
Gumbel–Hougaard	0.410 (0.46)	0.718 (0.79)	0.992 (1.00)	0.950 (0.96)	0.996 (1.00)	1.000 (1.00)	0.996 (1.00)	1.000 (1.00)	1.000 (1.00)
Gaussian	0.162 (0.18)	0.235 (0.27)	0.620 (0.67)	0.536 (0.56)	0.808 (0.80)	0.994 (0.98)	0.776 (0.79)	0.958 (0.96)	1.000 (1.00)
t_2	0.296 (0.34)	0.544 (0.60)	0.950 (0.96)	0.766 (0.78)	0.958 (0.95)	1.000 (1.00)	0.880 (0.87)	0.976 (0.97)	1.000 (1.00)

Tabella 4.3: Potenza empirica del test GOF (livello $\alpha = 0.05$) quando l'ipotesi nulla è Frank. Valori: replica ($R = 500$, $J = 350$) e, tra parentesi, valori del paper.

DGP	$d = 2$			$d = 3$			$d = 5$		
	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
Cook–Johnson	0.076 (0.07)	0.134 (0.10)	0.224 (0.23)	0.078 (0.08)	0.152 (0.16)	0.672 (0.70)	0.086 (0.09)	0.306 (0.30)	0.964 (0.96)
Gumbel–Hougaard	0.128 (0.14)	0.276 (0.28)	0.678 (0.72)	0.654 (0.64)	0.860 (0.87)	0.998 (1.00)	0.850 (0.87)	0.984 (0.98)	1.000 (1.00)
Gaussian	0.066 (0.07)	0.082 (0.08)	0.118 (0.13)	0.176 (0.16)	0.260 (0.25)	0.394 (0.39)	0.222 (0.23)	0.342 (0.34)	0.574 (0.59)
t_2	0.126 (0.14)	0.326 (0.27)	0.712 (0.73)	0.524 (0.50)	0.752 (0.75)	0.984 (0.98)	0.576 (0.58)	0.828 (0.81)	0.992 (0.99)

di violazione di H_0 al crescere della dimensione. Per quanto riguarda l'ipotesi nulla Gumbel–Hougard in dimensione $d = 5$, la dimensione empirica risulta conservativa (4.2) e questo comportamento tende a riflettersi anche sulla potenza contro alcune alternative come Gaussiana o t_2. In generale, per H_0 Gumbel–Hougaard la potenza (già bassa) può ridursi passando da $d = 3$ a $d = 5$ come evidenziato anche nei risultati di Savu e Trede (2008) [10].

Tabella 4.4: Potenza empirica del test GOF (livello $\alpha = 0.05$) quando l'ipotesi nulla è Gumbel–Hougaard. Valori: replica ($R = 500$, $J = 350$) e, tra parentesi, valori del paper.

DGP	$d = 2$			$d = 3$			$d = 5$		
	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
Cook–Johnson	0.122 (0.14)	0.242 (0.25)	0.708 (0.72)	0.274 (0.27)	0.606 (0.61)	1.000 (1.00)	0.254 (0.24)	0.650 (0.70)	1.000 (1.00)
Frank	0.108 (0.10)	0.154 (0.15)	0.384 (0.37)	0.140 (0.14)	0.276 (0.27)	0.780 (0.79)	0.125 (0.11)	0.251 (0.24)	0.780 (0.83)
Gaussian	0.084 (0.08)	0.102 (0.10)	0.146 (0.14)	0.092 (0.10)	0.140 (0.14)	0.334 (0.36)	0.049 (0.04)	0.071 (0.06)	0.234 (0.22)
t_2	0.068 (0.06)	0.055 (0.07)	0.086 (0.09)	0.074 (0.08)	0.096 (0.11)	0.256 (0.23)	0.039 (0.04)	0.055 (0.05)	0.169 (0.13)

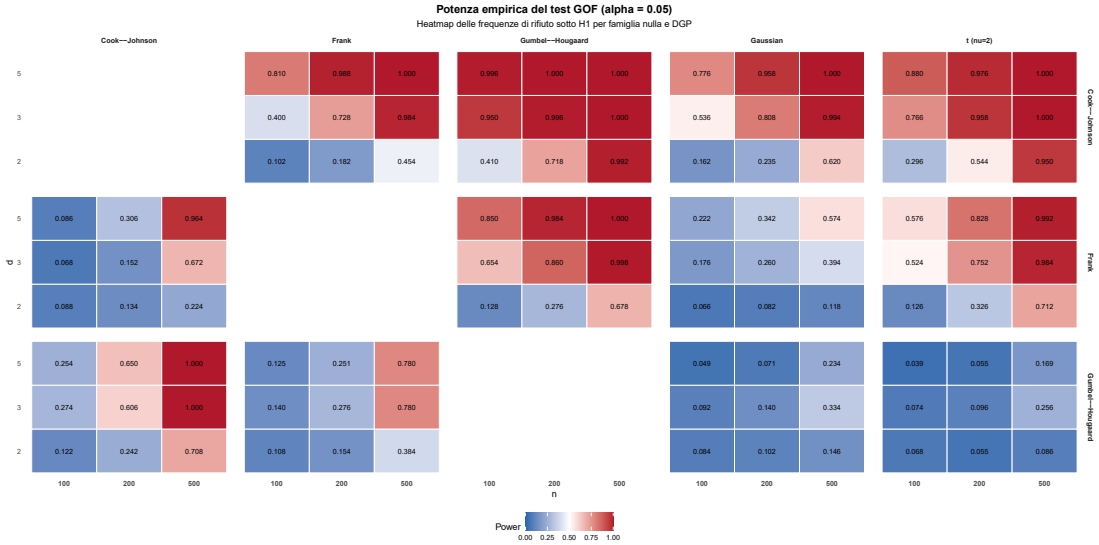

Figura 4.2: Heatmap della potenza empirica ($\alpha = 0.05$) per i diversi scenari di alternativa (DGP) e ipotesi nulla. I valori nelle celle sono le frequenze di rifiuto stimate. Il colore riassume la potenza (valori prossimi a 1 indicano alta capacità discriminante).

Capitolo 5

Applicazione empirica del GOF: rendimenti azionari e risultati

L'obiettivo di questo capitolo è applicare ai dati reali i test di bontà di adattamento descritti nei capitoli precedenti, replicando l'analisi empirica di Savu e Trede (2008) [10]. In particolare, lo studio originale analizza la struttura di dipendenza tra i rendimenti mensili di sei titoli appartenenti al settore chimico dell'indice S&P 500 nel periodo 1996–2006, utilizzando dati di mercato estratti da Datastream.
Poichè Datastream è un database professionale a licenza e poichè la replicazione dei risultati empirici avviene a distanza di oltre quindici anni dalla pubblicazione del paper, si sono presentate diverse sfide, legate principalmente alla natura dinamica dei mercati finanziari: molte delle aziende analizzate hanno subito fusioni, acquisizioni o ristrutturazioni che hanno reso le serie storiche originali difficilmente reperibili sui comuni provider finanziari moderni [16]. Si è reso pertanto necessario ricostruire un dataset coerente con quello originale attraverso fonti alternative e una procedura di pulizia e verifica più articolata.

5.1 Dati e preprocessing

Il dataset oggetto di analisi comprende le serie storiche dei prezzi di chiusura mensili per sei aziende statunitensi del settore chimico. Il periodo di osservazione si estende dal 1° gennaio 1996 al 1° febbraio 2006. In totale sono disponibili 122 osservazioni mensili sui prezzi; a partire da queste, si calcolano $n = 121$ rendimenti mensili, in linea con l'impostazione dello studio replicato.
Le aziende considerate sono:

1. Air Products & Chemicals (APD)
2. Dow Chemicals (DOW)
3. DuPont (DD)
4. Eastman Chemical (EMN)
5. Praxair (PX)
6. Rohm & Haas (ROH)

5.1.1 Frequenza mensile e convenzione sulle date

I dati sono stati scaricati selezionando la frequenza Monthly e l'intervallo 1996–2006. In questo modo, le fonti utilizzate restituiscono una sola osservazione per mese, riferita alla chiusura dell'ultimo giorno di mercato del mese. Le date sono trattate come indice mensile, e i rendimenti sono calcolati tra osservazioni mensili consecutive.

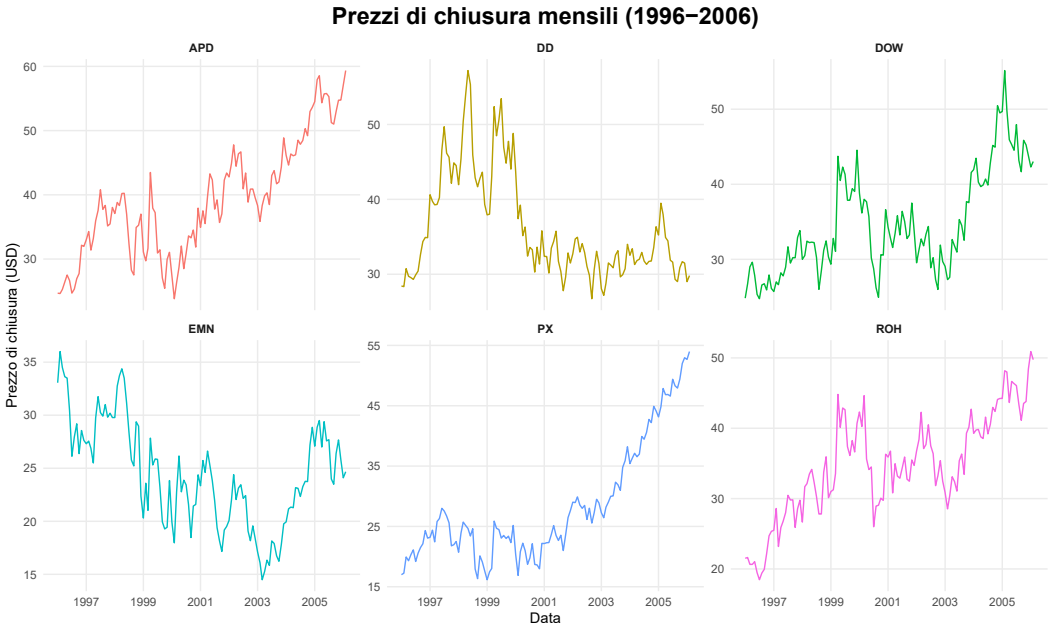

Figura 5.1: Prezzi di chiusura mensili nel periodo Gennaio 1996– Febbraio 2006 per ciascun titolo.

5.1.2 Problematiche nel reperimento dei dati storici

Il reperimento delle serie storiche per il periodo 1996-2006 ha richiesto l'integrazione di diverse fonti dati eterogenee. L'utilizzo di un'unica fonte standardizzata (come Yahoo Finance) si è rivelato inefficace a causa di eventi societari che hanno alterato la continuità dei ticker originali.
In una prima fase esplorativa, i dati sono stati estratti da Yahoo Finance ([22]) e Investing.com ([23]). Mentre per titoli come **APD** ed **EMN** le serie storiche estratte da Yahoo Finance risultavano coerenti con le statistiche descrittive del paper originale (media e deviazione standard), per altri titoli sono emerse criticità significative:

- **Il caso DuPont (DD) e Dow Chemicals (DOW):** Dow Chemical e DuPont hanno completato la fusione nel 2017 dando origine a DowDuPont (ticker DWDP), e i titoli Dow e DuPont hanno cessato di essere negoziati separatamente [17]. In seguito, Dow Inc. è tornata a essere una società indipendente e ha iniziato a negoziare con ticker **DOW** nel 2019 [18]. Analogamente, DuPont de Nemours ha iniziato a negoziare come entità indipendente con ticker **DD** nel 2019 [19]. Di conseguenza, sui provider che associano i ticker ai titoli attualmente negoziati, **DOW** e **DD** spesso non presentano una storia coerente per il periodo 1996–2006. In particolare, durante l'analisi preliminare, è stato riscontrato che le serie storiche scaricate per **DD** e **DOW** presentavano una correlazione di Kendall (τ) prossima a 1, un valore anomalo che suggeriva l'identicità dei dati. Per risolvere il problema, è stato pertanto necessario reperire un dataset storico specifico per la "vecchia" DuPont (pre-fusione) da archivi alternativi (Investing.com [23]), verificando che le statistiche di base (Media ≈ 0.36, SD ≈ 7.27) coincidessero con quelle riportate da Savu e Trede (2008) [10].

- **Il caso Rohm & Haas (ROH):** L'azienda è stata acquisita da Dow nel 2009 e ha cessato di essere una società quotata indipendente [21]. I dati storici sono stati recuperati tramite fonti alternative, in particolare ADVFN ([24]). Tuttavia, una prima analisi dei rendimenti ha evidenziato un valore anomalo pari a circa il 180% in un singolo mese. Un rendimento di tale entità per una società chimica stabile è indice quasi certo di un errore nei dati o di un evento tecnico non rettificato.

- **Praxair (PX):** La business combination tra Praxair e Linde AG è stata completata nel 2018 e le azioni Praxair sono state delistate [20]. Pertanto anche Praxair ha richiesto l'utilizzo di dati storicizzati provenienti da Investing.com ([23]) per coprire correttamente la finestra temporale richiesta.

5.1.3 Tracciabilità delle fonti

La Tabella 5.1 sintetizza, per ciascun titolo, la fonte dati effettivamente utilizzata.

Tabella 5.1: Tracciabilità del dataset replicato (Gennaio 1996– Febbraio 2006): fonte e note operative.

Ticker	Fonte	Note
APD	Yahoo Finance	Copertura completa nel periodo.
EMN	Yahoo Finance	Copertura completa nel periodo.
PX	Investing	Ticker non più negoziato dopo 2018: necessario archivio storico.
DOW	Investing	Ticker riattivato nel 2019: necessario archivio storico.
DD	Investing	Ticker riattivato nel 2019: sostituito con serie coerente 1996–2006.
ROH	ADVFN	Delist dopo acquisizione: necessario archivio storico.

5.1.4 Validazione dei dati e rilevazione delle anomalie

La ricostruzione del dataset non si è limitata alla raccolta delle serie, ma ha previsto controlli quantitativi per individuare inconsistenze. I criteri principali sono stati:

1. confronto di media e deviazione standard dei rendimenti con i valori riportati nello studio replicato;
2. controllo della matrice di dipendenza tramite τ di Kendall (Sez. 5.2);
3. ispezione di valori estremi e salti anomali nei rendimenti.

Due segnali hanno portato a intervenire sul reperimento dei dati e sulla pulizia:

- **DD quasi identico a DOW in una prima estrazione.** Come già detto, in una versione preliminare del dataset, la serie etichettata come **DD** risultava praticamente coincidente con **DOW**, con $\tau(\mathbf{DD}, \mathbf{DOW})$ prossima a 1 e statistiche descrittive sovrapponibili. Questo ha motivato la sostituzione della serie **DD** con un archivio storico specifico per la DuPont precedente alla fusione.
- **ROH con rendimento anomalo ($\sim 180\%$).** Particolare attenzione è stata dedicata alla correzione della serie storica di Rohm & Haas **(ROH)**. L'analisi dell'anomalia del 180% ha permesso di identificare la presenza di uno stock split (frazionamento azionario) avvenuto nell'Agosto 1998, non rettificato nel dataset grezzo [25]. Nello specifico, il prezzo è passato da circa 9.94 USD a 27.8 USD senza una giustificazione di mercato, suggerendo un rapporto di split inverso e una mancata rettifica dei prezzi antecedenti. E' stata pertanto

implementata una procedura di correzione che ha individuato la data del salto e applicato un fattore di rettifica ($ratio \approx 3$) a tutti i prezzi antecedenti al 1° Settembre 1998. Dopo questa correzione, la deviazione standard e la media dei rendimenti di **ROH** si sono allineate ai valori attesi dal paper.

Figura 5.2: ROH: prezzi mensili grezzi (sinistra) e corretti (destra) con indicazione della data del salto (09-1998). La correzione è applicata ai prezzi antecedenti tramite fattore di rettifica stimato.

Algorithm 7 ROH: Schema di correzione retrospettiva per split non rettificato

Require: Prezzi mensili P_t, $t = 1, \ldots, 122$
1: Calcola il rendimento $R_t = (P_t/P_{t-1} - 1) \cdot 100$ per $t = 2, \ldots, 122$
2: Identifica il mese anomalo t_0 con $|R_t| > 50\%$
3: Definisci $r = P_{t_0}/P_{t_0-1}$ (fattore di rettifica)
4: **for** $t = 1$ **to** $t_0 - 1$ **do**
5: $\quad P_t \leftarrow r \cdot P_t$
6: **end**
7: **return** Serie corretta $\{P_t\}$

5.1.5 Calcolo dei rendimenti mensili

A partire dai prezzi di chiusura mensili $P_{i,t}$, sono stati calcolati i rendimenti mensili percentuali per ogni titolo i e mese t come:

$$R_{i,t} = \left(\frac{P_{i,t}}{P_{i,t-1}} - 1 \right) \cdot 100, \tag{5.1}$$

con $t = 2, \ldots, 122$.

Versione	Media	SD	Min	Max
Grezzi (pre-correzione)	2.56	18.4	-24.6	180.0
Corretti (post-correzione)	1.07	8.71	-24.6	33.5

Tabella 5.2: ROH: statistiche descrittive dei rendimenti mensili prima e dopo la correzione dello split (1996–2006).

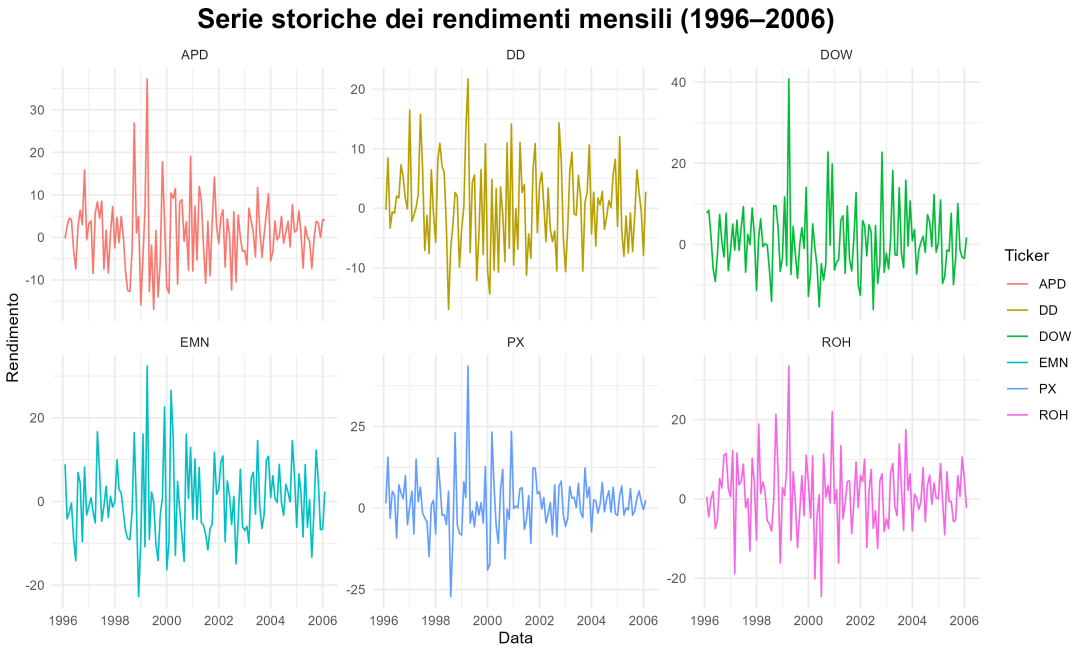

Figura 5.3: Serie temporali dei rendimenti percentuali mensili per i sei titoli (1996–2006) dopo la correzione di ROH. Le date rappresentano l'indice mensile associato alle osservazioni di chiusura.

5.1.6 Statistiche descrittive e confronto con lo studio originale

Dopo le operazioni di raccolta dati mirate e la correzione su **ROH**, le statistiche descrittive del dataset replicato risultano nel complesso allineate a quelle riportate nello studio originale, come sintetizzato in Tabella 5.3. Le discrepanze residue sono plausibilmente imputabili a differenze tra fonti (chiusure, aggiustamenti per dividendi o rettifiche), ma permettono di considerare il dataset coerente per la replica del test GOF.

Tabella 5.3: Statistiche descrittive dei rendimenti mensili (Periodo: 1996-2006). Confronto tra i dati replicati e quelli originali di Savu e Trede (2008).

Titolo	**Media (Replica)**	**SD (Replica)**	**Media (Paper)**	**SD (Paper)**	**Δ Media**	**Δ SD**
APD	1.05	8.22	1.030	8.304	0.020	-0.084
DOW	0.79	8.55	0.797	8.109	-0.007	0.441
DD	0.29	7.22	0.364	7.271	-0.074	-0.051
EMN	0.16	9.15	0.189	9.155	-0.029	-0.005
PX	1.33	8.76	1.324	8.969	0.006	-0.209
ROH	1.07	8.71	1.139	9.320	-0.069	-0.610

5.1.7 Trasformazione in pseudo-osservazioni

Poichè l'approccio semiparametrico non richiede un modello parametrico per le marginali, i rendimenti sono stati trasformati in pseudo-osservazioni su (0,1) tramite ranghi empirici riscalati:

$$U_{ij} = \frac{R_{ij}}{n+1}, \tag{5.2}$$

dove R_{ij} è il rango del rendimento del titolo j nel campione di ampiezza n [10]. Quindi, a partire dalla matrice dei rendimenti $\mathbf{X}$ di dimensione $n \times d$, ogni osservazione X_{ij} (rendimento del titolo j al tempo i) è stata trasformata utilizzando 5.2 con R_{ij} rango di X_{ij} tra le n osservazioni del titolo j. Questa trasformazione isola la struttura di dipendenza nelle osservazioni $\mathbf{U}_i$ (catturata dalla copula) dalle marginali e costituisce l'input del test di bontà di adattamento (Capitolo 3).

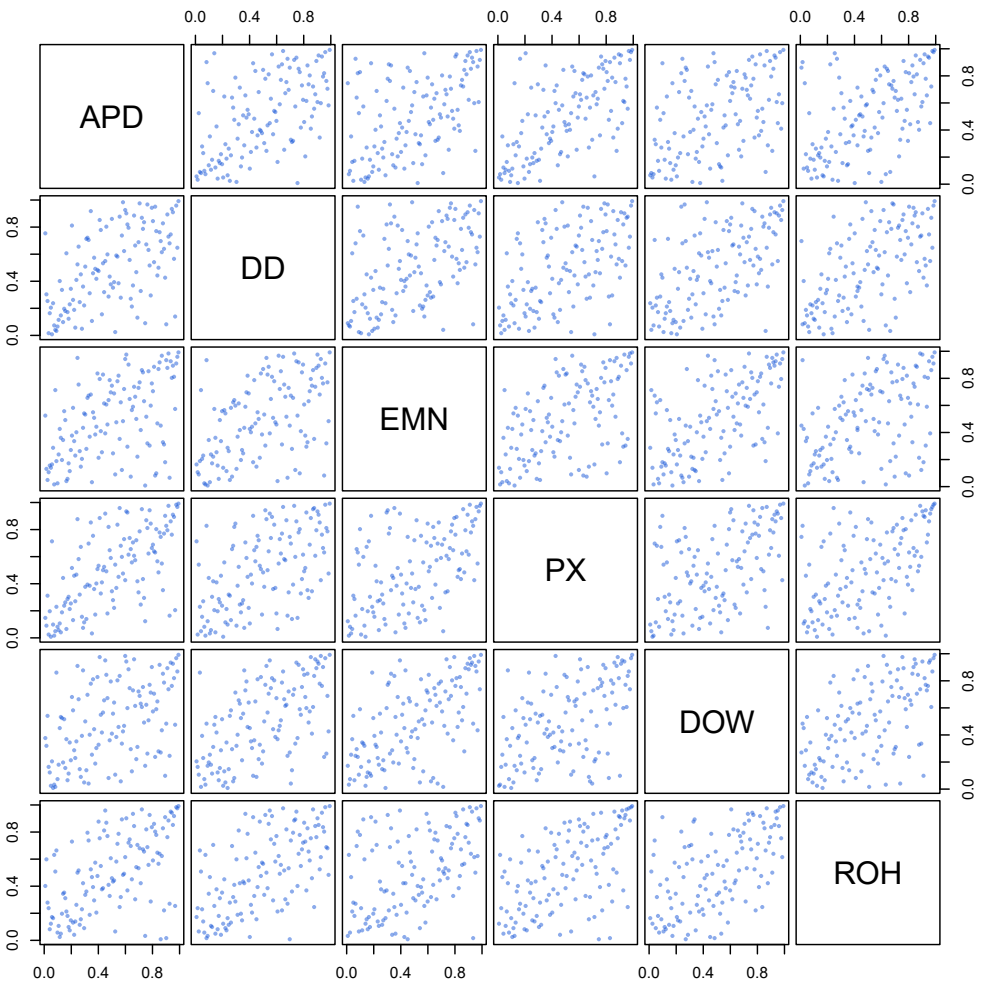

Figura 5.4: Pairplot delle pseudo-osservazioni: la dipendenza emerge nella struttura congiunta visualizzata dagli scatter plot bivariati.

5.2 Analisi esplorativa della dipendenza

Prima di procedere con i test formali, è utile descrivere la dipendenza tra i titoli con una misura rank-based coerente con la modellazione copula. Poichè si stanno testando famiglie di copule Archimedee ad un parametro, che implicano una struttura di dipendenza simmetrica e scambiabile, una misura di associazione naturale è la τ di Kendall: essa è invariante per trasformazioni monotone strettamente crescenti e può essere interpretata come misura di concordanza [1].
La Figura 5.5 riporta la heatmap della matrice di Kendall's τ stimata sul dataset

finale. Un controllo importante riguarda la coppia DD–DOW: nella versione preliminare errata τ risultava prossima a 1 (serie quasi identiche), mentre nel dataset finale corretto τ è pari a circa 0.41, valore plausibile per due aziende dello stesso settore ma distinte.

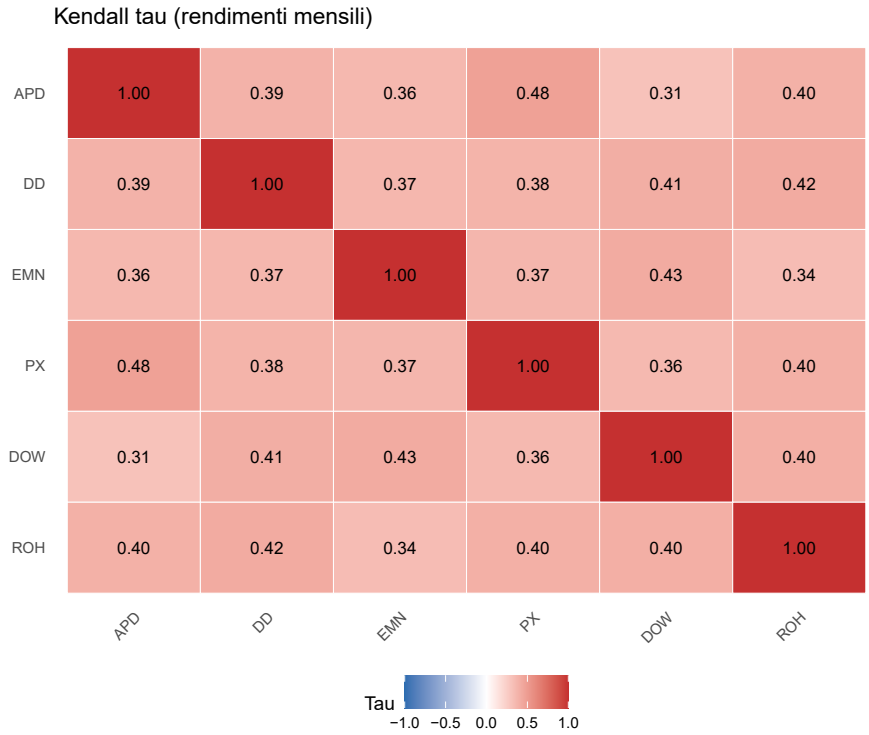

Figura 5.5: Heatmap della matrice di Kendall τ sui rendimenti mensili replicati.

I valori riportati nella Figura 5.5 mostrano una dipendenza positiva moderata tra tutte le coppie di titoli, con valori di τ compresi tra 0.31 e 0.48. Questo suggerisce che l'assunzione di dipendenza omogenea implicita nelle copule Archimedee ad un parametro non è palesemente violata, sebbene vi siano alcune variazioni tra le coppie. Inoltre, gli stessi autori notano che, essendo i vettori aleatori da copule Archimedee sempre equi-correlati, non è indicato considerare portafogli troppo eterogenei per paese/industria, mentre ha senso lavorare su titoli dello stesso settore [10].
La Tabella 5.4 confronta le Kendall's τ a coppie tra il dataset replicato e i valori del paper. Le differenze sono contenute (in valore assoluto al massimo circa 0.06) mentre la maggior parte delle coppie differisce di pochi centesimi. Queste discrepanze sono compatibili con differenze di fonte dati, convenzioni di chiusura mensile e criteri di rettifica, supportano la validità del dataset ricostruito e permettono di procedere con l'applicazione dei test di bontà di adattamento.

Tabella 5.4: Confronto delle Kendall's τ a coppie: replica vs paper [10]. $\Delta\tau = \tau_{\text{rep}} - \tau_{\text{paper}}$.

Coppia	τ **Replica**	τ **Paper**	$\Delta\tau$
APD-DOW	0.311	0.294	+0.017
APD-DD	0.390	0.428	-0.038
APD-EMN	0.360	0.336	+0.024
APD-PX	0.476	0.451	+0.025
APD-ROH	0.397	0.391	+0.006
DOW-DD	0.412	0.402	+0.010
DOW-EMN	0.426	0.427	-0.001
DOW-PX	0.356	0.298	+0.058
DOW-ROH	0.404	0.380	+0.024
DD-EMN	0.365	0.387	-0.022
DD-PX	0.382	0.391	-0.009
DD-ROH	0.424	0.431	-0.007
EMN-PX	0.371	0.321	+0.050
EMN-ROH	0.337	0.348	-0.011
PX-ROH	0.403	0.398	+0.005

5.3 Stima dei parametri delle famiglie candidate

Poichè nel test di bontà di adattamento considerato si ha a che fare con ipotesi composte, la copula esatta non è nota neanche sotto l'ipotesi nulla e, pertanto, richiede, per ciascuna famiglia Archimedea considerata, la stima del parametro di dipendenza θ del generatore φ_θ. Seguendo l'impostazione di Savu e Trede (2008) [10], la stima avviene tramite canonical maximum likelihood (CML) sulle pseudo-osservazioni $\mathbf{U}$, massimizzando la pseudo-log-verosimiglianza data da

$$\ln L(\theta; U_{ij}, i = 1, \ldots, d, j = 1, \ldots, n) = \sum_{j=1}^{n} \ln c_\theta(U_{1j}, \ldots, U_{dj})$$

con c_θ densità della copula

$$c(u_1, \ldots, u_d) = \varphi^{-1(d)}(\varphi(u_1) + \cdots + \varphi(u_d)) \cdot \prod_{i=1}^{d} \varphi'(u_i).$$

Quindi,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ln L(\theta)$$

La stima $\hat{\theta}$ consente di costruire la proiezione $\widehat{V}_j = C_{\hat{\theta}}(U_{1j}, \ldots, U_{dj})$ e la Kendall distribution stimata $K_{\hat{\theta}}$, che entrano direttamente nella statistica test e nel bootstrap [10].

Le famiglie candidate coincidono con quelle considerate nell'applicazione empirica del paper: Cook–Johnson, Frank e Gumbel–Hougaard [10]. Per inizializzare l'ottimizzazione numerica, è utile disporre di valori iniziali ragionevoli del parametro: una scelta naturale, adottata anche nella pratica applicativa, è ricavare una stima iniziale a partire dalla media $\bar{\tau}$ delle Kendall's τ calcolate dal campione originale (in modo coerente con l'impostazione exchangeable delle famiglie a un parametro) [6]. In particolare, sfruttando $\bar{\tau}$ si ricavano valori iniziali coerenti con le relazioni τ–θ delle famiglie Archimedee ($\theta_0^G = 1/(1-\bar{\tau})$ per Gumbel–Hougaard; $\theta_0^C = 2\bar{\tau}/(1-\bar{\tau})$ per Cook–Johnson).
L'ottimizzazione è eseguita con vincoli sui parametri: $\theta \geq 1$ per Gumbel–Hougaard, $\theta > 0$ per Cook–Johnson, mentre per Frank il parametro è ammesso su $\mathbb{R} \setminus \{0\}$ ma, in dimensione $d > 2$, si considera il dominio positivo in coerenza con le condizioni di completa monotonia della generatrice.
La Tabella 5.5 riporta le stime CML ottenute nella replica, il valore della log-pseudo-verosimiglianza e il confronto con i valori del paper [10]. Le differenze rispetto alle stime originali risultano contenute (ordine dei centesimi), coerentemente con quanto già osservato per le statistiche descrittive e per la matrice delle τ di Kendall.

Tabella 5.5: Stime CML del parametro $\hat{\theta}$ sulle pseudo-osservazioni ($d = 6$, $n = 121$) e confronto con Savu e Trede (2008) [10]. $\Delta = \hat{\theta}_{\text{rep}} - \hat{\theta}_{\text{paper}}$.

Famiglia	$\hat{\theta}$ (replica)	$\ln L(\hat{\theta})$	$\hat{\theta}$ (paper)	Δ
Cook–Johnson	0.701	119.984	0.693	+0.008
Frank	3.607	149.691	3.534	+0.073
Gumbel–Hougaard	1.541	156.694	1.500	+0.041

I valori di log-verosimiglianza sono riportati come diagnostica numerica. Come già sottolineato in Savu e Trede (2008) [10], la scelta finale del modello è basata sui risultati del test GOF con bootstrap, presentati nella sezione successiva.

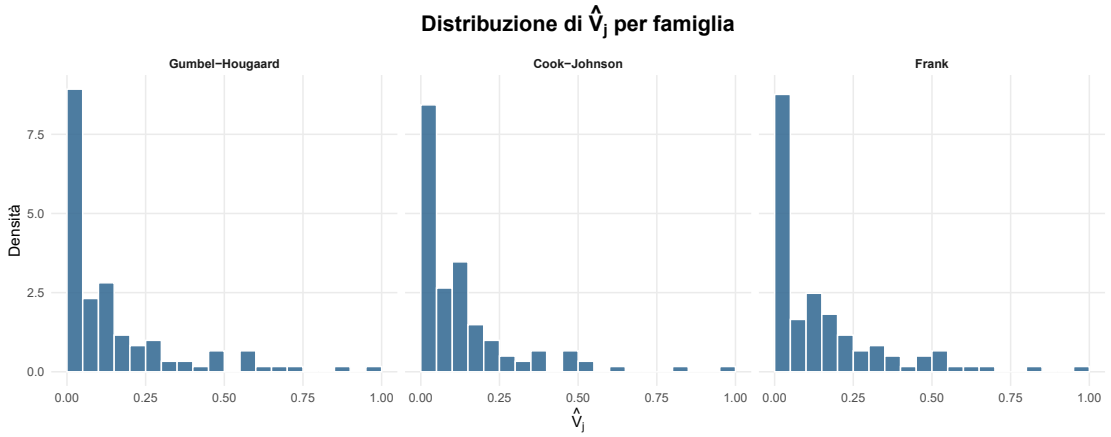

Figura 5.6: Distribuzione dei valori $\widehat{V}_j = C_{\hat{\theta}}(U_{1j}, \ldots, U_{dj})$ $(d = 6, n = 121)$ per ciascuna famiglia candidata. Il confronto è utile come diagnostica preliminare, poiché il test GOF si basa sul binning di $\widehat{V}$ e sui conteggi attesi calcolati tramite la Kendall distribution $K_{\hat{\theta}}$.

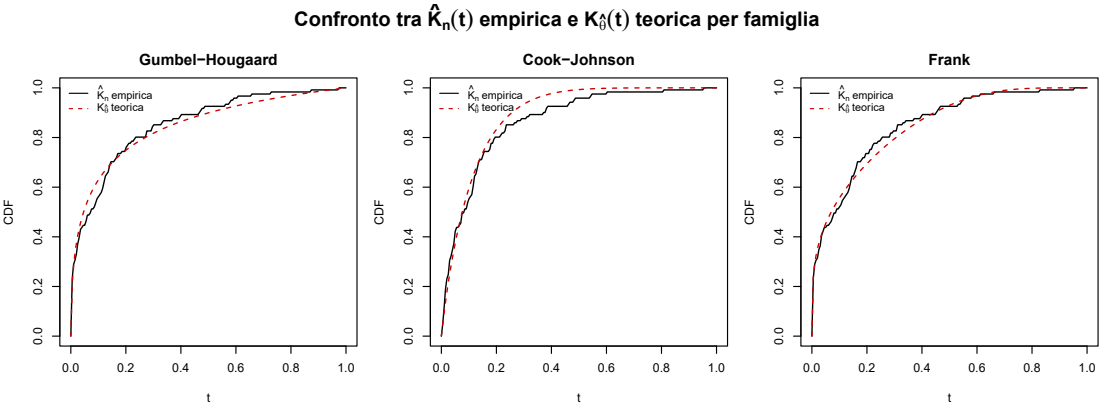

Figura 5.7: Diagnostica sulla proiezione: per ciascuna famiglia candidata, confronto tra $\widehat{K}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{\widehat{V}_j \leq t\}$ (linea nera) e la Kendall distribution teorica $K_{\hat{\theta}}(t) = \mathbb{P}(C_{\hat{\theta}}(\mathbf{U}) \leq t)$ (linea rossa tratteggiata). Scostamenti sistematici suggeriscono potenziale misspecificazione, che verrà valutata formalmente dal test GOF con bootstrap.

5.3.1 Stabilità numerica della stima

La massimizzazione della pseudo-verosimiglianza è un problema numerico standard e, in particolare, per verificare la stabilità della stima ottenuta, è utile analizzare il profilo della negativa pseudo-log-verosimiglianza $-\ell(\theta) = -\ln L(\theta)$ in un intorno di $\hat{\theta}$, controllando che presenti un minimo netto e ben definito. Infatti, la presenza di un

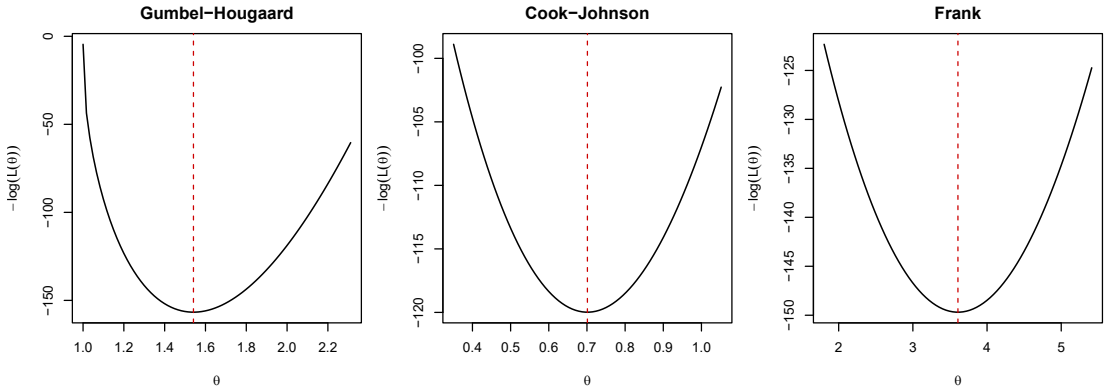

Figura 5.8: Profili della negativa log-pseudo-verosimiglianza $-\ell(\theta)$ attorno a $\hat{\theta}$ per le famiglie candidate. La linea verticale indica $\hat{\theta}$. Un minimo netto supporta la stabilità dell'ottimizzazione numerica nella stima CML.

minimo netto attorno a $\hat{\theta}$ indica che la stima è ben identificata e che l'ottimizzatore non converge a soluzioni di bordo o a minimi locali.

5.4 Goodness-of-fit: risultati e confronto con il paper

Questa sezione conclude l'applicazione empirica del test di bontà di adattamento per famiglie parametriche di copule Archimedee proposto da Savu e Trede (2008) [10]. L'oggetto del test è la struttura di dipendenza dei rendimenti mensili di sei titoli del settore chimico dell'indice S&P 500: come sottolineato dagli autori, né la media né la deviazione standard (e più in generale nessuna caratteristica delle marginali) entrano nella statistica del test, poiché le marginali vengono stimate non-parametricamente tramite la trasformazione in pseudo-osservazioni [10].
Il test è applicato alle tre famiglie Archimedee considerate anche nel paper (Cook–Johnson, Frank, Gumbel–Hougaard).
Per ciascuna famiglia, la procedura segue la pipeline descritta in [10], già formalizzata nel Capitolo 3 e parzialmente discussa nelle sezioni precedenti:
(i) pseudo-osservazioni via ranghi,
(ii) stima CML del parametro,
(iii) proiezione univariata tramite $\widehat{V}_j = C_{\widehat{\theta}}(U_j)$,
(iv) binning uniforme di [0,1] in $B = 20$ intervalli,
(v) statistica di tipo χ^2 sui conteggi osservati/attesi e
(vi) bootstrap parametrico per p-value e valori critici. [10]

5.4.1 Statistiche del test, bootstrap e p-value

Seguendo il paper di riferimento [10], l'intervallo unitario viene suddiviso in $B = 20$ intervalli uniformi [0,0.05], (0.05,0.10], ..., (0.95,1]. Indicando con $\{a_k\}_{k=0}^{B}$ i breakpoints, la statistica test è

$$T = \sum_{k=1}^{B} \frac{(O_k - E_k)^2}{E_k},$$

con

$$O_k = \#\{j : \widehat{V}_j \in (a_{k-1}, a_k]\}, \qquad E_k = n\Big(K_{\hat{\theta}}(a_k) - K_{\hat{\theta}}(a_{k-1})\Big).$$

dove O_k rappresenta i conteggi osservati mentre E_k quelli attesi. Come discusso nel Capitolo 3, T non segue una distribuzione χ^2 standard perché il parametro θ della copula viene stimato e interviene sia nella costruzione della proiezione univariata $\widehat{V}_j = C_{\hat{\theta}}(U_{1j}, \ldots, U_{dj})$ che nella Kendall distribution teorica $K_{\hat{\theta}}$ [10]. Per questo motivo la distribuzione sotto H_0 viene approssimata con bootstrap parametrico.
Il bootstrap ripete l'intera pipeline: per $b = 1, \ldots, J$ si genera un campione artificiale $U^{*(b)} \sim C_{\hat{\theta}}$, si ristima $\hat{\theta}^{*(b)}$ con CML e si ricalcola $T^{*(b)}$ con lo stesso binning. Nel paper vengono utilizzate 5000 repliche e nella replica si adotta lo stesso valore $J = 5000$ [10].
Il p-value bootstrap è calcolato come

$$\hat{p} = \frac{1 + \sum_{b=1}^{J} \mathbf{1}\{T^{*(b)} \geq T\}}{J + 1}.$$

Ne segue che, a parità di J, esiste un limite inferiore di risoluzione: con $J = 5000$, il p-value minimo riportabile è $1/(J+1) \approx 2 \cdot 10^{-4}$.

5.4.2 Risultati: T_{obs}, p-value e decisione

La Tabella 5.6 riporta i risultati finali del GOF per le tre famiglie candidate. L'esito è chiaro: la famiglia Gumbel–Hougaard non viene rifiutata (neppure al 10%), mentre Cook–Johnson e Frank sono rigettate a tutti i livelli considerati. Questo è coerente con la conclusione del paper, che nell'illustrazione empirica trova Gumbel–Hougaard come migliore descrizione della dipendenza, con rigetto di Cook–Johnson e Frank [10].

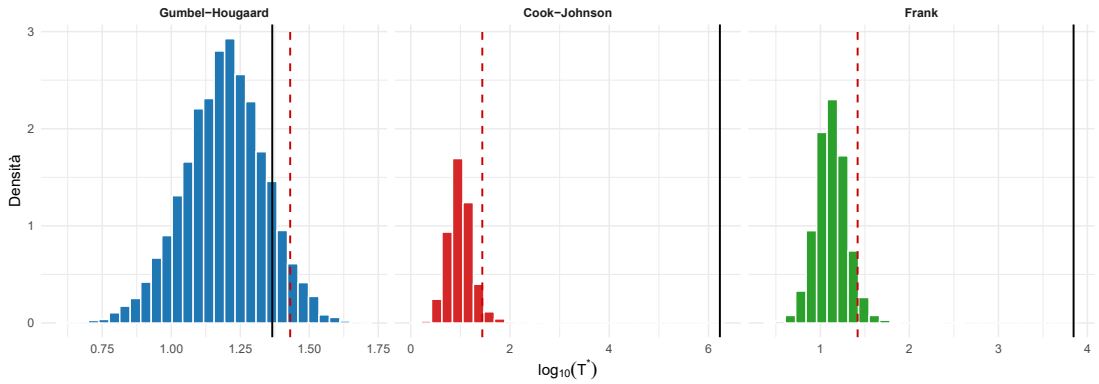

Figura 5.9: Distribuzione bootstrap di $\log_{10}(T^{*(b)})$ per ciascuna famiglia candidata ($B = 20$, $J = 5000$). La linea continua indica $\log_{10}(T_{\text{obs}})$, mentre la linea tratteggiata indica il quantile critico $\log_{10}(c_{0.95})$ ($\alpha = 0.05$). Per Cook–Johnson e Frank la statistica osservata è molto oltre la coda della distribuzione sotto H_0.

Tabella 5.6: Risultati del test GOF sul dataset empirico con $B = 20$ e bootstrap parametrico $J = 5000$. Si riportano T_{obs}, p-value bootstrap e i quantili critici $c_{1-\alpha}$ stimati dalla distribuzione bootstrap.

Famiglia	$\hat{\theta}$	T_{obs}	$\hat{p}$	$c_{0.90}$	$c_{0.95}$	$c_{0.99}$
Gumbel–Hougaard	1.541	23.225	0.115	23.864	26.969	33.373
Cook–Johnson	0.701	1 695 937	$2.0 \cdot 10^{-4}$	20.167	27.683	72.033
Frank	3.607	6 989.459	$2.0 \cdot 10^{-4}$	21.992	26.204	39.195

Per visualizzare il confronto tra la statistica test osservata T_{obs} e la distribuzione sotto H_0 stimata via bootstrap, la Figura 5.9 mostra gli istogrammi di $T^{*(b)}$ per ciascuna famiglia. Nel caso Gumbel–Hougaard, T_{obs} cade all'interno della parte alta ma non estrema della distribuzione bootstrap, coerentemente con un p-value intorno a 0.11. Nei casi Cook–Johnson e Frank, invece, T_{obs} risulta enormemente più grande di qualsiasi valore $T^{*(b)}$ generato sotto H_0, e ciò spiega perché il p-value si attesti al minimo $1/(J+1)$.
Questa è esattamente la conclusione evidenziata da Savu e Trede (2008) [10] nell'illustrazione empirica: Gumbel è la copula che si adatta megli ai dati, mentre Frank e Cook-Johnson non descrivono adeguatamente i dati.

5.4.3 Analisi dei risultati: celle critiche e contributi a T_{obs}

Per interpretare, è utile scendere al livello delle celle del binning. In particolare, per ciascuna famiglia si costruisce una tabella con breakpoints $(a_{k-1}, a_k]$, conteggi

Tabella 5.7: Diagnostica del bin finale (0.95,1] ($B = 20$). Il contributo contrib_{20} mostra come, per Cook–Johnson e Frank, la statistica T sia dominata dalla coda.

Famiglia	O_{20}	E_{20}	$1 - K_{\hat{\theta}}(0.95)$	contrib_{20}
Gumbel–Hougaard	1	$8.46 \cdot 10^{-1}$	$6.99 \cdot 10^{-3}$	$2.81 \cdot 10^{-2}$
Cook–Johnson	1	$5.90 \cdot 10^{-7}$	$4.87 \cdot 10^{-9}$	$1.695 \cdot 10^{6}$
Frank	1	$1.43 \cdot 10^{-4}$	$1.18 \cdot 10^{-6}$	$6.97 \cdot 10^{3}$

osservati O_k, conteggi attesi E_k e contributo

$$\text{contrib}_k = \frac{(O_k - E_k)^2}{E_k}, \qquad T_{\text{obs}} = \sum_{k=1}^{B} \text{contrib}_k.$$

Questo dettaglio è particolarmente importante perché, in una statistica χ^2-based, celle con E_k molto piccoli possono dominare la somma anche se O_k è piccolo (o addirittura pari a 1).
Nel dataset ricostruito si verifica un punto in coda alta per la proiezione $\widehat{V}$: per tutte le famiglie risulta $O_{20} = 1$ nel bin finale (0.95,1]. In particolare, in tutte le famiglie l'unico punto in coda ($\widehat{V}_j \geq 0.95$) corrisponde alla stessa data (1999-04-01), con $\widehat{V}_j$ compreso tra 0.952 e 0.974. La differenza è che l'atteso E_{20} varia enormemente tra le famiglie, perché dipende dalla Kendall distribution $K_{\hat{\theta}}$. In particolare, per Cook–Johnson e Frank la probabilità teorica della coda $1 - K_{\hat{\theta}}(0.95)$ è estremamente piccola, e di conseguenza E_{20} è quasi nullo. In questa situazione,

$$\text{contrib}_{20} = \frac{(1 - E_{20})^2}{E_{20}} \approx \frac{1}{E_{20}},$$

quindi basta un singolo punto osservato nel bin perché la statistica esploda.
La Tabella 5.7 sintetizza la diagnostica di coda prodotta dallo script: per Gumbel–Hougaard l'atteso nel bin finale è dell'ordine di 10^{-1}, e il contributo è trascurabile; per Frank l'atteso è dell'ordine di 10^{-4}, con contributo dell'ordine di 10^{3}; per Cook–Johnson l'atteso scende fino a 10^{-7}, con contributo dell'ordine di 10^{6}.

5.4.4 Confronto con i risultati di Savu–Trede (2008)

Per quanto riguarda i risultati del paper, gli autori concludono che Gumbel–Hougaard è l'unica famiglia non rigettata ($p = 0.710$), mentre Cook–Johnson e Frank risultano incompatibili con i dati ($p < 0.001$) [10].
Come già mostrato, le stime $\hat{\theta}$ sono molto vicine a quelle del paper (ordine dei centesimi), segno che la ricostruzione del dataset e la pipeline di stima sono coerenti con l'impostazione originale. Le statistiche T_{obs}, invece, non coincidono numericamente. Questo scostamento va letto alla luce della struttura stessa del

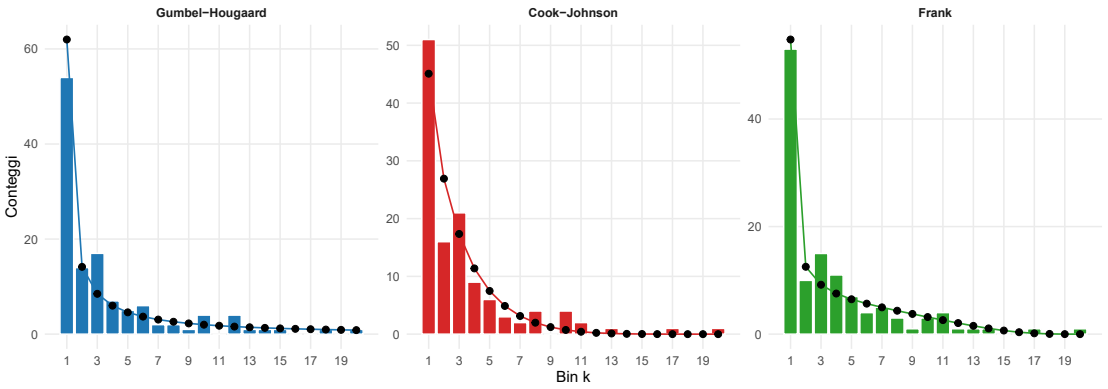

Figura 5.10: Binning su [0,1] con $B = 20$: conteggi osservati O_k (barre) e attesi E_k (punti) per ciascuna famiglia. Nei casi Cook–Johnson e Frank l'atteso negli ultimi bin si avvicina a zero, rendendo la statistica molto sensibile alla presenza di anche un solo punto in coda.

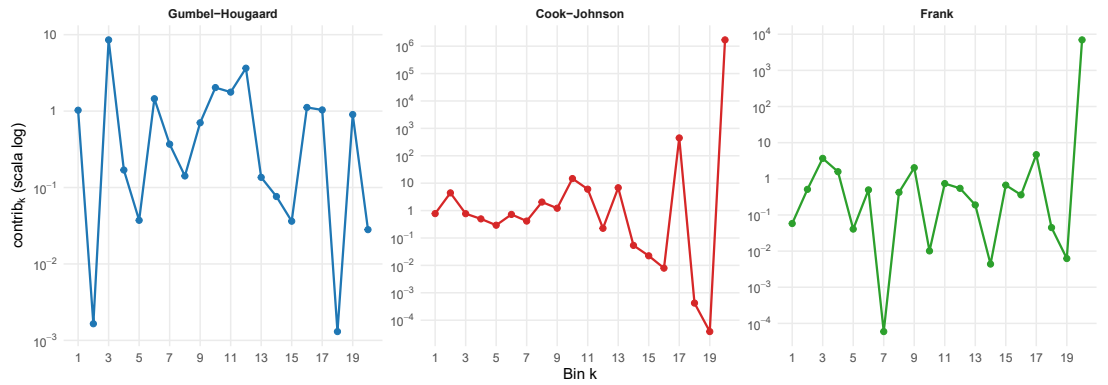

Figura 5.11: Contributi per bin $\text{contrib}_k = (O_k - E_k)^2/E_k$ nella statistica osservata T_{obs}. La figura evidenzia immediatamente se T_{obs} è distribuita su più intervalli oppure dominata da una singola cella di coda (come accade per Cook–Johnson e, in misura minore, per Frank).

test: T è una somma di contributi $(O_k - E_k)^2/E_k$ e può diventare molto sensibile a differenze localizzate nella regione di coda, soprattutto quando E_k è estremamente piccolo. Nel dataset ricostruito, la presenza di un singolo punto nel bin (0.95,1] porta, per Cook–Johnson e Frank, a contributi dominanti (Tab. 5.7) e quindi a valori di T molto elevati. Nonostante queste differenze numeriche, la parte sostanziale della replica — l'esito del test e la selezione del modello — risulta allineata con Savu e Trede (2008) [10].

Tabella 5.8: Confronto con Savu e Trede (2008) [10]. La replica riproduce la stessa conclusione inferenziale: Gumbel–Hougaard non viene rifiutata, mentre Cook–Johnson e Frank sono rigettate.

	Replica			Paper		
Famiglia	$\hat{\theta}$	T_{obs}	p-value	$\hat{\theta}$	T_{obs}	p-value
Cook–Johnson	0.701	1 695 937	$2.0 \cdot 10^{-4}$	0.693	29 580.2	< 0.001
Frank	3.607	6 989.4	$2.0 \cdot 10^{-4}$	3.534	211.4	< 0.001
Gumbel–Hougaard	1.541	23.225	0.110	1.500	14.4	0.710

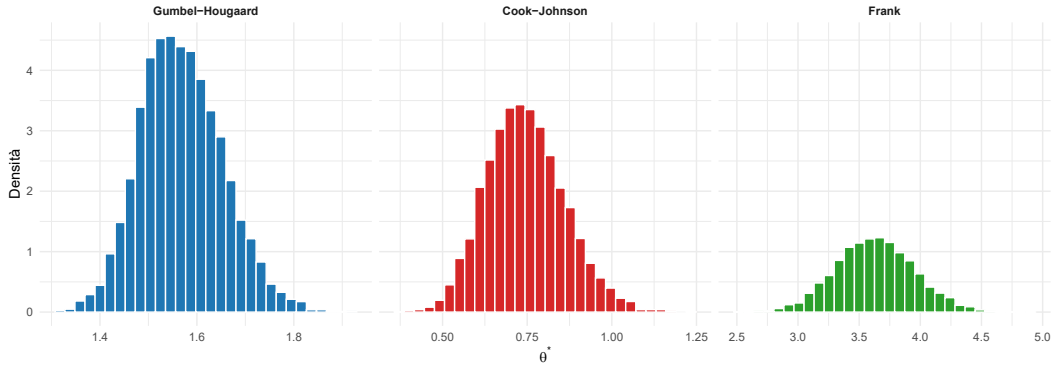

Figura 5.12: Distribuzione bootstrap di $\hat{\theta}^{*(b)}$ per ciascuna famiglia candidata ($B = 20$, $J = 5000$). La variabilità di $\hat{\theta}^*$ fornisce una misura empirica dell'incertezza di stima nella pipeline bootstrap.

5.5 Robustezza e verifiche supplementari

L'illustrazione empirica in Savu e Trede (2008) [10] è costruita con scelte operative fissate ($B = 20$, $J = 5000$). Nel contesto di una replica, sono state aggiunte alcune verifiche mirate, non per modificare la procedura originale, ma per mostrare che la conclusione (Gumbel-Hougaard descrive meglio la dipendenza, le altre famiglie no) non è un artefatto di una singola impostazione.

5.5.1 Stabilità del bootstrap: distribuzioni di T^* e $\hat{\theta}^*$

Un primo controllo è guardare la distribuzione delle repliche bootstrap. Per ciascuna famiglia si ottengono coppie $(\hat{\theta}^{*(b)}, T^{*(b)})$ per $b = 1, \ldots, 5000$. Oltre alla distribuzione bootstrap di $T^{*(b)}$ (Fig. 5.9), è utile riportare anche la distribuzione di $\hat{\theta}^{*(b)}$: in questo modo si ha un'indicazione empirica della variabilità della stima CML nel contesto del bootstrap.

Tabella 5.9: Analisi di sensibilità rispetto al binning uniforme: p-value bootstrap del GOF al variare del numero di bin B ($J = 5000$). Per Cook–Johnson il p-value coincide con il minimo risolvibile $1/(J+1) \approx 2 \cdot 10^{-4}$.

B	**Gumbel–Hougaard**	**Cook–Johnson**	**Frank**	**Esito (5%)**
10	0.156	$2.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	G: NR; CJ/F: R
15	0.243	$2.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	G: NR; CJ/F: R
20	0.115	$2.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	G: NR; CJ/F: R
25	0.656	$2.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	G: NR; CJ/F: R
40	0.572	$2.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	G: NR; CJ/F: R

In Figura 5.12 le distribuzioni bootstrap di $\hat{\theta}^{*(b)}$ risultano unimodali e concentrate attorno a $\hat{\theta}$. L'assenza di multimodalità o di masse sui valori estremi del parametro suggerisce una buona stabilità della stima CML nel bootstrap.

5.5.2 Sensibilità a B e alla gestione delle celle con E_k molto piccoli

Poiché per Cook–Johnson e Frank il rigetto è legato alla presenza di celle con conteggi attesi E_k estremamente piccoli nella coda alta, un controllo di robustezza naturale consiste nel variare il numero di bin B del binning uniforme su [0,1], mantenendo invariati lo schema bootstrap e il numero di repliche ($J = 5000$). In linea generale, l'esito qualitativo dovrebbe rimanere invariato; tuttavia, quando la statistica è dominata da una singola cella con $E_k \approx 0$, il valore numerico di T può cambiare anche di ordini di grandezza al variare di B.
I risultati ottenuti per $B \in \{10,15,20,25,40\}$ sono riassunti in Tabella 5.9. Per Cook–Johnson il p-value coincide sistematicamente con il minimo risolvibile $1/(J+1) \approx 2 \cdot 10^{-4}$, mentre per Frank risulta sempre molto piccolo; al contrario, per Gumbel–Hougaard il p-value resta stabilmente sopra la soglia del 5%. Questo conferma che la conclusione dell'analisi principale (Gumbel non rigettata: NR; Cook–Johnson e Frank rigettate: R) non dipende dalla scelta specifica $B = 20$. L'esito rimane invariato anche ai livelli $\alpha = 0.10$ e $\alpha = 0.01$.

Aggregazione della coda alta: ultimo bin (0.90,1]

Un controllo più mirato consiste nell'aggregare gli ultimi due intervalli del binning con $B = 20$, sostituendo (0.90,0.95] e (0.95,1] con un unico intervallo (0.90,1]. L'obiettivo è verificare se il rigetto delle famiglie Cook–Johnson e Frank dipenda in modo determinante dalla presenza di una singola cella di coda con E_k quasi nullo. La Tabella 5.10 confronta p-value e statistiche osservate tra la configurazione standard ($B = 20$) e il caso con coda aggregata. Per Cook–Johnson e Frank si

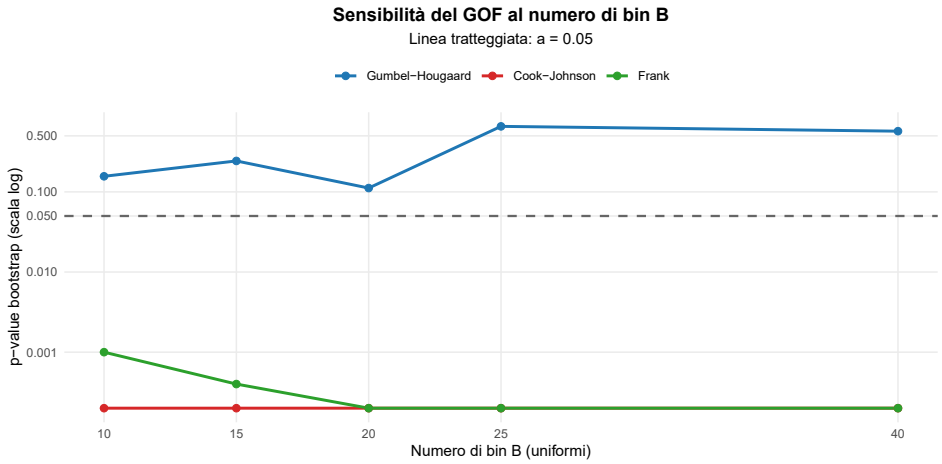

Figura 5.13: Sensibilità del p-value bootstrap al numero di bin B (binning uniforme). La linea tratteggiata indica la soglia $\alpha = 0.05$. Per Gumbel–Hougaard i p-value restano sopra α per tutti i valori di B considerati, mentre Cook–Johnson e Frank risultano sistematicamente rigettate (con p-value molto piccoli, spesso al minimo risolvibile con $J = 5000$).

Tabella 5.10: Confronto tra binning uniforme $B = 20$ e variante con aggregazione della coda alta (0.90,1] (bin effettivi $B = 19$). Si riportano T_{obs} e p-value bootstrap ($J = 5000$).

Famiglia	**B=20 (uniforme)**		**Tail merge (0.90,1]**	
	T_{obs}	p-value	T_{obs}	p-value
Gumbel–Hougaard	23.225	0.115	22.615	0.107
Cook–Johnson	$1.70 \cdot 10^{6}$	$2.0 \cdot 10^{-4}$	$2.61 \cdot 10^{4}$	$2.0 \cdot 10^{-4}$
Frank	$6.99 \cdot 10^{3}$	$2.0 \cdot 10^{-4}$	$1.71 \cdot 10^{2}$	$8.0 \cdot 10^{-4}$

osserva una forte riduzione di T_{obs} (coerente con l'aumento dell'atteso nel bin finale), ma l'esito inferenziale non cambia: le due famiglie restano rigettate, mentre Gumbel–Hougaard rimane la famiglia compatibile con i dati al livello considerato. Questo suggerisce che il rigetto non è un artefatto esclusivamente dovuto al bin estremo (0.95,1], ma riflette una misspecificazione più strutturale.
La Figura 5.14 evidenzia che, per Cook–Johnson e Frank, l'aggregazione della coda riduce l'amplificazione dovuta a celle con E_k estremamente piccoli (e infatti T_{obs} diminuisce di ordini di grandezza), ma la statistica osservata resta comunque ben oltre il quantile critico bootstrap. Al contrario, per Gumbel–Hougaard T_{obs} rimane vicino alla regione centrale della distribuzione sotto H_0, coerentemente con il mancato rigetto anche nel caso tail merge. In sintesi, il rigetto di Cook–Johnson e Frank non è attribuibile esclusivamente al bin estremo (0.95,1], ma persiste anche

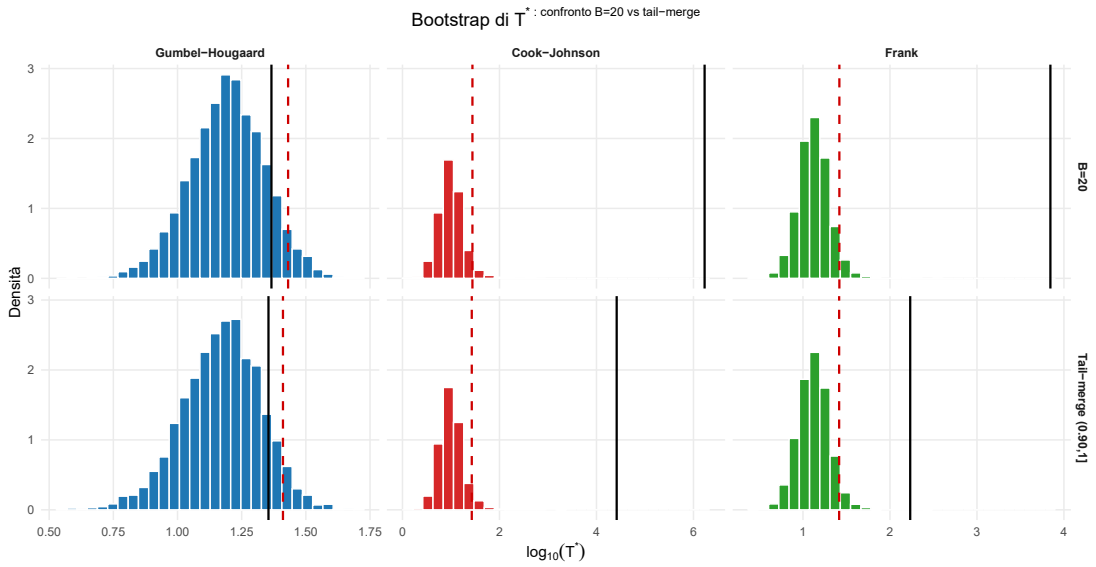

Figura 5.14: Distribuzione bootstrap di $\log_{10}(T^{*(b)})$ per ciascuna famiglia candidata, a confronto tra binning uniforme $B = 20$ e variante con aggregazione della coda alta (0.90,1] (tail merge). Le linee continue indicano $\log_{10}(T_{\text{obs}})$, mentre le linee tratteggiate indicano $\log_{10}(c_{0.95})$, calcolati nei due scenari.

dopo una gestione più conservativa della coda alta.

5.5.3 Sensibilità al bootstrap (J, seed, stabilità dei p-value)

Nel test GOF il p-value è stimato tramite bootstrap parametrico,

$$\hat{p} \;=\; \frac{1 + \sum_{b=1}^{J} \mathbb{I}\{T^{*(b)} \geq T_{\text{obs}}\}}{J+1},$$

per cui la risoluzione minima è $1/(J+1)$. Di conseguenza, per Cook–Johnson e Frank (dove $\hat{p}$ è già al minimo con $J = 5000$) un aumento di J non può modificare l'esito del test e avrebbe solo il ruolo di documentare ulteriormente che T_{obs} si colloca ben oltre la coda della distribuzione sotto H_0.
Per Gumbel–Hougaard, invece, il p-value osservato è intermedio ($\hat{p} \approx 0.11$): in questo caso ha senso verificare la stabilità numerica del risultato rispetto a
(i) un aumento di J,
(ii) seed diversi.
La Tabella 5.11 riporta il confronto tra $J = 5000$ (configurazione di riferimento) e $J = 20000$. Il p-value e i quantili critici cambiano in modo trascurabile e la decisione rimane invariata (non rigetto per Gumbel–Hougaard).

Tabella 5.11: Stabilità del bootstrap per Gumbel–Hougaard al variare di J (binning $B = 20$). Si riportano T_{obs}, p-value bootstrap e quantili critici $c_{1-\alpha}$ (bootstrap).

J	T_{obs}	p-value	$c_{0.90}$	$c_{0.95}$	$c_{0.99}$	**Esito (5%)**
5000	23.225	0.115	23.864	26.969	33.373	NR
20000	23.225	0.112	23.745	26.501	32.508	NR

Tabella 5.12: Sensibilità al seed per Gumbel–Hougaard ($B = 20$, $J = 5000$). Si riportano p-value bootstrap e quantili critici $c_{1-\alpha}$. In tutti i casi si osserva mancato rigetto.

seed	p-value	$c_{0.90}$	$c_{0.95}$	$c_{0.99}$	**Esito (5%)**
111	0.1142	23.7160	26.7926	32.8630	NR
222	0.1062	23.4195	26.0792	31.9250	NR
333	0.1196	24.0659	26.7370	32.8703	NR
444	0.1148	23.8163	26.5236	33.1380	NR
555	0.1152	23.7455	26.0819	33.1005	NR

Questo conferma che il mancato rigetto non dipende dalla particolare realizzazione del bootstrap.

Sensibilità al seed: stabilità del p-value per Gumbel–Hougaard

Per un'ulteriore verifica relativamente alla famiglia Gumbel–Hougaard, il test è stato ripetuto con $J = 5000$ mantenendo invariati $B = 20$ e $\hat{\theta}$, ma variando il seed del generatore pseudo-casuale. I risultati per cinque seed distinti sono riportati in Tabella 5.12.
Il p-value bootstrap oscilla in un intervallo ristretto (circa 0.106–0.120) ed è sempre ben al di sopra della soglia $\alpha = 0.05$. Anche i valori critici $c_{0.95}$ variano moderatamente tra i seed, ma restano sempre superiori a T_{obs}, per cui l'esito del test è invariato (nessun rigetto al 10%, 5% e 1% per tutti i seed considerati). Questa variabilità è coerente con l'errore Monte Carlo atteso per la stima p-value, dell'ordine di con $J = 5000$.

5.5.4 Trattamento dei ties nelle pseudo-osservazioni

La trasformazione in pseudo-osservazioni è costruita a partire dai ranghi marginali $U_{ij} = \text{rank}(X_{ij})/(n+1)$. Quando non ci sono osservazioni ripetute (ties), ogni marginale produce n ranghi distinti e quindi valori U_{ij} su una griglia discreta

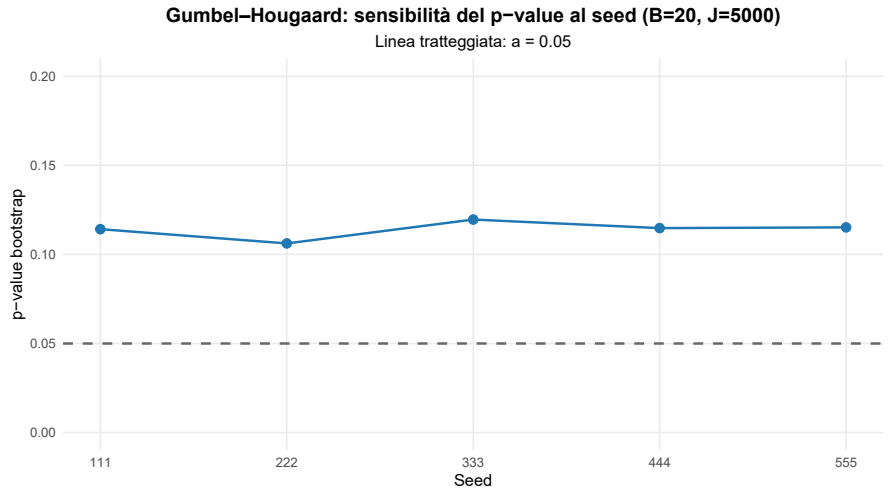

Figura 5.15: Stabilità del p-value bootstrap per Gumbel–Hougaard al variare del seed ($B = 20$, $J = 5000$). La linea tratteggiata indica $\alpha = 0.05$.

Tabella 5.13: Controllo sui ties nelle pseudo-osservazioni: numero di valori distinti per marginale ($n = 121$). Solo ROH presenta un tie (due osservazioni coincidenti).

Titolo	Valori distinti	Ties
APD	121	0
DD	121	0
DOW	121	0
EMN	121	0
PX	121	0
ROH	120	1

$\{1/(n+1), \ldots, n/(n+1)\}$. In presenza di ties, invece, la gestione dei ranghi può produrre valori non perfettamente allineati alla griglia [6].
Nel dataset finale ricostruito, il controllo sui ties mostra che cinque titoli non presentano valori ripetuti, mentre per **ROH** si osserva un solo tie (due osservazioni con lo stesso valore), verosimilmente dovuto a arrotondamenti nella fonte storica. Questo implica che l'effetto dei ties sulla costruzione di **U** è estremamente limitato ed infatti l'impatto sulla stima CML e sul test GOF risulta trascurabile.

5.5.5 Sensibilità alla correzione dello stock split di ROH

Nel preprocessing (Sez. 5.1) la serie **ROH** richiede una correzione retrospettiva per uno stock split non rettificato. Nella versione principale del dataset, il fattore di rettifica viene stimato direttamente dai prezzi mensili attorno al salto

Tabella 5.14: Robustezza rispetto alla correzione dello split di ROH (binning $B = 20$, bootstrap $J = 5000$). Confronto tra correzione stimata (ratio ≈ 2.80) e correzione imposta ($\times 3$).

	Correzione stimata (ratio ≈ 2.80)			**Split $\times 3$**		
Famiglia	$\hat{\theta}$	T_{obs}	p-value	$\hat{\theta}$	T_{obs}	p-value
Gumbel–Hougaard	1.541	23.225	0.115	1.542	23.170	0.110
Cook–Johnson	0.701	1 695 937	2.0×10^{-4}	0.703	1 682 397	2.0×10^{-4}
Frank	3.607	6 989.459	2.0×10^{-4}	3.617	6 915.030	2.0×10^{-4}

(ratio $\approx P_{\text{post}}/P_{\text{pre}} \simeq 2.80$), e applicato ai prezzi antecedenti, prima di ricalcolare i rendimenti. In alternativa, si può imporre il fattore 3 coerente con la natura dell'evento societario. Questa seconda scelta è particolarmente utile come controllo: ci si aspetta che i risultati GOF non cambino in modo qualitativo.
La Tabella 5.14 confronta la stima del parametro e i risultati del GOF nelle due varianti:

- **correzione stimata** (ratio ≈ 2.80);
- **correzione imposta** (factor $= 3$).

Le differenze numeriche sono contenute e, soprattutto, la conclusione inferenziale rimane invariata: Gumbel–Hougaard non viene rigettata, mentre Cook–Johnson e Frank restano incompatibili con i dati (con p-value al limite di risoluzione del bootstrap nel caso $J = 5000$).
In sintesi, la scelta del fattore di rettifica per lo split di ROH modifica leggermente la serie dei rendimenti e quindi le pseudo-osservazioni, ma non altera la selezione del modello né l'esito del GOF. Questo supporta l'interpretazione che la conclusione della replica (Gumbel–Hougaard come unica famiglia compatibile) non sia un artefatto di una singola scelta di preprocessing.

5.6 Applicazione empirica su un periodo più recente (2019–2025)

Oltre alla replica letterale dell'illustrazione empirica di Savu e Trede (2008) ([10]) sul periodo 1996–2006, è stata svolta un'applicazione aggiuntiva su dati più recenti con un duplice obiettivo: (i) verificare se, a distanza di anni e su un diverso regime di mercato, emergano conclusioni qualitativamente simili sulla compatibilità delle famiglie candidate, e (ii) documentare come gli eventi societari intervenuti nel tempo rendano complesso riproporre lo stesso identico set di titoli.

5.6.1 Scelta dei titoli e finestra temporale

L'intenzione iniziale era riutilizzare gli stessi sei titoli dell'analisi 1996–2006 (APD, DOW, DD, EMN, PX, ROH). Tuttavia, nel periodo più recente considerato intervengono vincoli strutturali dovuti a fusioni, acquisizioni e delisting (già discussi in Sez. 5.1):

- **DD e DOW:** dopo la fusione del 2017 e la successiva separazione, i ticker **DD** e **DOW** tornano ad essere negoziati come entità indipendenti solo dal 2019; per avere due serie distinte e confrontabili è quindi stato necessario partire almeno da tale data.
- **PX:** Praxair è stata delistata a seguito della business combination con Linde (2018). Per mantenere un universo di titoli comparabile (settore e dimensione), PX viene sostituita con **LIN** (Linde plc), disponibile come serie storica recente.
- **ROH:** Rohm & Haas è stata acquisita e non è più quotata come titolo indipendente. Non è quindi disponibile una serie recente direttamente confrontabile. Ne consegue che la dimensione del vettore passa da $d = 6$ a $d = 5$.

Alla luce di queste considerazioni, il campione è costituito da cinque titoli:

$$\text{APD},\ \text{EMN},\ \text{DD},\ \text{DOW},\ \text{LIN},$$

con prezzi mensili dal giugno 2019 al dicembre 2025 in quanto questa finestra garantisce copertura comune per tutti i titoli. I dati sono stati scaricati da Yahoo Finance ([22]).

5.6.2 Dati, rendimenti e controlli preliminari

I prezzi mensili (chiusura di fine mese) sono stati trasformati in rendimenti percentuali mensili secondo (5.1). Nel periodo considerato sono disponibili 79 osservazioni mensili sui prezzi e, di conseguenza, $n = 78$ rendimenti mensili comuni a tutti i titoli.
Le statistiche descrittive dei rendimenti sono riportate in Tabella 5.15. In questa finestra considerata non emergono outlier estremi (nessun mese con $|R| \geq 40\%$).
La dipendenza è descritta tramite Kendall's τ (Tabella/heatmap in Fig. 5.17). In particolare, i valori sono moderatamente positivi però variano sensibilmente tra le coppie (ad esempio $\tau(\text{DOW}, \text{APD}) \approx 0.26$ vs $\tau(\text{DOW}, \text{EMN}) \approx 0.55$). Poiché le copule Archimedee a un parametro implicano dipendenza omogenea tra tutte le coppie, l'applicazione del GOF va interpretata come verifica della compatibilità con un modello exchangeable che approssima la dipendenza complessiva.
Anche in questa applicazione, i rendimenti vengono trasformati in pseudo-osservazioni tramite ranghi riscalati (Sez. 5.1.7), così da isolare la dipendenza dalle marginali.

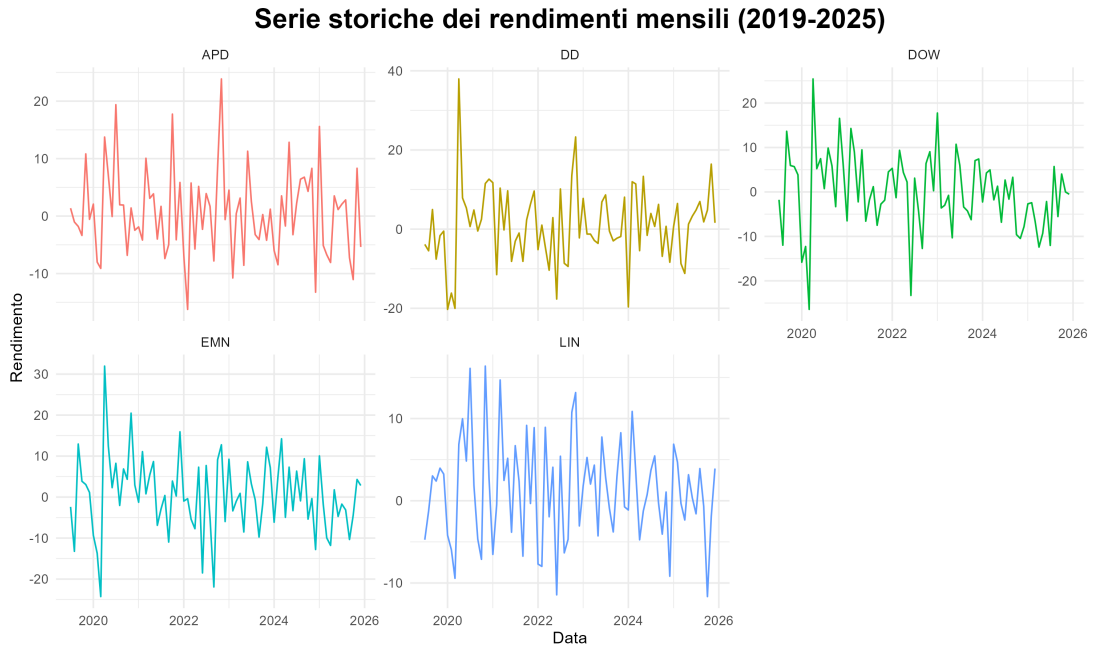

Figura 5.16: Serie temporali dei rendimenti percentuali mensili per i cinque titoli (2019–2025).

Tabella 5.15: Statistiche descrittive dei rendimenti mensili (2019–2025), $n = 78$ mesi comuni.

Ticker	n	Media	SD	Min	Max
APD	78	0.585	7.57	-16.2	23.9
DD	78	0.932	9.65	-20.3	38.0
DOW	78	-0.100	8.78	-26.4	25.4
EMN	78	0.470	9.36	-24.3	32.0
LIN	78	1.270	6.20	-11.7	16.4

5.6.3 Stima CML dei parametri

Come nell'analisi principale, per ciascuna famiglia candidata (Cook–Johnson, Frank, Gumbel–Hougaard) il parametro θ è stimato via CML sulle pseudo-osservazioni. Le stime ottenute (con log-pseudo-verosimiglianza come diagnostica numerica) sono riportate in Tabella 5.16.

5.6.4 Test GOF con bootstrap (scelta di B e risultati)

Rispetto al caso 1996–2006, il numero di osservazioni è più piccolo ($n = 78$ invece di $n = 121$). Per ridurre il rischio di celle con attesi molto piccoli (che possono

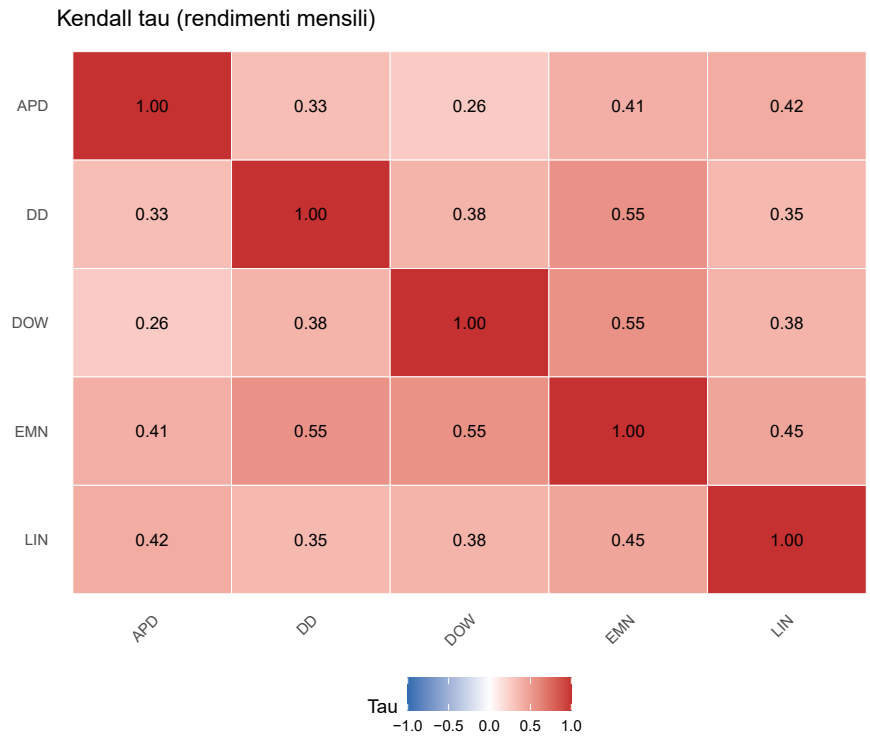

Figura 5.17: Heatmap della matrice di Kendall τ sui rendimenti mensili (2019–2025).

Tabella 5.16: Stime CML su pseudo-osservazioni (2019–2025), $d = 5$, $n = 78$.

Famiglia	$\hat{\theta}$	$\log L(\hat{\theta})$
Gumbel–Hougaard	1.562108	73.455
Cook–Johnson	0.937294	76.432
Frank	3.836856	77.071

dominare una statistica di tipo χ^2), si adotta un binning uniforme con $B = 10$, mantenendo $J = 5000$ repliche bootstrap come nell'analisi principale.
La Tabella 5.17 riporta T_{obs}, p-value bootstrap e quantili critici stimati.
In questa finestra temporale, l'evidenza è diversa rispetto al 1996–2006: con $B = 10$, Cook–Johnson è rigettata al 10% e al 5% ($\hat{p} \approx 0.0116$), mentre Gumbel–Hougaard e Frank non sono rigettate ai livelli usuali. Questo risultato non contraddice la replica del paper (che riguarda rendimenti di un periodo differente, numerosità maggiore e $d = 6$), ma mostra che la selezione della famiglia che si adatta meglio ai dati può dipendere dal periodo osservato e dall'universo effettivamente disponibile.

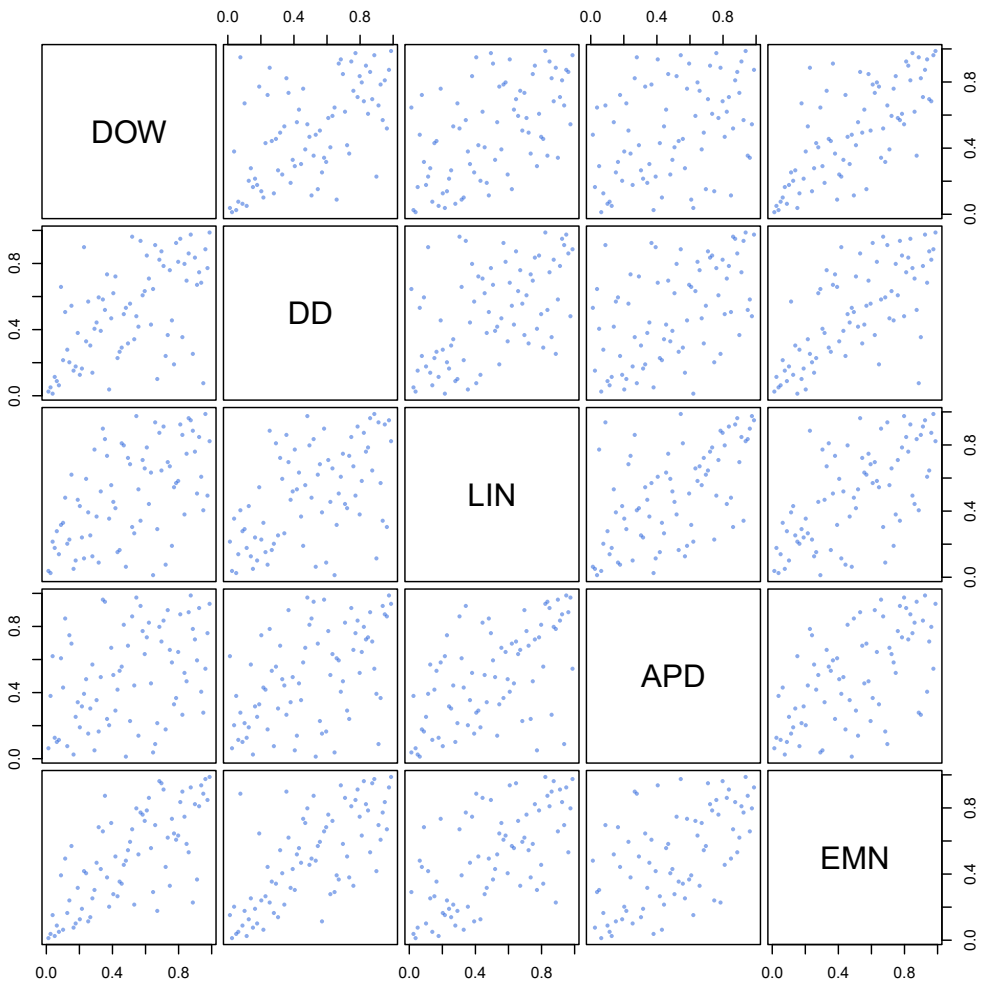

Figura 5.18: (Opzionale) Pairplot delle pseudo-osservazioni (2019–2025).

Tabella 5.17: Risultati del GOF sul periodo 2019–2025 ($d = 5$, $n = 78$), binning uniforme $B = 10$, bootstrap $J = 5000$.

Famiglia	$\hat{\theta}$	T_{obs}	p-value	$c_{0.90}$	$c_{0.95}$	$c_{0.99}$
Gumbel–Hougaard	1.562	9.634	0.185	11.485	13.303	18.288
Cook–Johnson	0.937	27.283	0.0116	9.922	12.555	30.362
Frank	3.837	4.705	0.617	10.536	12.755	20.141

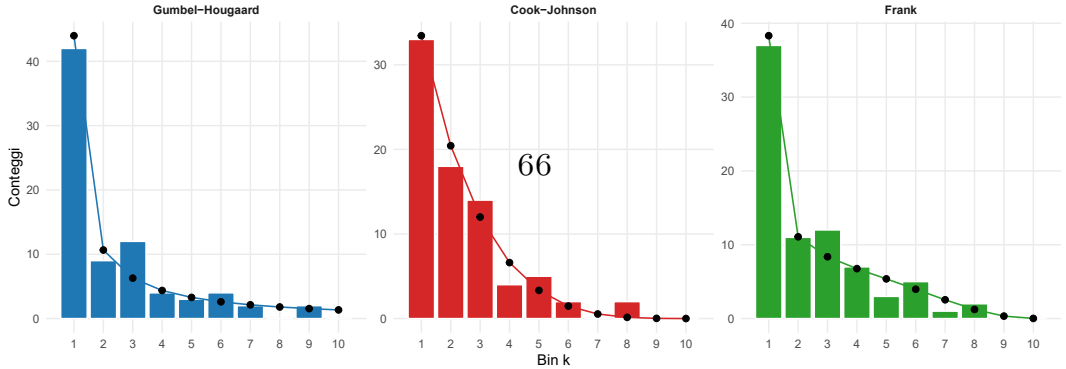

Figura 5.20: Binning su [0,1] con $B = 10$: conteggi osservati O_k (barre) e attesi E_k (punti) (2019–2025).

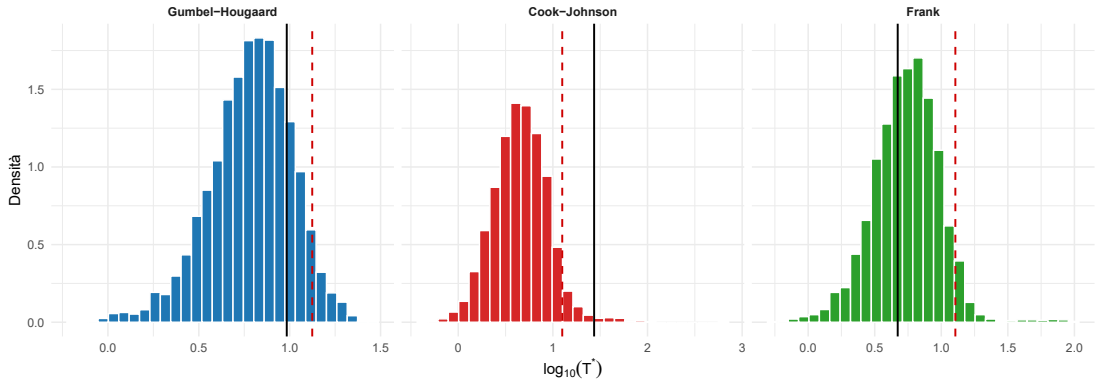

Figura 5.19: Distribuzione bootstrap di $\log_{10}(T^{*(b)})$ per ciascuna famiglia (2019–2025, $B = 10$, $J = 5000$). Linea continua: T_{obs}; linea tratteggiata: $c_{0.95}$.

Tabella 5.18: Robustezza rispetto al binning nel periodo 2019–2025 ($d = 5$, $n = 78$, $J = 5000$). Confronto dei risultati del test GOF tra $B = 10$ (configurazione principale) e $B = 15$. Si riportano T_{obs}, p-value bootstrap e il valore critico $c_{0.95}$.

	$B = 10$			$B = 15$		
Famiglia	T_{obs}	p-value	$c_{0.95}$	T_{obs}	p-value	$c_{0.95}$
Gumbel–Hougaard	9.634	0.185	13.303	10.584	0.544	20.001
Cook–Johnson	27.283	0.0116	12.555	63.680	0.00360	19.556
Frank	4.705	0.617	12.755	15.900	0.117	19.671

Robustezza rispetto alla scelta del binning: confronto $B = 10$ vs $B = 15$

Per verificare la stabilità delle conclusioni, oltre alla configurazione principale con $B = 10$ si è ripetuto il test con $B = 15$, mantenendo invariato il bootstrap ($J = 5000$).
La Tabella 5.18 mostra che l'esito qualitativo rimane invariato: Cook–Johnson è rigettata anche con $B = 15$ (in questo caso anche per $\alpha = 0.01$), mentre Gumbel–Hougaard e Frank non vengono rifiutate a nessun livello considerato. In particolare, passando da $B = 10$ a $B = 15$, il p-value di Cook–Johnson diminuisce (da $\hat{p} \approx 0.012$ a $\hat{p} \approx 0.0036$), mentre per Gumbel–Hougaard e Frank i p-value restano sopra $\alpha = 0.10$, pur variando per effetto della diversa discretizzazione.

Capitolo 6

Conclusioni

Questa tesi ha affrontato il problema della validazione di modelli di dipendenza basati su copule in un contesto tipicamente finanziario, dove ciò che conta non è solo la correlazione lineare, ma la struttura congiunta dei rendimenti e, in particolare, il comportamento nelle code. Nel quadro delle copule (Teorema di Sklar), la dipendenza viene separata dalle marginali e può essere modellata con famiglie parametriche interpretabili. Tuttavia, proprio perché la copula impone una forma specifica alla dipendenza, diventa essenziale chiedersi se tale forma sia compatibile con i dati [15]. In questo scenario si colloca il test di bontà di adattamento proposto da Savu e Trede (2008) per copule Archimedee multivariate a un parametro [10]: un test costruito su una proiezione univariata e sulla Kendall distribution, con valori critici ottenuti via bootstrap parametrico per gestire correttamente l'ipotesi nulla composta e la presenza di parametri stimati.
Il primo contributo del lavoro è stato implementare in modo riproducibile l'intera pipeline del test (pseudo-osservazioni, stima CML, costruzione della statistica basata su binning, bootstrap parametrico) e verificarne il comportamento tramite simulazione Monte Carlo, replicando le evidenze riportate nel paper in termini di dimensione empirica e potenza empirica. A causa del costo computazionale elevato della procedura annidata (Monte Carlo esterno e bootstrap interno), la replica è stata eseguita con valori di R e J ridotti rispetto allo studio originale. Questo aumenta la variabilità simulativa, ma consente comunque di verificare se l'implementazione riproduce i pattern qualitativi attesi. Nel complesso, i risultati ottenuti sono coerenti con l'obiettivo della replica: sotto H_0 la frequenza di rifiuto si colloca nell'intorno del livello nominale, e lo studio di potenza mostra un incremento della probabilità di rifiuto al crescere di n e, in molti casi, anche di d.
Queste evidenze supportano la correttezza operativa dell'implementazione e confermano che il test, pur essendo costruito su una statistica in stile χ^2, richiede bootstrap parametrico per approssimare la distribuzione della statistica test sotto l'ipotesi nulla per via della stima del parametro e della dipendenza della Kendall

distribution da $\hat{\theta}$.
Il secondo contributo centrale è stata la replica dell'illustrazione empirica di Savu e Trede (2008) su rendimenti mensili di titoli del settore chimico S&P 500 nel periodo 1996–2006 [10]. Poiché i dati originali provenivano da Datastream (database professionale a licenza) e la distanza temporale dalla pubblicazione rende non banale ricostruire le serie storiche (fusioni, delisting, cambi di ticker, rettifiche), la replica ha richiesto anche un lavoro sostanziale di reperimento dei dati, tracciabilità delle fonti e controlli di coerenza (statistiche descrittive, Kendall's τ, identificazione e correzione di anomalie tecniche come lo split di ROH). Nonostante queste difficoltà, l'esito inferenziale della replica risulta allineato con quello del paper: tra le famiglie candidate (Cook–Johnson, Frank e Gumbel–Hougaard), la copula Gumbel–Hougaard è l'unica non rigettata ai livelli usuali, mentre Cook–Johnson e Frank risultano incompatibili con i dati. Le differenze numeriche nelle statistiche osservate rispetto ai valori del paper possono essere lette alla luce della struttura della statistica, che somma contributi $(O_k - E_k)^2/E_k$ e può amplificare fortemente scostamenti localizzati quando alcuni attesi E_k diventano estremamente piccoli. In altre parole, la replica mostra anche un punto metodologicamente importante: in campioni finiti la statistica test può essere molto sensibile alla regione di coda (e alle celle con $E_k \approx 0$), per cui sono state affiancate ai risultati del test anche diagnostiche sui conteggi attesi/osservati e sui contributi per bin.
Per rendere la replica più informativa, sono state aggiunte verifiche mirate di robustezza che non modificano la procedura del paper, ma ne controllano la stabilità rispetto a scelte operative e variabilità simulativa: (i) sensibilità al binning (variazione di B e gestione più conservativa della coda alta), (ii) stabilità del bootstrap rispetto al numero di repliche J e al seed, (iii) controllo dell'impatto della correzione dello split di ROH. Nel caso empirico principale, queste verifiche confermano che la conclusione sostanziale non è un artefatto di una singola impostazione: il non rifiuto di Gumbel–Hougaard e il rigetto delle altre famiglie persistono sotto scelte ragionevoli di binning e sotto repliche bootstrap indipendenti, mentre i valori numerici (p-value e quantili critici) mostrano fluttuazioni compatibili con la variabilità Monte Carlo del bootstrap.
Infine, è stata svolta un'applicazione aggiuntiva su un periodo più recente per valutare se, su un diverso regime di mercato e con un universo di titoli necessariamente modificato da eventi societari (fusioni, delisting, acquisizioni), emergano conclusioni qualitativamente simili o diverse. Questa estensione ha anche un significato pratico per la replica: mostra che riprodurre un'analisi empirica a distanza di anni non è solo un esercizio statistico, ma implica vincoli reali sulla disponibilità e comparabilità delle serie storiche. Nel periodo 2019–2025, con un campione più corto e d ridotta, è stata adottata una scelta di binning più parsimoniosa e sono state considerate verifiche di robustezza rispetto a B. L'esito osservato suggerisce che, in questa finestra, la compatibilità delle famiglie candidate può differire rispetto

al 1996–2006: alcune famiglie risultano plausibili, mentre altre vengono rigettate, evidenziando come la struttura di dipendenza possa cambiare nel tempo e dipendere dall'universo effettivamente osservato.

Il lavoro ha alcuni limiti naturali. In primo luogo, l'analisi è focalizzata su copule Archimedee a un parametro, che implicano scambiabilità ed omogeneità della dipendenza tra tutte le coppie: assunzioni comode e interpretabili, ma potenzialmente restrittive in portafogli reali. In secondo luogo, il test ha solo potere banale contro copule non Archimedee con la stessa $K(t)$ e inoltre tende ad essere meno efficace nell'individuare copule misspecificate quando il grado di dipendenza presente nei dati è basso. Infine, come evidenziato dall'analisi empirica, la qualità del dato (rettifiche, discontinuità, coerenza temporale) è una componente determinante: differenze di fonte o di trattamento possono riflettersi in modo significativo su statistiche sensibili alle code.

A partire da questi risultati, diverse estensioni risultano naturali: (i) considerare famiglie di copule più flessibili (ad esempio modelli nested/gerarchici o costruzioni vine) per rilassare l'ipotesi exchangeable; (ii) confrontare il test basato su Kendall distribution con altre classi di GOF (processo della copula empirica, trasformazioni tipo Rosenblatt) e studiarne il trade-off tra potenza, stabilità e costo computazionale; (iii) esplorare modelli di dipendenza time-varying o a regimi, più adatti a dati finanziari che attraversano fasi di mercato differenti; (iv) ampliare l'applicazione empirica a universi di titoli più grandi e a frequenze diverse, mantenendo controlli rigorosi di tracciabilità e qualità del dato.

Bibliografia

[1] Roger B. Nelsen. *An introduction to copulas*. Second. Springer Series in Statistics. New York: Springer, 2006.

[2] Paul Embrechts, Filip Lindskog e Alexander McNeil. «Modelling dependence with copulas and applications to risk management». In: *Handbook of heavy tailed distributions in finance* 8.1 (2003).

[3] Fabrizio Durante e Carlo Sempi. «Copula theory: an introduction». In: *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*. Springer. 2010.

[4] Fabrizio Durante, Carlo Sempi et al. *Principles of copula theory*. Vol. 474. CRC press Boca Raton, FL, 2016.

[5] Piotr Jaworski, Fabrizio Durante, Wolfgang Karl Hardle e Tomasz Rychlik. *Copula theory and its applications*. Vol. 198. Springer, 2010.

[6] Marius Hofert, Ivan Kojadinovic, Martin Mächler e Jun Yan. *Elements of copula modeling with R*. Springer, 2018.

[7] Pravin K Trivedi, PK Trivedi e David M Zimmer. *Copula modeling: an introduction for practitioners*. Now Publishers Inc, 2007.

[8] Arkady Shemyakin e Alexander Kniazev. *Introduction to Bayesian estimation and copula models of dependence*. John Wiley & Sons, 2017.

[9] Ostap Okhrin, Alexander Ristig e Ya-Fei Xu. «Copulae in high dimensions: an introduction». In: *Applied quantitative finance* (2017).

[10] Cornelia Savu e Mark Trede. «Goodness-of-fit tests for parametric families of Archimedean copulas». In: *Quantitative Finance* 8.2 (2008).

[11] Christian Genest, Kilani Ghoudi e L-P Rivest. «A semiparametric estimation procedure of dependence parameters in multivariate families of distributions». In: *Biometrika* 82.3 (1995).

[12] Christian Genest e Louis-Paul Rivest. «Statistical inference procedures for bivariate Archimedean copulas». In: *Journal of the American statistical Association* 88.423 (1993).

[13] Christian Genest, Bruno Rémillard e David Beaudoin. «Goodness-of-fit tests for copulas: A review and a power study». In: *Insurance: Mathematics and economics* 44.2 (2009).

[14] Philippe Barbe, Christian Genest, Kilani Ghoudi e Bruno Rémillard. «On Kendall's process». In: *Journal of multivariate analysis* 58.2 (1996).

[15] Valdo Durrleman, Ashkan Nikeghbali e Thierry Roncalli. «Which copula is the right one?» In: *Available at SSRN 1032545* (2000).

[16] *LSEG Datastream Web Service / Datastream overview*. LSEG Developer Portal. Accesso tramite account Datastream autorizzato; servizio soggetto a licenza/abbonamento. n.d. URL: https://developers.lseg.com/eikon-apis/datastream-web-service/docs.

[17] *DowDuPont Merger Successfully Completed*. DuPont Investor Relations press release. Merger effective Aug 31, 2017; Dow and DuPont shares ceased trading separately. 2017. URL: https://www.investors.dupont.com/news-and-media/press-release-details/2017/DowDuPont-Merger-Successfully-Completed/default.aspx.

[18] *Dow completes separation from DowDuPont*. Dow Investor Relations press release. Dow begins regular way trading under the "DOW" ticker on April 2, 2019. 2019. URL: https://investors.dow.com/en/news/news-details/2019/Dow-completes-separation-from-DowDuPont/default.aspx.

[19] *DuPont Becomes Independent Company; begins trading under "DD" ticker symbol*. DuPont Investor Relations press release. DuPont common stock begins trading under "DD" on June 3, 2019. 2019. URL: https://www.investors.dupont.com/news-and-media/press-release-details/2019/DuPont-Becomes-Independent-Company-Uniquely-Positioned-to-Drive-Innovation-Led-Growth-and-Shareholder-Value/default.aspx.

[20] *Business Combination Between Praxair and Linde AG Successfully Completed*. Linde press release. Completion Oct 31, 2018; Praxair shares delisted; new ticker LIN. 2018. URL: https://www.linde.com/news-and-media/2018/business-combination-between-praxair-and-linde-ag-successfully-completed.

[21] *Dow 10-K filing: Acquisition of Rohm and Haas Company (completed April 1, 2009)*. U.S. SEC EDGAR filing. Disclosure of completion date and transaction details. 2012. URL: https://www.sec.gov/Archives/edgar/data/29915/000002991510000024/tdccform10-k2009.htm.

[22] *Yahoo Finance*. Yahoo Finance. Portale di dati finanziari e quotazioni di mercato. n.d. URL: https://finance.yahoo.com/.

[23] *Investing.com*. Investing.com. Portale di dati finanziari e notizie di mercato. n.d. URL: `https://www.investing.com/`.

[24] *ADVFN*. ADVFN. Portale di dati finanziari e quotazioni. n.d. URL: `https://www.advfn.com/`.

[25] *Company News; Rohm & Haas Announces Executive Succession Plans*. The New York Times. Riferimento citato per l'evento di stock split nel 1998. 1998. URL: `https://www.nytimes.com/1998/07/29/business/company-news-rohm-haas-announces-executive-succession-plans.html`.