Dipartimento di Economia e Finanza

Corso di Laurea: Banche e Intermediari Finanziari

Cattedra: Risk Management and Compliance

Stimare la volatilità dei mercati finanziari azionari: un confronto tra modelli tradizionali e di machine learning

Prof. Giancarlo Mazzoni

RELATORE

Prof.ssa Marta Catalano

CORRELATRICE

Stefano Squarcia

CANDIDATO

Anno Accademico 2024/2025

"*In physics you're playing against God, and he doesn't change his laws very often. In finance, you're playing against God's creatures, agents who value assets based on their ephemeral opinions*."

Emanuel Derman

Indice

Introduzione

La volatilità è un argomento centrale nella letteratura finanziaria e la sua fondamentale rilevanza risiede nella vasta gamma delle sue applicazioni, non limitandosi esclusivamente a tematiche di *risk management*, ma estendendosi ad aspetti quali la costruzione di portafogli efficienti (*asset allocation*) e il *pricing* di strumenti finanziari, tra cui assumono particolare importanza i contratti di opzione.

Un embrionale concetto di volatilità, in statistica definita deviazione standard, si può rinvenire in un'opera del 1806 di François Arago[1] sulla dispersione degli errori nelle misure geodetiche. Questi concetti furono ulteriormente sviluppati nel 1809 da Carl Friedrich Gauss[2], che sviluppò il metodo dei minimi quadrati per minimizzare gli errori nelle osservazioni astronomiche, introducendo implicitamente la varianza come somma dei quadrati delle deviazioni. Adolphe Quetelet[3] nel 1835 applicò concetti di dispersione alla statistica sociale, analizzando le deviazioni dalla media nelle altezze e nei pesi delle persone, mentre bisognerà attendere sessant'anni prima che Karl Pearson[4] nel 1894 definisca formalmente la varianza come misura di dispersione dei dati intorno alla media e la sua radice, la deviazione standard, meglio nota in finanza come volatilità.

La volatilità può determinare vaste conseguenze sull'economia nel suo complesso, ed è continuamente stimata e valutata dalle Autorità Finanziarie di tutto il mondo, come Banche Centrali e Governi. Per esempio, sia la Federal Reserve che la Banca Centrale Europea prendono in considerazione la volatilità dei titoli nel processo di formazione delle politiche monetarie. Inoltre, la volatilità influisce negativamente sulla liquidità del mercato, poiché empiricamente quando la prima aumenta, la seconda tende a diminuire, ed è fondamentale anche per le strategie di copertura. Infatti, in condizioni di stress del mercato, non solo la volatilità aumenta, ma anche le correlazioni tra i diversi titoli. In queste circostanze, gli strumenti derivati possono svolgere il ruolo di assicurazione contro le improvvise flessioni del mercato. Per tutte queste ragioni la volatilità è universalmente

[1] F. Arago, *Mémoire sur la dispersion des erreurs dans les mesures géodésiques*, 1806.
[2] C. F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, Hamburg: Perthes et Besser, 1809.
[3] A. Quetelet, *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*, Paris: Bachelier, 1835.
[4] K. Pearson, *Contributions to the Mathematical Theory of Evolution*, London: Dulau & Co., 1894.

riconosciuta come il principale e basilare indicatore del rischio dei mercati finanziari e l'importanza della sua previsione è una conseguenza naturale.[5]

Più precisamente, per rischio di mercato si intende il rischio di variazioni del valore di mercato di uno strumento o di un portafoglio di strumenti finanziari in relazione a variazioni inattese delle condizioni di mercato, come ad esempio prezzi azionari, tassi di interesse, tassi di cambio e le volatilità di tali variabili; esso include dunque i rischi su posizioni in valuta, in titoli obbligazionari e azionari, così come su tutte le altre attività e passività finanziarie scambiate da una società o da un individuo.

Ne consegue che a seconda del tipo di prezzo cui si fa riferimento, i rischi di mercato possano assumere connotazioni differenti. In particolare questo lavoro si concentra sul rischio azionario, che si verifica quando il valore di mercato delle posizioni assunte è sensibile all'andamento dei mercati azionari, come ad esempio nel caso di titoli azionari, *stock-index future* o *stock option*.

Può anche assumere rilevanza il cosiddetto rischio di volatilità, che si verifica quando il valore di mercato delle posizioni assunte è sensibile a variazioni della volatilità di una delle variabili sopra considerate, nel nostro caso riferite al mercato azionario. Si può pensare ad esempio ai contratti di opzione, il cui valore dipende dal livello della volatilità.

L'analisi del rischio azionario, su cui si concentra questo lavoro, implica un approccio complesso, in quanto la variabilità dei risultati dipende da una molteplicità di circostanze. In prima battuta possiamo affermare che i corsi azionari sono il principale fattore di rischio per il prezzo degli strumenti finanziari derivati scritti su di essi, mentre i fattori di rischio relativi ai prezzi delle azioni sono costituiti principalmente dalle *news* che riguardano gli emittenti. Risulta dunque arduo identificare relazioni matematiche simili a quella che nel caso dei mercati obbligazionari collega i prezzi dei titoli e i tassi d'interesse di mercato. Nonostante la letteratura sull'argomento sia vasta e siano stati proposti molti modelli, negli ultimi decenni si è assistito a un aumento delle applicazioni delle tecniche

[5] Sulla differenza tra predizione e previsione: «'Previsione' significa un'attribuzione di probabilità alle varie alternative possibili; si tratta di cosa del tutto diversa dalla 'predizione', intesa come indicazione di un'unica tra le alternative possibili, come 'quella che si verificherà'.»
F. de Finetti, L. Nicotra, *Bruno de Finetti: un matematico scomodo*, Livorno: Salomone Belforte & C., 2008.

di apprendimento automatico (*machine learning*) in molteplici settori, tra cui rivestono particolare rilevanza i mercati finanziari.

L'obiettivo della presente tesi è analizzare comparativamente l'applicazione di modelli econometrici tradizionali e di tecniche di apprendimento automatico per la stima della volatilità realizzata nei mercati azionari.

Il lavoro è organizzato come segue: nel primo capitolo, partendo dalla distinzione tra *risk-neutral world*, utilizzato per scopi di *pricing*, e *real world*, utilizzato per finalità di *risk management*, si fornirà un inquadramento teorico sul concetto di volatilità, visualizzandola sia da un punto di vista probabilistico che statistico.

Il secondo capitolo introduce il concetto di volatilità implicita e degli indici di mercato ad essa collegati, analizzando inoltre i principali fatti stilizzati che hanno favorito lo sviluppo dei modelli econometrici impiegati nella previsione della volatilità. In particolare, saranno presentati il modello a media mobile ponderata esponenziale, detto EWMA (*Exponential Weighted Moving Average*), il modello ARCH (*AutoRegressive Conditional Heteroscedasticity*) e la sua generalizzazione, il modello GARCH (*Generalized AutoRegressive Conditional Heteroscedasticity*), oltre che le sue varianti.

Nel terzo capitolo, saranno introdotti i principali modelli di apprendimento automatico (*machine learning*). In particolare, l'analisi verrà circoscritta ai metodi di apprendimento supervisionato, ritenuti maggiormente idonei per le analisi di regressione. Inizialmente tratteremo il metodo XGBoost, un'evoluzione metodologica dei modelli tradizionali di *Gradient Boosting*, per poi affrontare il vasto mondo delle Reti Neurali Artificiali (RNA). Nel dettaglio osserveremo le Reti Neurali *Feedforward* (FNN), la Reti Neurali Ricorrenti (RNN) e in particolare le reti con lunga memoria a breve termine (LSTM).

Il quarto e ultimo capitolo sarà dedicato alla stima della volatilità realizzata di tre indici rappresentativi dei mercati statunitense, europeo e asiatico. A tal fine verranno posti a confronto i modelli di machine learning XGBoost e LSTM con alcuni tra i principali modelli econometrici tradizionali, in particolare l'EWMA, il GARCH(1,1) e una delle sue estensioni, l'EGARCH(1,1,1). Per questi ultimi modelli sarà illustrata la stima dei parametri attraverso il metodo della massima verosimiglianza.

Capitolo 1: Il concetto di volatilità, il pricing e il risk management

1.1 Volatilità attesa e storica

Il rischio di mercato può essere formalmente ricondotto alla variabilità del rendimento associato a un determinato investimento. Tale variabilità può essere considerata in un'ottica probabilistica *ex ante*, ossia come dispersione del rendimento atteso rispetto alla propria media ponderata per la probabilità dei diversi scenari, oppure in un'ottica statistica *ex post*, in cui la volatilità è definita come variabilità dei rendimenti effettivamente osservati in un determinato orizzonte temporale.

Definendo il rendimento di un'attività finanziaria come una variabile aleatoria discreta X, con insieme immagine $Im(X)$, la varianza, o più precisamente il tasso di varianza atteso, può essere espresso come:

$$\sigma_x^2 = \sum_{x \in Im(X)} (x - \mathbb{E}[X])^2 \cdot P(X = x) \qquad (1.1)$$

$\mathbb{E}[X] = \sum_{x \in Im(X)} x \cdot P(X = x)$ è il valore atteso della variabile aleatoria, cioè la media pesata dei valori che possono essere ipoteticamente assunti in futuro dalla variabile, ognuno pesato per la relativa probabilità di accadimento. In base alle nostre ipotesi il valore atteso rappresenta il tasso di rendimento atteso del nostro investimento.

$P(X = x) \coloneqq p_X(x)$ è la densità discreta, una funzione che definisce la distribuzione di probabilità della variabile aleatoria discreta X, associando a ciascun valore discreto $x \in Im(X)$, la probabilità che la variabile X assuma proprio quel valore. Questa è delineata graficamente da una funzione a gradini con altezza data dalla probabilità associata alle osservazioni.

Quando si passa dal caso discreto al caso continuo, si ipotizza che la variabile aleatoria X possa assumere valori in un insieme continuo di numeri reali. In tale contesto, la densità discreta $p_X(x)$ viene sostituita dalla funzione di densità $f_X(x)$, definita positiva e continua.

Il tasso di varianza atteso di una variabile aleatoria continua *X* sarà calcolato come:

$$\sigma_x^2 = \int_a^b (x - \mathbb{E}[X])^2 \, f_X(x) dx$$

(1. 2)

In questo caso *a* e *b* sono gli estremi dell'insieme dei valori che possono essere assunti dalla variabile aleatoria, mentre $\mathbb{E}[X] = \int_a^b xf(x)dx$.

Sia nel caso discreto sia in quello continuo, la deviazione standard (o scarto quadratico medio) è definita come la radice quadrata della varianza, come riportato di seguito:

$$\sigma_X = \sqrt{\sigma_X^2}$$

(1. 3)

In ambito finanziario la deviazione standard prende il nome di volatilità e viene preferita operativamente alla varianza poiché, da un punto di vista quantitativo, è espressa nella stessa unità di misura dei rendimenti. Questo la rende più agevolmente interpretabile, soprattutto quando la si confronta con i valori medi.

Esiste inoltre una seconda formulazione della varianza, basata sulla funzione generatrice dei momenti, che risulta valida tanto nel caso discreto quanto in quello continuo. Tale definizione consente di calcolare la varianza (e, di conseguenza, la deviazione standard) come differenza tra il secondo momento della variabile aleatoria e il quadrato del suo primo momento. Formalmente:

$$\sigma_X^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

(1. 4)

Nonostante in ambito finanziario la volatilità *ex ante* rispecchi la filosofia probabilistica sottostante il processo di investimento, questa richiede valutazioni tutt'altro che agevoli in relazione alla previsione dei rendimenti e alla stima delle probabilità di manifestazione dei diversi scenari. Per queste ragioni dal punto di vista operativo è preferibile fare riferimento a stime *ex post*, quantificate a partire dalla distribuzione empirica dell'attività considerata, cioè utilizzando la serie storica dei suoi tassi di variazione, cioè i rendimenti.

Operativamente, se si dispone di una serie storica di rendimenti percentuali $r_i = \frac{S_i - S_{i-1}}{S_{i-1}}$, dove *S* è il prezzo di chiusura dell'attività finanziaria considerata, la deviazione standard è data dalla media aritmetica dei rendimenti storici realizzati, cioè effettivamente prodotti dall'investimento in tutta la sua storia. Formalmente:

$$\sigma_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \mu)^2}$$

(1. 4)

dove μ è la media dell'intera serie storica dei rendimenti.

È importante affermare che non esiste una definizione univoca di volatilità. Infatti, a seconda della frequenza di osservazione dei rendimenti si otterranno diverse stime della volatilità e in particolare si osserva che queste stime tendono ad aumentare man mano che l'intervallo temporale considerato diventa più piccolo. Di solito il prezzo di un'azione viene rilevato a intervalli di tempo fissi, ad esempio una tipica finestra temporale utilizzata nel risk management, che sarà riproposta nell'analisi empirica di questo lavoro, è quella giornaliera. Chiaramente la serie dei rendimenti utilizzata dovrà presentare la stessa frequenza di quella dei prezzi.

Si può osservare che i rendimenti logaritmici $r_i = \ln(\frac{S_i}{S_{i-1}})$ rappresentano un'ottima approssimazione dei rendimenti percentuali, man mano che la frequenza di osservazione dei dati viene ridotta. Nell'analisi empirica di questo lavoro utilizzeremo rendimenti logaritmici, in quanto risultano additivi nel tempo e sono facilmente modellabili da un punto di vista statistico, rendendo i calcoli di varianza e volatilità più lineari.

Nella pratica per stimare la volatilità ci si affida ad un campione di dati, non all'intera serie storica dell'azione considerata. La volatilità di un intero set di dati può essere definita in termini statistici come la deviazione standard della popolazione. La scelta di utilizzare un campione di una certa numerosità serve ad evitare di includere nella nostra stima dei dati che non riflettano le attuali condizioni di mercato. Operativamente ciò implica due modifiche rispetto alla formula precedente, che diventa:

$$\hat{\sigma}_t = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r}_t)^2}$$

(1. 5)

In primo luogo μ, la media della popolazione, viene sostituita con il suo naturale stimatore, cioè la media campionaria $\bar{r}_t$, che nel nostro caso è rappresentata dal rendimento medio realizzato del campione nel periodo di tempo considerato.

La seconda modifica consiste nell'utilizzo di n-1 al posto di n come divisore: ciò avviene perché stimando la media della popolazione con la media campionaria e sostituendola nell'equazione della volatilità, si sottostima la varianza di un fattore pari a $(n-1)/n$.

L'aggiustamento di questo termine, che risulta nella divisione della somma degli scostamenti quadratici attorno alla media campionaria per $n-1$ invece di n, è chiamata correzione di Bessel[6] (o correzione per i gradi di libertà) e si applica in quanto stimare la media consuma una parte dell'informazione contenuta nei dati, cioè un grado di libertà. Tale correzione garantisce che la varianza campionaria sia uno stimatore non distorto della varianza della popolazione.

È importante sottolineare che queste modifiche valgono sotto l'ipotesi che i rendimenti siano indipendenti, identicamente distribuiti e con media e varianza finite.

Per modellare i rendimenti tipicamente si utilizza la distribuzione normale, coerente con il *framework* di log-normalità dei prezzi delle azioni introdotto da Black-Scholes-Merton. Tuttavia, come osserveremo, l'evidenza empirica ha rigettato questa ipotesi, suggerendo l'utilizzo di distribuzioni probabilistiche alternative, come ad esempio la t di Student.[7]

Ad ogni modo, sotto le ipotesi formulate, la Legge Debole dei Grandi Numeri assicura che, al crescere della numerosità campionaria, la media campionaria, e di conseguenza la varianza campionaria, convergano in probabilità ai corrispondenti valori della

[6] F. W. Bessel, *Über die Wahrscheinlichkeit der Beobachtungsfehler*, Astronomische Nachrichten, vol. 1, 1818, pp. 255–267.

[7] La distribuzione venne descritta nel 1908 da William Sealy Gosset, che pubblicò il suo risultato sotto lo pseudonimo "Student", in quanto la fabbrica di birra Guinness presso la quale era impiegato vietava ai propri dipendenti di pubblicare articoli affinché questi non divulgassero segreti di produzione.
Student [W. S. Gosset], *The probable error of a mean*, Biometrika, vol. 6, n. 1, 1908, pp. 1–25

popolazione. Ciò implica che la varianza campionaria costituisca uno stimatore consistente della varianza della popolazione: al crescere di *n* la stima della varianza si avvicina con probabilità sempre maggiore al reale valore della popolazione. Formalmente:

$$\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2 \text{ per } n \to +\infty$$

(1. 6)

Tuttavia, la deviazione standard campionaria non costituisce uno stimatore consistente della deviazione standard della popolazione a causa della disuguaglianza di Jensen.[8] Quest'ultima afferma che, data una funzione convessa o concava $g(\cdot)$, vale sempre:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \text{ se } g \text{ è convessa,}$$

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]) \text{ se } g \text{ è concava.}$$

(1. 7)

Poiché la radice quadrata è una funzione concava, si ottiene che:

$$\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$$

(1. 8)

Applicando questo risultato al caso della varianza stimata, segue che:

$$\mathbb{E}[\hat{\sigma}_n] \leq \sqrt{\mathbb{E}[\hat{\sigma}_n^2]}$$

(1. 9)

Il che implica che la deviazione standard stimata risulti sistematicamente sottostimata rispetto al valore atteso corretto. Si tratta dunque di uno stimatore distorto (*biased*) verso il basso.

[8] J. L. W. V. Jensen, *Sur les fonctions convexes et les inégalités entre les valeurs moyennes*, Acta Mathematica, 30(1), 1906, pp. 175–193

Sotto le stesse ipotesi formulate precedentemente si può introdurre uno stimatore consistente della deviazione standard della popolazione, l'errore standard (*standard error*), pari al rapporto tra deviazione standard del campione e radice quadrata del numero di osservazioni:

$$SE = \frac{\hat{\sigma}}{\sqrt{n}}$$

(1. 10)

Questa misura permette di quantificare la variabilità delle stime campionarie della deviazione standard, in funzione della dimensione campionaria n. All'aumentare di n, l'errore standard decresce, garantendo maggiore precisione nelle stime.

Tutte le statistiche campionarie menzionate variano da campione a campione e dunque sono variabili aleatorie, mentre le statistiche della popolazione sono dei numeri.

Un'ulteriore questione riguarda l'unità di misura del tempo, in particolare occorre stabilire se il tempo vada misurato in giorni di calendario o in giorni lavorativi. Empiricamente diversi studi hanno osservato che la volatilità è molto più alta quando la borsa è aperta rispetto a quando è chiusa, pertanto nella nostra analisi considereremo solo giorni lavorativi.[9]

Poiché la maggior parte delle variazioni nei rendimenti azionari è attribuibile all'attività di trading, nella pratica professionale si considera convenzionalmente l'anno composto da 252 giorni di contrattazione, corrispondenti al numero medio di giorni di apertura dei mercati finanziari. La volatilità, quando utilizzata per il pricing delle opzioni, viene solitamente espressa in termini annuali e definita come la deviazione standard del tasso di rendimento giornaliero, ipotizzando una composizione continua.

Se infatti si ipotizza che i rendimenti giornalieri siano rappresentati da variabili aleatorie indipendenti e identicamente distribuite, con media μ e varianza σ^2, allora il rendimento

[9] Si possono menzionare a riguardo i seguenti studi:
E. F. Fama, *The Behavior of Stock-Market Prices*, Journal of Business, 38(1), 1965, pp. 34–105.
K. R. French, *Stock Returns and the Weekend Effect*, Journal of Financial Economics, 8(1), 1980, pp. 55–69.
R. Roll, *Orange Juice and Weather*, American Economic Review, 74(5), 1984, pp. 861–880.
K. R. French & R. Roll, *Stock Return Variances: The Arrival of Information and the Reaction of Traders*, Journal of Financial Economics, 17(1), 1986, pp. 5–26.

relativo a un periodo di T giorni, è anch'esso distribuito normalmente con media $\mu_T = \mu_g T$ e varianza $\sigma_T^2 = \sigma_g^2 T$. La deviazione standard relativa a T giorni può essere dunque ottenuta utilizzando un'approssimazione nota come "regola della radice quadrata". Secondo questo criterio la varianza dei rendimenti cresce in maniera proporzionale all'orizzonte temporale, mentre di conseguenza la deviazione standard (ossia la volatilità) aumenta con la radice quadrata di tale orizzonte.

In questo contesto, la volatilità riferita a un certo orizzonte temporale può essere ottenuta a partire da quella giornaliera attraverso la moltiplicazione per la radice quadrata del numero di giorni di contrattazione:

$$\sigma_T = \sigma_g \sqrt{T}$$

(1. 11)

Al contrario per finalità di risk management, il tempo viene di solito misurato in giorni e la volatilità delle attività, di solito quotata come giornaliera, può essere agevolmente ricavata a partire da volatilità relative a orizzonti temporali più lunghi nel seguente modo:

$$\sigma_g = \frac{\sigma_T}{\sqrt{T}}$$

(1. 12)

Nell'analisi finora condotta la volatilità è stata considerata come un parametro costante, ipotesi che equivale a ritenere che i rendimenti dei fattori di mercato (ossia i prezzi) siano caratterizzati da una distribuzione temporalmente stabile. Tale assunzione risulta tuttavia in contrasto con l'evidenza empirica, la quale mostra come le stime della volatilità siano soggette a variazioni nel tempo e come si alternino fasi di elevata turbolenza a periodi di relativa stabilità.

Questa caratteristica si riflette direttamente sulla quantità e qualità dei dati utilizzati nelle stime, incidendo sulla scelta dell'orizzonte temporale e della frequenza di osservazione. Se la volatilità fosse costante, l'accuratezza delle stime crescerebbe proporzionalmente al numero di osservazioni disponibili in un determinato intervallo temporale. Tuttavia, dati troppo remoti potrebbero risultare scarsamente rappresentativi delle condizioni di mercato correnti, riducendo l'affidabilità delle stime.

Per questo motivo, nella prassi operativa si adotta spesso la regola secondo cui l'orizzonte temporale delle stime deve corrispondere a quello dell'analisi da effettuare: ad esempio, la valutazione di un investimento annuale richiede tipicamente l'impiego di un anno di dati storici. L'evidenza empirica, tuttavia, ha messo in luce un fenomeno ricorrente noto come *volatility clustering*, tale per cui i periodi di elevata volatilità tendono a concentrarsi e ad alternarsi a fasi relativamente stabili.

Tale osservazione ha motivato lo sviluppo di modelli che consentono di descrivere volatilità e correlazioni come variabili dinamiche, in grado di evolvere nel tempo. Tra questi, particolare rilievo assumono i modelli della famiglia GARCH (*Generalized AutoRegressive Conditional Heteroscedasticity*), che saranno oggetto di approfondimento nel capitolo successivo.

1.2 Volatilità di portafoglio, rischio specifico e rischio sistematico

Le fluttuazioni nei prezzi azionari possono essere ricondotte, in via generale, a due categorie di fattori.

La prima riguarda eventi e informazioni specifiche dell'impresa, quali, ad esempio, la conquista di nuove quote di mercato o l'introduzione di un'innovazione di prodotto. Tali fonti di rischio vengono definite specifiche, caratteristiche o idiosincratiche, in quanto influenzano esclusivamente la singola impresa o un numero ristretto di società.

La seconda categoria concerne invece notizie e shock che interessano l'intero sistema economico e che, pertanto, hanno ripercussioni su tutte le imprese. Esempi tipici sono le decisioni di politica monetaria di una banca centrale, come la variazione dei tassi di interesse. Questo tipo di rischio viene detto sistematico, non diversificabile o di mercato.

La costruzione di un portafoglio composto da un numero elevato di titoli consente di ridurre l'impatto del rischio specifico, poiché gli effetti idiosincratici tendono a compensarsi reciprocamente. Il rischio sistematico, invece, non può essere eliminato attraverso la diversificazione, poiché coinvolge in maniera congiunta l'intero insieme dei titoli. L'ammontare complessivo di rischio che permane in un portafoglio dipende dunque dal grado di esposizione dei singoli titoli ai fattori comuni di mercato.

Dal punto di vista statistico, all'aumentare del numero di titoli in portafoglio, la volatilità complessiva risulta determinata in misura crescente dalla covarianza media tra i rendimenti azionari, la cui incidenza cresce più che proporzionalmente rispetto a quella delle varianze individuali. Il vantaggio della diversificazione è all'inizio significativo, mentre diminuisce gradualmente all'aumentare delle azioni in portafoglio, fino a convergere ad un certo valore di volatilità, che rappresenta il rischio di mercato non diversificabile.

La rappresentazione grafica della relazione tra volatilità del portafoglio e numero di titoli mostra chiaramente come la diversificazione riduca progressivamente il rischio: la curva della volatilità decresce rapidamente con l'aggiunta dei primi titoli, per poi appiattirsi man mano che la riduzione del rischio specifico si esaurisce e rimane prevalente il rischio sistematico.

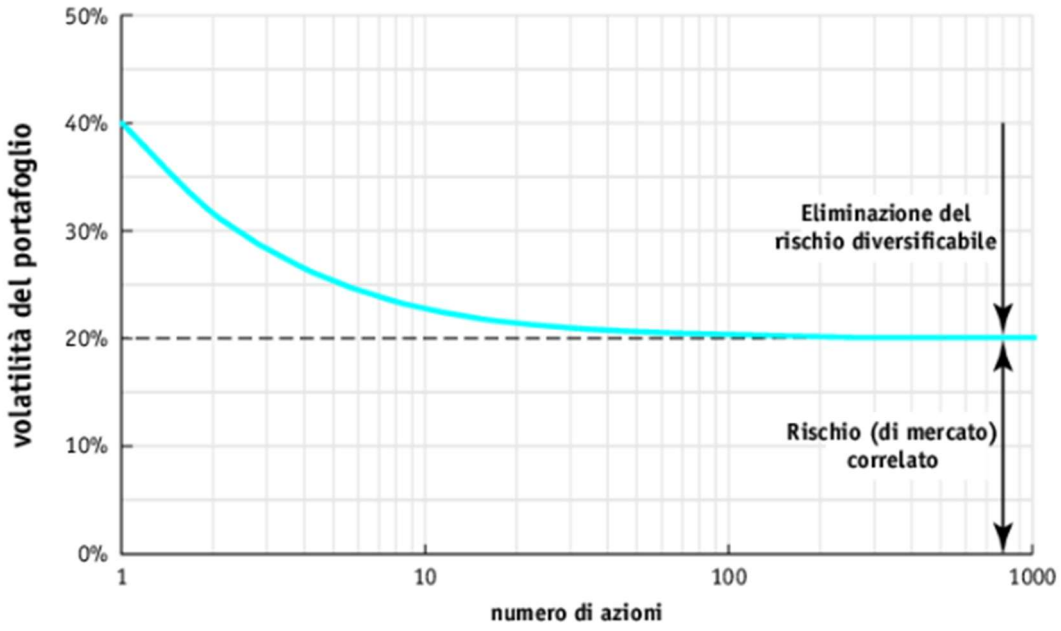

Figura 1.1: Volatilità di un portafoglio al ridursi del rischio diversificabile

Come precedentemente osservato, la combinazione di più titoli all'interno di un portafoglio con pesi positivi, salvo il caso teorico in cui tutte le azioni presentino una perfetta correlazione positiva pari a +1, determina una volatilità complessiva inferiore rispetto alla media ponderata delle volatilità dei singoli titoli. Di conseguenza, la volatilità di un indice azionario ampio, quale lo Standard & Poor's 500, risulta sistematicamente inferiore alla media delle volatilità dei 500 titoli azionari che lo compongono.

Lo S&P 500 è un indice *value weighted*[10], costituito dalle 500 principali società statunitensi per capitalizzazione di mercato corretta per il flottante, calcolata come prodotto tra il prezzo di ciascuna azione e il numero di azioni effettivamente disponibili per la negoziazione. Grazie a questa metodologia, l'indice rispecchia in maniera fedele l'andamento del mercato azionario statunitense e viene comunemente utilizzato come *proxy* del mercato nei modelli di equilibrio.

La sua capacità di includere le società più rilevanti in termini di capitalizzazione consente allo S&P 500 di rappresentare circa l'80% del mercato azionario statunitense, nonostante copra soltanto 500 delle oltre 5000 imprese quotate negli Stati Uniti.

Accanto allo S&P 500, nell'analisi empirica di questo lavoro sarà stimata la volatilità di altri due indici azionari ampi e diversificati, costruiti secondo la metodologia *value weighted*. Per il mercato europeo la scelta è ricaduta sullo STOXX Europe 600, lanciato da STOXX Ltd. nel 1998, che includendo 600 società quotate provenienti da 17 Paesi europei, rappresenta in maniera equilibrata i diversi segmenti di capitalizzazione (*large*, *mid* e *small cap*) e costituisce il benchmark più completo per il mercato azionario europeo. Parallelamente, l'MSCI AC (*All Country*) Asia Pacific, lanciato anch'esso nel 1998 e parte della serie *MSCI All Country Indexes*, sviluppata da *Morgan Stanley Capital International*, comprende circa 1240 titoli provenienti da 15 paesi della regione Asia-Pacifico, includendo sia i mercati sviluppati, come ad esempio il Giappone e l'Australia, sia quelli emergenti della regione. Coprendo approssimativamente l'85 % della

[10] Un'importante distinzione rispetto agli indici *value weighted* è rappresentata dagli indici *price weighted*, nei quali il peso attribuito a ciascun titolo dipende esclusivamente dal suo prezzo, indipendentemente dalla capitalizzazione complessiva. Il più noto esempio è il Dow Jones Industrial Average (DJIA), introdotto nel 1896 da Charles H. Dow, che rappresenta ancora oggi uno degli indici di riferimento del mercato azionario statunitense.

capitalizzazione di mercato *free-float adjusted* di ciascun paese rappresentato risulta essere l'indice più ampio dell'area geografica considerata.

In Italia, l'indice azionario *value weighted* di riferimento è invece il FTSE MIB, che raccoglie i 40 titoli a maggiore capitalizzazione del mercato azionario italiano, rappresentando circa l'80% della capitalizzazione complessiva. Una sua estensione è costituita dal FTSE Italia All-Share, che include anche i segmenti *mid cap* e *small cap*, per un totale di circa 220 società quotate, fornendo così una rappresentazione più completa del mercato domestico.[11]

Le argomentazioni trattate in questo paragrafo sono all'origine della moderna teoria della selezione di portafoglio, proposta da Harry Markowitz[12] in un noto articolo del 1952 e nei lavori correlati di Andrew Roy (1952)[13] e, nel settore assicurativo, di Bruno De Finetti (1940).[14]

Attraverso questi studi è stato introdotto il passaggio dalla logica della valutazione dei singoli titoli a quella della valutazione dei portafogli. Inoltre la *Modern Portfolio Selection* si pone alla base dei contributi in tema di formazione dei prezzi di equilibrio nei mercati finanziari forniti dal Capital Asset Pricing Model (CAPM)[15] e dall'Arbitrage Pricing Theory (APT).[16]

La trattazione della volatilità che abbiamo effettuato in questo paragrafo si riferisce a una logica di asset allocation. A questo punto affrontiamo due ulteriori finalità di interesse per cui è necessario stimare e prevedere la volatilità: il pricing e il risk management.

---

[11] Per approfondimenti sulle metodologie di calcolo degli indici si vedano:
Standard & Poor's, *S&P 500 Index Methodology*, S&P Dow Jones Indices, ultima versione aggiornata.
STOXX Ltd., *STOXX Europe 600 Index Guide*, Qontigo/Deutsche Börse Group, ultima versione aggiornata
MSCI, *MSCI AC Asia Pacific Index Methodology*, MSCI Barra, ultima versione aggiornata.
FTSE Russell, *FTSE MIB Index Series Ground Rules*, ultima versione aggiornata.
FTSE Russell, *FTSE Italia All-Share Index Series Ground Rules*, ultima versione aggiornata

[12] H. M. Markowitz, *Portfolio Selection*, *The Journal of Finance*, 7(1), 1952, pp. 77–91.

[13] A. D. Roy, *Safety First and the Holding of Assets*, *Econometrica*, 20(3), 1952, pp. 431–449.

[14] B. De Finetti, *Il problema dei pieni*, *Giornale dell'Istituto Italiano degli Attuari*, 11, 1940, pp. 1–88.

[15] W. F. Sharpe, *Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk*, *The Journal of Finance*, 19(3), 1964, pp. 425–442.

[16] S. A. Ross, *The Arbitrage Theory of Capital Asset Pricing*, *Journal of Economic Theory*, 13(3), 1976, pp. 341–360.

1.3 Il Pricing

Nel pricing, l'obiettivo è stimare il valore attuale dei flussi di cassa prodotti da un certo strumento finanziario. Per realizzarlo è necessario inizialmente stimare la distribuzione dei prezzi futuri, successivamente calcolare i valori attesi/medi dei pagamenti futuri e attualizzarli.

Qualora il flusso di cassa da valutare non fosse esposto a rischio, cioè esente da qualsiasi tipo di incertezza, il tasso di sconto sarebbe uguale al tasso risk-free, che rappresenta il *price of time*, ispirato al postulato del rendimento del danaro introdotto da Bruno de Finetti.[17]

Se invece il flusso di cassa da valutare è rischioso, dunque esposto a un fattore di incertezza, non sarà noto l'importo effettivamente realizzato. L'individuo che valuta il flusso di cassa attribuirà una distribuzione di probabilità soggettiva alle possibili realizzazioni future (*subjective view*). Occorre definire una legge di equivalenza tra posizioni incerte, che definisca il trade-off tra gli obiettivi contrastanti di rendimento atteso e rischio.

È opportuno distinguere tra valore atteso e valore stimato di un flusso di cassa rischioso.

Il primo, noto anche come previsione o speranza matematica, è una *probability-weighted average of possible values*, cioè un valore medio calcolato su tutti i possibili valori futuri. Il secondo è genericamente riferito a quantificazioni numeriche basate su un ventaglio di ipotesi più limitato, ottenute trascurando alcuni dei valori possibili. Poiché il valore stimato contiene una distorsione (*bias*) rispetto al valore atteso, si può dunque affermare che un valore atteso è un valore stimato con distorsione nulla.

La valutazione di un flusso di cassa rischioso deve tenere conto anche del livello di variabilità del flusso di cassa: il suo valore attuale deve incorporare la remunerazione monetaria del tempo (*price of time*) e deve essere aggiustato per la remunerazione del rischio (*price of risk*), concetto che nasce dall'ipotesi di avversione al rischio degli agenti

[17] B. de Finetti, *Sulle operazioni finanziarie: Note di matematica finanziaria*, Istituto Italiano degli Attuari, Roma, 1935.

economici: un individuo avverso al rischio richiederà una remunerazione per passare da una posizione certa, che viene investita oggi, a una incerta che verrà incassata in futuro.

Per un agente avverso al rischio un'attività rischiosa ha un valore minore del suo valore atteso, mentre una passività rischiosa ha un valore maggiore.

In letteratura vengono generalmente distinti tre principali approcci di valutazione:

1. DCE/RL: attualizzazione dell'equivalente certo con caricamento esplicito del rischio.

Questo approccio prevede che il premio per il rischio sia calcolato ed esplicitato in termini monetari, piuttosto che incorporato implicitamente nel valore atteso, come avviene negli altri due metodi. In caso di individuo avverso al rischio il premio sarà positivo, mentre sarà nullo nel caso di un individuo neutrale al rischio. Questo approccio risulta coerente con il modello di Markowitz, in quanto si ipotizza che l'individuo decisore sia avverso al rischio e in particolare dotato di una funzione di utilità quadratica, ed è tipicamente utilizzato quando il flusso di cassa rischioso non risulta replicabile tramite attività finanziarie negoziate sui mercati (*unhedgeable*).

2. RAD: attualizzazione al tasso rischioso.

In questo metodo il premio per il rischio è incorporato direttamente nel tasso di sconto applicato per scontare i flussi di cassa futuri. Considerando un agente avverso al rischio, nel caso di un'attività il tasso aggiustato per il rischio dovrà includere un'adeguata maggiorazione rispetto al tasso privo di rischio, cioè il premio. All'aumentare della rischiosità del flusso di cassa cresce il tasso di attualizzazione corretto per il rischio, con conseguente riduzione del valore attuale stimato.

Viceversa, nel caso di una passività, il tasso aggiustato per il rischio dovrà includere un'adeguata riduzione (sconto) del tasso privo di rischio (relativo al periodo di differimento considerato). I premi al rischio e dunque i tassi aggiustati per il rischio, sono *entity specific*, cioè anche a parità di rischio percepito, dipendono dal livello di avversione al rischio caratteristico delle parti coinvolte nella valutazione. In una valutazione che adotti la prospettiva dei partecipanti al mercato, andrà fatto il miglior sforzo per garantire alle stime dei premi al rischio livelli adeguati di oggettività, utilizzando opportuni modelli equilibrio di mercato, di cui l'esempio più celebre è sicuramente il Capital Asset Pricing Model.

3. DCE/RN: attualizzazione dell'equivalente certo con risk loading implicito.

La caratteristica chiave di quest'ultimo metodo, tipico dei modelli stocastici di valutazione delle opzioni, è di considerare tutti gli agenti economici come neutrali al rischio. Il flusso di cassa da valutare è il payoff di un'opzione finanziaria, o più in generale di un derivato, e la sua valutazione ha la finalità specifica di determinare un prezzo di mercato coerente con l'assenza di opportunità di arbitraggio. Una volta scelto il modello di *option pricing*, calibrando il modello sui dati correnti di mercato, come ad esempio prezzi tassi e volatilità, possiamo individuare la struttura della probabilità che utilizziamo per calcolare il valore atteso del flusso di cassa. Queste sono denominate *risk neutral* e ci consentono di affermare che l'equivalente certo del payoff può essere inteso come un equivalente certo di mercato, uguale per tutti gli individui.

Il presupposto per l'utilizzo di questo metodo è che il flusso di cassa da valutare sia *hedgeable*, ossia replicabile con un portafoglio di prodotti quotati sul mercato e come hanno mostrato Black and Scholes, i payoff dei contratti derivati risultano, almeno in linea teorica, replicabili

In questo approccio, il portafoglio replicante è di solito costruibile solo sotto stringenti ipotesi di perfezione del mercato. Inoltre si tratta di un portafoglio dinamico, in quanto la sua struttura deve essere costantemente ribilanciata nel tempo. Si realizza dunque la cosiddetta *dynamic hedging strategy*.

Il portafoglio replicante è generalmente individuato sulla base di un modello stocastico di pricing e dunque possiamo affermare che attraverso il metodo DCE/RN si realizza una valutazione *marked-to-model*. Se, almeno dal punto di vista teorico, viene risolta la soggettività dell'avversione al rischio individuale, restano margini di discrezionalità nella valutazione, dovuti sia alla scelta del modello che alle tecniche statistiche e ai dati utilizzati per la sua calibratura.

In questo caso, rispetto al metodo DCE/RL, si ha un risk loading implicito, il cui segno è determinato dal valore delle probabilità risk neutral. Nel mondo risk neutral, il tasso di rendimento atteso di tutti i titoli è uguale al tasso di interesse privo di rischio r e inoltre, il valore attuale di ogni futuro pagamento si calcola attualizzandone il valore atteso in base al tasso privo di rischio.

Quando passiamo dal mondo risk neutral al mondo reale, che ipotizziamo essere popolato da individui avversi al rischio, varia (in caso di avversione al rischio aumenta) sia il tasso di interesse che utilizziamo per attualizzare il valore atteso, sia il suo stesso valore atteso. Per riequilibrare la relazione nel mondo reale è necessario ridurre il valore atteso del derivato e ciò implica un rialzo delle probabilità effettive (*P-measure*) utilizzate per calcolarlo, che risulteranno appunto più elevate se paragonate alle rispettive probabilità risk neutral (*Q-measure*). Il punto chiave è che quando valutiamo un derivato, ne calcoliamo il valore in termini relativi, rispetto all'attività sottostante, ed è proprio il prezzo di quest'ultima che ingloba il trade-off tra rischio e rendimento, che si riflette direttamente nel prezzo del derivato. Una diretta conseguenza è che la volatilità rimane invariata tra mondo reale e mondo neutrale al rischio, come dimostrato dal Teorema di Girsanov.[18]

Se gli investitori richiedessero un tasso di rendimento più elevato, in funzione delle loro percezioni circa il rischio che si assumono, il prezzo del sottostante scenderebbe, e viceversa nel caso contrario. Dunque il passaggio dal problema della valutazione, ad esempio di un derivato (mondo risk neutral) a quello delle analisi di scenario (mondo naturale o reale), su cui si basa il risk management, avviene cambiando la misura di probabilità, tra proiezioni neutrali al rischio, utilizzate per la valutazione dei derivati, e le proiezioni nel mondo reale, adottate invece per le analisi di scenario.

[18] I. V. Girsanov, On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. Theory of Probability & Its Applications, 5(3), 1960, pp. 285–301.

1.4 Il Risk Management

Nell'ambito del risk management, l'analisi si concentra sul mondo reale, in contrapposizione al mondo neutrale rispetto al rischio. In questo caso l'obiettivo non è la valutazione, pertanto i futuri pagamenti non vengono attualizzati, in quanto non vi è interesse per i futuri eventi in un mondo ipotetico in cui tutti sono neutrali verso il rischio.

Tuttavia le analisi di scenario richiedono talvolta che venga utilizzato non solo il mondo reale, per generare gli scenari fino alla fine dell'orizzonte temporale prescelto, ma anche il mondo neutrale verso il rischio, per valutare i contratti che a tale data sono ancora in essere. L'esigenza di misurare e controllare in modo adeguato i rischi assunti da una società, e in particolare per la sua natura, da una banca, è particolarmente rilevante nell'attività di investimento e negoziazione di titoli, che risulta esposta alla volatilità dei prezzi delle attività scambiate. Per le istituzioni che assumono posizioni speculative in valute, obbligazioni o azioni, esiste infatti una concreta possibilità che le perdite associate a una singola posizione annullino, nell'arco di un breve intervallo temporale, i profitti realizzati nel corso di periodi ben più lunghi.

I rischi di mercato vengono generalmente identificati, anche dalle autorità di vigilanza, con i rischi inerenti il solo portafoglio di negoziazione (*trading book*), inteso come l'insieme di posizioni assunte per un periodo di tempo breve o brevissimo, con l'obiettivo di trarre profitto dalle variazioni dei prezzi di mercato; in realtà, essi riguardano tutte le attività/passività finanziarie detenute da una banca, comprese quelle acquistate per finalità di investimento, e destinate a essere conservate in bilancio per un lungo arco di tempo (*banking book*).[19]

I rischi di mercato sono venuti assumendo, nell'ambito dei mercati finanziari internazionali, una rilevanza crescente nel corso dell'ultimo decennio, principalmente a causa dei seguenti fenomeni: il processo di *securitization* che ha portato alla progressiva sostituzione di attività illiquide (prestiti, mutui) con attività dotate di un mercato

[19] Tale distinzione fra portafoglio di negoziazione e altre attività/passività, per quanto artificiale, risponde alla classificazione introdotta dalla direttiva CEE n. 93/6 e dalle proposte del Comitato di Basilea dell'aprile 1993, entrambe relative all'estensione dei coefficienti patrimoniali ai rischi di mercato.
Consiglio delle Comunità Europee, *Direttiva 93/6/CEE del Consiglio, del 15 marzo 1993, sull'adeguatezza patrimoniale delle imprese di investimento e degli enti creditizi.*
Comitato di Basilea 1993: *Comitato di Basilea per la Vigilanza Bancaria, The Supervisory Treatment of Market Risks, aprile 1993*.

secondario liquido e dunque di un prezzo; la progressiva crescita del mercato degli strumenti finanziari derivati, il cui principale profilo di rischio per gli intermediari finanziari che li negoziano è rappresentato dalla variazione del relativo valore di mercato causata da variazioni dei prezzi delle attività sottostanti e/o dalle condizioni di volatilità degli stessi; la crescente diffusione di nuovi standard contabili, come ad esempio l'IFRS 9,[20] che prevedono l'iscrizione in bilancio del valore di mercato (e non più del costo storico di acquisto) per una vasta gamma di attività e passività finanziarie. Questi nuovi standard hanno contribuito a rendere maggiormente visibili gli effetti del rischio di mercato, accentuandone l'importanza nella letteratura accademica e regolamentare.[21]

La crescente attenzione ai rischi di mercato non ha riguardato esclusivamente gli intermediari finanziari e il mondo accademico, ma si è naturalmente estesa anche alle autorità di vigilanza.

Già a partire dall'aprile del 1993 il Comitato di Basilea per la Vigilanza Bancaria ha formulato proposte volte all'estensione dei requisiti patrimoniali a copertura del rischio di mercato. Analoga raccomandazione è giunta dalla direttiva comunitaria n. 93/6 del marzo 1993 relativa all'adeguatezza patrimoniale delle imprese di investimento e degli enti creditizi. Tali proposte sono state successivamente recepite dalle autorità di vigilanza dei principali paesi economicamente sviluppati. Successivamente, il primo emendamento al Concordato di Basilea (1996)[22] ha introdotto requisiti patrimoniali specifici per il rischio di mercato, ammettendo l'uso di modelli interni (VaR). Con Basilea II (2004)[23] sono stati rafforzati i controlli, introducendo pratiche di *backtesting* e *stress testing* per verificare la robustezza dei modelli interni. Dopo la crisi del 2008, Basilea 2.5[24] ha imposto misure come il VaR stressato e il cosiddetto VaR stressato e l'*Incremental Risk Charge* (IRC).

---

[20] International Accounting Standards Board (IASB), *International Financial Reporting Standard 9 – Financial Instruments (IFRS 9)*, 2014, ultima versione aggiornata.

[21] A. Sironi, A. Resti, *Rischio e valore nelle banche. Misura, regolamentazione, gestione*, Milano: Egea, ristampa aggiornata a gennaio 2021, cap. 6, p. 143.

[22] Comitato di Basilea per la Vigilanza Bancaria, *Amendment to the Capital Accord to Incorporate Market Risks*, gennaio 1996.

[23] Comitato di Basilea per la Vigilanza Bancaria, *International Convergence of Capital Measurement and Capital Standards: A Revised Framework (Basilea II)*, giugno 2004.

[24] Comitato di Basilea per la Vigilanza Bancaria, *Revisions to the Basel II Market Risk Framework (Basilea 2.5)*, luglio 2009.

Basilea III[25] ha successivamente rafforzato i requisiti qualitativi del capitale e introdotto i *buffer*. Più recentemente, il FRTB[26] (*Fundamental Review of the Trading Book*), nell'ambito delle più recenti riforme di Basilea III, talvolta indicate come Basilea IV), ha rivoluzionato l'approccio al portafoglio di negoziazione delle banche, secondo la quale il capitale regolamentare a fronte dei rischi di mercato va determinato in base all' Expected Shortfall al 97,5% invece che al VaR al 99%.

1.4.1 Value at Risk (VaR) ed Expected Shortfall (ES)

Nel risk management, gli obiettivi principali sono l'identificazione, la misurazione e la gestione del rischio. In tale contesto, un aspetto centrale riguarda la misurazione dell'esposizione al rischio e la conseguente quantificazione della massima perdita potenziale in un determinato orizzonte temporale. A tal fine è necessario condurre analisi di scenario, esplorando l'insieme delle possibili situazioni che potrebbero verificarsi in una certa data futura, ordinare queste previsioni secondo una distribuzione di profitti e perdite e fissare un livello di confidenza, statisticamente un quantile, sotto il quale siamo interessati a non scendere.

Questa è la logica del VaR (*Value at Risk*), una famiglia di modelli che permette di quantificare, confrontare e aggregare il rischio connesso a posizioni e portafogli differenti. In precedenza si ricorreva a misure di rischio specifiche per le diverse tipologie di posizioni, come ad esempio la *duration* e il *basis point value* per i titoli obbligazionari, il *beta* per i titoli azionari, le *greche* per le opzioni. L'utilizzo delle misure di sensibilità presentava tuttavia diversi limiti: in particolare posizioni di natura diversa erano quantificate con misure diverse, impedendo di confrontare e aggregare tra loro i rischi assunti in diverse aree dell'attività di negoziazione, ostacolando sia la comunicazione orizzontale tra le diverse aree, sia quella verticale, con il top management della banca. Inoltre, l'utilizzo di queste misure non prendeva in considerazione il diverso grado di volatilità e di correlazione dei fattori di rischio.

[25] Comitato di Basilea per la Vigilanza Bancaria, *Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems*, dicembre 2010 (rev. giugno 2011).
[26] Comitato di Basilea per la Vigilanza Bancaria, *Minimum Capital Requirements for Market Risk (Fundamental Review of the Trading Book – FRTB)*, gennaio 2016 (rev. gennaio 2019).

Una delle prime istituzioni a sviluppare un modello VaR e la prima a renderlo pubblico è stata la banca statunitense J.P. Morgan, autrice del modello RiskMetrics.[27] Nel 1998 il gruppo dei ricercatori autori del modello ha creato una società indipendente, in parte controllata dalla stessa J.P. Morgan, la quale ha assunto la stessa denominazione del modello (RiskMetrics Group).

Alla fine degli anni ottanta l'allora presidente della banca, Dennis Weatherstone, chiese di ricevere un'informazione sintetica, racchiusa in un singolo valore monetario, relativa ai rischi di mercato dell'intera banca nei diversi segmenti di mercato (azionario, obbligazionario, valute, derivati, commodity ecc.) e nelle diverse aree geografiche di operatività della banca. Nacque così il VaR, una misura della massima perdita che una posizione o un portafoglio può subire, dato un certo livello di confidenza, entro un predeterminato orizzonte temporale.

Il VaR rappresenta dunque una misura di tipo probabilistico, ed essendo funzione di due parametri, l'orizzonte temporale (in giorni) e il livello di confidenza, assume valori diversi in corrispondenza di differenti livelli di queste variabili.

Indicando con L la perdita sull'orizzonte temporale prescelto, con $\mathrm{P}(L > VaR)$ la probabilità che la perdita superi il VaR e con X il livello di confidenza, si ha:

$$\mathrm{P}(L > VaR) = 1 - X \qquad (1.\ 13)$$

La scelta dell'orizzonte temporale appropriato dipende dalla rapidità con cui un portafoglio può essere liquidato. A tal proposito nel 1996 il Comitato di Basilea, con il primo emendamento al Concordato di Basilea, attuato nel 1998, ha imposto alle banche l'obbligo di detenere capitale a fronte non solo dei rischi di credito, ma anche dei rischi di mercato. In particolare, l'Emendamento del 1996 ha distinto tra *banking book* e *trading book*, stabilendo che le banche calcolassero i requisiti patrimoniali relativi ai rischi di mercato utilizzando il *Value at Risk* (VaR) con un orizzonte temporale di dieci giorni e un livello di confidenza (X) pari al 99%.

[27] J.P. Morgan/Reuters, *RiskMetrics — Technical Document*, New York: J.P. Morgan, 1996 (seconda edizione).

Il calcolo del VaR può essere condotto attraverso diverse metodologie; le più diffuse sono le seguenti:

1. Approccio parametrico
2. Metodo delle simulazioni storiche
3. Metodo delle simulazioni mediante il metodo Montecarlo

Il metodo delle simulazioni storiche si basa sull'impiego diretto delle serie storiche dei rendimenti: ciò consente di evitare ipotesi distributive teoriche e di stimare la perdita potenziale a un determinato livello di confidenza ricorrendo esclusivamente ai dati effettivamente osservati.

Il metodo delle simulazioni Montecarlo, invece, genera un elevato numero di scenari ipotetici al fine di riprodurre l'evoluzione stocastica dei fattori di mercato, consentendo così di stimare la distribuzione dei rendimenti di portafoglio.

Tuttavia per la nostra trattazione, rileva il primo approccio, in quanto è l'unico che richiede una stima della volatilità dei rendimenti dei fattori di mercato. Tale procedura, nota come metodo delle varianze-covarianze, si fonda sull'ipotesi, frequentemente contestata, che le variazioni di valore dei fattori di mercato seguano una distribuzione normale caratterizzata da media μ e deviazione standard σ.

In tale contesto, il calcolo del VaR è espresso dalla seguente equazione

$$VaR = \mu + \sigma N^{-1}(X)$$

(1. 14)

dove $N^{-1}(X)$, talvolta indicato con z_α rappresenta la funzione inversa della distribuzione cumulata di una distribuzione normale standard, cioè il quantile associato al livello di confidenza X (rispettivamente $1 - \alpha$).

Utilizzare la funzione di ripartizione normale standard è vantaggioso in quanto permette una corrispondenza biunivoca tra diversi valori di X e i corrispondenti livelli della probabilità, che resta valida indipendentemente dai valori assunti dalla media e dalla deviazione standard della variabile considerata. Per brevità, la probabilità associata a

valori della normale standard inferiori o uguali a $N^{-1}(X)$, cioè $N(X)$, viene spesso indicata semplicemente con *X*.

La possibilità di associare intervalli di confidenza a multipli della deviazione standard non è un risultato esclusivo della distribuzione normale. Per ogni variabile aleatoria, discreta o continua, con media μ e deviazione standard σ finite, vale infatti la disuguaglianza di Chebyshev[28], secondo cui:

$$P[|r - \mu| > \alpha\sigma] \leq \frac{1}{\alpha^2}$$

(1. 15)

Naturalmente gli intervalli che si ottengono ipotizzando la normalità sono più informativi rispetto a quelli derivanti dalla disuguaglianza di Chebyshev.

Per determinare il VaR con questo approccio, si suppone di solito che il tasso di variazione atteso delle variabili di mercato (μ) sia nullo. Come vedremo l'ipotesi è ragionevole, in quanto, in un breve intervallo di tempo, il valore atteso è piccolo rispetto alla deviazione standard e ciò rispetta le ipotesi di efficienza dei mercati. Nella pratica, quando si calcola il VaR per misurare i rischi di mercato, spesso si utilizza un orizzonte giornaliero, poiché di norma non esistono dati sufficienti per stimare il comportamento delle variabili di mercato per periodi più lunghi di un giorno.

Le variazioni di valore delle posizioni azionarie in portafoglio sono generalmente derivate da quelle dei fattori di rischio mediante coefficienti di sensibilità lineari, che esprimono la reattività del valore di mercato delle posizioni rispetto alle variazioni dei fattori stessi.

Tale impostazione definisce l'approccio *delta-normal*, il quale assume che i fattori di rischio seguano una distribuzione normale e che il legame con le posizioni di portafoglio sia di natura lineare. Ne consegue che anche le variazioni complessive del portafoglio risulteranno distribuite normalmente, consentendo il calcolo del VaR come multiplo della deviazione standard della distribuzione stimata.

[28] P. L. Chebyshev, "Des valeurs moyennes," *Journal de Mathématiques Pures et Appliquées*, 12, 1867, pp. 177–184.

Un'alternativa a tale metodo è rappresentata dall'approccio *asset-normal*, che utilizza come fattori di rischio le variazioni logaritmiche dei prezzi degli strumenti finanziari inclusi nel portafoglio, imponendo che esse siano distribuite normalmente. Ciò equivale ad assumere che i prezzi seguano una distribuzione log-normale. Questo approccio, introdotto originariamente da J.P. Morgan nella prima versione del modello RiskMetrics, fornisce una rappresentazione alternativa ma coerente con l'ipotesi di normalità dei rendimenti. Quando le variazioni di valore del portafoglio in giorni successivi, cioè i rendimenti giornalieri, sono variabili casuali normali identiche, indipendenti, a media nulla, si può applicare nuovamente la regola della radice, in quanto il VaR è concepito come un multiplo della deviazione standard:

$$VaR_T = VaR_g\sqrt{T}$$

(1. 16)

In tutti gli altri casi, tale formulazione deve essere intesa come una prima approssimazione.

Nel caso di un portafoglio, oltre alla stima delle varianze, è necessario considerare anche le stime delle covarianze o delle correlazioni tra le diverse posizioni. I metodi di stima più utilizzati per le covarianze sono gli stessi delle varianze e saranno trattati nel prossimo capitolo.

È opportuno osservare che la definizione di Value at Risk ammette la possibilità che si verifichino perdite superiori al livello stesso del VaR, con probabilità pari a $\alpha = 1 - X$. Nel caso in cui tale evento si realizzi, tuttavia, il modello non fornisce alcuna indicazione circa l'entità della perdita eccedente. Le due figure seguenti permettono di visualizzare questo limite in modo immediato.

Nella prima figura (1.2) il VaR individua la soglia di perdita corrispondente al livello di confidenza stabilito. Nella seconda figura (1.3), pur mantenendo invariato il valore del VaR, la distribuzione presenta una coda sinistra molto più pronunciata: ciò implica che rispetto al primo caso, la probabilità di incorrere in perdite gravi (cioè oltre il livello del VaR), sia ben più elevata. Questo esempio evidenzia come il VaR, pur rappresentando un utile indicatore di rischio, non sia in grado di catturare l'effettiva esposizione a eventi

estremi e lasci scoperta l'informazione sulla dimensione delle perdite oltre la soglia fissata.

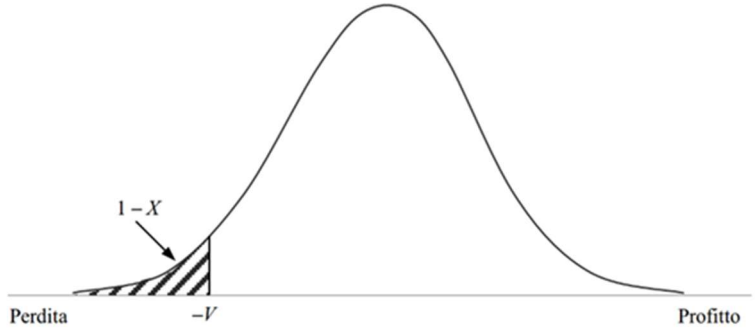

Figura 1.2: VaR sotto ipotesi di normalità delle variazioni di valore

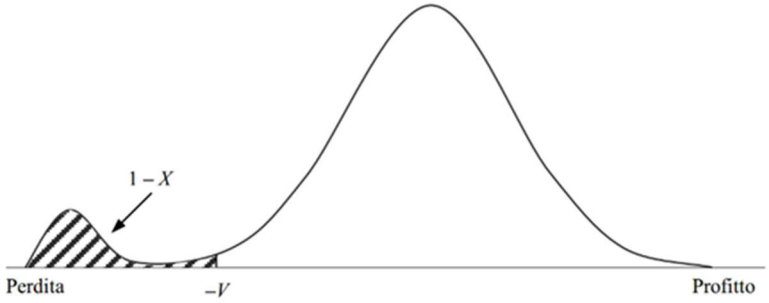

Figura 1.3: Perdite estreme eccedenti a parità di VaR

Per ovviare a tale limite è stata introdotta una misura di rischio alternativa, l'*Expected Shortfall* (ES), talvolta indicato anche come *Conditional VaR* (C-VaR) o *tail loss*, che rappresenta la perdita media attesa nei prossimi T giorni, ipotizzando di trovarci nella coda della distribuzione, a sinistra del quantile rappresentato dal VaR.[29]

Recuperando l'equazione con cui abbiamo definito il VaR (1.14), possiamo esprimere l'Expected Shortfall nel seguente modo:

$$ES = \mathbb{E}[L|L > VaR]$$

(1. 17)

[29] Per degli approfondimenti sull'Expected Shortfall si vedano:
P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, *Coherent Measures of Risk*, *Mathematical Finance*, 9(3), 1999, pp. 203-228.
C. Acerbi, D. Tasche, *Expected Shortfall: A Natural Coherent Alternative to Value at Risk*, *Economic Notes*, 31(2), 2002, pp. 379-388

Questa misura di rischio estende il concetto di VaR poiché non si limita a considerare la soglia critica di perdita, ma tiene conto anche della dimensione media delle perdite che si verificano oltre tale soglia. L'Expected Shortfall può essere calcolato nelle stesse tre modalità che abbiamo menzionato per il VaR. Nel caso dell'approccio parametrico, ipotizzando che le variazioni giornaliere del valore del portafoglio si distribuiscano in modo normale, valgono le seguenti relazioni:

$$ES = \mu + \sigma \frac{e^{-\frac{[N^{-1}(X)]^2}{2}}}{(1-X)\sqrt{2\pi}}$$

(1. 18)

$$ES_T = ES_g\sqrt{T}$$

(1. 19)

Il VaR, a differenza dell'ES, non è una misura di rischio coerente in quanto se calcolato con il metodo delle simulazioni (sia storiche che Montecarlo) potrebbe non rispettare la condizione di subadditività, cioè che la misura di rischio di un portafoglio risultante dalla fusione di altri due portafogli o titoli non deve essere maggiore delle misure di rischio relative ai due portafogli o titoli originari. Questa proprietà risulta invece verificata quando il VaR viene calcolato mediante l'approccio parametrico.

I modelli di misurazione del valore a rischio non si esauriscono in una singola metodologia, ma piuttosto comprendono una famiglia di tecniche finalizzate a conseguire i seguenti tre obiettivi:

1. Definire i fattori di rischio che possono influenzare il valore del portafoglio della banca e rappresentarne la possibile evoluzione futura assegnando loro una distribuzione di probabilità;

2. Costruire la distribuzione di probabilità dei possibili valori futuri del portafoglio della banca associata a ognuno dei possibili valori assunti dai fattori di rischio;

3. Sintetizzare la distribuzione di probabilità dei possibili valori futuri del portafoglio della banca in una o più misure di rischio, così da renderle comprensibili al *top management*.

1.4.2 Asimmetria negativa e leptocurtosi

Come abbiamo già accennato, l'ipotesi di normalità della distribuzione dei rendimenti è spesso rigettata empiricamente, a causa dei fenomeni di leptocurtosi e *negative skewness*.

L'asimmetria negativa si riferisce alla maggiore presenza di osservazioni nella parte a sinistra della media rispetto che al lato destro. Matematicamente ciò si traduce nella negatività dell'indice di asimmetria (S), che si basa sul momento terzo centrato della distribuzione:

$$S = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^3}{\hat{\sigma}^3(n-1)}$$

(1. 20)

La leptocurtosi invece indica che le distribuzioni empiriche dei rendimenti delle attività finanziarie presentano una punta più alta e delle code più spesse di quelle proprie di una distribuzione normale (*fat-tailness*).[30] Analiticamente ciò implica che il momento quarto della distribuzione, cioè la curtosi (K), sia superiore a 3, o analogamente, che l'indice di curtosi in eccesso (*excess kurtosis*, $EK = K - 3$) sia positivo.

$$K = \frac{\sum_{i=1}^{n}(r_i - \bar{r})}{\hat{\sigma}^4(n-1)}$$

(1.21)

Ne consegue che in presenza di leptocurtosi, la probabilità di osservare sia rendimenti molto vicini alla media che rendimenti estremi è maggiore rispetto a quanto suggerirebbe una distribuzione normale con uguale media e uguale deviazione standard, rischiando dunque di sottovalutare la frequenza e l'entità dei *crash*.

La presenza di asimmetria negativa e di leptocurtosi ha un'ampia rilevanza nel risk management, in quanto indica che è più probabile osservare rendimenti negativi nell'estremo sinistro della distribuzione, cioè fortemente inferiori alla media, che non

[30] Per un esame del problema delle *fat tails* si vedano:
L.K. Chew, *Fundamentals of Probability Theory for Finance*. World Scientific, 1994, pp. 63–70.
D. Duffie, J. Pan, An Overview of Value at Risk. *Journal of Derivatives*, 4(3), 1997, pp. 7–49.

nell'estremo destro, quindi rendimenti positivi e fortemente superiori alla media. Ciò implica che perdite particolarmente elevate si verificano più frequentemente di quanto implicito in una distribuzione normale. Di conseguenza la probabilità di conseguire perdite superiori al VaR parametrico calcolato, per esempio, con livello di confidenza del 99% è in realtà superiore all'1%.

Tuttavia, anche se la distribuzione dei rendimenti di singoli fattori di mercato non fosse normale, i rendimenti di un portafoglio diversificato il cui valore dipende da un numero elevato di fattori di mercato fra loro indipendenti sono comunque distribuiti secondo una normale per il teorema del limite centrale. L'evidenza empirica mostra però che i fattori di mercato tendono a non essere indipendenti, ma a muoversi in modo correlato proprio in caso di forti perdite dovute a eventi catastrofici.

La leptocurtosi è una delle ragioni principali per l'adozione di ipotesi alternative rispetto alla distribuzione normale, che può essere sostituita ad esempio con la t di Student. Questa presenta media zero, varianza unitaria, interamente definita da un parametro *v*, denominato *gradi di libertà*, che controlla il grado di leptocurtosi della distribuzione. Quanto minori sono i gradi di libertà, tanto maggiore è lo spessore delle code. Al crescere di *v*, la distribuzione t di Student converge verso la distribuzione normale standard.

Questa distribuzione è caratterizzata da code più spesse rispetto a quelle di una normale, riflettendo più adeguatamente la probabilità associata a movimenti estremi dei fattori di mercato e conducendo spesso a una migliore approssimazione dei movimenti del mercato.[31] Ipotizzando che le variazioni di valore dei fattori di mercato si distribuiscano secondo una t di Student con *v* gradi di libertà, le espressioni analitiche del VaR (1.14) e dell'ES (1.18) diventano rispettivamente:

$$VaR = \mu + \sigma t_v^{-1}(X) \qquad (1.22)$$

[31] Si vedano a tal proposito:
R. Blattberg, N. Gonedes, A Comparison of the Stable and Student Distributions as Statistical Models for Stock Prices. *Journal of Business*, 47(2), 1974, pp. 244–280.
R.J. Rogalski, D.E. Vinso, Empirical Properties of Foreign Exchange Rates. *Journal of International Economics*, 8(3), 1978, pp. 379–396.
T. Wilson, Value at Risk. *Risk Magazine*, 6(10), 1993, pp. 111–117

$$ES = \mu - \sigma \frac{f_{t,\nu}[t_\nu^{-1}(X)]}{1-X} \cdot \frac{\nu + [t_\nu^{-1}(X)]^2}{\nu - 1}$$

(1.23)

dove $t_\nu^{-1}(X)$ è il quantile della distribuzione t di Student con *v* gradi di libertà, corrispondente al livello di confidenza *X*, mentre $f_{t,\nu}(\cdot)$ è la funzione di densità della t di Student con *v* gradi di libertà.

Un approccio alternativo consiste nel correggere le misure di VaR parametrico basate sulla distribuzione normale, introducendo aggiustamenti per tenere conto della *skewness* e della curtosi della distribuzione empirica dei rendimenti.

Tale approccio si basa sulla metodologia originariamente proposta da Cornish e Fisher[32] e richiede che il percentile $z_\alpha = N^{-1}(X)$, utilizzato nel calcolo del VaR e basato sulla normale standard venga corretto come segue:

$$z_\alpha^* = z_\alpha + \frac{1}{6}(z_\alpha^2 - 1)S + \frac{1}{24}(z_\alpha^3 - 3z_\alpha)EK - \frac{1}{36}(2z_\alpha^3 - 5z_\alpha)S^2$$

(1.24)

dove *S* indica l'indice di asimmetria, mentre *EK* la curtosi in eccesso.

Tuttavia, è bene affermare in conclusione che la correzione di Cornish e Fisher risulta essere un'approssimazione valida solo quando la reale distribuzione dei rendimenti non si discosta eccessivamente da una distribuzione normale.[33]

Le considerazioni sin qui sviluppate evidenziano la centralità della volatilità nei mercati finanziari e la necessità di modelli in grado di descriverne l'evoluzione dinamica. Nel capitolo successivo analizzeremo quindi i principali approcci econometrici alla stima della volatilità, partendo dalle misure basate su dati impliciti e storici, tra cui i modelli GARCH.

[32] E.A. Cornish, R.A. Fisher, The Percentile Points of Distributions Having Known Cumulants. *Proceedings of the Cambridge Philosophical Society*, 34(2), 1937, pp. 335–341.
[33] Per maggiori dettagli si veda: K. Dowd, *Measuring Market Risk*, John Wiley & Sons, 2002.

Capitolo 2: Metodi di stima della volatilità

La stima della volatilità dei rendimenti dei fattori di mercato, rappresenta un problema complesso e di grande rilevanza sia in ambito teorico che empirico. La letteratura ha affrontato tale tema in più occasioni, proponendo approcci metodologici differenti che, in linea generale, possono essere ricondotti a due categorie principali, distinte sul piano logico. La prima categoria comprende i modelli che si basano sull'analisi storica dei dati di volatilità e di correlazione, utilizzati per formulare previsioni circa il loro andamento futuro. I modelli più elementari di questa classe assumono volatilità e correlazioni come parametri costanti. Tale ipotesi, tuttavia, è stata progressivamente superata, in quanto l'evidenza empirica mostra come la volatilità sia caratterizzate dall'alternanza di fasi di elevata turbolenza e periodi di relativa stabilità. In questo contesto, la volatilità storica riflette il comportamento osservato del mercato e rappresenta pertanto una misura di tipo *backward looking*, fondata sull'assunto che l'analisi del passato possa costituire, uno strumento utile per l'interpretazione e la previsione del futuro. Un secondo approccio alla stima della volatilità e delle correlazioni consiste invece nell'utilizzo delle informazioni implicite nei prezzi delle opzioni. In questo caso, l'orizzonte temporale della volatilità stimata coincide con la vita residua dell'opzione stessa. La cosiddetta volatilità implicita riflette le aspettative del mercato riguardo all'evoluzione futura della volatilità e costituisce, quindi, una misura interamente *forward looking*.

Oltre a questa distinzione di natura metodologica, è opportuno considerare anche le diverse finalità d'uso delle misure di volatilità. In generale, la volatilità storica trova maggiore applicazione nell'ambito del *risk management*, mentre la volatilità implicita è impiegata prevalentemente ai fini di *pricing*. Ciò non esclude, tuttavia, utilizzi incrociati: la volatilità storica può infatti essere applicata anche a modelli di valutazione dei derivati, come nel caso dei modelli GARCH[34], mentre la volatilità implicita è frequentemente utilizzata anche a fini di gestione del rischio, in quanto svolge il ruolo di indicatore del *sentiment* di mercato. Un esempio può essere l'utilizzo dell'indice VIX, basato sulla volatilità implicite delle opzioni scritte sull'S&P 500, per anticipare possibili fasi di turbolenza nei mercati finanziari.

[34] Si veda a tal proposito: Peter F. Christoffersen, Kris Jacobs, "Which GARCH Model for Option Valuation?", *Management Science*, 50(9), 2004, pp. 1204–1221

2.1 La volatilità implicita

La volatilità implicita rappresenta un concetto di fondamentale rilevanza nell'ambito della teoria delle opzioni e inoltre si pone come diretta conseguenza del *framework* introdotto nell'approccio DCE/RN.

Una volta specificato un modello di pricing per le opzioni, la volatilità implicita, anche nota come IV (*implied volatility*) è definita come il valore che equipara il prezzo teorico dell'opzione a quello effettivamente osservato sul mercato.

Il modello maggiormente impiegato per la stima della volatilità implicita è il modello di Black-Scholes-Merton (BSM), sviluppato per opzioni europee su azioni prive di dividendi.[35] Le formule di valutazione di una *call* europea (*c*) e di una *put* europea (*p*), presentano rispettivamente le seguenti formulazioni:

$$c = S_0 N(d_1) - K e^{-rT} N(d_2) \tag{2.1}$$

$$p = K e^{-r} \ N(-d_2) - S_0 N(-d_1) \tag{2.2}$$

I parametri d_1 e d_2 sono rispettivamente uguali a:

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \tag{2.3}$$

$$d_2 = \frac{ln\left(\frac{S_0}{K}\right) + \left(r - \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T} \tag{2.4}$$

[35] F. Black e M. Scholes, "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*, vol. 81, n. 3, 1973, pp. 637-654.
R. C. Merton, "Theory of Rational Option Pricing", *Bell Journal of Economics and Management Science*, vol. 4, n. 1, 1973, pp. 141-183.

dove S_0è il prezzo dell'azione al tempo zero, K è il prezzo d'esercizio, r è il tasso privo di rischio composto continuamente, σ è la volatilità del prezzo dell'azione, T è il tempo di scadenza dell'opzione e $N(X)$ rappresenta la funzione di distribuzione cumulata di una variabile normale standard, cioè la probabilità che questa assuma un valore inferiore a X, analogamente a quanto abbiamo osservato nel paragrafo relativo al VaR.

In particolare $N(d_2)$ rappresenta la probabilità in un mondo neutrale al rischio che la call venga esercitata, mentre $N(d_1)$ equivale al *delta* della call, cioè la derivata del prezzo dell'opzione rispetto al prezzo dell'attività sottostante.

$$\Delta_c = \frac{\partial c}{\partial S} = N(d_1)$$

(2.5)

Pertanto, per coprire una posizione corta su una call europea, occorre avere una posizione lunga su $N(d_1)$ azioni. Analogamente, per coprire una posizione lunga su una call europea, occorre avere una posizione corta su $N(d_1)$ azioni.

Nel caso di un'opzione put europea scritta su un titolo che non paga dividendi, il delta della formula Black-Scholes-Merton è:

$$\Delta_p = \frac{\partial p}{\partial S} = N(d_1) - 1$$

(2.6)

Tale grandezza assume un valore negativo. Pertanto, per coprire una posizione lunga su un'opzione put si deve assumere una posizione lunga sull'azione sottostante, mentre per coprire una posizione corta su una put si deve assumere una posizione corta sull'azione sottostante. Dall'osservazione delle due equazioni del modello di Black-Scholes-Merton si deduce che il prezzo di un'opzione risulta funzione delle cinque variabili che abbiamo definito sopra.[36]

[36] È opportuno ricordare che, nello stesso anno della formulazione originaria del modello, Merton (1973) ne propose un'estensione che includeva formalmente i dividendi. In tale contesto, qualora l'attività sottostante generi un dividendo, il *dividend yield* $q = \frac{DIV}{S_0}$ rappresenta una variabile esplicativa aggiuntiva all'interno del modello BSM.

Inserendo queste variabili nel modello di pricing prescelto, è possibile ricavare il valore dell'opzione. Se tuttavia si conosce il prezzo di mercato dell'opzione, poiché tutte le variabili sono note ad eccezione della volatilità, è possibile utilizzare il modello di pricing a ritroso per calcolare la volatilità implicita nel prezzo stesso dell'opzione, proprio come il rendimento implicito a scadenza di un'obbligazione può essere calcolato equiparando il prezzo di mercato di un'obbligazione alla sua formula di valutazione.

Inoltre, a partire dalla *put-call parity* si può dimostrare che, dati un certo prezzo d'esercizio e una certa scadenza, la volatilità implicita da utilizzare nel modello Black-Scholes-Merton per valutare una call europea deve essere sempre uguale a quella utilizzata per valutare la corrispondente put. Ciò vale, in prima approssimazione, anche per le opzioni americane, che presentano la caratteristica di poter essere esercitate in ogni istante.

Poiché le formule di pricing di un'opzione come quella di Black e Scholes non possono essere invertite analiticamente, il calcolo della volatilità implicita si basa generalmente su un processo iterativo; tra gli algoritmi più usati vi è, ad esempio, il metodo di Newton-Raphson, o delle tangenti[37], processo analogo a quello seguito per il calcolo del tasso di rendimento effettivo (*IRR*) di un titolo obbligazionario.

A questo punto, la volatilità implicita potrebbe sembrare l'indicatore più efficace per prevedere la volatilità futura, poiché, essendo basata sui prezzi di mercato, riflette in tempo reale le aspettative degli operatori finanziari. Tuttavia, questo apparente vantaggio è subordinato a due condizioni fondamentali: l'affidabilità e la comune accettazione del modello di valutazione adottato dagli operatori di mercato e l'assenza di distorsioni strutturali (squilibri temporanei tra domanda e offerta, interruzioni delle contrattazioni o ridotta liquidità e profondità di mercato) nei mercati in cui l'opzione è negoziata.

In conclusione, la volatilità implicita presenta due limiti rilevanti come strumento di previsione nell'ambito del risk management: l'esistenza di un contratto di opzione avente come sottostante l'attività per la quale si intende stimare la volatilità del rendimento e che la durata residua dell'opzione coincida con il periodo di riferimento considerato.

[37] Per un interessante approfondimento si veda:
W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1992.

Le condizioni sopra descritte rendono la sola volatilità implicita poco adatta a prevedere la volatilità delle variabili finanziarie a fini di risk management. È invece più plausibile che essa possa integrare le stime fornite da altre metodologie e segnalare, per le attività per le quali esiste un mercato di opzioni efficiente e regolamentato, eventuali discordanze rispetto a tali previsioni. Queste discrepanze, sebbene non particolarmente frequenti, indicherebbero una significativa divergenza tra le attese del mercato e le previsioni basate sui dati passati.[38]

La CBOE (Chicago Board Options Exchange) calcola e diffonde in tempo reale il prezzo di vari indici costruiti sulla volatilità implicita delle opzioni. Tra questi, il più noto è il VIX (CBOE Volatility Index), nato dagli studi di Menachem Brenner e Dan Galai[39] e costruito sulla volatilità implicita delle opzioni a 30 giorni scritte sull'indice S&P 500. Il VIX è comunemente definito "indice della paura" (*fear index*) e costituisce una misura sintetica e largamente utilizzata delle aspettative di volatilità del mercato azionario.

A differenza di altri strumenti finanziari, il VIX non può essere acquistato o venduto direttamente, ma viene negoziato prevalentemente tramite strumenti derivati (*futures* e opzioni), ETF ed ETN[40], che consentono agli operatori di assumere posizioni speculative o di copertura sulla volatilità attesa del mercato.

Il VIX viene calcolato sulla base di numerose opzioni call e put negoziate sulla CBOE e la sua formulazione fornisce una misura della volatilità attesa dal mercato, sulla base della quale possono formarsi le aspettative sulle possibili variazioni di volatilità del mercato azionario nel prossimo futuro.[41]

---

[38] A. Sironi, A. Resti, *Rischio e valore nelle banche. Misura, regolamentazione, gestione*, Milano: Egea, ristampa aggiornata a gennaio 2021, cap. 7, pp. 230-231.

[39] M. Brenner, D. Galai, *New Financial Instruments for Hedging Changes in Volatility*, Financial Analysts Journal, vol. 45, n. 4, 1989, pp. 61-65

[40] Un *exchange-traded fund* è un titolo rappresentato da una quota di un portafoglio. che si scambia direttamente sul mercato. Un exchange-traded notes è uno strumento finanziario che replica passivamente la performance di sottostanti o indici diversi dalle materie prime come, ad esempio, indici di valute, indici azionari o obbligazionari.

[41] Per i dettagli sulla metodologia di calcolo del VIX si veda:
Cboe Global Indices, *Cboe Volatility Index® Methodology*, Chicago Board Options Exchange, 2022.

Di seguito possiamo osservare il grafico della serie storica dei prezzi di chiusura giornalieri del VIX a partire dalla sua data di introduzione.

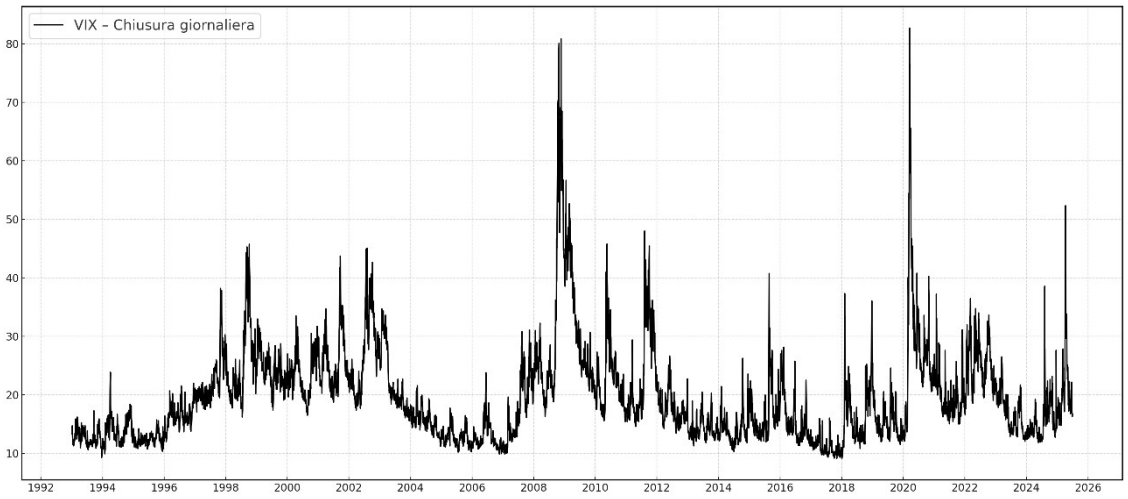

Figura 2.1: Serie storica dei prezzi di chiusura giornalieri del VIX (01/1993-06/2025)

Oltre al VIX, il CBOE utilizza la stessa metodologia per calcolare i seguenti strumenti:

- Indice di volatilità giornaliera (VIX1D), introdotto a maggio 2023
- Indice di volatilità a 9 giorni (VIX9DSM), introdotto a marzo 2013
- Indice di volatilità a 3 mesi (VIX3MSM), introdotto a novembre 2007
- Indice di volatilità a 6 mesi (VIX6MSM), introdotto a gennaio 2011
- Indice di volatilità a 1 anno (VIX1YSM), introdotto a luglio 2011

Ognuno di questi indici riflette la volatilità annualizzata prevista per un certo orizzonte temporale dell'indice S&P 500.

2.2 Fatti stilizzati nei mercati azionari

Tornando al mondo reale, si osserva che gran parte delle serie storiche azionarie, rilevate a intervalli di tempo brevi, ad esempio giornalieri o settimanali, presentano empiricamente le seguenti regolarità statistiche:

1. Non stazionarietà dei prezzi. I prezzi delle azioni non sono stazionari, ma esplosivi. Per stazionarietà si intende la proprietà di un processo stocastico le cui caratteristiche distribuzionali restano costanti nel tempo. Ne consegue che il processo avrà media e varianza costanti e finite. Un tipico esempio di processo stocastico stazionario è quello che rappresenta la serie storica dei rendimenti.

2. Non normalità dei rendimenti. L'ipotesi di normalità della distribuzione dei rendimenti è spesso rigettata empiricamente, a causa di fenomeni di leptocurtosi e negative skewness.

3. Media dei rendimenti prossima a zero e bassa autocorrelazione. I rendimenti presentano una media prossima allo zero e un'autocorrelazione molto contenuta.

4. Autocorrelazione positiva dei rendimenti al quadrato. I quadrati dei rendimenti presentano una dipendenza seriale positiva. Poiché i rendimenti hanno media prossima allo zero, possiamo approssimare il valore atteso del quadrato di un generico rendimento come la sua varianza. Pertanto, l'autocorrelazione positiva dei rendimenti al quadrato implica empiricamente la presenza di periodi di alta volatilità alternati a periodi di bassa volatilità, cioè il fenomeno noto come *volatility clustering*. In termini econometrici ciò si traduce nel raggruppamento della varianza dell'errore del modello autoregressivo utilizzato per descrivere e prevedere la serie storica.

5. *Leverage effect*. La natura delle *news* è correlata in modo asimmetrico alla volatilità. In particolare, shock negativi tendono a generare aumenti significativi della volatilità, mentre shock positivi hanno un effetto meno marcato.

Queste regolarità statistiche, ricorrenti nelle serie storiche, sono dette "fatti stilizzati" e hanno motivato l'esigenza di una modellistica in grado di catturare la dinamica della volatilità. Per spiegare le variazioni di volatilità sono state avanzate diverse spiegazioni teoriche, ma nessuna fornisce una caratterizzazione esaustiva. Tra le più rilevanti troviamo:

1. Annunci di notizie: l'arrivo di notizie impreviste induce gli operatori a riformulare le proprie aspettative. Ne consegue un ribilanciamento dei portafogli, con aumenti di volatilità nei periodi in cui le notizie hanno un impatto diretto sui prezzi degli asset.

2. Leva finanziaria: Quando un'impresa si finanzia sia con il debito che con il capitale, solo il capitale riflette la volatilità dei flussi di cassa dell'impresa. Tuttavia, quando il prezzo del capitale proprio diminuisce, questo deve riflettere la stessa volatilità dei flussi di cassa dell'impresa e quindi i rendimenti negativi dovrebbero portare a un aumento della volatilità del capitale proprio. L'effetto leva è pervasivo nei rendimenti azionari, soprattutto negli indici azionari ampi.[42]

3. Feedback della volatilità: La retroazione della volatilità è motivata da un modello in cui la volatilità di un'attività è prezzata. Un calo del prezzo di un'attività induce gli investitori a richiedere un premio al rischio più elevato, generando ulteriore volatilità.[43] Questo effetto è particolarmente rilevante in fasi di transizione economica.

4. Illiquidità: episodi di illiquidità temporanea possono amplificare la volatilità dei prezzi. Shock di domanda o offerta in mercati sottili producono effetti sproporzionati sulle quotazioni.[44]

5. Incertezza macroeconomica: variazioni nelle aspettative sullo stato dell'economia possono causare variazioni della volatilità attraverso un ciclo di feedback tra disponibilità di portafoglio, convinzioni sullo stato generale dell'economia e prezzi degli asset. Anche questo effetto dovrebbe essere più consistente quando l'economia è in transizione.[45]

Le cause economiche della variazione temporale della volatilità includono tutte quelle menzionate, oltre ad alcune non ancora identificate, come le cause comportamentali.

---

[42] Christie, A. A. (1982). *The stochastic behavior of common stock variances: Value, leverage and interest rate effects*. Journal of Financial Economics, 10(4), 407–432.

[43] Bekaert, G., & Wu, G. (2000). *Asymmetric volatility and risk in equity markets*. Review of Financial Studies, 13(1), 1–42.

[44] Intuitivamente, se il mercato è ipervenduto (comprato), un piccolo shock negativo (positivo) produce una piccola diminuzione (aumento) della domanda. Tuttavia, poiché pochi partecipanti sono disposti a comprare (vendere), questo shock ha un effetto sproporzionato sui prezzi.

[45] Veronesi, P. (1999). *Stock market overreaction to bad news in good times: A rational expectations equilibrium model*. Review of Financial Studies, 12(5), 975–1007.
Collard, F., Dellas, H., Diba, B., & Loisel, O. (2018). *Optimal monetary and fiscal policy in large crises*. Journal of Monetary Economics, 97, 42–63.

2.3 La stima della volatilità con dati storici: le medie mobili semplici (SMA)

Iniziamo a questo punto la trattazione dei modelli di stima della volatilità che utilizzano dati storici. Recuperiamo inizialmente la formula della stima della volatilità storica introdotta nel primo capitolo (1.5), sempre considerando una frequenza di osservazione dei rendimenti giornaliera ma indicando per semplicità $\hat{\sigma}_t$ con σ_t.

$$\sigma_t = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_{t-i} - \bar{r}_t)^2}$$

(2.6)

Ricordando che, per frequenze di osservazione dei rendimenti molto ridotte si ha:

$$r_i = \ln(\frac{S_i}{S_{i-1}}) \cong \frac{S_i - S_{i-1}}{S_{i-1}}$$

(2.7)

Ipotizziamo a questo punto di stimare la volatilità per il periodo successivo $(t + 1)$. La stima sarà effettuata sulla base dei dati osservabili dal secondo rendimento del campione fino ad oggi, spostando in avanti di un periodo la finestra temporale del campione: tale approccio è detto metodo delle medie mobili semplici (*Simple Moving Average*, SMA) e costituisce il metodo più elementare per la stima della volatilità storica. Una media mobile può essere definita come una media relativa a un numero fisso di dati che scorrono nel tempo: il trascorrere del tempo permette che il dato più lontano venga sostituito da quello più recente, lasciando immutata la dimensione del campione.

Questo criterio può essere utilizzato come previsione della volatilità futura nell'ambito del risk management, al fine di calcolare misure di rischio come VaR ed Expected Shortfall, utilizzando l'approccio parametrico. Per queste finalità l'equazione della volatilità viene modificata sostituendo il denominatore $n - 1$ con n, così da passare da una stima corretta della volatilità a una stima di massima verosomiglianza e inoltre si suppone che la media dei rendimenti sia nulla, assunzione ragionevole in caso di una rilevazione ad alta frequenza dei rendimenti, oltre che coerente con le ipotesi di efficienza dei mercati.

Effettuando le citate modifiche, la formula del tasso di varianza sarà pari a:

$$\sigma_t^2 = \frac{1}{n}\sum_{i=1}^{n} {r_{t-i}}^2$$

(2.8)

L'utilizzo del criterio delle medie mobili semplici per la stima della volatilità storica, usata come previsione della volatilità futura in un'ottica di risk management, porta con sé due problemi:

Il primo riguarda il trade-off fra contenuto informativo e reattività alle condizioni più recenti. A parità di altre condizioni, la scelta di un numero di osservazioni (n) più elevato conduce a una stima di volatilità più stabile, in quanto gli eventuali shock del fattore di mercato considerato incidono in misura proporzionalmente inferiore sulla stima della volatilità. Inoltre, un arco temporale molto ampio offre un elevato contenuto informativo, in quanto il campione utilizzato per la stima della volatilità è più ampio. Tuttavia, allo stesso tempo, la stima della volatilità prodotta risponderà in modo lento a variazioni improvvise delle condizioni di mercato, risultando dunque poco aggiornata.

In presenza di un sensibile e improvviso incremento della volatilità dei fattori di mercato, per esempio, l'utilizzo della volatilità storica calcolata su un ampio campione passato conduce infatti ad attribuire un peso marginale alle condizioni più recenti, che invece potrebbero riflettere meglio le condizioni di mercato future. Ciò è dovuto al fatto che, nello stimare la volatilità storica, si attribuiscono ai singoli scostamenti quadratici dalla media pesi uniformi.

Ne segue che, se la stima della volatilità storica è funzionale alla previsione della volatilità futura, la scelta di un numero elevato di osservazioni passate può condurre a risultati meno affidabili. Per questo, la maggioranza delle istituzioni che utilizzano modelli VaR, nella stima della volatilità giornaliera utilizzata per misurare i rischi dell'attività di trading, scelgono intervalli temporali relativamente contenuti, generalmente fra 20 e 50 giorni.

Di seguito possiamo osservare gli effetti di un diverso numero di osservazioni sulla reattività della volatilità stimata dei rendimenti logaritmici giornalieri dell'indice

S&P 500. Ipotizzando che il mercato sia aperto 252 giorni in un anno, possiamo calcolare la volatilità storica utilizzando finestre mobili di ampiezza mensile, bimestrale, trimestrale, semestrale e annuale, ponendo alternativamente $n = 21$, $n = 42$, $n = 63$, $n = 126$ e $n = 252$, nel periodo luglio 2019-giugno 2025.

Figura 2.3: Stima della volatilità giornaliera dello S&P 500 tramite medie mobili semplici

Il secondo problema, noto in letteratura come *echo effect*[46] o *ghost features*[47], si manifesta quando la stima della volatilità subisce una variazione (tanto più pronunciata quanto minore è il numero di osservazioni nel campione) non solo nel momento in cui il fattore di mercato considerato subisce una variazione pronunciata, ma anche quando il dato relativo a questo shock esce dal campione e viene sostituito da un dato più recente.

Mentre la prima variazione nella stima della volatilità è pienamente giustificata, la seconda (che è sempre di segno contrario alla prima) non lo è per il semplice fatto che quando il dato di shock esce dal campione è probabile che nessuna novità rilevante abbia interessato il fattore di mercato considerato.

La figura sottostante riporta un esempio, riguardante il comportamento della volatilità dell'S&P 500 nel periodo da gennaio a luglio 2020.

[46] Figlewski, S. (1994), *Forecasting Volatility Using Historical Data*, *Financial Markets, Institutions & Instruments*, vol. 3, n. 2, pp. 1-88.
[47] Alexander, C. (1996), *The Handbook of Risk Management and Analysis*, Chichester: Wiley, pp. 235-237.

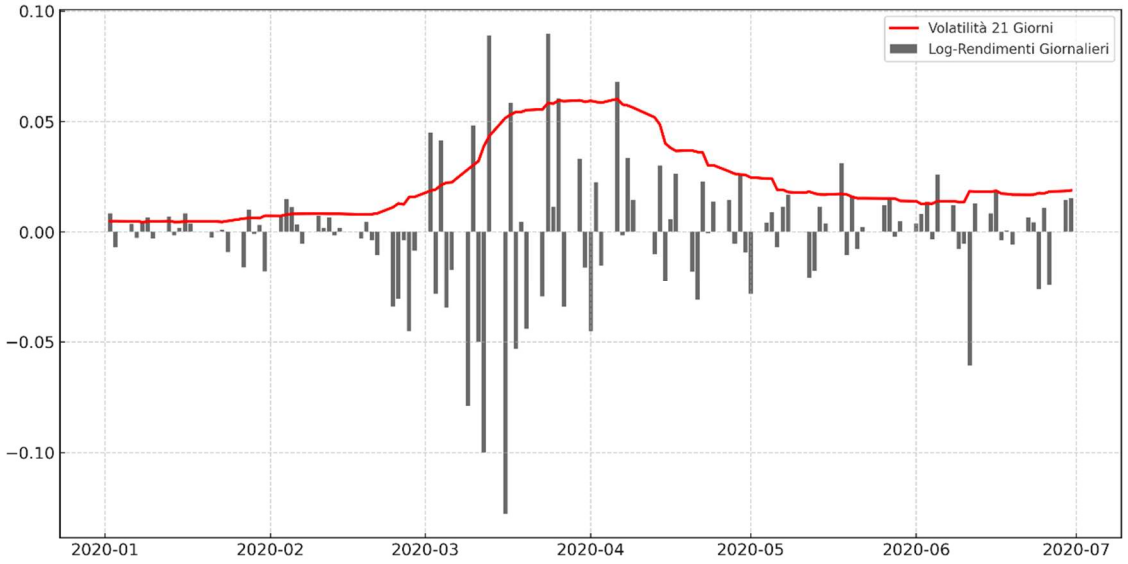

Figura 2.4: Effetto eco della volatilità dello S&P 500 durante lo scoppio della pandemia

Tale volatilità mostra una variazione verso la metà di marzo a seguito del *lockdown* globale causato dalla pandemia di COVID-19, che ha provocato alcuni giorni di rendimenti fortemente negativi, come mostrato dagli istogrammi grigi in figura; inoltre, la volatilità si riduce sensibilmente già a partire dalla metà di aprile, pur in presenza di rendimenti giornalieri non estremi, e ciò avviene unicamente perché le osservazioni di metà marzo sono uscite dal campione di calcolo della deviazione standard.

2.4 Il modello a media mobile esponenziale (EWMA)

L'esigenza di superare questi due problemi ha condotto allo sviluppo di un secondo metodo di stima della volatilità storica, introdotto nel 1994 dalla banca d'investimento statunitense J.P. Morgan all'interno di RiskMetrics. Questo modello è basato sull'utilizzo della media mobile esponenziale e si differenzia dal metodo della media mobile semplice in quanto, nel calcolare la media degli scarti quadratici dal rendimento medio, attribuisce loro una ponderazione non uniforme.

La logica è relativamente lineare: utilizzando un numero elevato di osservazioni passate, ma attribuendo un peso maggiore a quelle più recenti, si ottiene una stima della volatilità caratterizzata da elevato contenuto informativo e maggiore sensibilità agli shock recenti. Questo approccio presenta due vantaggi principali: da un lato, la stima reagisce più

rapidamente a shock del fattore di mercato e, dall'altro, uno shock pronunciato del fattore di mercato esce in modo graduale dalla stima della volatilità, evitando l'effetto eco.

Recuperiamo l'ultima formula della varianza che abbiamo definito, ma questa volta anziché effettuare una ponderazione equa (pari a $1/n$), consideriamo un peso diverso per ogni osservazione passata.

$$\sigma_t^2 = \sum_{i=1}^{n} w_i {r_{t-i}}^2$$

(2.9)

Dove w_i rappresenta il peso assegnato all'osservazione di i giorni fa e deve essere strettamente positivo. Dato che vogliamo dare meno peso alle osservazioni meno recenti, si avrà che $w_i < w_j$ con $i < j$. La somma dei pesi deve essere pari all'unità, per cui:

$$\sum_{i=i}^{n} w_i = 1$$

(2.10)

Il modello a media mobile esponenziale (detta anche *exponentially-weighted moving average*, o EWMA) è un particolare tipo di media ponderata, in cui i pesi associati alle diverse osservazioni (w_i) diminuiscono esponenzialmente tornando indietro nel tempo. Queste ponderazioni possono essere considerate potenze diverse di una medesima costante λ, con $0 < \lambda < 1$. In particolare si ha che:

$$w_{i+1} = w_i \cdot \lambda$$

(2.11)

La costante λ, denominata *decay factor*, indica il grado di persistenza attribuito alle osservazioni campionarie passate. Infatti, se la costante λ è più vicina a uno, le sue potenze successive (λ^2, λ^3, λ^4…), che rappresentano i pesi associati alle rispettive osservazioni passate, si avvicinano a zero molto lentamente, adeguando meno rapidamente la media alle condizioni più recenti. Se invece λ fosse più piccolo, le sue potenze successive tendono più rapidamente a zero e le osservazioni passate escono più

rapidamente dalla stima di σ. Per questo, $(1 - \lambda)$ rappresenta la velocità di decadimento delle osservazioni passate.

Se ponessimo $\lambda = 1$, la ponderazione sarebbe uguale per tutte le osservazioni passate e la media ponderata esponenziale coinciderebbe con la media semplice. Viceversa, per valori di λ più distanti da 1 la media esponenziale si differenzia in modo più marcato da quella semplice, attribuendo pesi sensibilmente differenziati a osservazioni relative a momenti diversi.

Considerando i rendimenti logaritmici r_t di un fattore di mercato a media nulla, possiamo ottenere la stima della varianza basata sulle medie mobili esponenziali nella seguente modalità:

$$\sigma_t^2 = (1 - \lambda) r_{t-1}^2 + \lambda \sigma_{t-1}^2$$

(2.12)

La stima della varianza relativa al giorno t (ottenuta alla fine del giorno $t - 1$) viene dunque calcolata attraverso un meccanismo ricorsivo utilizzando la stima della varianza ottenuta alla fine del giorno precedente (σ_{t-1}^2) e il quadrato del tasso di variazione della variabile di mercato osservato tra il giorno $t - 2$ e il giorno $t - 1$, indicato con r_{t-1}^2.[48]

La figura sottostante ripropone la volatilità giornaliera dell'indice S&P500 già mostrata precedentemente, cioè calcolata come la deviazione standard basata su una media mobile semplice a 21 giorni secondo l'equazione (2.8) posta sotto radice quadrata. A questa è affiancata la stima ottenuta con il metodo delle medie mobili esponenziali, con λ arbitrariamente posto pari a 0,94, valore che in base alle ricerche effettuate da J.P. Morgan, consente previsioni del tasso di varianza molto vicine ai valori osservati.[49]

[48] Per una dimostrazione dettagliata si veda: A. Sironi, A. Resti, *Rischio e valore nelle banche. Misura, regolamentazione, gestione*, Milano: Egea, ristampa aggiornata a gennaio 2021, cap. 7, pp. 214-215

[49] J.P. Morgan, *RiskMetrics Monitor, Fourth Quarter 1995*, New York, J.P. Morgan, 1995.

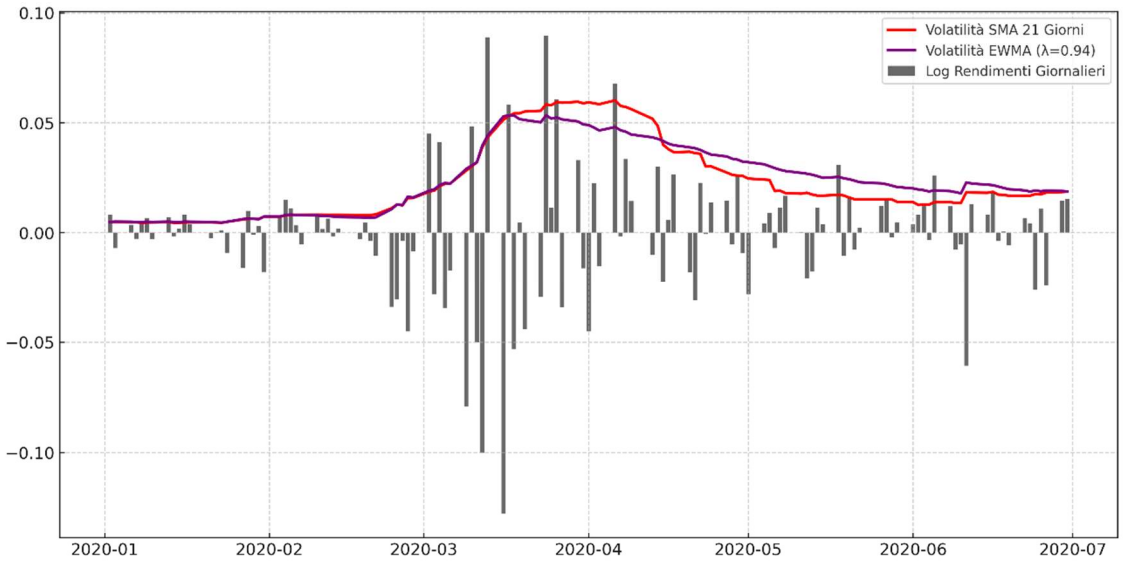

Figura 2.5: Riduzione dell'effetto eco della volatilità dello S&P 500 tramite il modello EWMA

Si nota come lo shock di marzo 2020 ha generato un immediato, sensibile aumento nella volatilità stimata; tuttavia, diversamente da quanto accade con le medie mobili semplici, utilizzando il metodo delle medie mobili esponenziali non si osserva una repentina diminuzione della volatilità al momento in cui il dato esce dal campione (21 giorni dopo), poiché il suo peso risulta già progressivamente ridotto, generando un calo più graduale della volatilità stimata.

Aldilà dei vantaggi conseguiti rispetto alle medie mobili semplici, anche questo modello presenta alcune criticità pratiche, principalmente connesse alla scelta del parametro di decadimento (λ), che dovrebbe dipendere dalla velocità con la quale si ritiene che la volatilità del rendimento del fattore di mercato vari nel tempo. Se si crede che questa si modifichi lentamente nel tempo, sarebbe più corretto scegliere un fattore λ prossimo a uno, in modo da attribuire un peso elevato anche alle osservazioni più lontane. Viceversa, se si ritiene che la volatilità del fattore di mercato vari spesso, anche in modo repentino, sarebbe più corretto utilizzare un fattore λ più ridotto. La scelta di λ dovrebbe inoltre dipendere, in un'ottica di *risk-measurement* e dunque di misurazione della perdita potenziale, dall'orizzonte temporale di detenzione della posizione (*holding period*). Quanto minore è tale orizzonte, tanto maggiore è infatti, a parità di altre condizioni, l'esigenza che la stima della volatilità, e dunque della perdita potenziale, si adegui più rapidamente alle nuove condizioni di mercato, e dunque tanto minore dovrebbe essere λ.

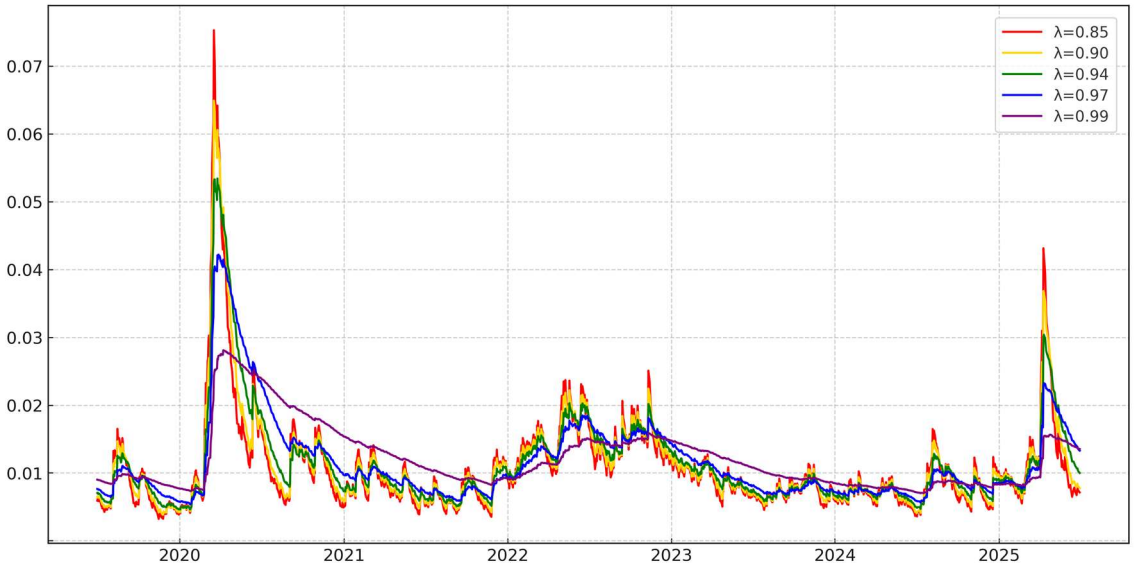

Figura 2.6: Stima della volatilità giornaliera dello S&P 500 tramite il modello EWMA

La figura mostra come, utilizzando un valore di λ più elevato (0,99), si ottenga una stima della volatilità più stabile e dunque meno reattiva alle condizioni più recenti del mercato. Viceversa, utilizzando un valore di λ più ridotto (0,9) la stima diviene a sua volta più volatile, in quanto le nuove osservazioni ricevono un peso proporzionalmente più elevato.

Il valore di λ utilizzato dalla banca dati di RiskMetrics è pari a 0,94 per la volatilità giornaliera (rischio relativo alle posizioni del portafoglio di negoziazione) e a 0,97 per la volatilità mensile (rischio relativo alle posizioni del portafoglio immobilizzato).

Tuttavia, la scelta di un parametro λ costante si scontra con l'evidenza empirica: in alcuni mercati un aumento di volatilità tende a persistere più giorni, in altri si manifesta come shock temporaneo. In linea teorica, sarebbe quindi preferibile aggiornare frequentemente λ in base ai dati di mercato, piuttosto che mantenerlo esogenamente invariato. La determinazione del valore ottimale di λ avviene solitamente tramite il metodo della massima verosimiglianza, assumendo una certa distribuzione dei rendimenti, o attraverso i minimi quadrati ordinari (OLS), minimizzando la somma dei quadrati degli errori (*Sum of Squared Errors*, SSE); i due approcci forniscono risultati coincidenti qualora si utilizzi lo stesso campione di dati e nel caso della massima verosomiglianza, si ipotizzi che gli errori siano distribuiti come una normale a media nulla.

Naturalmente il valore ottimale del *decay factor* dipende strettamente dal campione di dati prescelto e richiede dunque un aggiornamento periodico.

2.5 Il modello ARCH(q)

Il modello ARCH(*q*), cioè *autoregressivo ad eteroschedasticità condizionata di ordine q* (o analogamente "a *q* ritardi") è stato proposto originariamente da Robert Engle nel 1982[50]. Questo è stato il primo modello a contemplare esplicitamente l'esistenza del *volatility clustering*, assumendo che la varianza condizionata dipenda linearmente dai quadrati dei rendimenti passati fino a un tempo passato *q*. Per i suoi studi Engle è stato insignito del Premio Nobel per l'Economia nel 2003.

Dunque questo modello risulta antecedente rispetto all'EWMA di cui abbiamo appena discusso, e mentre quest'ultimo proviene dall'industria, il modello ARCH (q) è nato in ambito accademico. assume che la varianza condizionata dipenda linearmente dai quadrati dei rendimenti passati fino a un lag q

Prima di introdurre analiticamente l'ARCH (q), analizziamo singolarmente le parole che descrivono questo modello:

• Eteroschedasticità: significa varianza che muta nel tempo, e si contrappone all'ipotesi di omoschedasticità, cioè varianza costante nel tempo. Una serie storica che presenta eteroschedasticità è caratterizzata dal già citato *volatility clustering*, cioè dall'alternanza tra periodi di elevata volatilità e periodi di relativa tranquillità; tale fenomeno risulta particolarmente evidente nelle serie storiche di natura finanziaria, soprattutto se rilevate a intervalli di tempo ristretti, ad esempio frequenza giornaliera.

• Condizionata (o condizionale): indica che le previsioni ottenute sono basate sulle informazioni disponibili fino al momento in cui si effettua la stima[51]: in pratica, le stime della volatilità condizionale riflettono il livello corrente di incertezza generato dagli shock passati e dunque si possono definire una funzione dell'informazione raccolta fino all'istante temporale di stima. Al contrario la volatilità storica, una misura ex-post della variabilità dei rendimenti, viene anche detta non condizionata e in questo senso riassume gli shock dei prezzi occorsi nel periodo di stima. La volatilità condizionale tende a

[50] R. F. Engle, *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation*, *Econometrica*, vol. 50, n. 4, 1982, pp. 987-1007.

[51] In letteratura, il concetto di *information set* indica l'insieme delle informazioni disponibili fino a un certo istante temporale e utilizzabili per la formulazione delle previsioni. Nel contesto dei modelli ARCH, esso rappresenta la base informativa su cui si condiziona la stima della varianza dei rendimenti.

catturare il tasso di persistenza di questi shock e riassume così l'effetto della volatilità passata sul livello corrente di incertezza circa gli eventi futuri.

• Autoregressivo: si riferisce al metodo utilizzato per modellare l'eteroschedasticità condizionale, il quale si basa appunto su una regressione della varianza su sé stessa. In pratica, ciò indica che i livelli passati della volatilità hanno un certo grado di influenza sui valori futuri.

Per comprendere ulteriormente la logica sottostante a questo modello, ma anche dei successivi modelli GARCH che osserveremo, è necessario chiarire in generale la differenza fra una stima non condizionata e una condizionata. La prima è tipicamente ottenuta utilizzando un campione più o meno ampio di dati storici: così, per esempio, la media non condizionale di una serie di dati r_t (con t = 1, 2, …, T), può essere stimata con la media campionaria $\overline{y_t} = \frac{1}{T}\sum_{t=1}^{T} y_t$

Viceversa, se la variabile in questione dipende da una o più altre grandezze, è possibile stimare una media condizionata ricorrendo, per esempio, a una regressione dei minimi quadrati. Così, se *y* dipendesse linearmente, al netto di un fattore di disturbo, dal valore di una certa *x*, potremmo stimare la sua media condizionata come:

$$\bar{y}_{t|x=\bar{x}_t} = \alpha + \beta\bar{x}_t$$

(2.13)

Se invece risultasse che *y* dipende linearmente dal suo valore passato, il modello per la stima della media condizionata diventerebbe:

$$\bar{y}_{t|y_{t-1}} = \alpha + \beta y_{t-1}$$

(2.14)

cioè un modello autoregressivo del primo ordine, anche chiamato AR(1).

La differenza fondamentale tra media non condizionata e media condizionata è dunque legata al fatto che la prima è una costante, mentre la seconda necessita di un modello di specificazione, da stimare mediante una tecnica di regressione, e dà luogo a una stima che varia nel tempo.

I modelli econometrici a serie storica sono per lo più volti, mediante l'utilizzo di un'equazione da stimare come l'ultima osservata, a cercare di prevedere la media di una certa variabile aleatoria. Al contrario, i modelli GARCH, tra cui rientra anche l'ARCH(q), hanno come scopo prioritario quello di modellare e prevedere la varianza di una variabile aleatoria. La stessa logica trova applicazione nel caso della varianza non condizionata e di quella condizionata. La prima è una costante stimata come varianza campionaria. La seconda è invece variabile nel tempo e viene stimata sulla base di un determinato modello. In questo senso, le diverse versioni dei modelli GARCH rappresentano diverse specificazioni del modello che spiega la varianza condizionata.

La costruzione di un modello a eteroschedasticità condizionata autoregressiva comporterebbe, in effetti, due specificazioni distinte: una per la media e una per la varianza. Nelle applicazioni finanziarie, tuttavia, la media è spesso posta uguale a zero o a una costante. È bene considerare che, mentre un eventuale errore di specificazione della varianza non danneggia in modo significativo le previsioni della media, una scorretta specificazione della media influenzerebbe e dunque renderebbe distorte anche le previsioni della varianza.

Per coerenza con la nostra trattazione, possiamo derivare l'equazione dell'ARCH(q) a partire dalla formula della varianza modificata alla base della derivazione dell'EWMA.

Come abbiamo osservato, anche l'EWMA cattura il raggruppamento della volatilità, ma in più il modello ARCH(q) tiene conto di un'ulteriore tendenza empirica, cioè l'attitudine della volatilità a tendere verso la sua media nel lungo termine. Definendo con q il numero di osservazioni passate dei rendimenti al quadrato, con ω il parametro della varianza di lungo termine pari al prodotto tra quest'ultima, indicata con V_L e il suo peso γ, la stima della varianza condizionata secondo il modello ARCH(q) è la seguente:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \, r_{t-i}^2$$

(2.15)

I parametri ω e α_i vengono stimati mediante il metodo della massima verosomiglianza, o alternativamente con il metodo dei minimi quadrati ordinari (OLS), ipotizzando una distribuzione normale per i rendimenti.

I singoli pesi devono essere tutti maggiori di zero per evitare che la varianza risulti negativa. Inoltre, poiché la loro somma deve essere pari all'unità, si ha che:

$$\gamma + \sum_{i=1}^{q} \alpha_i = 1$$

(2.16)

Il modello stima dunque la varianza come una media mobile di q rendimenti passati elevati al quadrato.

Poiché la media dei rendimenti giornalieri è prossima allo zero, in termini econometrici i rendimenti azionari possono essere approssimati con i termini di errore di un modello autoregressivo utilizzato per rappresentare la serie storica dei prezzi.[52]

In pratica, se si verifica un improvviso shock della variabile considerata, ciò determina un errore di previsione, il quale a sua volta genera, se il suo coefficiente α è positivo, un rialzo della previsione della volatilità relativa ai q periodi futuri.

Il fatto che gli errori di previsione vengano considerati al quadrato fa sì che il rialzo si verifichi indipendentemente dal segno dell'errore. Si ritiene dunque probabile che a una prima variazione pronunciata della variabile considerata ne seguano altre; ciò è coerente con l'evidenza empirica secondo la quale una variazione significativa dei prezzi tende a essere seguita da altrettante variazioni significative.

In questo senso il modello ARCH(q) riconosce esplicitamente la differenza fra volatilità non condizionata e volatilità condizionata, ammettendo che quest'ultima vari nel tempo in relazione agli errori di previsione passati; inoltre esso modella adeguatamente il fatto che una variazione pronunciata di un certo fattore di mercato tende a persistere nel tempo, generando il fenomeno di *volatility clustering* descritto in precedenza.

[52] In formule, un rendimento r_i può essere suddiviso in una componente nota μ e in un termine aleatorio ε_t nel seguente modo: $r_i = \mu + \varepsilon_i$. Se si assume $\mu \cong 0$, si avrà semplicemente $r_i = \varepsilon_i$.

La varianza non condizionata nel modello ARCH(q) è uguale a:

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i}$$

(2.17)

Questo modello, pur avendo rappresentato un approccio pionieristico alla modellizzazione della volatilità condizionata, presenta limiti empirici: richiede infatti un numero elevato di osservazioni passate dei rendimenti al quadrato per descrivere correttamente la dinamica della varianza, risultando poco flessibile e oneroso dal punto di vista applicativo.

2.6 Il modello GARCH(*p*,*q*)

Per superare tali criticità, Tim Bollerslev ha introdotto nel 1986 il modello GARCH(p,q) (*Generalized ARCH*)[53], che estende la formulazione originaria proposta da Engle nel 1982, includendo, oltre ai q termini ARCH relativi ai rendimenti passati degli errori di previsione, anche i p ritardi relativi ai valori passati della varianza condizionata. Questa estensione, consente al modello di adattarsi efficacemente anche con un basso ordine, laddove un ARCH avrebbe invece richiesto un numero ben maggiore di osservazioni.[54]

La sua formulazione analitica è la seguente:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \, r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-i}^2$$

(2.18)

Anche in questo caso $\omega = \gamma V_L$ e $r_i = \varepsilon_i$, mentre i singoli pesi devono essere positivi e la loro somma deve essere pari a 1. In questo modello i parametri ω, α_i e β_i sono stimati attraverso il metodo della massima verosomiglianza ipotizzando una certa distribuzione.

[53] T. Bollerslev, *Generalized Autoregressive Conditional Heteroskedasticity*, Journal of Econometrics, vol. 31, n. 3, 1986, pp. 307-327.

[54] Peter R. Hansen, Asger Lunde, "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?", *Journal of Applied Econometrics*, 20(7), 2005, pp. 873–889.

Nel GARCH(p,q) la varianza non condizionale è pari a:

$$\sigma^2 = \frac{\omega}{\sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}$$

(2.19)

Il modello GARCH(1,1) costituisce la specificazione più utilizzata, poiché combina parsimonia parametrica ed elevata capacità descrittiva, risultando particolarmente efficace nel rappresentare il *clustering* della volatilità osservato nei mercati finanziari:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$$

(2.20)

In questa versione del modello, una volta stimati i parametri, è possibile ricavare in seguito γ come $1 - \alpha - \beta$.

Per garantire la stazionarietà del GARCH(1,1) è necessario che $\alpha + \beta < 1$. In caso contrario, la varianza di lungo periodo risulterebbe negativa, rendendo il modello non applicabile

Quando $\alpha + \beta > 1$, il peso γ assegnato alla varianza di lungo termine è negativo e il processo tende a divergere, invece di convergere verso la media.

Se infine $\alpha + \beta = 1$, e dunque γ e ω sono uguali a zero, si ricade nel caso specifico del modello IGARCH (*Integrated GARCH*). In questa situazione, la varianza non condizionale, ossia la varianza di lungo periodo risulta indefinita. Da qui la denominazione di modello integrato nella varianza. Questo modello introduce un necessario trade-off tra persistenza della varianza ed effetto dei rendimenti più recenti. L'IGARCH risulta interessante in quanto, ipotizzando un valore nullo di gamma e dunque di omega, si ottiene il modello delle medie mobili esponenziali (EWMA) semplicemente ponendo $\beta = \lambda$ e di conseguenza $\alpha = 1 - \lambda$.

Sostituendo ricorsivamente σ_{t-1}^2 nell'equazione del GARCH (1,1) si può osservare che il peso assegnato al quadrato del rendimento è $\alpha\beta^{i-1}$. Poiché dunque i pesi diminuiscono esponenzialmente al tasso β, questo parametro può essere interpretato analogamente al *decay factor* dell'EWMA. La principale differenza tra questi due modelli è che il

GARCH(1,1), pur trattando la varianza come processo stocastico, include un meccanismo di ritorno verso la varianza media di lungo periodo, verso il quale viene sospinta nel lungo termine.

Un approccio alternativo alla stima dei parametri del GARCH (1,1), è il cosiddetto *variance targeting*[55], che vincola la varianza incondizionata del modello (V_L) a coincidere con la varianza campionaria dei dati, o con un valore teoricamente prestabilito. Tale vincolo riduce il numero di parametri da stimare ed empiricamente risulta essere più stabile.

In tal caso $\omega = V_L(1 - \alpha - \beta)$ e i soli parametri da stimare sono α e β.

2.7 Il modello EGARCH(*p*,*o*,*q*) e altre versioni alternative al GARCH(*p*,*q*)

A fianco dei vantaggi esplicati, il modello di Bollersev presenta tuttavia alcuni limiti. Oltre a risultare più complesso e oneroso rispetto al semplice utilizzo di una media mobile esponenziale, specie nel caso in cui si consideri più di un solo ritardo per ogni variabile, il GARCH conserva l'ipotesi di normalità degli errori di previsione/rendimenti, tendendo a sottostimare la probabilità di osservare eventi estremi. Inoltre questo modello non è in grado di catturare gli effetti asimmetrici della volatilità (*leverage effect*), in quanto nelle equazioni che sono state presentate i rendimenti sono elevati al quadrato e dunque l'impatto di uno shock di mercato sulla previsione della volatilità risulta indipendente dal suo segno.

Per superare questo limite sono state proposte numerose versioni alternative all'originale generalizzazione proposta da Bollersev. Oltre al già citato Integrated GARCH (IGARCH), le più diffuse per la previsione della volatilità dei rendimenti a scopo di risk management sono l'Exponential GARCH (EGARCH) e il GJR-GARCH (dai cognomi dei tre autori che l'hanno introdotto nel 1993).

Questi modelli si riferiscono all'ultimo problema menzionato, cioè la non considerazione del segno degli errori di previsione, e ipotizzano asimmetrica la cosiddetta *News Impact Curve* (NIC), che misura l'effetto di uno shock nel periodo corrente sulla varianza

[55] R. F. Engle, J. Mezrich, *GARCH for Groups*, Risk, 1996, pp. 36-40.

condizionata nel periodo successivo, e permette così di facilitare il confronto tra i diversi modelli.

L'EGARCH(*p*,*o*,*q*), modella il logaritmo naturale della varianza condizionata, garantendo che rimanga sempre positiva, anche qualora la parte destra dell'equazione assuma valori negativi.[56]

$$\ln(\sigma_t^2) = \omega + \sum_{i=1}^{p} \alpha_i \left(\left|\frac{r_{t-i}}{\sigma_{t-i}}\right| - \sqrt{\frac{2}{\pi}}\right) + \sum_{h=1}^{o} \gamma_h \frac{r_{t-h}}{\sigma_{t-h}} + \sum_{j=1}^{q} \beta_j \ln(\sigma_{t-j}^2)$$

(2.21)

In questo modello, gli errori di previsione vengono considerati sia in valore assoluto sia nel loro segno, evitando la trasformazione al quadrato tipica del GARCH standard e permettendo di considerare la diversa reazione degli operatori alle notizie positive piuttosto che a quelle negative.

L'equazione del modello EGARCH(1,1,1) è la seguente:

$$\ln(\sigma_t^2) = \omega + \alpha\left(\left|\frac{r_{t-1}}{\sigma_{t-1}}\right| - \sqrt{\frac{2}{\pi}}\right) + \gamma \frac{r_{t-1}}{\sigma_{t-1}} + \beta \ln(\sigma_{t-1}^2)$$

(2.22)

Dove $\left|\frac{r_{t-1}}{\sigma_{t-1}}\right| = z_{t-1}$ è il valore assoluto di una variabile casuale normale standard, mentre meno il suo valore atteso, cioè $\sqrt{\frac{2}{\pi}} = \mathbb{E}[|z_t|] \cong 0.7979$ rappresenta il valore atteso del valore assoluto di una variabile normale standard.

I due shock presenti nella formula si comportano in modo diverso: il valore assoluto del primo produce un aumento simmetrico della log-varianza per il periodo di tempo considerato, mentre il valore del secondo produce un effetto asimmetrico. Il parametro γ tipicamente assume valori negativi, riflettendo il fatto che shock negativi tendono a

[56] Nelson, D. B., *Conditional heteroskedasticity in asset returns: A new approach*. Econometrica, 59(2), 1991, pp. 347–370.

generare incrementi di volatilità maggiori rispetto a shock positivi di pari entità. Porre $\gamma = 0$ equivale ad accettare una risposta simmetrica della volatilità agli shock di mercato e dunque assenza di leverage effect.

Nel caso usuale in cui $\gamma < 0$, l'intensità dello shock può essere scomposta a seconda del segno dell'errore: in particolare lo *shock coefficient* sarà $\alpha + \gamma$ se $\frac{r_{t-i}}{\sigma_{t-i}} < 0$, mentre se $\frac{r_{t-i}}{\sigma_{t-i}} > 0$ sarà pari a $\alpha - \gamma$.

I modelli EGARCH presentano in genere una migliore capacità di adattamento rispetto ai GARCH standard, grazie al termine asimmetrico che consente di catturare l'effetto leva. Inoltre, l'uso degli *shock* standardizzati nella dinamica della log-varianza riduce l'effetto degli *outlier*.

Il modello GJR-GARCH, introdotto da Glosten, Jagannathan e Runkle nel 1993[57], estende il modello GARCH modellando diversamente rispetto all'EGARCH l'asimmetria dell'impatto degli errori sulla varianza condizionata dei titoli azionari.

L'equazione di un GJR-GARCH (1,1,1) è la seguente:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 I_{|r_{t-1}<0|} + \beta \sigma_{t-1}^2$$

(2.23)

Dove *I* è una funzione indicatrice che vale 1 se $r_{t-1} < 0$ e 0 altrimenti.

In questo caso il coefficiente γ deve essere positivo, mentre in caso sia uguale a zero il modello coincide con il GARCH(1,1).

Esistono due versioni molto simili al GJR-GARCH, il TGARCH e l'aGARCH.

Il TGARCH modella direttamente la deviazione standard condizionata della serie storica, anziché la sua varianza.[58] Ciò permette al modello di reagire più rapidamente agli *shock* rispetto al GJR, anche se la mancata elevazione al quadrato complica l'interpretazione dei coefficienti.

[57] Glosten, L. R., Jagannathan, R., & Runkle, D. E., *On the relation between the expected value and the volatility of the nominal excess return on stocks*. Journal of Finance, 48(5), 1993, pp. 1779–1801.
[58] J. M. Zakoïan, *Threshold heteroskedastic models*. Journal of Economic Dynamics and Control, 18(5), 1994, pp. 931–955.

La formulazione di un TGARCH (1,1,1) è la seguente:

$$\sigma_t = \omega + \alpha|r_{t-1}| + \gamma|r_{t-1}|I_{|r_{t-1}<0|} + \beta\sigma_{t-1}$$

(2.24)

L'aGARCH invece modifica direttamente il termine di errore al quadrato, sottraendogli il coefficiente γ. Questo modello introduce un effetto asimmetrico più attenuato, risultando più adatto ai mercati caratterizzati da volatilità moderata. Tuttavia risulta più difficile da stimare rispetto agli altri due modelli.[59] L'equazione di un aGARCH (1,1,1) nella formulazione proposta da Ding, Granger & Engle nel 1993 è:

$$\sigma_t^2 = \omega + \alpha(r_{t-1} - \gamma I_{t-1})^2 + \beta\sigma_{t-1}^2$$

(2.25)

In tutti questi modelli la stima dei parametri avviene utilizzando il metodo della massima verosomiglianza o, qualora il modello non sia ben specificato, mediante la quasi-massima verosomiglianza (QML)

Nel corso degli anni la letteratura econometrica ha prodotto una sconfinata varietà di ulteriori versioni del GARCH, che non affrontiamo nella nostra analisi in quanto non sono concretamente utilizzate per la previsione della volatilità delle variabili finanziarie a scopo di risk management. La maggior parte di questi ulteriori modelli appartengono alla famiglia dei cosiddetti GARCH-X, cioè dei modelli GARCH cui vengono aggiunte delle variabili esogene, come ad esempio il volume degli scambi, la differenza tra il numero di ordini di acquisto e di vendita (*order flow)*, la liquidità o altre *dummy variables* che catturano l'effetto delle notizie sul mercato.

I modelli esaminati in questo capitolo, appartenenti alla famiglia GARCH, costituiscono il quadro di riferimento tradizionale per la stima della volatilità. Tuttavia negli ultimi anni, i progressi dell'intelligenza artificiale e del *machine learning* hanno aperto nuove prospettive di ricerca, fondate sulla capacità di apprendere strutture complesse e non lineari direttamente dai dati.

[59] Ding, Z., Granger, C. W. J., & Engle, R. F., *A long memory property of stock market returns and a new model*. Journal of Empirical Finance, 1(1), 1993, pp. 83–106.

Capitolo 3: Machine Learning per la stima della volatilità

3.1 Intelligenza artificiale e machine learning

La terza rivoluzione industriale, avviata negli anni '50, è stata caratterizzata dalla digitalizzazione, con l'introduzione dei *mainframe computer*, dei *personal computer* e, successivamente, di Internet**.** Tutto ciò ha trasformato fortemente il mondo del lavoro, permettendo a programmi informatici di gestire le attività di routine, senza la necessità di eseguirle manualmente.

L'intelligenza artificiale è considerata una delle tecnologie abilitanti della quarta rivoluzione industriale[60], in quanto favorisce l'introduzione di nuove soluzioni produttive, il miglioramento delle condizioni di lavoro, la creazione di modelli di business innovativi, l'aumento della produttività e il miglioramento della qualità dei prodotti.

L'intelligenza artificiale, oggi al centro del dibattito internazionale, rappresenta nel suo significato più ampio lo sviluppo di metodi che consentano a un sistema artificiale, tipicamente informatico o di automazione, di replicare e simulare una forma di intelligenza umana.

Secondo lo standard ISO/IEC 42001:2023[61], l'intelligenza artificiale è definita come «la capacità di un sistema informatico di mostrare abilità tipicamente umane, quali il ragionamento, l'apprendimento, la pianificazione e la creatività».

Il *machine learning* costituisce una branca dell'intelligenza artificiale. Sebbene i due termini vengano talvolta usati in modo intercambiabile, il machine learning si distingue in quanto consente ai sistemi informatici di apprendere e migliorare le proprie prestazioni in modo autonomo. In particolare si concentra sullo sviluppo di algoritmi in grado di apprendere da insiemi di dati, identificandone delle sistematicità, così da produrre delle previsioni, utilizzando metodi di analisi statistica.

Intuitivamente un sistema è in grado di apprendere se è in grado di migliorare le proprie prestazioni attraverso la sua stessa attività. Nel machine learning ciò avviene innanzitutto

[60] Boston Consulting Group, *Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries*, 2015.
[61] ISO/IEC (2023). *ISO/IEC 42001:2023 – Information technology – Artificial intelligence – Management system*. International Organization for Standardization.

grazie ad una grande quantità di dati a disposizione (*big data*), che permette di accrescere la comprensione dei fenomeni, rendendo stabili le condizioni a partire dalle quali effettuare le previsioni. Una conseguenza diretta riguarda l'evoluzione tecnologica dei computer, i quali devono garantire capacità di memorizzazione elevate e rapidità nell'elaborazione dei dati.

Arthur Samuel, informatico statunitense che introdusse nel 1959 il termine in relazione al gioco della dama, definì il machine learning come la capacità dei computer di apprendere da insiemi di dati senza essere esplicitamente programmati.[62]

Secondo Herbert Simon l'apprendimento automatico riguarda qualsiasi cambiamento in un sistema che gli permetta di avere prestazioni migliori nella ripetizione di uno stesso compito o di un altro compito, nella stessa popolazione.[63]

Tom Mitchell nel 1997 ha formalizzato la definizione di apprendimento automatico come segue: «Un programma informatico apprende dall'esperienza E, rispetto a una certa classe di compiti T e a una misura di prestazione P, se le sue prestazioni nei compiti di T, misurate attraverso P, migliorano con l'esperienza E».[64]

I compiti di apprendimento automatico T riguardano l'elaborazione di un esempio, considerato come un insieme di caratteristiche che possono essere quantificate.

La misura di performance P è una valutazione quantitativa delle capacità dell'algoritmo di apprendimento automatico. Di solito la prestazione P è specifica per il tipo di compito T e in particolare, la scelta di P si basa sulla natura della variabile di output del modello: se questa è categorica (cioè assume valori discreti) si utilizzerà un metodo di classificazione, se deve essere quantificata (e dunque è continua nel tempo) si utilizzeranno metodi di regressione. Queste sono le due principali divisioni dell'apprendimento automatico con supervisione, che rileverà per la nostra analisi.

[62] A. L. Samuel, *Some studies in machine learning using the game of checkers*. IBM Journal of Research and Development, 3(3), 1959, pp. 210–229.
[63] A. Simon, *On the behavioral and rational foundations of economic dynamics*. Journal of Economic Behavior and Organization, 5(1), 1984, pp. 35–55.
[64] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

Di solito ci interessa sapere come si comporta un sistema di apprendimento automatico su dati che non ha mai incontrato prima, poiché è questo che ne determina il funzionamento nel momento in cui viene utilizzato nel mondo reale.

È fondamentale inoltre testare e confrontare diversi modelli, al fine di valutarne l'affidabilità e selezionare quello con le migliori prestazioni.

L'addestramento degli algoritmi avviene inizialmente su un *training set* (insieme di dati di apprendimento) per stimarne i parametri; successivamente il modello viene validato su un *validation set*, al fine di verificarne la capacità di generalizzare su dati non osservati. Infine, una volta fissati i parametri si allena nuovamente il modello sia sul training sia sul validation set e si verifica su un ulteriore set di dati disgiunto dai precedenti (*test set*), valutandone l'accuratezza rispetto a un *benchmark* attraverso misure di performance.

Nella pratica, in presenza di dataset di grandi dimensioni, i dati vengono generalmente suddivisi in circa 60% per il *training set*, 20–30% per il *validation set* e 10–20% per il *test* set. Un elemento cruciale dell'apprendimento automatico è la valutazione delle prestazioni, che consente di misurare la capacità del modello di generalizzare su dati non osservati. Senza una valutazione, non sarebbe possibile parlare di miglioramento, mentre a sua volta, la valutazione delle prestazioni richiede la capacità di accettare un certo tipo di informazioni dall'ambiente esterno.

Nelle applicazioni finanziarie da tempo gli investitori utilizzano le serie storiche per prevedere i valori futuri e basano le loro scelte su tali risultati. Con la disponibilità di nuovi strumenti e tecnologie, ma anche a causa della crescente complessità e volatilità dei mercati finanziari si è passati da semplici statistiche univariate a modelli di regressione e, successivamente, a modelli più complessi. L'intelligenza artificiale e l'apprendimento automatico (ML) possono essere considerati come la fase successiva di questo processo.

I modelli econometrici osservati nello scorso capitolo, seppur solidi da un punto di vista teorico, presuppongono una struttura specifica della volatilità condizionata, di tipo lineare e parametricamente definita. Inoltre, tendono a sottostimare eventi estremi e a rispondere lentamente a shock di mercato non convenzionali, motivo per cui risultano talvolta

insufficienti in contesti reali, caratterizzati da dati ad alta frequenza e forte non linearità, come ad esempio la volatilità nei mercati finanziari.[65]

La principale innovazione apportata dall'apprendimento automatico è che le regole che gli algoritmi utilizzano per prendere decisioni sono apprese dall'insieme dei dati e non sono programmate dall'uomo. L'esperienza viene quindi presentata sotto forma di dati di addestramento.

Seppur collegate, esiste una differenza metodologica fondamentale che distingue la statistica dal machine learning: nella prima, il processo parte dalla formulazione di un'ipotesi da verificare attraverso la raccolta e l'analisi dei dati; nel secondo, al contrario, si raccolgono inizialmente i dati e successivamente si testano diversi modelli, selezionando infine quello che meglio descrive i fenomeni osservati. Anche la terminologia di base risulta differente: mentre statistici ed econometrici trattano variabili indipendenti e dipendenti, chi si occupa di *data science* affronta caratteristiche (*features*) e obiettivi (*targets*).

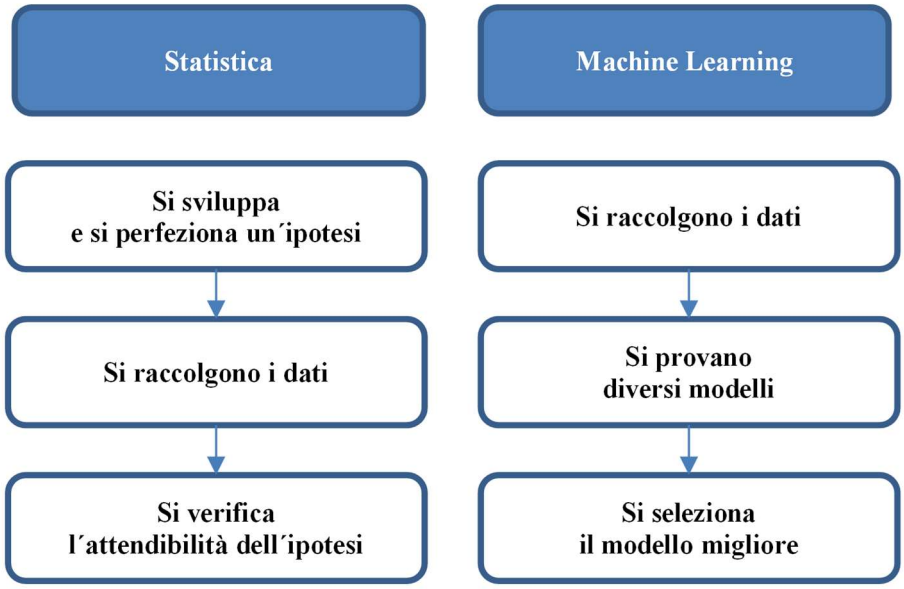

Figura 3.1: Metodologie utilizzate in statistica e nel machine learning

[65] S.-H. Poon, C. W. J. Granger. *Forecasting volatility in financial markets: A review*. Journal of Economic Literature, 41(2), 2003, pp.478–539.

3.2 Categorie di modelli di Machine Learning

Esiste un'ampia gamma di modelli di machine learning da utilizzare a seconda dei dati disponibili e del compito previsto. Questi modelli sono raggruppati in quattro categorie principali:

1. Apprendimento senza supervisione (*unsupervised learning*); non si occupa di prevedere il valore di una certa variabile, ma di comprendere meglio l'ambiente descritto dai dati. In particolare, il compito dell'algoritmo è riconoscere delle sistematicità presenti nei dati in input (le descrizioni), come ad esempio strutture latenti, pattern ricorrenti o raggruppamenti significativi (clustering), senza alcuna indicazione specifica sui valori corretti in uscita (le soluzioni). Le tecniche di clustering, volte a formare gruppi di osservazioni con caratteristiche omogenee, rappresentano lo strumento principale utilizzato in questa branca del machine learning. Gli algoritmi non supervisionati cercano di trovare una struttura nell'input che consenta una nuova rappresentazione delle informazioni disponibili nei dati, piuttosto che fare previsioni sul futuro. Di conseguenza, in assenza di un riferimento noto (*ground truth*), la valutazione dei modelli non supervisionati risulta meno oggettiva e spesso dipende da criteri di interpretazione qualitativa. L'apprendimento non supervisionato è utilizzato per l'analisi esplorativa dei dati, per effettuare tattiche di *cross-selling*, per la segmentazione dei consumatori e per l'identificazione delle immagini, finalità la cui spiegazione esula gli scopi di questo lavoro.

2. Apprendimento con supervisione (*supervised learning*); riguarda i modelli previsivi, utilizzati per stimare i valori futuri di variabili sia discrete che continue. In particolare, consiste nell'addestramento di un modello su un insieme di dati etichettati, ovvero in cui è noto il valore target (*output*) associato a ogni osservazione (*input*). In questo contesto, l'obiettivo dell'algoritmo è apprendere una regola generale, in particolare una funzione predittiva che identifichi la relazione tra dati in entrata e in uscita. La funzione stimata viene successivamente applicata a nuovi dati, al fine di prevedere correttamente il valore della variabile di output. Per una nuova osservazione in input, l'algoritmo deve essere in grado di stimare correttamente il valore atteso dell'output. Questo tipo di apprendimento è largamente utilizzato in problemi di regressione (come la previsione della volatilità) e classificazione (come la solvibilità di un soggetto). Le due tipologie sono influenzate

dalla variabile di output: se questa è categorica si utilizzerà un metodo di classificazione, se deve essere quantificata si utilizzeranno metodi di regressione.

Per raggiungere questi obiettivi è necessario supervisionare la procedura di apprendimento e tra i dati utilizzati per questa categoria del machine learning figurano le etichette (*labels*), cioè i valori numerici della variabile obiettivo che si vuole prevedere e le caratteristiche (*features*), cioè tutte le variabili indipendenti utilizzate per le previsioni. Mentre le *labels* sono un'esclusiva dell'apprendimento supervisionato, le *features*, riguardano anche l'*unsupervised learning*.

3. Apprendimento con supervisione parziale (*semi-supervised learning*); rappresenta un approccio ibrido tra i primi due metodi osservati e consiste nella ricerca di una previsione avendo a disposizione sia dati con labels per la variabile obiettivo, sia senza *labels* (di solito in numero di gran lunga superiore), questi ultimi utili per formare dei *cluster* utili per le previsioni.

In questo caso, poiché solo una frazione dei dati è etichettata, le metodologie di apprendimento non supervisionato (dati senza *labels*) vengono utilizzate per rilevare la struttura delle caratteristiche, mentre l'apprendimento supervisionato (dati con *labels*) viene impiegato per creare modelli per fare previsioni.

Questo metodo è praticamente utilizzato dagli esseri umani ed è replicato da molti algoritmi di machine learning che si basano sullo studio dei modi in cui i nostri cervelli elaborano i dati.

4. Apprendimento per rinforzo (*reinforcement learning*); a differenza dei primi tre metodi, che riguardano decisioni indipendenti tra di loro, quest'ultimo si applica a contesti dinamici, in cui l'agente deve prendere decisioni sequenziali interagendo con l'ambiente e ricevendo ricompense (*rewards*) o penalità (*costs*) in funzione delle proprie azioni.

L'obiettivo dell'algoritmo è massimizzare il valore atteso dei ricavi futuri, apprendendo esclusivamente per tentativi ed errori all'interno dell'ambiente considerato.

Questo settore è considerato uno tra i più interessanti e attivi dell'apprendimento automatico. Viene utilizzato per la guida autonoma e per l'addestramento dei robot, mentre nel campo finanziario può essere impiegato per la copertura dinamica delle opzioni e la gestione del portafoglio.

La famiglia dei modelli di apprendimento supervisionato è la più utilizzata nell'ambito dell'apprendimento automatico, nonché quella che occorrerà per la nostra analisi.

I modelli più sofisticati, che utilizzano una serie di variabili, sono in grado di descrivere le relazioni con maggiore profondità. Tuttavia, è anche più probabile che imparino un rumore casuale appartenente a uno specifico campione di addestramento. Quando ciò si verifica, si parla di sovradattamento (*overfitting*) del modello ai dati di addestramento. Inoltre, i modelli complessi potrebbero essere più difficili da interpretare, rendendo più ardua la comprensione della natura delle relazioni alla base delle previsioni, mentre al contrario i modelli più semplici potrebbero non essere sufficientemente flessibili e produrre risultati distorti.

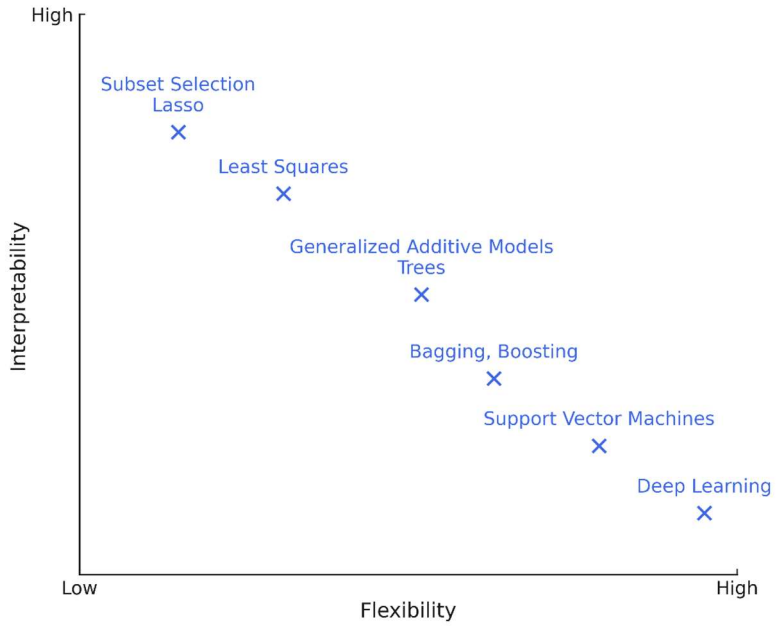

Figura 3.2: Trade-off tra flessibilità e interpretabilità nei modelli di machine learning

Ad oggi, il machine learning è largamente utilizzato non solo da *hedge fund* quantitativi e banche d'investimento, ma anche da autorità di vigilanza, gestori patrimoniali e imprese fintech, per applicazioni che spaziano dal risk management all'allocazione di portafoglio[66], fino all'asset pricing.[67]

[66] Si veda a riguardo: Y. Ma, R. Han, W. Wang, *Portfolio optimization with return prediction using deep learning and machine learning*, Expert Systems with Applications, 165, 2021

[67] Per un'applicazione dei metodi di machine learning all'asset pricing si vedA: S. Gu, B. Kelly, D. Xiu, *Empirical Asset Pricing via Machine Learning*, Review of Financial Studies, 33(5), 2020, pp. 2223–2273.

3.3 Il modello XGBoost

Il primo modello di machine learning impiegato in questo lavoro appartiene alla categoria dell'apprendimento supervisionato, in particolare alla famiglia degli algoritmi di *gradient boosting*, e prende il nome di XGBoost **(***Extreme Gradient Boosting)*. Questo metodo si fonda sulla combinazione (*ensemble***)** di diversi alberi decisionali addestrati in sequenza, utilizzando la tecnica del *gradient boosting*, con l'obiettivo di correggere progressivamente gli errori commessi dai modelli precedenti.

Gli alberi decisionali rappresentano uno dei modelli più semplici e interpretabili del machine learning supervisionato, e vengono utilizzati sia per compiti di regressione che di classificazione. Il modello prende il nome dalla sua struttura ad albero, nella quale ogni nodo interno rappresenta una condizione logica applicata a una variabile esplicativa (detta *feature*), ogni ramo indica il risultato di tale condizione, e ogni foglia fornisce un output: un valore stimato nel caso di regressione, oppure una classe nel caso di classificazione. Il nodo radice contiene la caratteristica con il più elevato contenuto informativo (*information content*).

Un albero decisionale suddivide ricorsivamente il dataset in sottoinsiemi via via più omogenei rispetto alla variabile target. Ad ogni passo si selezionano la variabile e la soglia di *split* che massimizzano l'omogeneità dei gruppi risultanti. Gli alberi si applicano sia a problemi di classificazione sia di regressione e sono apprezzati per la loro interpretabilità.

Nei problemi di classificazione, l'omogeneità è misurata attraverso funzioni obiettivo come l'indice di Gini o l'entropia:

L'indice di Gini quantifica il grado di impurità di un nodo come:

$$Gini(p) = 1 - \sum_{i=1}^{n} p_i^2 \qquad (3.1)$$

dove p_i è la probabilità di osservazioni nella i-esima classe.

L’entropia è una misura del disordine informativo ed è definita come:

$$Entropia(p) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

(3.2)

Entrambe le misure sono minime quando un nodo contiene osservazioni di una sola classe, e massime quando le classi sono equamente distribuite. Nei problemi di regressione, si utilizza tipicamente la varianza come funzione da minimizzare.

Sebbene siano semplici da interpretare, gli alberi decisionali presentano alcuni limiti. In particolare, sono modelli ad alta varianza, quindi molto sensibili ai dati su cui sono addestrati, e tendono a sovradattarsi se non sono opportunamente regolati. Per questo motivo, nella pratica si preferisce spesso utilizzarli all’interno di tecniche di *ensemble*, come il *Random Forest*, cioè la combinazione delle previsioni generate da diversi alberi decisionali, il *Boosting* o il *Gradient Boosting*, al fine di migliorarne la stabilità e la capacità predittiva.

In generale, il *boosting* è una tecnica di potenziamento che ha l’obiettivo di correggere progressivamente gli errori commessi da un modello, costruendo una sequenza di modelli deboli (*weak learners*) in modo tale che ciascun nuovo modello si concentri sugli esempi non correttamente predetti da quelli precedenti. Le previsioni finali non sono ottenute tramite una semplice media, ma attraverso una combinazione pesata delle uscite prodotte da ciascun modello. In particolare, i modelli che hanno ottenuto errori più bassi durante l’addestramento vengono valorizzati maggiormente, ricevendo pesi più alti, mentre quelli meno precisi hanno un peso inferiore. Questo meccanismo consente di ridurre l'influenza di modelli poco performanti, migliorando la robustezza dell’insieme.

Un esempio classico di questo approccio è l’*AdaBoost*, nel quale, dopo ogni iterazione, vengono aggiornati non solo i pesi dei modelli, ma anche i pesi delle osservazioni del dataset: gli esempi classificati erroneamente acquisiscono un peso maggiore, spingendo così i modelli successivi a porre maggiore attenzione su di essi.

Il *Gradient Boosting*, invece, rappresenta un'evoluzione più sofisticata di questa strategia. Anche in questo caso i modelli deboli (tipicamente alberi decisionali poco profondi)

vengono addestrati in sequenza, ma l'aggiornamento avviene sulla base del gradiente della funzione di perdita. In pratica, ogni nuovo albero è costruito per approssimare il residuo, ovvero l'errore commesso dal modello precedente, e quindi per ridurre progressivamente l'errore totale.

Il principio alla base di questa tecnica è un'ottimizzazione iterativa guidata dal gradiente della funzione di perdita, da cui il metodo prende il nome. A ogni iterazione, il modello aggiorna le sue previsioni nella direzione del gradiente negativo della funzione di perdita, seguendo un procedimento analogo a quello della discesa del gradiente (*gradient descent*) che osserveremo nel prossimo paragrafo, ma applicato a modelli compositi non lineari come gli alberi.

Formalmente, l'aggiornamento del modello può essere rappresentato nel seguente modo:

$$F_t(x_i) = F_{t-1}(x_i) + \eta f_t(x_i)$$

(3.3)

dove x_i è il vettore delle caratteristiche (*feature*) in input su cui il modello fa la previsione, mentre $f_t(x)$ è il nuovo albero decisionale addestrato al round t. Il parametro $\eta \in (0,1]$ è il cosiddetto *learning rate* che controlla l'intensità dell'aggiornamento: valori più piccoli richiedono un maggior numero di alberi rallentando l'aggiornamento, ma tendono a favorire la generalizzazione; valori più grandi accelerano l'apprendimento ma aumentano il rischio di overfitting. La previsione per il campione i al passo t sarà dunque $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$.

Il *Gradient Boosting* è molto flessibile e può adattarsi a funzioni di perdita personalizzate, rendendolo ideale per una vasta gamma di problemi. Tuttavia, è più suscettibile all'*overfitting* rispetto ad altri metodi, soprattutto quando viene lasciato senza regolarizzazione o su set di dati molto variabili. Tra le implementazioni più avanzate e diffuse del metodo Gradient Boosting spicca il modello XGBoost (eXtreme Gradient Boosting), sviluppato da Tianqi Chen nel 2016.[68] Il suo successo è legato alla capacità di coniugare accuratezza predittiva, capacità di adattamento dimensionale (scalabilità) e

[68] T. Chen, C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), ACM, New York, 2016, pp. 785–794.

efficienza computazionale (robustezza), tanto da diventare uno degli algoritmi di riferimento nelle competizioni di *data science* e nelle applicazioni industriali, inclusi i settori bancario e finanziario.

Il modello XGBoost rappresenta un'ulteriore evoluzione del Gradient Boosting tradizionale, rispetto al quale introduce diverse ottimizzazioni sia sul piano teorico che su quello pratico. La principale è l'introduzione esplicita all'interno della funzione obiettivo di tre termini di regolarizzazione: L1 (*lasso*, con coefficiente α) che aggiunge la somma dei valori assoluti dei valori delle foglie $|w_j|$ e tende ad azzerarne molti rendendo il modello più sparso (*feature selection* implicita); L2 (*ridge*, con coefficiente λ) che aggiunge la somma dei quadrati dei valori delle foglie w_j^2 riducendone l'ampiezza ma senza azzerarli (*shrinkage*), così da stabilizzare il modello; γT, un termine proporzionale al numero di foglie T (con coefficiente γ), che penalizza alberi eccessivamente complessi. Posti insieme, questi termini limitano l'overfitting e migliorano la capacità di generalizzazione, soprattutto in presenza di molte *feature*.

L'addestramento dell'XGBoost minimizza una funzione di perdita globale regolarizzata:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{m} \Omega\left(f_k\right)$$

(3.4)

dove y_i è il valore reale dell'osservazione i, $\hat{y}_i^{(t)} = F_t(x_i)$ è la predizione effettuata dal modello al passo t, $l(y_i, \hat{y}_i)$ è la funzione di perdita (ad esempio nei problemi di regressione è tipico utilizzare l'errore quadratico medio), f_k è l'albero che viene aggiunto alla *k*-esima iterazione, $\Omega(f_k)$ è il termine di regolarizzazione che penalizza la complessità del *k*-esimo albero e ϕ rappresenta l'insieme di tutti i parametri del modello, cioè le strutture degli alberi, i valori delle foglie e tutti gli iperparametri impliciti.

La regolarizzazione $\Omega(f_k)$ assume la seguente forma:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j|$$

(3.5)

dove $L1 = \alpha \sum_{j=1}^{t} |w_j|$ e $L2 = \frac{1}{2}\lambda \sum_{j=1}^{t} w_j^2$. Il valore della *j*-esima foglia w_j, anche detto *score leaf*, rappresenta la stima del valore target per tutti i campioni che terminano in quella foglia, ed è su di esso che il modello XGBoost applica le penalizzazioni L1 e L2. Poiché il modello XGBoost è costruito in modo incrementale, a ogni iterazione t si mantiene fisso quanto appreso fino al *round* $\mathrm{t}-1$, cioè la previsione corrente $F_t(x)$ e si stima il miglior albero f_t da aggiungere, da aggiungere al modello, così che l'aggiornamento avvenga secondo l'equazione (3.3). Poiché al passo t sono ottimizzati soltanto i parametri del nuovo albero, la parte di regolarizzazione relativa agli alberi precedenti resta costante: per questo motivo è necessario passare da una visione globale $\mathcal{L}(\phi)$ a una visione locale $\tilde{\mathcal{L}}^{(t)}$, considerando solo il termine di regolarizzazione dell'albero al passo t. Indicando con $\hat{y}_i^{(t-1)}$ le previsioni correnti, risulta la seguente funzione obiettivo locale:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Omega(f_t)$$

(3.6)

Durante l'addestramento, l'algoritmo non può risolvere esattamente questa funzione in quanto risulta non differenziabile. Dunque si ricorre a un'approssimazione della funzione di perdita usando una espansione di Taylor al secondo ordine, attorno al vettore delle previsioni correnti $\hat{\boldsymbol{y}}^{(t-1)}$, nella quale compaiono il gradiente g_i e l'hessiano h_i della perdita valutati in $\hat{y}_i^{(t-1)}$:

$$l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \cong\ l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2$$

(3.7)

Scartando il primo termine, che essendo costante non influenza l'ottimizzazione, la funzione da minimizzare diventa:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{I=1}^{n} [\, g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \Omega(f_t)$$

(3.8)

dove $\Omega(f_t)$ assume la stessa forma della (3.5).

Così facendo abbiamo la possibilità di incorporare sia il gradiente g_i che misura l'errore, sia il secondo derivato h_i che misura la convessità della funzione di perdita, rendendo l'ottimizzazione più precisa ed efficiente.

Dato che ogni osservazione x_i cade in una sola foglia j del nuovo albero f_t ed ogni foglia è associata ad un valore costante w_j, si avrà che:

$$f_t(x_i) = w_j$$

(3.9)

per ogni $i \in I_j$, dove I_j è l'insieme di osservazioni assegnate alla foglia j.

Poiché un albero restituisce un valore di foglia costante per tutte le osservazioni che vi ricadono, possiamo riscrivere la funzione obiettivo locale in termini di foglie j:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2} H_j w_j^2] + \Omega(f_t)$$

(3.10)

dove T è sempre il numero totale di foglie, $G_j = \sum_{i \in I_j} g_i$ è la somma dei gradienti nella foglia j e $H_J = \sum_{i \in I_j} h_i$ è la somma degli hessiani nella foglia j.

È importante ricordare che, durante la derivazione del valore ottimale delle foglie w_j, viene temporaneamente omesso il termine L1, ponendo il peso $\alpha = 0$, in quanto il termine $|w_j|$ non è derivabile in 0 e rende difficile ottenere una soluzione analitica chiusa.

Considerando solo la parte L2 della regolarizzazione e unificando i termini elevati al quadrato si ha:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$

(3.11)

Ogni foglia dell'albero ha un output ottimale w_j^* calcolato in forma chiusa e utilizzato per aggiornare il modello globale. Per calcolarlo, è necessario derivare i termini rispetto a w_j e imporre la condizione di minimo:

$$\frac{\partial \tilde{\mathcal{L}}^{(t)}}{\partial w_j} = G_j + \big(H_j + \lambda\big) w_j = 0$$

(3.12)

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

(3.13)

Una volta calcolato il valore ottimale delle foglie w_j^* e riformulata la funzione obiettivo per foglia, il modello XGBoost utilizza queste quantità per valutare se conviene effettuare uno *split* su un determinato nodo dell'albero. Questa valutazione si basa sul guadagno informativo (*gain*), ovvero sulla riduzione della funzione obiettivo ottenuta suddividendo un nodo in due sottoinsiemi.

Recuperando la funzione obiettivo approssimata per foglia e il valore ottimale di ogni foglia, possiamo sostituire w_j^* nella funzione obiettivo per calcolare la perdita ottimizzata associata a un nodo intero (non ancora diviso):

$$\tilde{\mathcal{L}}_{nodo} = -\frac{1}{2} \cdot \frac{G^2}{H + \lambda}$$

(3.14)

dove $G = \sum_{i \in I} g_i$, $H = \sum_{i \in I} h_i$ e I è l'insieme delle osservazioni che ricadono all'interno del singolo nodo.

A questo punto supponiamo di voler dividere questo nodo in due sottogruppi, sinistro (L) e destro (R). Otteniamo le rispettive somme: $G_L = \sum_{i \in I_L} g_i$, $H_L = \sum_{i \in I_L} h_i$, $G_R = \sum_{i \in I_R} g_i$ e $H_R = \sum_{i \in I_R} h_i$ dove I_L e I_R sono gli insiemi delle osservazioni che cadono rispettivamente nel sotto-nodo sinistro e destro.

La funzione obiettivo dopo lo *split* sarà la somma delle perdite ottimizzate di ciascuna foglia:

$$\tilde{\mathcal{L}}_{split} = -\frac{1}{2} \cdot \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda}\right)$$

(3.15)

Il guadagno informativo (*gain*) prodotto dallo split è definito come la riduzione della funzione obiettivo, al netto della penalità strutturale γ per l'aggiunta di una nuova foglia:

$$Gain = \frac{1}{2} \cdot \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + L}\right) - \gamma$$

(3.15)

dove $G = G_L + G_R$ e $H = H_L + H_R$.

Il *gain* misura la riduzione della funzione obiettivo, cioè l'errore totale più la regolarizzazione ottenuta dallo *split*. Se il *gain* è positivo e sufficientemente grande, lo *split* è considerato vantaggioso. Il parametro γ rappresenta il costo di aggiungere una nuova foglia ed è sottratta una sola volta per ciascuno *split* candidato, che viene effettuato solo se la riduzione dell'errore al netto di γ risulta positiva. Il parametro η, cioè il *learning rate* non compare mai nel calcolo del *gain*: lo *shrinkage* viene applicato dopo la scelta dello split, in fase di aggiornamento delle previsioni e quindi non altera il criterio di selezione. Operativamente l'algoritmo calcola inizialmente i gradienti g_i e gli hessiani h_i per ciascuna osservazione e di seguito valuta ogni possibile *split* calcolando il *gain* associato. Successivamente si effettua lo *split* solo se il *gain* supera γ, così da contenere

l'*overfitting*. È importante ricordare che l'intero albero viene costruito *greedy*, scegliendo a ogni passo lo *split* con massimo *gain* senza osservare le possibili combinazioni future.

Altri ulteriori vantaggi del modello XGBoost rispetto al generico *gradient boosting* sono i seguenti:

-Crescita dell'albero decisionale per impostazione predefinita: XGBoost fa crescere gli alberi per livello (*level-wise* o *depthwise*): a ogni profondità vengono valutati e divisi i nodi dello stesso livello prima di passare al successivo. Questo schema consente di espandere in profondità solo quando lo *split* produce un guadagno sufficiente sulla funzione obiettivo, generando strutture più bilanciate e stabili a fronte di una minore aggressività rispetto agli approcci *leaf-wise*, cioè espandendo solo la foglia col gain più alto, anche se più lontana dalla radice dell'albero. È un approccio flessibile che, se ben regolarizzato, permette di apprendere relazioni complesse, pur richiedendo in ogni caso attenzione per evitare *overfitting*.

- Gestione avanzata dei valori mancanti: l'algoritmo è in grado di gestire automaticamente la presenza di dati mancanti, determinando in fase di addestramento il percorso ottimale da seguire nei nodi decisionali anche in assenza di un valore osservato. Questo consente di sfruttare al meglio l'informazione contenuta nella distribuzione dei dati, senza la necessità di interventi manuali di imputazione o rimozione degli stessi.

- Strategia di potatura ottimizzata (*pruning*): Al termine della costruzione dell'albero, i rami che non contribuiscono in modo significativo alla riduzione della funzione di perdita vengono rimossi. Questo meccanismo consente di migliorare la generalizzazione del modello, eliminando complessità strutturali non necessarie e riducendo il rischio di *overfitting*.

- Parallellizzazione del calcolo: grazie a una struttura dati interna altamente ottimizzata, nota come *DMatrix*, l'algoritmo è in grado di rendere paralleli l'elaborazione dei dati e la costruzione degli alberi in modo efficiente. Ciò rende il modello XGBoost adatto anche all'analisi di dataset di grandi dimensioni, tipici in contesti finanziari o industriali.

3.4 Le reti neurali artificiali

Oltre ai modelli basati sul *boost* del gradiente, esiste un'ulteriore famiglia di algoritmi che ben si presta a descrivere e prevedere l'evoluzione di variabili finanziarie: le reti neurali artificiali (*artificial neural networks*, ANN), note anche come percettroni multi-strato (*multi-layer perceptrons*), modelli computazionali ispirati alla struttura e al funzionamento dei neuroni biologici del cervello umano.

Il primo lavoro accademico sull'argomento risale al 1943, quando due neurofisiologi, McCulloch e Pitts, scrissero un articolo presentando un modello semplificato che descriveva il funzionamento dei neuroni e la struttura di una rete neurale.[69]

Successivamente, nel 1958, Rosenblatt pubblicò la versione aggiornata dell'articolo in cui introdusse formalmente il concetto di percettrone, cioè l'equivalente di un neurone ma considerando una rete neurale artificiale.[70]

Le reti neurali artificiali hanno iniziato a diffondersi alla fine degli anni Ottanta, e da subito se ne è compresa la forte predisposizione per scopi di *forecast*, in particolare in ambito finanziario. I pionieri in questo campo utilizzarono le prime versioni di modelli *black-box* per intercettare pattern non lineari nei dati di mercato, seppur con capacità computazionali molto limitate.[71]

Questo approccio suscitò da subito un grande entusiasmo, ma cadde in disuso a causa dello sviluppo di altri modelli di apprendimento automatico, come le SVM (*Support Vector Machines*), il *boosting* e le foreste casuali. Ciò avvenne in parte a causa della richiesta delle reti neurali di molti interventi manuali, mentre i nuovi metodi risultavano più automatici, superando in molti problemi le reti neurali poco addestrate.

Nel frattempo, nel corso del primo decennio del 2000, complici la crescente disponibilità di dati ad alta frequenza e il progresso nei sistemi di calcolo (GPU, architetture parallele, *cloud computing*), diversi gruppi di appassionati di reti neurali si sono impegnati ad

[69] McCulloch, W. S.; Pitts, W. "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics*, 5, 1943, pp.115–133.
[70] Rosenblatt, F. "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, 65(6), 1958, pp. 386–408.
[71] Hutchinson, J. M.; Lo, A. W.; Poggio, T. "A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks", *The Journal of Finance*, 49(3), 1994, 851–889.

applicare questa tecnologia su architetture di calcolo sempre più sofisticate e su set di dati sempre più ampi, migliorando sensibilmente la potenza degli algoritmi.

Le reti neurali artificiali sono riemerse prepotentemente dopo il 2010, caratterizzate da nuove architetture, ulteriori funzionalità e una serie di successi su alcuni problemi come la classificazione di immagini e video e la modellazione del parlato e del testo, ponendosi come pietra miliare di una nuova branca del machine learning, nota con il nome *deep learning* (apprendimento profondo). La ragione principale del successo di questi algoritmi è sicuramente la disponibilità di insiemi di dati di addestramento sempre più ampi, resa possibile dall'uso su larga scala della digitalizzazione nella scienza e nell'industria.

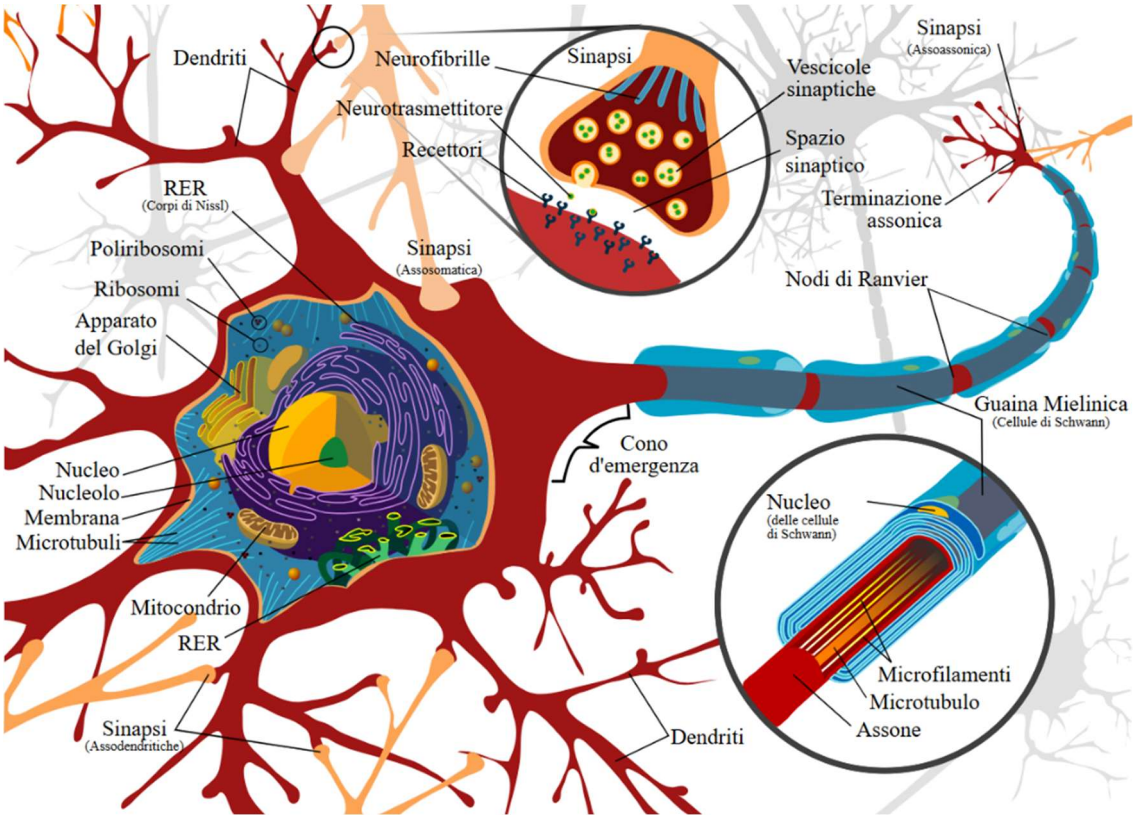

Figura 3.3: Struttura biologica di una rete neurale

3.4.1 Struttura di una rete neurale artificiale

Una rete neurale con un singolo strato nascosto è una forma semplice, seppur potente, di modello predittivo non lineare. È composta da tre strati principali: uno strato di input (*input layer*), composto dalle variabili esplicative, cioè le caratteristiche; uno strato nascosto (*hidden layer*) contenente i percettroni, che hanno il compito di elaborare le informazioni attraverso trasformazioni non lineari; e uno strato di output (*output layer*), che produce la previsione finale. Dunque la variabile obiettivo non è direttamente collegata alle caratteristiche, ma entrambe sono collegate ai neuroni, che svolgono un ruolo di ponte.

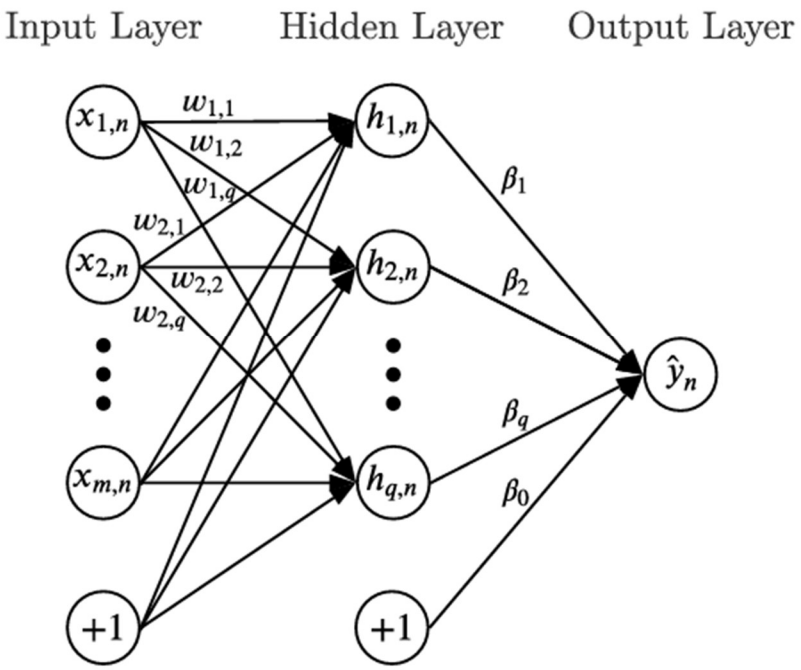

Figura 3.4: Struttura di una rete neurale artificiale feedforward

Partendo dall'immagine, possiamo definire i parametri e le variabili della rete neurale:

Il parametro $w_{m,q}$ rappresenta il peso che lega la m-esima caratteristica (*input layer*) con il q-esimo neurone. Il numero complessivo dei pesi è uguale al prodotto tra le caratteristiche e i neuroni. Il parametro β_q è invece il peso che lega il q-esimo neurone alla variabile obiettivo, cioè la previsione finale calcolata dal modello, $\hat{y}_n$.

Per calcolare il valore di un generico percettrone è necessario moltiplicare ogni caratteristica per il rispettivo peso che lo collega al percettrone e sommare i prodotti.

Contestualmente si aggiunge una costante, definita *bias* del modello, che risulta diversa per ogni collegamento e infine si applica alla quantità ottenuta una funzione, detta di attivazione (*activation function*).

Le reti neurali richiedono infatti che vengano specificate le funzioni che collegano le caratteristiche ai percettroni e i percettroni alla variabile obiettivo. In caso si abbiano m caratteristiche il valore di un generico neurone, è:

$$h_q = f(\alpha_q + \sum_{j=1}^{M} w_{j,q} x_j)$$

(3.16)

dove α_q è il *bias* e *f* è la funzione di attivazione prescelta.

La variabile obiettivo invece, sempre nel caso di M caratteristiche e H percettroni, sarà pari a:

$$\hat{y}_n = F(\beta_0 + \sum_{q=1}^{H} \beta_q h_q)$$

(3.17)

La funzione di attivazione deve essere specificata *ex ante* e deve presentare la caratteristica di essere non-lineare. Ipotizzare una funzione lineare, detta anche *identity activation function*, del tipo $f(y) = y$, equivarrebbe a far collassare il modello in una regressione lineare multipla. Ogni percettrone può essere dunque considerato come una trasformazione non lineare di una combinazione lineare delle caratteristiche.

Le funzioni di attivazione più diffuse sono le seguenti:

• Funzione logistica (sigmoidea), che per qualsiasi valore di y giace tra 0 e 1.

$$f(y) = \frac{1}{1 + e^{-y}} = \frac{e^{y}}{1 + e^{y}}$$

(3.18)

• Funzione tangente iperbolica (*hyperbolic tangent function*), o *tanh*. Questa funzione è simile alla logistica ma presenta valori compresi tra -1 e +1, anziché tra 0 e +1.

$$f(y) = \frac{e^{2y} - 1}{e^{2y} + 1} = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}$$

(3.19)

• Funzione di unità lineare rettificata (*rectified linear unit* – ReLU)

$$f(y) = \max(y, 0)$$

(3.20)

Pur essendo lineare nel suo tratto positivo, questa funzione presenta un punto di rottura (*kink*), consentendo alle reti che la utilizzano di andare oltre la semplice combinazione lineare degli input. Inoltre, la ReLU è molto efficiente computazionalmente, favorisce un apprendimento stabile e rapido, e come osserveremo evita problemi comuni come il *vanishing gradient*, risultando particolarmente adatta per compiti di regressione come la previsione della volatilità. Esiste anche un'ulteriore versione di questa funzione, la cosiddetta ReLU bucata (*leaky ReLU function*), che per $y < 0$ è uguale a αy, mentre coincide con la ReLU per $\alpha \geq 0$.

In ogni caso le funzioni di attivazioni non lineari sono principalmente utilizzate per collegare lo strato di input con lo strato nascosto, mentre nei compiti di regressione, per collegare quest'ultimo con lo stato di output, si tendono ad impiegare funzioni lineari. Tuttavia nei problemi di classificazione, in cui l'output è una probabilità, dunque un valore compreso tra 0 e 1, risulterebbe naturale utilizzare anche in uscita una funzione di attivazione logistica.

Una rete neurale può anche presentare diversi strati nascosti e in questo caso si parla di *multi-layer Artificial Neural Network*. Anche se avessimo T uscite multiple nello strato di output, ci sarebbe comunque una sola funzione-obiettivo, eventualmente posta uguale alla somma ponderata delle singole funzioni-obiettivo utilizzate per ognuno dei T obiettivi. Il numero ottimale degli strati nascosti e dei neuroni varia da problema a problema, ma in genere si procede per tentativi, aumentando gli strati e i neuroni finché si realizza che ulteriori aumenti comportano dei benefici molto ridotti in termini di

accuratezza. Tuttavia, il teorema di approssimazione universale di Hornik[72] afferma che le reti neurali artificiali con un singolo strato nascosto possono approssimare qualsiasi funzione continua in modo arbitrario. Per questo motivo, senza perdita di generalità, in questo lavoro si assumeranno reti neurali con un singolo strato nascosto.

3.4.2 Algoritmo di discesa del gradiente e retropropagazione

I parametri della rete neurale devono essere scelti in modo che le previsioni fornite dalla rete in base al training set siano il più vicino possibile ai valori del target. Per realizzare questo obiettivo è necessario definire una funzione-obiettivo da minimizzare, solitamente rappresentata dall'errore quadratico medio (MSE) o, se non si ha interesse nel penalizzare gli errori più consistenti, dall'errore assoluto medio (MAE). Considerando una sola osservazione, questa funzione prende il nome di *loss function*, mentre la somma delle funzioni di perdita è detta *cost function*.

Ad esempio, considerando l'errore quadratico medio, avremmo che:

$$Loss(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \tag{3.21}$$

$$Cost(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, \hat{y}_i) \tag{3.22}$$

Quando si costruisce una rete neurale, l'errore quadratico medio viene solitamente minimizzato utilizzando l'algoritmo di discesa del gradiente (*gradient descent algorithm*), che consente di definire iterativamente il punto di minimo di una funzione.

L'algoritmo consiste, dopo aver fissato dei valori iniziali del *bias* e dei pesi, nella determinazione della loro modifica, così da migliorare gradualmente la funzione obiettivo, scegliendo il sentiero che permette la discesa più rapida possibile verso il punto

[72] K.Hornik, Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), pp. 251–257.

di minimo (*steepest descent*). Si ottengono così nuove stime dei parametri, che vengono succesivamente modificate scegliendo nuovamente la via più rapida.

La lunghezza del passo è determinata dal tasso di apprendimento η (*learning rate*). In ogni iterazione si calcola il gradiente, cioè il tasso di variazione della funzione-obiettivo in funzione dei parametri del modello.

Definendo la *cost function* come $E(\theta)$, con $\theta = (w, \beta, \alpha)$, possiamo indicare gli aggiornamenti dei pesi di uscita β e dei pesi dello strato nascosto w nel seguente modo:

$$\Delta w = -\eta \nabla_w E(\theta)$$

(3.23)

$$\Delta \beta = -\eta \nabla_\beta E(\theta)$$

(3.24)

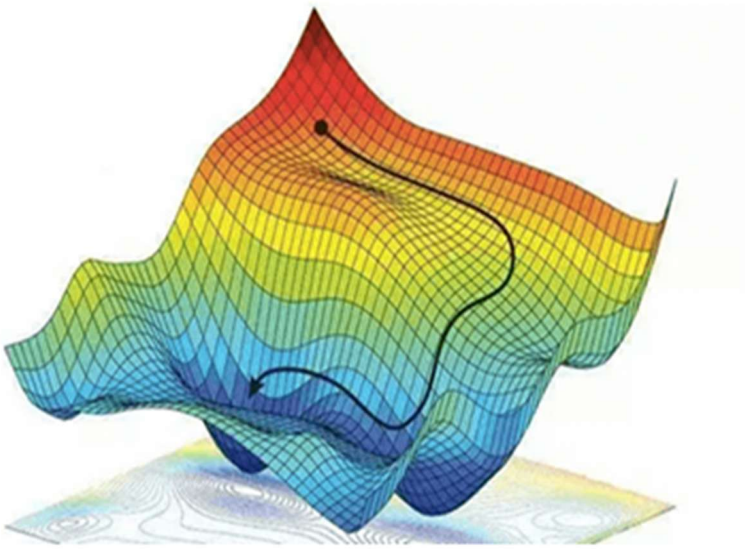

Figura 3.5: Rappresentazione grafica del gradient descent algorithm

Chiaramente, nel momento in cui si devono stimare più parametri, tutti i loro valori cambiano ad ogni iterazione. Quando il numero dei parametri da stimare risulta molto elevato, la determinazione del gradiente per ogni parametro potrebbe risultare proibitiva in termini di tempi di calcolo. Perciò è stata ideata una scorciatoia nota come retro-propagazione (*back-propagation*), un metodo che utilizzando la cosiddetta regola della

catena (*chain rule*)[73], permettemdo di tornare indietro dalla fine della rete neurale al suo inizio, calcolando le derivate parziali dell'errore quadratico medio (o di un'altra funzione-obiettivo) rispetto ai valori dei parametri della rete neurale.

Durante l'addestramento della rete tramite *backpropagation,* alcune funzioni di attivazione, come la sigmoide o la tanh, possono causare il cosiddetto problema del gradiente nullo (*vanishing gradient*), che si verifica quando le derivate calcolate per aggiornare i pesi diventano progressivamente più piccole attraversando gli strati, fino a diventare quasi nulle. All'aumentare del numero di passi di propagazione in avanti in una rete, ad esempio a causa di una maggiore profondità di rete, i gradienti dei pesi precedenti vengono calcolati con un numero crescente di moltiplicazioni. Queste moltiplicazioni riducono l'ampiezza del gradiente. Di conseguenza, i gradienti dei pesi precedenti saranno esponenzialmente inferiori ai gradienti dei pesi successivi.

In particolare, la funzione logistica tende a saturarsi per valori di input molto grandi o molto piccoli, rendendo le sue derivate prossime allo zero. Ciò potrebbe causare instabilità nel processo di apprendimento, soprattutto nei primi strati della rete, rischiando di rallentarlo o arrestarlo completamente. Per questo motivo, nelle reti neurali si preferisce utilizzare funzioni di attivazione come la ReLU, che non soffrono dello stesso problema nei valori positivi, rendendo l'addestramento più stabile ed efficiente.

Questo problema, oltre a caratterizzare le reti neurali semplici a più strati, riguarda anche le reti neurali ricorrenti, una classe di reti neurali artificiali che osserveremo nel prossimo paragrafo.

Nel momento in cui le reti neurali sono composte da decine di migliaia di parametri, continuando a modificare i parametri finché non minimizzano l'errore E nel *training set*, si avrebbe un modello molto complesso e si rischierebbe l'*overfitting*.

Una buona prassi è quella di continuare a migliorare il modello di machine learning, adattandolo ai dati e rendendolo più sofisticato, finché i risultati del validation set non iniziano a peggiorare. Pertanto, quando si implementa una rete neurale, è necessario calcolare la funzione di costo dopo ogni periodo per entrambi i set di dati, interrompendo

[73] Attraverso questo metodo si calcola la derivata delle funzioni composte come prodotto tra la derivata della funzione più esterna, con argomento invariato, e la derivata della funzione più interna.

l'addestramento (*early stopping*) quando la funzione di costo per il secondo set inizia ad aumentare. In particolare è necessario fissare un certo numero di epoche, detto *patience*, entro il quale si attende che la funzione di costo migliori. Se entro il valore fissato della pazienza non si osservano diminuzioni della funzione di costo, l'algoritmo deve tornare indietro e fermarsi sul modello con la minima funzione di costo per il *validation set*. Dunque, oltre ad addestrare la rete neurale, il software deve memorizzare i parametri che minimizzano la funzione di costo nel *validation set*.

3.5 Modelli di reti neurali per dati sequenziali

Fino a questo punto abbiamo trattato le reti neurali semplici (*plain vanilla*), dette anche *feedforward*. In questa tipologia di rete neurale le informazioni si muovono solo in una direzione rispetto a nodi d'ingresso, cioè esclusivamente in avanti attraverso i nodi nascosti fino ai nodi d'uscita e sono raffigurate in un grafico aciclico diretto. Quando l'algoritmo effettua i calcoli per un'osservazione, non ha memoria calcoli effettuati per gli obiettivi nei periodi precedenti, dunque gli input sono elaborati in modo indipendente per ogni previsione. Se venisse modificata la sequenza mediante la quale le caratteristiche alimentano la rete neurale, le stime non cambierebbero più di tanto.

3.5.1 Le reti neurali ricorrenti (RNN)

Introduciamo a questo punto le reti neurali ricorrenti (*recurrent neural networks*, RNN), in cui è rilevante la sequenza temporale con cui vengono osservate e presentate le caratteristiche e questo è importante nei casi in cui la relazione tra l'obiettivo e le caratteristiche tende a cambiare nel tempo. Mentre le reti *feedforward* non considerano gli input avvenuti a tempi precedenti, per cui l'output è determinato solamente dall'attuale input, le RNN sono costruite in modo da avere una memoria che si estende per un certo periodo di tempo: gli input della rete al tempo *t* sono sia i valori delle caratteristiche sia le previsioni compiute al tempo $t - 1$.

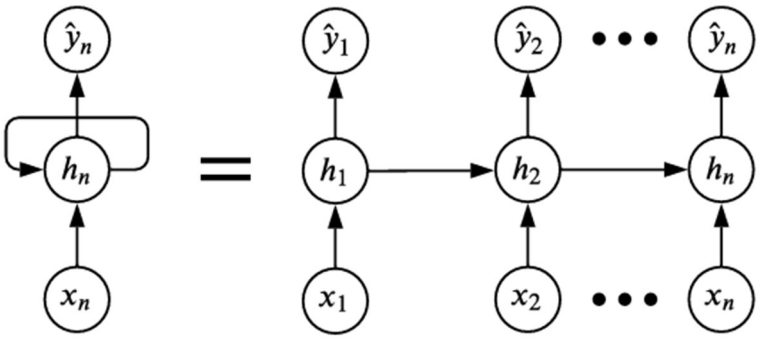

Figura 3.6: Struttura di una rete neurale ricorrente

Le reti neurali ricorrenti sono molto utilizzate nell'ambito della previsione delle serie storiche, in particolare in ambito finanziario. Ipotizzando di voler prevedere la prossima osservazione di una serie storica basandosi sull'ultimo valore osservato come sola caratteristica. Se considerassimo una funzione di attivazione lineare, la previsione sarebbe una combinazione lineare dell'osservazione più recente e della previsione di quell'osservazione, rendendo il modello molto simile all'EWMA, con l'aggiunta di un drift determinato dal *bias*. Questa semplice RNN è una generalizzazione non-lineare dell'EWMA.

Oltre che nel campo della meccanica statistica, l'origine delle reti neurali ricorrenti è da ricercarsi al pari delle reti neurali classiche, nel settore della neuroscienza, in cui il termine "ricorrente" viene utilizzato per descrivere strutture ad anello in anatomia.[74]

Durante gli anni '40, diversi studiosi proposero l'esistenza di un feedback nel cervello, anche se ciò risultava in contrasto con la precedente comprensione del sistema neurale come una struttura puramente *feedforward*. Anche il già citato documento di McCulloch e Pitts del 1943 considerava l'ipotesi di reti contenenti cicli e in cui la loro attività attuale poteva essere influenzata da attività indefinitamente lontane nel passato.

[74] Si vedano a tal proposito:
S. Ramón y Cajal, *Histologie du système nerveux de l'homme & des vertébrés*, Vol. II, A. Maloine, Paris, 1909, p. 149.
R. Lorente de Nó, "Vestibulo-Ocular Reflex Arc", *Archives of Neurology and Psychiatry*, 30(2), 1933, pp. 245–291.

Tuttavia, al pari delle classiche reti neurali artificiali a più strati, le RNN tradizionali soffrono del problema del *vanishing gradient*, che limita la loro capacità di apprendere dipendenze a lungo termine.

3.5.2 Le reti con lunga memoria a breve termine (LSTM)

Questo problema è stato affrontato e superato a partire dalla metà degli anni '90 con lo sviluppo dell'architettura a "lunga memoria a breve termine" (*Long Short-Term Memory*, LSTM) [75], rendendola la variante RNN standard per la gestione delle dipendenze a lungo termine. L'introduzione delle LSTM ha segnato un punto di svolta, consentendo di gestire con maggiore efficacia la memoria a lungo termine in serie storiche altamente autocorrelate, che come abbiamo osservato nello scorso capitolo risulta essere una caratteristica chiave nelle dinamiche della volatilità finanziaria.

Due anni dopo sono state introdotte le reti neurali ricorrenti bidirezionali (BRNN)[76], ottenute dalla combinazione di due RNN che elaborano lo stesso input in direzioni opposte, dando origine all'architettura LSTM bidirezionale. Questa tecnologia ha rivoluzionato il riconoscimento vocale, permettendo un enorme miglioramento nella traduzione automatica, nella modellazione del linguaggio e dell'elaborazione multilingue del linguaggio.

Le LSTM mirano a fornire una memoria a breve termine per le reti neurali ricorrenti, che può durare migliaia di intervalli temporali (da qui "memoria a *lungo* termine"). Il nome è stato fornito in analogia con la relazione tra memoria a lungo e a breve termine, studiata dagli psicologi cognitivi dall'inizio del XX secolo.

Le reti LSTM presentano diverse porte (*gates*) che determinano se mantenere un'informazione e, in caso affermativo, quanta informazione conservare nel sistema.

[75] S. Hochreiter; J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, 9(8), 1997, pp. 1735–1780.
[76] M. Schuster; K. K. Paliwal, "Bidirectional recurrent neural networks". *IEEE Transactions on Signal Processing*, 45(11), 1997, pp. 2673–2681.

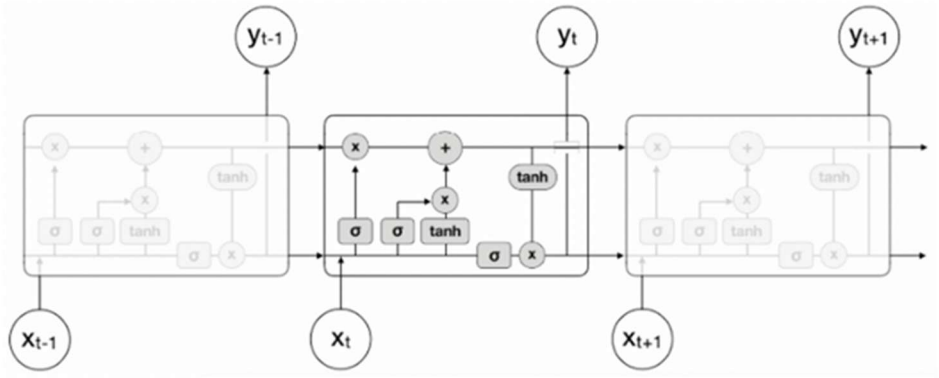

Figura 3.7: Struttura esterna di una LSTM

Se osservate dall'esterno, i nodi delle reti LSTM sembrano uguali a quelli delle RNN. Tuttavia gli strati nascosti delle LSTM risultano molto più complesse al suo interno, in quanto sono composte, oltre che da una cella di memoria (*cell state*, c_t), da tre porte (*gates*), una di ingresso (*input gate*, i_t), una di uscita (output gate, o_t) e una di oblio o dimenticanza (*forget gate,* f_t). Queste regolano l'informazione contenuta nella *cell state*, determinando se debba essere mantenuta e, in caso affermativo, quanta conservarne all'interno del sistema.

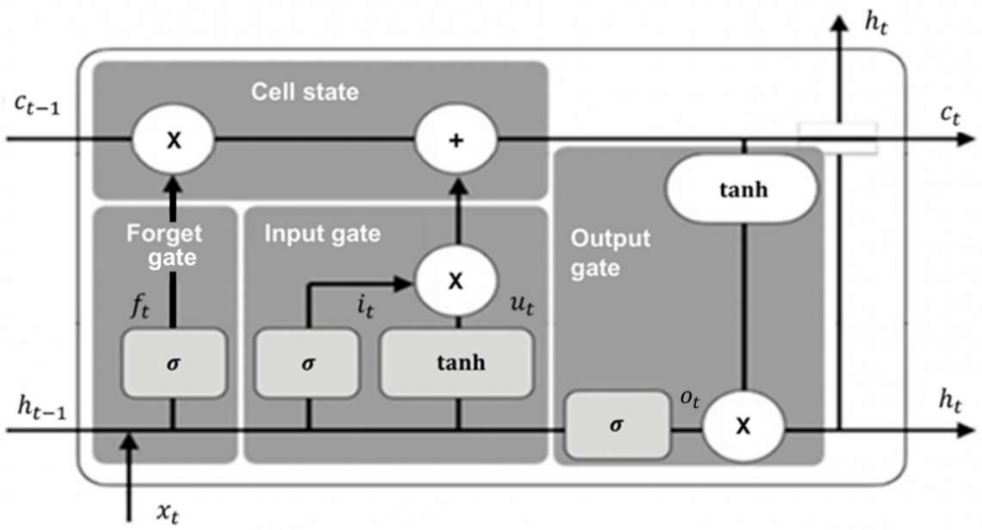

Figura 3.8: Struttura interna di una LSTM

L'idea alla base delle unità LSTM consiste dunque nell'utilizzare diverse strutture di *gating* che assumono la forma di unità moltiplicative elemento per elemento (utilizzando il prodotto di Hadamard, indicato con ⊙), per controllare il flusso di informazioni nella rete.

Ogni porta di ingresso determina quando un input è abbastanza significativo da dover essere memorizzato, se deve essere mantenuto o quando deve essere dimenticato. Analogamente, ogni porta di uscita determina se l'informazione deve essere trasmessa all'unità successiva o meno. Inoltre, il *cell state*, che si propaga nel tempo insieme all'output dell'unità LSTM, subisce cambiamenti minimi durante i diversi *step* temporali, a differenza dall'output stesso. Ciò significa che la derivata della funzione-obiettivo rispetto al *cell state* non decade né cresce esponenzialmente, permettendo al gradiente di mantenersi stabile. Questo garantisce che esista almeno un percorso in cui il gradiente non svanisce o esplode, rendendo la cosiddetta tecnica del *back-propagation through time* (BPTT) efficace nell'addestramento delle reti LSTM su sequenze molto lunghe.

Le informazioni scorrono sia verticalmente (dall'ingresso allo strato nascosto) sia orizzontalmente (tra osservazioni successive). Nella figura (3.8), i rettangoli grigio scuro rappresentano gli strati in cui i parametri vengono appresi durante l'addestramento, mentre i rettangoli grigio chiaro le funzioni di attivazione applicate. I cerchi e le ellissi bianche simboleggiano invece operazioni elementari moltiplicazioni ($\times$), somme ($+$) e applicazioni di funzioni non lineari (σ per la logistica, $tanh$ per la tangente iperbolica).[77] Il *cell state* è situato nella parte superiore dell'unità LSTM e presenta connessioni con tutti gli altri *gate*.

Data un'unita LSTM di dimensione q, possono essere definire formalmente l'aggiornamento dello strato interno come segue:

La porta di memoria (*forget gate*) è la parte che decide quanta memoria del passato conservare: se il contesto è cambiato, spinge a rimuovere le informazioni, se invece il *pattern* continua, invita a ricordarle.

$$f_t = sigm(W_f x_t + U_f h_{t-1)}$$

(3.25)

dove $x_t \in \mathbb{R}^p$ è il vettore di input al tempo t, $h_{t-1} \in \mathbb{R}^q$ è lo strato nascosto precedente,

[77] Nell'immagine si può osservare come la funzione $tanh$ è rappresentata una volta come rettangolo quando indica un'operazione di trasformazione a livello di layer, e in un'ellisse quando rappresenta l'applicazione elemento per elemento della funzione.

cioè l'informazione che la rete portava con sé dal passo $t-1$, $W_f \in \mathbb{R}^{q \times p}$ è la matrice dei pesi che collega l'input x_t al forget gate.

Elementi di f_t prossimi a 1 indicano di conservare la corrispondente informazione in c_{n-1}, mentre valori vicini a 0 portano a rimuovere l'informazione contenuta all'interno di tali componenti.

La porta di ingresso (*input gate*) controlla l'aggiunta di nuove informazioni alla memoria della cella. Essa decide quali valori dal dato di input (insieme agli stati passati) devono essere utilizzati per aggiornare lo stato interno della cella. In particolare è necessario calcolare due diversi termini, il candidato di cella ($\tilde{c}_t$) che propone l'informazione da aggiungere e il vero e proprio input gate (i_t) che stabilisce quanta informazione aggiungere effettivamente. Le relazioni che li definiscono sono le seguenti:

$$i_t = sigm(W_i x_t + U_i h_{t-1})$$

(3.26)

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1}\right)$$

(3.27)

Il *cell state* (c_t) viene aggiornato combinando l'informazione da ricordare dal passato e quella che si vuole aggiungere dal presente, attraverso il prodotto di Hadamard:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_{t-1}$$

(3.28)

La porta di uscita (*output gate*) infine determina quale porzione di memoria della cella utilizzare per produrre l'output della rete, cioè lo strato nascosto h_t, che diventerà l'input ricorrente al passo successivo:

$$o_t = sigm(W_o x_n + U_o h_{t-1})$$

(3.29)

$$h_t = o_t \odot \tanh\left(c_t\right)$$

(3.30)

Grazie alla struttura a porte, le LSTM mantengono una memoria interna stabile e duratura, attenuando il problema del vanishing gradient e rendendo possibile l'apprendimento su sequenze lunghe. In questo modo la rete intercetta sia le dipendenze di breve periodo sia relazioni più articolate che si sviluppano su orizzonti temporali estesi. Nel contesto della previsione su serie storiche, ciò si traduce nella capacità di collegare eventi anche lontani nel tempo: per esempio, in ambito azionario, una LSTM può mettere in relazione il prezzo corrente con il comportamento del mercato osservato giorni, settimane o mesi prima, anche quando tali segnali non risultano immediatamente evidenti.

Nel 2014 è stata introdotta una versione semplificata delle celle LSTM, rappresentata dalle GRU[78] (Gated Recurrent Units). A differenza delle LSTM tradizionali, che utilizzano tre porte distinte (input, forget e output gate), le GRU fondono la porta di input e di oblio in un'unica porta di aggiornamento (update gate) e utilizzano una porta di reset per gestire l'integrazione delle informazioni passate. Questa struttura più semplice riduce il numero di parametri e rende le GRU più leggere e veloci da addestrare, pur mantenendo prestazioni paragonabili a quelle delle LSTM, soprattutto in compiti come la modellazione del linguaggio, la traduzione automatica e altre applicazioni sequenziali.

[78] K. Cho; B. van Merriënboer; D. Bahdanau; Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

Capitolo 4: Analisi empirica

4.1 Scelta del dataset

A questo punto abbiamo tutti gli elementi teorici necessari per poter iniziare la nostra analisi empirica.

Consideriamo tre indici differenti, rappresentativi di tre aree geografiche nel mondo: lo Standard and Poor's 500 per gli Stati Uniti (SPX), lo Stoxx 600 Europe per l'Europa (STOXX) e l'MSCI AC Asia Pacific per Asia e Oceania (MXAP).[79] Questi indici risultano comparabili in quanto risultano ampi e diversificati; inoltre presentano tutti e tre la caratteristica di essere ponderati per la capitalizzazione di mercato (*value weighted*).

Le serie storiche dei prezzi di chiusura giornalieri di questi tre indici sono state estratte da LSEG Workspace (già Refinitiv). L'S&P 500 e l'MSCI AC Asia Pacific sono quotati in dollari, mentre lo Stoxx 600 Europe in euro.

Dopo aver eliminato eventuali duplicati per data, è stata calcolata l'intersezione delle date valide, così da garantire la coerenza temporale dell'analisi. In particolare, i dati relativi ai tre indici considerati sono stati allineati sulla base delle sole date in cui tutti e tre presentano un valore di chiusura valido. Questa operazione ha comportato l'eliminazione delle osservazioni relative a giornate in cui almeno uno degli indici non era negoziato (festività locali o assenze di dati).

Il risultato è un dataset armonizzato composto da osservazioni giornaliere, distribuite uniformemente tra i tre indici, utilizzato come base per tutta l'analisi successiva.[80]

Per effettuare la nostra analisi, il campione dei prezzi è stato suddiviso in tre *dataset* parzialmente sovrapposti tra di loro, così da poter analizzare tre esperimenti distinti.

[79] In particolare è stata scelta la versione di quest'indice comprensiva dei titoli azionari giapponesi. L'indice conta un totale di 1233 titoli.

[80] Sia nell'ambito dello sviluppo di modelli econometrici che di machine learning, è fondamentale il *data cleaning*: questo consiste nell'identificare e modificare o eliminare le rilevazioni incoerenti, i dati non rilevanti, le duplicazioni, i valori anomali, i dati mancanti.

Tabella 4.1: Suddivisione dei tre esperimenti in training, validation e test set

Experiment	Training Set	Validation Set	Test Set
1	01/1998 - 12/2013	01/2014 - 12/2019	01/2020 - 06/2025
2	01/2007 - 12/2017	01/2018 - 12/2021	01/2022 - 06/2025
3	01/2017 - 12/2021	01/2022 - 12/2023	01/2024 - 06/2025

Ipotizzando osservazioni giornaliere e considerando solo quelle comuni ai tre indici, possiamo suddividere la distribuzione delle osservazioni secondo il criterio comunemente adottato del 60/20/20, come evidenziato nella tabella seguente:

Tabella 4.2: Numero delle osservazioni per training, validation e test set

Experiment	Training Obs	Training %	Validation Obs	Validation %	Test Obs	Test %	Total Obs	Total %
1	3520	56.1%	1411	22.5%	1339	21.4%	6270	100.0%
2	2430	56.5%	1002	23.3%	867	20.2%	4299	68.5%
3	1215	58.4%	498	23.9%	369	17.7%	2082	33.2%

Nella pagina successiva può essere osservato il numero di osservazioni giornaliere disponibili per ogni anno in tutti gli anni considerati nell'analisi. (Tabella 4.3)

A partire dai prezzi giornalieri di chiusura disponibili risulta agevole calcolare i rendimenti logaritmici giornalieri per ogni indice in base a quanto abbiamo osservato nei precedenti capitoli. Nella pagina seguente sono riportate le statistiche descrittive dei rendimenti logaritmici giornalieri per i tre indici considerati nell'analisi: S&P 500 (SPX), STOXX Europe 600 (STOXX) e MSCI AC Asia Pacific (MXAP). (Tabella 4.4)

Queste sono calcolate sul totale delle osservazioni, che coincide con il primo dei tre dataset, cioè da gennaio 1998 a giugno 2025.

Tabella 4.3: Numero di osservazioni giornaliere per anno (01/1998–06/2025)

Anno	Numero di osservazioni	Anno	Numero di osservazioni
1998	223	**2012**	222
1999	217	**2013**	222
2000	213	**2014**	222
2001	218	**2015**	224
2002	222	**2016**	222
2003	222	**2017**	213
2004	221	**2018**	249
2005	218	**2019**	250
2006	216	**2020**	252
2007	220	**2021**	251
2008	223	**2022**	250
2009	225	**2023**	248
2010	219	**2024**	250
2011	218	**2025**	120

Tabella 4.4: Statistiche descrittive dei log-rendimenti giornalieri (01/1998-06/2025)

Summary Statistics	S&P 500	STOXX Europe 600	MSCI AC Asia Pacific
Count	6270	6270	6270
Mean	0.03%	0.01%	0.01%
Std	1.29%	1.25%	1.22%
Skewness	-0.4163	-0.4895	-0.3887
Excess Kurtosis	9.7635	6.8102	6.7633
Max	10.96%	9.41%	10.00%
Min	-12.77%	-12.19%	-10.57%
25%	-0.52%	-0.54%	-0.59%
50%	0.07%	0.06%	0.05%
75%	0.64%	0.63%	0.66%

I risultati mostrano una media giornaliera prossima allo zero per tutti gli indici e una deviazione standard di circa l'1,2–1,3%, con evidenza di asimmetrie lievemente negative e una marcata leptocurtosi. Quest'ultima conferma la presenza di code pesanti e ciò potrebbe contraddire le ipotesi di normalità su cui si basano i modelli econometrici tradizionali che abbiamo analizzato nello scorso capitolo. Oltre alla normalità dei dati nel prossimo paragrafo testeremo anche la loro stazionarietà.

4.2 Test di normalità e stazionarietà per i log-rendimenti

Per verificare formalmente la normalità della distribuzione dei log-rendimenti, ricorriamo a un test statistico largamente utilizzato in letteratura: il Jarque-Bera Test (JB), che applichiamo per livelli di confidenza dell'1, 5 e 10%.

Il Jarque-Bera Test si basa sui momenti centrali di ordine superiore, cioè sulla simmetria (skewness) e sulla curtosi (kurtosis). La sua statistica è definita come:

$$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4})$$

(4.1)

dove n rappresenta il numero di osservazioni, S la skewness campionaria e K la kurtosis campionaria. Sotto l'ipotesi nulla di normalità, la statistica JB segue asintoticamente una distribuzione χ^2 con 2 gradi di libertà. Valori elevati della statistica indicano una significativa deviazione dalla normalità dovuta a eccessiva asimmetria e/o curtosi. Il test è noto per la sua sensibilità in campioni ampi: anche deviazioni lievi dalla normalità possono portare al rigetto dell'ipotesi nulla.

I risultati evidenziano valori della statistica decisamente elevati e p-value prossimi a zero, confermando il rigetto dell'ipotesi di normalità a tutti i livelli di significatività considerati.

Tabella 4.5: Jarque-Bera Normality Test per l'intero dataset (1998-2025)

Index	JB Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	25085.15	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
STOXX	12366.81	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
MXAP	12108.25	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes

Questi risultati trovano conferma anche da un punto di vista grafico: osservando infatti le distribuzioni empiriche dei log-rendimenti dei tre indici in analisi, si osserva chiaramente che si discostano molto dalle rispettive distribuzioni normali con stessa media e varianza, in quanto presentano una maggiore concentrazione nei valori centrali e delle code molto più spesse, risultando leptocurtiche.

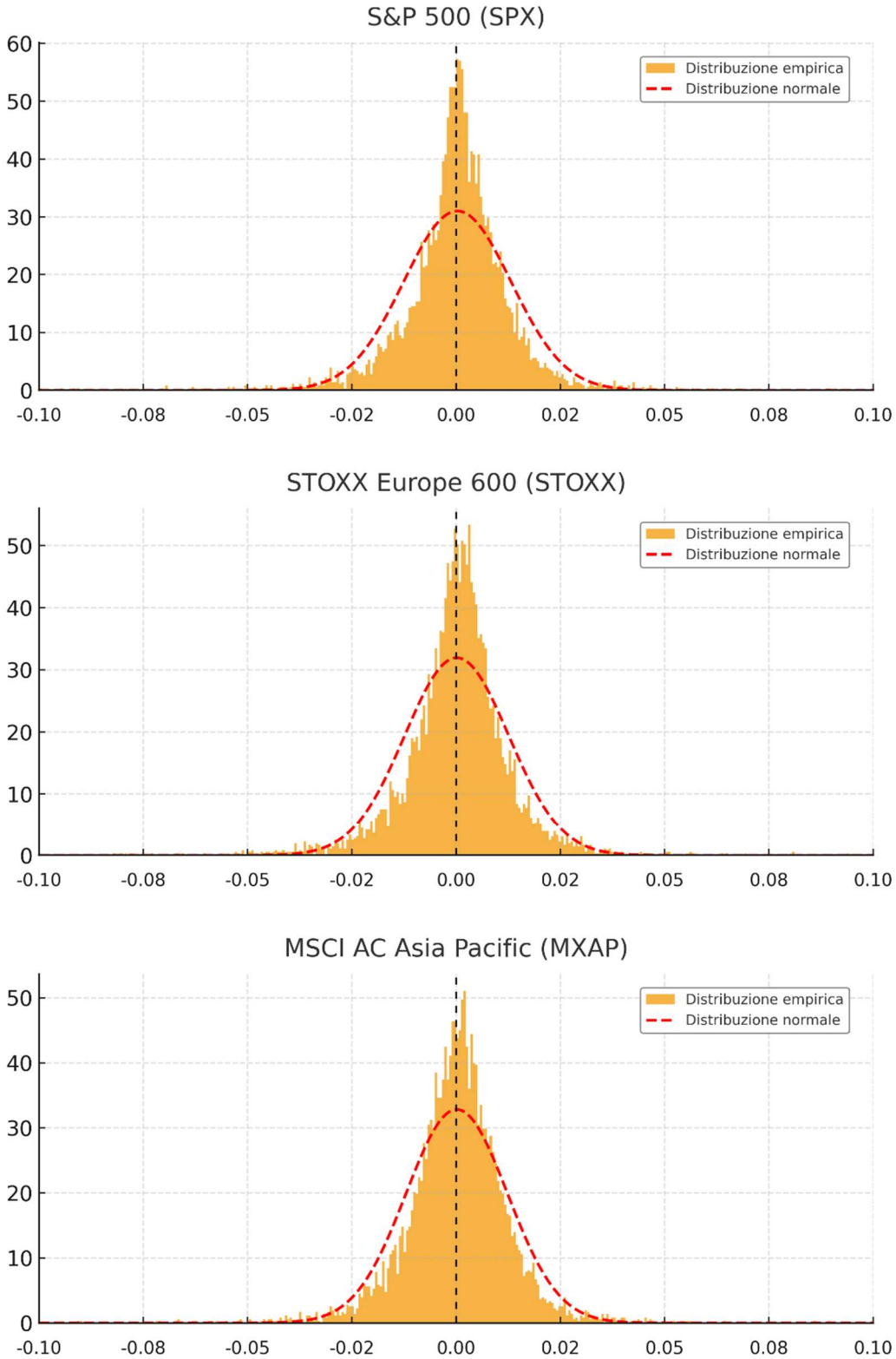

Figura 4.1: Confronto tra distribuzione normale ed empirica sull'intero dataset (1998-2025)

La presenza di leptocurtosi è ulteriormente confermata graficamente attraverso i QQ plot, che evidenziano deviazioni sistematiche rispetto alla linea rossa, che rappresenta la distribuzione normale, specialmente nelle code.

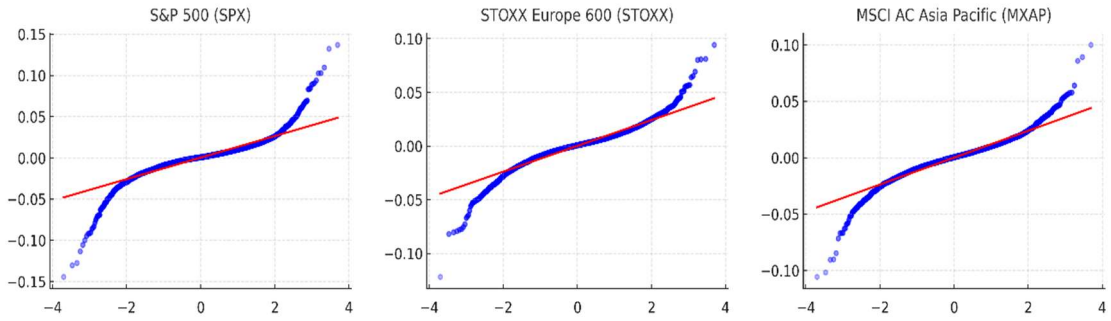

Figura 4.2: QQ Plot dei log-rendimenti sull'intero dataset (1998-2025)

Inoltre, per valutare la stazionarietà delle serie, è stato eseguito l'Augmented Dickey-Fuller Test (ADF), un'estensione del classico Dickey-Fuller Test che include ritardi supplementari della variabile dipendente per correggere l'eventuale autocorrelazione residua. In particolare il test verifica l'ipotesi nulla di non stazionarietà ($H_0: \gamma = 0$), contro l'alternativa che implica stazionarietà.

Il modello ADF può essere espresso nella forma:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{n} \delta_i \Delta y_{t-i} + \varepsilon_t$$

(4.2)

dove y_t è la serie in analisi, t una tendenza deterministica (opzionale), $\Delta y = y_t - y_{t-1}$ l'operatore differenza, e i termini anticipati delle differenze servono a correggere l'autocorrelazione residua. La statistica ADF è definita come il rapporto tra la stima del coefficiente γ e il suo errore standard:

$$ADF\ Stat = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

(4.3)

Questa statistica non segue una distribuzione t classica, ma viene confrontata con valori critici tabulati da Dickey e Fuller. Valori inferiori (cioè più negativi) rispetto ai valori soglia portano al rigetto dell'ipotesi nulla di non stazionarietà (H_0: $\gamma = 0$), suggerendo che la serie sia stazionaria. Nel caso in esame, le tre serie di log-rendimenti analizzate (S&P 500, STOXX Europe 600, MSCI AC Asia Pacific) riportano statistiche decisamente inferiori ai valori critici ai livelli di significatività convenzionali, accompagnate da p-value estremamente bassi. Le evidenze portano al rigetto dell'ipotesi nulla per tutte le serie, consentendo di concludere che i log-rendimenti sono stazionari nel periodo analizzato.

Tabella 4.6: Augmented Dickey-Fuller Stationarity Test per l'intero dataset (1998-2025)

Index	ADF Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	-13.83	< 0.001	-2.5672	Yes	-2.8624	Yes	-3.4323	Yes
STOXX	-20.18	< 0.001	-2.5672	Yes	-2.8624	Yes	-3.4323	Yes
MXAP	-26.93	< 0.001	-2.5672	Yes	-2.8624	Yes	-3.4323	Yes

Una volta assunto che la distribuzione normale non è la più adatta a descrivere i rendimenti logaritmici dei tre indici è necessario trovare un'alternativa. Ottenere una distribuzione che fitti al meglio i dati disponibili è fondamentale nella stima dei parametri dei modelli econometrici, mediante il metodo della massima verosomiglianza, che prevede di ipotizzare una distribuzione per i rendimenti. Stimare i parametri ipotizzando una distribuzione normale causerebbe il rischio di sottostimare osservazioni estreme, che qualora fossero negative rischierebbero di causare seri problemi in un'ottica di gestione del rischio di mercato. Viceversa i modelli di machine learning sono completamente non parametrici, e dunque non richiedono ipotesi sulla distribuzione dei rendimenti.

Per modellare la distribuzione empirica dei log-rendimenti, abbiamo scelto di adottare la distribuzione *t* di Student, parametrizzata da media μ, deviazione standard σ e gradi di libertà ν.

La funzione di densità corrispondente è definita come:

$$f(x;\mu,\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma}}[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2]^{-\frac{\nu+1}{2}}$$

(4.4)

dove $\Gamma(\cdot)$ rappresenta la funzione *gamma*. Per stimare i parametri del modello, si è adottato il metodo della massima verosimiglianza, che consiste nel massimizzare la funzione di log-verosimiglianza rispetto a μ, σ e ν. Tale funzione, data un campione di osservazioni $\{xi\}_{i=1}^{n}$, è espressa come:

$$\ell(\mu,\sigma,\nu) = n\left[\ln\Gamma\left(\frac{\nu+1}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\ln(\nu\pi) - \ln(\sigma)\right] - \frac{\nu+1}{2}\sum_{t=1}^{T}\ln(1+\frac{1}{\nu}(\frac{x_t-\mu}{\sigma})^2)$$

(4.5)

Data la complessità della forma analitica, l'ottimizzazione è stata condotta numericamente, consentendo di ottenere la stima di ν, la quale regola la pesantezza delle code della distribuzione.

Più ν è basso, più la distribuzione si discosta dalla normale e più frequentemente ci si aspetta rendimenti estremi (positivi o negativi), che la distribuzione normale sottostimerebbe.

Iniziamo osservando il numero di gradi di libertà ottimale per i tre indici, nell'intero campione di dati che abbiamo a disposizione:

Le stime di massima verosomiglianza dei gradi di libertà della *t* di Student sul totale delle osservazioni sono rispettivamente 2.78 per l'S&P 500, 2.93 per lo STOXX Europe 600 e 3.68 per l'MSCI AC Asia Pacific. Mentre l'indice asiatico è quello che tende maggiormente alla distribuzione normale, quello statunitense è quello che presenta un maggior livello di leptocurtosi. Nella pagina successiva possiamo osservare graficamente le distribuzioni dei log-rendimenti dei tre indici con le distribuzioni *t* di Student con i rispettivi gradi di libertà che meglio le approssimano.

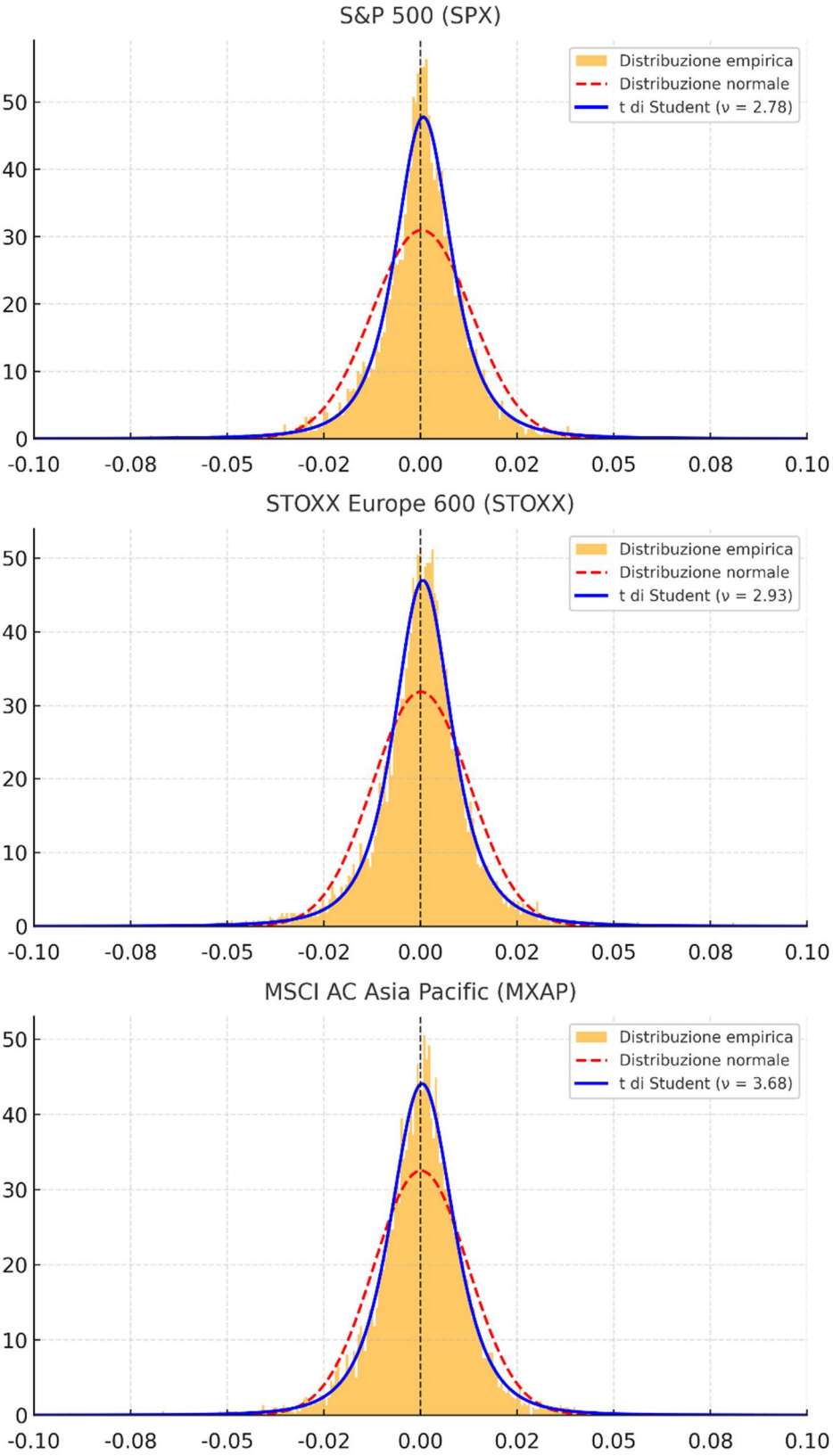

Figura 4.3: Confronto tra distribuzione normale e t di Student sull'intero dataset (1998-2025)

A questo punto possiamo iniziare la nostra analisi parallela dei tre indici sui tre diversi dataset. Per ogni esperimento, aggreghiamo il training e il validation set, in modo che diventino un unico training set per stimare i parametri dei modelli econometrici. Inizialmente cerchiamo la conferma dei test statistici per la stazionarietà e la normalità che abbiamo introdotto. Possiamo osservare come le indicazioni fornite dai test calcolati sull'intero campione di osservazioni sono confermate per tutti e tre i dataset: i log-rendimenti di training e validation set per ogni sottoperiodo risultano essere stazionari e non normali.

Tabella 4.7: JB Test, training+validation set primo esperimento (01/1998-12/2019)

Index	JB Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	13663.54	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
STOXX	5834.55	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
MXAP	8969.69	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes

Tabella 4.8: ADF Test, training+validation set primo esperimento (01/1998-12/2019)

Index	ADF Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	-13.28	< 0.001	-2.5671	Yes	-2.8621	Yes	-3.4317	Yes
STOXX	-16.26	< 0.001	-2.5671	Yes	-2.8621	Yes	-3.4317	Yes
MXAP	-15.75	< 0.001	-2.5671	Yes	-2.8621	Yes	-3.4317	Yes

Tabella 4.9: JB Test, training+validation set secondo esperimento (01/2007-12/2021)

Index	JB Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	6900.97	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
STOXX	3071.34	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
MXAP	4192.13	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes

Tabella 4.10: ADF Test, training+validation set secondo esperimento (01/2007-12/2021)

Index	ADF Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	-11.56	< 0.001	-2.5672	Yes	-2.8624	Yes	-3.4323	Yes
STOXX	-12.80	< 0.001	-2.5672	Yes	-2.8624	Yes	-3.4323	Yes
MXAP	-19.95	< 0.001	-2.5672	Yes	-2.8624	Yes	-3.4323	Yes

Tabella 4.11: JB Test, training+validation set terzo esperimento (01/2017-12/2023)

Index	JB Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	628.94	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
STOXX	531.53	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes
MXAP	1672.17	< 0.001	4.6052	Yes	5.9915	Yes	9.2103	Yes

Tabella 4.12: ADF Test, training+validation set terzo esperimento (01/2017-12/2023)

Index	ADF Stat	p-value	CV 10%	Reject (10%)	CV 5%	Reject (5%)	CV 1%	Reject (1%)
SPX	-43.52	< 0.001	-2.5677	Yes	-2.8632	Yes	-3.4342	Yes
STOXX	-19.74	< 0.001	-2.5677	Yes	-2.8632	Yes	-3.4342	Yes
MXAP	-15.85	< 0.001	-2.5677	Yes	-2.8632	Yes	-3.4342	Yes

Per completezza possiamo applicare la stima di massima verosomiglianza per stimare i gradi di libertà ottimali della *t* di Student per i log-rendimenti dei tre indici nei sottocampioni considerati. Come vedremo queste sottostime non saranno necessarie per la nostra analisi, ma possono essere utili al lettore per avere un quadro più completo.

Le stime sono riportate di seguito:

Tabella 4.13: Gradi di libertà ottimali per t di Student tramite MLE

Experiment	SPX	STOXX	MXAP
1	2.72	3.01	3.64
2	2.24	2.67	3.08
3	2.57	2.99	4.66

Di seguito riportiamo le distribuzioni dei tre dataset, sovrapponendo alle distribuzioni empiriche dei rendimenti logaritmici, oltre alle corrispondenti normali, le distribuzioni t di Student con i rispettivi gradi di libertà stimati:

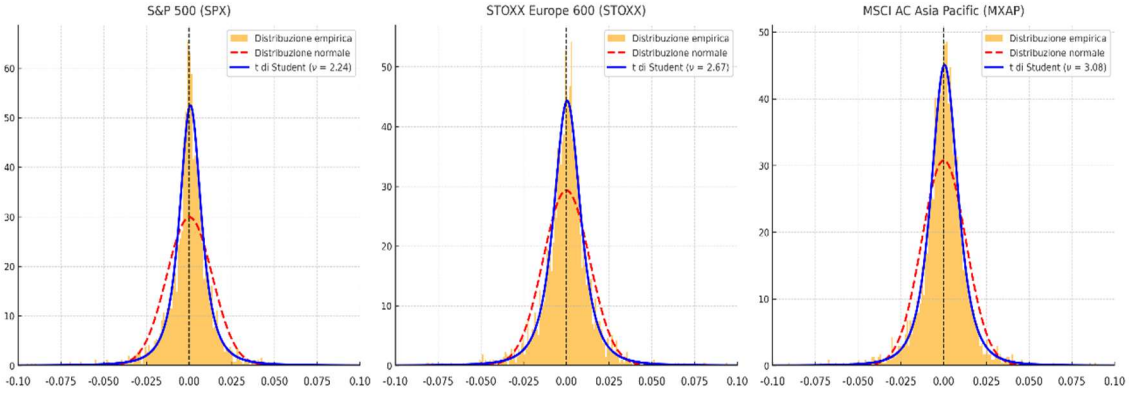

Figura 4.4: Normale vs t di Student, training+validation set primo dataset (1998-2019)

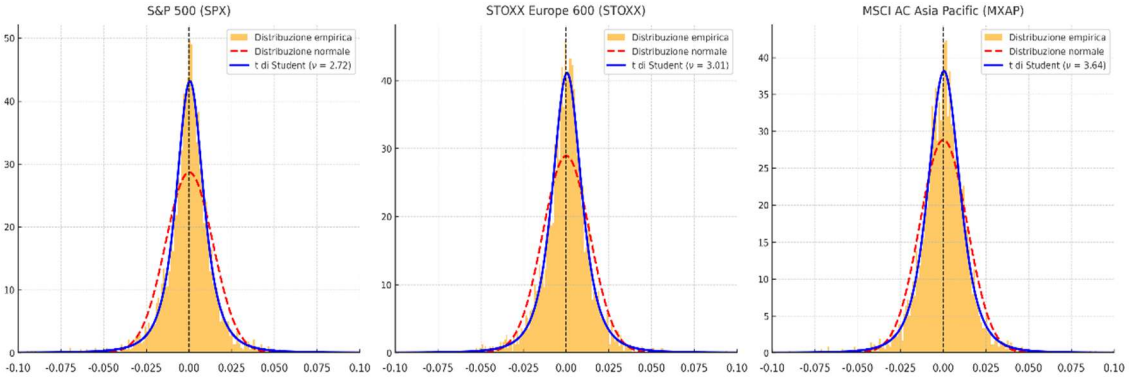

Figura 4.5: Normale vs t di Student, training+validation set secondo dataset (2007-2021)

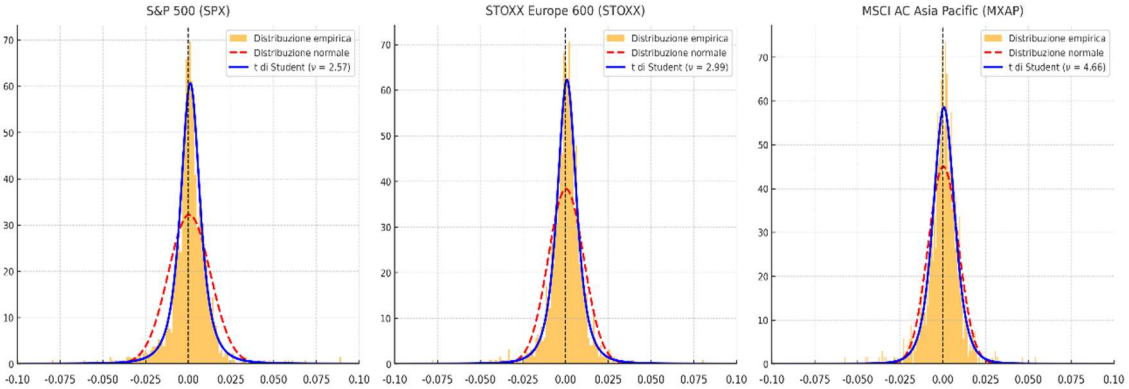

Figura 4.6: Normale vs t di Student, training+validation set secondo dataset (2007-2021)

4.3: Stima dei parametri dei modelli econometrici tradizionali

Si procede ora alla stima dei parametri dei modelli econometrici utilizzati nell'analisi: EWMA, GARCH(1,1) ed EGARCH(1,1,1), stimati tramite il metodo della massima verosimiglianza. L'EWMA con innovazioni distribuite come una t di Student, sebbene teoricamente attraente per modellare code pesanti, presenta limiti pratici rilevanti. La densità della t richiede infatti un fattore di scala, definito nel seguente modo:

$$scale_t = \sqrt{\sigma_t^2 \frac{\nu - 2}{\nu}}$$

(4.6)

che, per gradi di libertà prossimi a 2, tende a valori quasi nulli o indefiniti, rendendo la funzione di verosimiglianza piatta. Ciò spinge sistematicamente la stima di λ al suo valore massimo, cioè 1, producendo risultati degenerati. In secondo luogo, poiché la varianza di una t di Student è finita soltanto per $\nu > 2$, l'ottimizzazione congiunta di λ e ν risulta numericamente instabile; il vincolo $\nu > 2$ è essenziale ma spesso l'estremo ottimo ricade in corrispondenza di valori di ν troppo bassi o al limite della fattibilità, aggravando la degenerazione della stima. Dal punto di vista computazionale, il filtro EWMA-t richiede il calcolo ricorsivo di una scala non lineare e l'ottimizzazione di più parametri vincolati, operazione decisamente meno efficiente rispetto all'EWMA con innovazioni gaussiane, il cui *update* è invece banale e altamente stabile. Infine, quando l'obiettivo è spiegare le code pesanti e il *clustering* di volatilità sono preferibili modelli GARCH(1,1) con residui distribuiti secondo una *t* di Student, in grado di combinare in modo naturale la memoria di breve termine, tramite i parametri α e β, e le code pesanti, tramite ν (parametro che viene stimato contestualmente agli altri), senza il rischio di una stima degenerata di λ. Per questi motivi, nella pratica finanziaria si utilizza quasi esclusivamente l'EWMA classico (spesso con λ fissato a 0.94 come in RiskMetrics) e si ricorre ai GARCH-t per gestire le code pesanti.

Dunque nella nostra analisi la stima del parametro λ del modello EWMA avviene ipotizzando una distribuzione normale dei log-rendimenti.

Assumendo che i log-rendimenti r_t dato l'information set $\mathcal{F}_{t-1}$ siano distribuiti normalmente con media nulla e varianza σ_t^2, che si evolve nel tempo secondo l'equazione (2.12). Il valore iniziale σ_0^2 è fissato e nella nostra analisi è posto uguale al quadrato del rendimento del giorno precedente.

Per ottenere la stima di massima verosimiglianza del parametro $\lambda \in [0,1)$, si costruisce la seguente funzione di verosimiglianza congiunta:

$$L(\lambda) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2(\lambda)}} \exp\left(-\frac{r_t^2}{2\sigma_t^2(\lambda)}\right)$$

(4.7)

che, applicando il logaritmo, diventa:

$$\ell(\lambda) = -\frac{1}{2}\sum_{t=1}^{T}[\ln(2\pi) + \ln\left(\sigma_t^2(\lambda)\right) + \frac{r_t^2}{\sigma_t^2(\lambda)}]$$

(4.8)

In pratica si usa minimizzare la log-verosimiglianza negativa, detta appunto Negative Log-Likelihood:

$$NLL(\lambda) = -\ell(\lambda) = -\frac{1}{2}\sum_{t=1}^{T}[\ln\left(\sigma_t^2(\lambda)\right) + \frac{r_t^2}{\sigma_t^2(\lambda)}] + \frac{T}{2}\ln(2\pi)$$

(4.9)

Minimizzare NLL(λ) è equivalente a massimizzare L(λ), ma evita di manipolare prodotti potenzialmente instabili in precisione. Poiché la relazione tra NLL e λ non ammette una soluzione analitica (la dipendenza è infatti ricorsiva), si adotta una procedura di ottimizzazione numerica *bounded* su [0,1). Nel nostro caso abbiamo impiegato l'algoritmo di Brent, implementato tramite la funzione minimize scalar di SciPy con vincoli su λ. Questo metodo valuta in modo efficiente la NLL in diversi punti della griglia interna a [0,1) e converge rapidamente all'argomento λ che minimizza la funzione obiettivo, garantendo un compromesso ottimale tra stabilità numerica e accuratezza della stima.

Le stime del parametro λ del modello EWMA per i tre indici sono le seguenti:

Tabella 4.14: Stima del parametro λ del modello EWMA

Experiment	SPX	STOXX	MXAP
1	0.9210	0.9259	0.9334
2	0.9263	0.9217	0.9203
3	0.9096	0.9152	0.9235

Adesso è il momento di stimare i parametri dei modelli GARCH(1,1) ed EGARCH (1,1,1). In questi casi possiamo ipotizzare una distribuzione *t* di Student senza complicazioni, poiché nei modelli GARCH la distribuzione dell'innovazione entra esplicitamente nella funzione di verosimiglianza ed è quindi possibile stimarne i parametri, come i gradi di libertà della t, insieme a quelli della volatilità.

Al contrario, nel modello EWMA la volatilità è calcolata con una formula fissa e non derivata da una procedura di massima verosimiglianza: ciò implica che l'ipotesi di normalità dei rendimenti è incorporata e non modificabile senza ridefinire completamente il modello, rendendo di fatto impraticabile la stima di una distribuzione alternativa come la *t* di Student.

Iniziamo con il GARCH-t(1,1), ipotizzando che i rendimenti si distribuiscano come una *t* di Student con ν gradi di libertà. Nella stima di massima verosomiglianza si possono compiere due scelte differenti relativamente al parametro dei gradi di libertà. La prima è mantenerlo fisso in base a quanto abbiamo calcolato precedentemente, la seconda è stimarlo congiuntamente agli altri parametri del GARCH-t(1,1). In questa analisi si è scelto di stimare il parametro dei gradi di libertà congiuntamente agli altri parametri, ottenendo più flessibilità. Così facendo il modello tende a imparare la coda ottimale per i dati, adattandosi automaticamente a quanto siano intensi gli shock.

Dunque assumendo che i log-rendimenti soddisfino $r_t \sim t_\nu(0, \sigma_t^2)$, dove $t_\nu(0, \sigma_t^2)$ indica la *t* di Student standardizzata con media zero, gradi di libertà $\nu > 2$, e riscalata in modo da ottenere una varianza condizionata σ_t^2. In particolare, poiché la *t* di Student standardizzata ha varianza $\nu/(\nu - 2)$, è necessario moltiplicare la variabile casuale t_ν per

il fattore di scala $\sqrt{\frac{\nu-2}{\nu}}\sigma_t$. In base a quanto detto, l'innovazione standardizzata $z_t = \frac{r_t}{\sigma_t}$ può essere scritta, specificando che la sua distribuzione teorica sia una *t* di Student standardizzata con ν gradi di libertà e varianza pari a 1, nel seguente modo:

$$z_t^* = \frac{t_\nu}{\sqrt{\nu/(\nu-2)}}$$

(4.10)

Ponendo $r_t = \sigma_t z_t^*$, seguirà automaticamente $z_t \sim z_t^*$.

Possiamo recuperare la formula (2.20), che descrive la dinamica della varianza condizionata secondo il modello GARCH(1,1), ma ipotizzando che i rendimenti seguano il processo che abbiamo descritto. Si avrà dunque:

$$\sigma_t^2 = \omega + \alpha(\sigma_{t-1}^2 z_{t-1}^2) + \beta\sigma_{t-1}^2$$

(4.11)

con i vincoli $\omega > 0, \alpha \geq 0, \beta \geq 0, \alpha + \beta < 1$, per garantire positività e stazionarietà.

La densità della t standardizzata per $z_t = \frac{r_t}{\sigma_t}$ è:

$$f(z;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})}[1 + \frac{z^2}{\nu-2}]^{-\frac{\nu+1}{2}}$$

(4.12)

Assumendo indipendenza condizionata nel tempo, la funzione di verosomiglianza congiunta sui dati $\{r_t\}_{t=1}^T$ sarà:

$$L(\omega,\alpha,\beta,\nu) = \prod_{t=1}^{T} \frac{1}{\sigma_t} f(\frac{r_t}{\sigma_t};\nu)$$

(4.13)

Applicando il logaritmo otteniamo:

$$\ell(\theta) = \sum_{t=1}^{T}\left[\ln\Gamma\left(\frac{\nu+1}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\ln[(\nu-2)\pi] - \frac{1}{2}\ln(\sigma_t^2) - \frac{\nu+1}{2}\ln(1+\frac{r_t^2}{(\nu-2)\sigma_t^2})\right]$$

(4.14)

dove θ = (ω, α, β, ν).

Come nel caso dell'EWMA è usuale minimizzare la Negative Log-Likelihood $NLL(\lambda) = -\ell(\lambda)$ in quanto evita prodotti numericamente instabili. Poiché la relazione tra NLL e i parametri è ricorsiva (la varianza σ_t^2 dipende da σ_{t-1}^2 e da r_{t-1}^2), non esiste soluzione chiusa; si adotta quindi un'ottimizzazione numerica con i vincoli tipici del GARCH(1,1) L'algoritmo effettivamente impiegato è stato BFGS (Broyden–Fletcher–Goldfarb–Shanno), tipico per trovare minimi di funzioni differenziabili e con gradiente continuo.

Le stime dei parametri ω, α, β per i tre indici sono le seguenti:

Tabella 4.14: Stima dei parametri del GARCH-t(1,1), primo esperimento (1998-2019)

Index	ω	α	β	ν
SPX	0.00000211	0.1125	0.8747	5.91
STOXX	0.00000240	0.1123	0.8723	6.46
MXAP	0.00000191	0.0712	0.9159	6.00

Tabella 4.15: Stima dei parametri del GARCH-t(1,1), secondo esperimento (2007-2021)

Index	ω	α	β	ν
SPX	0.00000258	0.1378	0.8470	5.86
STOXX	0.00000302	0.1291	0.8510	5.96
MXAP	0.00000231	0.0917	0.8915	6.57

Tabella 4.16: Stima dei parametri del GARCH-t(1,1), terzo esperimento (2017-2023)

Index	ω	α	β	ν
SPX	0.00000237	0.1473	0.8365	6.06
STOXX	0.00000466	0.1620	0.7912	5.44
MXAP	0.00000432	0.1078	0.8451	6.45

Passiamo ora alla stima dei parametri del modello EGARCH (1,1,1), anche in questo caso ipotizzando che le innovazioni siano distribuite secondo una *t* di Student standardizzata con ν gradi di libertà e varianza unitaria.

Ricordando la relazione caratteristica di questo modello:

$$\ln(\sigma_t^2) = \omega + \alpha(|z_{t-1}| - \mathbb{E}[|z|]) + \gamma z_{t-1} + \beta \ln(\sigma_{t-1}^2)$$

(4.15)

Questa equazione risulta uguale alla (2.22) con la differenza che in questo caso $\mathbb{E}[|z|]$ è il valore atteso della *t* standardizzata, cioè:

$$\mathbb{E}[|z|] = \sqrt{\frac{\nu - 2}{\pi}} \frac{\Gamma\left(\frac{\nu - 1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$$

(4.16)

La log-verosomiglianza ha la stessa forma del caso base del GARCH-t, con la differenza che σ_t^2 è ottenuta dalla ricorsione logaritmica. L'ottimizzazione MLE procede analogamente al caso precedente, stimando $\theta = (\omega, \alpha, \beta, \gamma, \nu)$, soggetto ai vincoli $\alpha, \beta \geq 0, \nu > 2$. È bene ricordare che nel modello EGARCH (1,1,1) il coefficiente γ cattura l'asimmetria dell'impatto delle innovazioni sulla volatilità condizionale. Come possiamo osservare dall'equazione, un valore negativo di γ indica che shock negativi (*bad news*) incrementano la volatilità in misura maggiore rispetto a shock positivi (*good news*) di pari entità, in linea con il noto effetto leva (*leverage effect*) documentato nei mercati finanziari.

A differenza del GARCH classico, dove l'intercetta ω deve essere vincolata positiva per garantire σ_t^2, nell'EGARCH la positività della varianza è assicurata dalla funzione esponenziale e non è più necessario imporre $\omega \geq 0$. Un valore di $\omega < 0$ indica semplicemente che, in assenza di shock ($z_{t-1} = 0$), il *baseline* di log-varianza è negativo, ossia il livello di volatilità di lungo periodo risulta inferiore all'unità (in scala lineare). In termini operativi, $\omega < 0$ significa che il modello stima un livello medio di volatilità contenuto e non rappresenta un difetto di stima, ma piuttosto una naturale conseguenza della specificazione logaritmica del processo EGARCH.

Le stime dei parametri ω, α, β, γ e ν per i tre indici sono le seguenti:

Tabella 4.17: Stima dei parametri dell'EGARCH-t(1,1), primo esperimento (1998-2019)

Index	ω	α	γ	β	ν
SPX	-0.2120	0.1524	-0.1571	0.9757	5.74
STOXX	-0.1389	0.1614	-0.1327	0.9756	6.90
MXAP	-0.5213	0.2262	-0.1102	0.9405	5.53

Tabella 4.18: Stima dei parametri dell'EGARCH-t(1,1), secondo esperimento (2007-2021)

Index	ω	α	γ	β	ν
SPX	-0.3120	0.1784	-0.2093	0.9646	5.65
STOXX	-0.2355	0.1500	-0.1792	0.9734	5.94
MXAP	-0.1757	0.1473	-0.0998	0.9802	7.10

Tabella 4.19: Stima dei parametri dell'EGARCH-t(1,1), terzo esperimento (2017-2023)

Index	ω	α	γ	β	ν
SPX	-0.3265	0.1899	-0.1789	0.9635	6.65
STOXX	-0.3491	0.1350	-0.1929	0.9626	5.71
MXAP	-0.3638	0.1643	-0.0960	0.9611	8.22

Un aspetto che merita particolare attenzione riguarda l'interpretazione del parametro dei gradi di libertà ν associato alla distribuzione *t* di Student. Nel caso della stima diretta sui rendimenti, ν descrive la pesantezza delle code della distribuzione marginale dell'intera serie storica, risultando pertanto tipicamente più basso, in quanto ingloba sia l'eteroschedasticità sia gli effetti di *clustering* della volatilità. Nei modelli GARCH-t, invece, il parametro ν è stimato condizionatamente alla dinamica della volatilità: in questo contesto la distribuzione *t* non approssima più i rendimenti grezzi, ma i residui standardizzati (z_t), che sono già corretti per la componente autoregressiva della varianza. Di conseguenza, il valore stimato di ν tende ad aumentare, indicando code meno pesanti rispetto al caso marginale. Tale effetto risulta ancora più evidente nei modelli EGARCH-t, nei quali la struttura asimmetrica e la capacità di catturare il *leverage effect* riducono ulteriormente l'eccesso di curtosi lasciato nei residui. In sintesi, i tre valori stimati di ν non sono direttamente confrontabili tra loro, in quanto riflettono livelli diversi di modellizzazione: dalla distribuzione empirica dei rendimenti grezzi (*t* marginale), agli shock condizionati in un GARCH, fino ai residui più normalizzati dell'EGARCH.

4.4 Precisazioni sulla definizione di volatilità realizzata

Prima di introdurre i modelli di machine learning, è opportuno chiarire il concetto di volatilità realizzata, che sarà utilizzata sia come *benchmark* per confrontare le previsioni, sia come variabile *target* per l'addestramento. Come già osservato, la varianza (al pari della volatilità) è una variabile latente, dunque non ha una definizione univoca ma dipende sia dalla frequenza dei rendimenti al quadrato considerati (giornaliera, settimanale, mensile, ecc.) che dal numero di osservazioni che forniscono la base di calcolo. In tal senso, come abbiamo osservato nel capitolo 2, si può avere una varianza giornaliera calcolata su una certa base temporale, semplicemente considerando i quadrati dei rendimenti su quell'orizzonte temporale e calcolandone la media, nel seguente modo:

$$RV_t = \frac{1}{n}\sum_{i=0}^{n} {r_{t-i}}^2$$

(4.17)

Questa relazione equivale alla formula (2.8), con la differenza che il primo rendimento al quadrato considerato non è quello del periodo precedente, ma coincide con l'istante di valutazione della varianza realizzata. anziché con il periodo precedente. Negli ultimi anni, la crescente disponibilità di dati ad alta frequenza ha favorito lo sviluppo di nuove definizioni di varianza realizzata e di specifici modelli per la sua stima.[81] Considerando M rendimenti infragiornalieri è possibile definire la varianza realizzata giornaliera come:

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

(4.18)

Questa definizione può essere estesa mediante la media di più varianze consecutive, al fine di ridurre il rumore e rendere la misura più informativa. La volatilità realizzata si ottiene semplicemente estraendo la radice quadrata della varianza realizzata.

[81] Si vedano a tal proposito: M. Puke e K. Schweikert, *Coherent Forecasting of Realized Volatility*, working paper, University of Hohenheim, 2025.
T.G. Andersen, T. Bollerslev, F.X. Diebold, C. Vega, "Real-Time Price Discovery in Global Stock, Bond and Foreign Exchange Markets", *Journal of International Economics*, 73(2), 2007, pp. 251–277

4.5 Addestramento e validazione dei modelli di machine learning

Per la modellizzazione e previsione della varianza realizzata (Realized Variance, RV) dei tre indici, è stato implementato un *framework* in linguaggio Python, concepito per garantire flessibilità, scalabilità e replicabilità nell'applicazione del modello a differenti combinazioni di indici di mercato, orizzonti temporali e definizioni di variabile target.

Il quadro generale di modellizzazione e validazione è stato progettato per essere coerente tra i due approcci, XGBoost e LSTM. In entrambi i casi la metrica di riferimento è l'errore quadratico medio (MSE) e si applica l'*early stopping* sul validation set per evitare *overfitting* e selezionare in modo oggettivo la configurazione ottimale del modello. In entrambi i casi la capacità del modello è regolata da un insieme mirato di iperparametri. Mentre quelli con maggiore potere predittivo sono ottimizzati attraverso una *grid search* compatta, così da ridurre la complessità combinatoria e i tempi di calcolo senza sacrificare la qualità della selezione, gli iperparametri consolidati in letteratura sono mantenuti fissi per mantenere i modelli stabili.

Come è stato spiegato nello scorso capitolo, la natura dei modelli è ben diversa: XGBoost è un metodo di *gradient boosting* su alberi decisionali, ideato per ottenere robustezza, parsimonia e rapidità; LSTM al contrario è una rete neurale ricorrente (RNN), governata da iperparametri tipici del *deep learning* e dotata di porte interne e memoria di stato, che le consentono di apprendere pattern sequenziali latenti nei dati.

Mentre XGBoost privilegia robustezza e parsimonia con un *boosting* poco profondo e forte regolarizzazione stocastica, LSTM, pur rimanendo essenziale (un solo strato), sfrutta la memoria della sequenza e una regolarizzazione mirata. Le diverse scelte di *early stopping* (50/1200 rispetto a 10/120) riflettono la diversa granularità delle iterazioni (*round* vs epoche) e la differente rumorosità delle curve di validazione, mantenendo comunque coerenza metodologica e confrontabilità dei risultati.

In sintesi, mentre XGBoost cattura non linearità sulle *feature* secondo un approccio non sequenziale, la LSTM sfrutta la propria architettura sequenziale per modellare *pattern* dinamici latenti lungo il tempo, mantenendo però lo stesso protocollo di valutazione e selezione degli iperparametri adottato per XGBoost.

• XGBoost

L’architettura (pipeline) sviluppata si basa sull’integrazione di librerie open-source ampiamente consolidate nella letteratura scientifica, tra cui:

-pandas, per la gestione di strutture dati eterogenee di tipo tabellare e per la manipolazione efficiente di dataset di grandi dimensioni;

-NumPy, per il calcolo numerico ad alte prestazioni e per l’elaborazione vettoriale;

-scikit-learn, per funzioni ausiliarie di *pre-processing*, metriche di valutazione e gestione della validazione temporale;

xgboost, libreria cardine per l’implementazione dell’algoritmo XGBoost Regressor, ossia un modello di *gradient boosting* su alberi decisionali caratterizzato da elevata efficienza computazionale, capacità predittiva e robustezza nella modellazione di relazioni non lineari.

Oltre alle librerie principali dedicate alla modellizzazione, sono state utilizzate anche alcune librerie standard di Python, quali *os*, *re*, *warnings* e *itertools*. In particolare, *os* è stata utilizzata per la gestione dei percorsi e dei file di input/output, *re* per la normalizzazione dei nomi delle colonne tramite espressioni regolari, *warnings* per la soppressione di messaggi non rilevanti durante l’esecuzione, e *itertools* (nello specifico la funzione *product*) per generare in maniera efficiente tutte le combinazioni di iperparametri nella *grid search*. Tali librerie, pur non avendo un ruolo diretto nella definizione del modello, sono risultate fondamentali per assicurare la robustezza e l’automazione dell’intero flusso sperimentale.

Il modello adottato, XGBoost Regressor, come spiegato nello scorso capitolo, appartiene alla famiglia degli *ensemble methods* e si basa sulla logica del *gradient boosting*. Questo approccio consiste nella costruzione sequenziale di più alberi decisionali deboli, ciascuno dei quali viene addestrato per correggere gli errori residui lasciati dagli alberi precedenti. L’insieme finale costituisce una combinazione pesata di alberi che, collettivamente, ottimizzano la funzione obiettivo e riducono progressivamente l’errore di previsione. La funzione obiettivo adottata è di tipo quadratico (*reg:squarederror*), così da mantenere coerenza sia con il modello LSTM, allenato sull’MSE sia con la variabile target (RV), anch’essa espressa al quadrato e dunque necessariamente definita positiva.

Un aspetto cruciale del framework riguarda la configurazione degli iperparametri, che regolano la complessità del modello, la capacità di generalizzazione e la stabilità numerica. La scelta di alcuni iperparametri è stata effettuata tramite una *grid search* compatta, progettata per ridurre i tempi computazionali pur mantenendo un buon grado di esplorazione dello spazio dei parametri. In particolare, la griglia ha incluso:

-*max_depth*: [2, 3]

Definisce la profondità massima degli alberi. Valori ridotti consentono di catturare relazioni semplici e riducono il rischio di overfitting, privilegiando la generalizzazione.

-*learning_rate* (η): [0.05, 0.1]

Determina il contributo di ciascun albero alla previsione complessiva. Valori contenuti garantiscono una convergenza più stabile, sebbene richiedano un numero maggiore di *boosting rounds*.

-*reg_lambda* (λ): [0.0, 1.0]

Coefficiente di regolarizzazione L2 sui pesi dei nodi, che contribuisce a ridurre la complessità e a migliorare la robustezza

Inoltre all'interno del disegno sperimentale, parte degli iperparametri di XGBoost è stata mantenuta fissa. Questi sono in particolare:

-*subsample*: [0.8]

Stabilisce la proporzione di osservazioni campionate casualmente per ciascun albero. L'introduzione di campionamento stocastico per alberoc ontribuisce a ridurre il rischio di *overfitting*.

-*colsample_bytree*: [0.8]

Indica la quota di variabili utilizzate per la costruzione di ciascun albero. Analogamente a *subsample*, introduce variabilità e riduce la varianza delle stime.

-min_child_weight: [1.0]

Impone un vincolo minimo sulla somma dei pesi delle istanze in un nodo, contribuendo alla regolarizzazione e alla stabilità del modello.

La griglia di ricerca è stata concentrata sugli iperparametri che esercitano il maggiore impatto sulla capacità funzionale del modello e sulla sua stabilità (in particolare *max depth*, *learning rate* e *reg lambda*). Questa scelta risponde a tre esigenze fondamentali, che meritano di essere approfondite.

In primo luogo, vi è una rilevanza statistica differenziale degli iperparametri. Nel *boosting*, la complessità della funzione appresa è regolata principalmente dalla profondità degli alberi (*max depth*), che determina il grado di non linearità modellabile, dal passo di aggiornamento (*learning rate*), che governa la velocità e la stabilità del processo di apprendimento, e dalla regolarizzazione L2 (*reg lambda*), che contribuisce a stabilizzare i pesi delle foglie. Parametri come *subsample* e *colsample bytree* introducono invece forme di regolarizzazione stocastica (bagging su righe e colonne), utili in scenari con feature numerose e fortemente correlate. Tuttavia, nel presente caso le feature a disposizione sono poche e specifiche (RV e rendimenti al quadrato osservati in $t-1$), riducendo l'impatto marginale di tali iperparametri sulla generalizzazione. Per analoghe ragioni, *min child weight* è stato mantenuto fisso a un valore unitario: un vincolo più severo avrebbe introdotto rigidità eccessiva, imponendo soglie troppo elevate di informatività per variabili già limitate.

In secondo luogo, si è perseguita un'efficienza computazionale accompagnata da un controllo della complessità. L'introduzione di più livelli per ciascun iperparametro avrebbe moltiplicato il numero di combinazioni possibili all'interno della griglia (effetto combinatorio), comportando costi di calcolo crescenti in modo non lineare. Considerato che ogni addestramento richiede centinaia di *boosting rounds* e che la procedura deve essere ripetuta per 27 serie (tre indici, tre esperimenti e tre orizzonti), concentrare la ricerca sui parametri più influenti ha consentito di ottenere un migliore equilibrio tra accuratezza e tempo computazionale, mantenendo al contempo la ripetibilità dell'esperimento.

Infine, si è posta particolare attenzione alla stabilità, riproducibilità e coerenza temporale. Nei problemi di *time series*, l'uso di *subsample* introduce randomizzazione a ogni round, ma in contesti a bassa dimensionalità come il presente può aumentare artificialmente la varianza tra *run* e compromettere la confrontabilità dei risultati. Fissare *subsample* e *colsample bytree* a 0.8, valori consolidati in letteratura, ha fornito regolarizzazione

sufficiente senza introdurre variabilità eccessiva, garantendo invece un comportamento coerente tra i vari esperimenti. La scelta di mantenere costanti tali parametri, così come *min child weight*, ha permesso di focalizzare la ricerca sui driver principali della performance predittiva, migliorando la robustezza e la generalizzabilità delle stime.

In sintesi, la fissazione di alcuni iperparametri a valori standard e ragionevoli ha consentito di ridurre drasticamente lo spazio di ricerca e i costi computazionali, senza sacrificare la qualità delle previsioni. L'impiego di un numero elevato di alberi (*n estimators* = 1200), combinato con una tecnica di *early stopping* che prevede una *patience* pari a 50 iterazioni senza miglioramenti sul validation set, ha reso superflua una ricerca esaustiva sul numero di *boosting rounds*, permettendo al modello di selezionare automaticamente in maniera ottimale la complessità effettiva in funzione del *learning rate*.

Un'ulteriore riflessione riguarda infine la scelta di non includere nella griglia di ricerca alcuni iperparametri disponibili in XGBoost, quali il termine di regolarizzazione L1 (*alpha o reg alpha*) e il parametro di regolarizzazione sulla complessità delle partizioni (*gamma*). Il primo, che penalizza i pesi delle foglie spingendoli verso zero, è particolarmente utile in scenari con feature numerose e potenzialmente ridondanti, condizione non presente nel nostro caso, dove le variabili esplicative sono poche, selezionate e direttamente collegate al fenomeno di interesse. Analogamente, l'introduzione di *gamma* avrebbe imposto penalizzazioni addizionali sulla creazione di nuove partizioni, con il rischio di eliminare informazioni utili in presenza di dataset relativamente semplici. Per questo motivo, tali parametri sono stati mantenuti ai valori di default, ritenuti più appropriati rispetto a un'esplorazione sistematica che avrebbe comportato costi elevati senza benefici proporzionati.

• LSTM

Per la fase di modellizzazione e previsione è stato implementato un framework in linguaggio Python basato su reti neurali ricorrenti (*Recurrent Neural Networks*, RNN), nella variante *Long Short-Term Memory* (LSTM). Tale architettura è stata scelta poiché particolarmente adatta alla modellizzazione di serie temporali, grazie alla capacità di catturare dipendenze di lungo periodo e relazioni non lineari tra osservazioni sequenziali. Analogamente all'approccio sviluppato per XGBoost, anche in questo caso il framework è stato concepito con finalità di flessibilità e replicabilità, consentendo l'applicazione uniforme a differenti combinazioni di indici di mercato, configurazioni di campionamento e orizzonti temporali della *Realized Variance* (RV).

L'implementazione si è avvalsa di un insieme di librerie ampiamente consolidate nella comunità scientifica, tra cui:

-*pandas* e *NumPy*, per la gestione, manipolazione e trasformazione delle strutture dati;

-*scikit-learn*, per le funzioni ausiliarie di *pre-processing* e di valutazione;

-*TensorFlow*/*Keras*, per la definizione, l'addestramento e la validazione delle LSTM.

In particolare *Keras* ha fornito l'interfaccia ad alto livello per la definizione e l'addestramento dei modelli, mentre *TensorFlow* ha gestito il *backend* computazionale, garantendo efficienza e compatibilità con l'hardware (GPU) e strumenti di *deployment*. Questa combinazione consente di mantenere la chiarezza e compattezza del codice, senza rinunciare alla robustezza di *TensorFlow*.

Anche in questo caso, oltre alle librerie principali dedicate alla modellizzazione, sono state utilizzate le stesse librerie standard di Python menzionate nel caso dell'XGBoost e pur non concorrendo direttamente alla costruzione del modello predittivo, hanno svolto un ruolo essenziale nel garantire l'automazione, la pulizia e la stabilità del framework sperimentale. Come per il caso di XGBoost, l'input al framework è stato fornito in formato Excel e suddiviso in fogli distinti, ciascuno rappresentativo di una combinazione specifica tra indice di mercato, orizzonte temporale e finestra di calcolo della varianza realizzata. Inoltre, per migliorare la stabilità numerica durante l'addestramento della rete neurale e per garantire che le variabili esplicative potessero essere confrontate in modo coerente, queste sono state scalate attraverso una trasformazione standardizzata (*z-score*).

Infine, per rispettare l'input tridimensionale richiesto dalle LSTM, i dataset sono stati ristrutturati in sequenze: ogni osservazione è stata rappresentata come una finestra temporale (*lookback window*) contenente un numero prefissato di input, utilizzati per prevedere il valore futuro della volatilità realizzata. Il partizionamento del dataset in training, validation e test set è stato effettuato preservando rigorosamente la sequenzialità temporale, come già avvenuto per XGBoost, al fine di evitare fenomeni di *look-ahead bias* e *data leakage*. In ciascun esperimento, sono state rispettate cardinalità specifiche per ciascun insieme, in modo da assicurare coerenza temporale e omogeneità di trattamento tra indici e orizzonti.

La configurazione degli iperparametri ha rappresentato un aspetto cruciale per garantire un compromesso ottimale tra capacità predittiva e sostenibilità computazionale. Il modello LSTM è stato implementato come una rete sequenziale composta da un unico strato LSTM, seguito da uno strato di *dropout* e da uno strato denso con attivazione ReLU, quindi da un neurone di uscita a attivazione lineare per la previsione scalare. La scelta di limitarsi a un solo strato LSTM, pur lasciando variabile il numero di unità (32 o 64), risponde alla necessità di mantenere il modello sufficientemente flessibile da catturare le dipendenze temporali presenti nelle serie, ma al contempo snello e stabile, così da ridurre il rischio di overfitting e contenere i costi computazionali. In particolare, gli iperparametri principali che regolano la capacità predittiva del modello LSTM includono:

-*units*: [32, 64]

Definisce il numero di unità nascoste nello strato LSTM, ossia la dimensionalità dello spazio latente appreso. Valori più bassi (32) favoriscono un modello leggero e rapido, mentre un numero più elevato (64) consente di catturare relazioni temporali più complesse, al costo di una maggiore complessità computazionale.

-*lookback*: [10, 20, 30]

Stabilisce il numero di osservazioni passate considerate come input della rete, così da valutare la sensibilità del modello rispetto a input più brevi o più estesi.

-*batch size*: [64, 128]

Determina il numero di sequenze processate simultaneamente durante l'addestramento;

-*learning rate*: [0.0005, 0.001]

Controlla la velocità con cui i pesi della rete vengono aggiornati durante la fase di ottimizzazione. Un valore più basso (0.0005) rende l'apprendimento più graduale e stabile, mentre uno più alto (0.001) accelera la convergenza ma può aumentare il rischio di oscillazioni o overfitting.

-*epochs*: [120]

Definisce il limite massimo di iterazioni consentite sull'intero dataset di addestramento, ossia il numero massimo di volte in cui la rete può rielaborare tutti i dati a disposizione durante la fase di training. Per evitare sprechi computazionali e fenomeni di sovrallenamento, è stata implementata la tecnica di *early stopping*, con *patience* pari a 10, che interrompe l'allenamento in assenza di miglioramenti significativi sul MSE del validation set per dieci epoche consecutive.

-*dropout*: [0.2]

Introduce un meccanismo di regolarizzazione, disattivando in modo casuale una frazione delle connessioni tra neuroni durante il training. In questo modo si riduce il rischio di *overfitting* e si favorisce la capacità del modello di generalizzare su dati non osservati.

L'ottimizzazione degli iperparametri è stata condotta mediante una *grid search* compatta, focalizzata esclusivamente su quelli a maggior impatto sulla capacità predittiva (*units*, *lookback* e *learning rate*). Tale scelta ha consentito di ridurre drasticamente la complessità combinatoria, bilanciando la necessità di esplorare diverse configurazioni con i vincoli di tempo di calcolo imposti dal numero elevato di esperimenti (27 serie complessive). Parametri consolidati in letteratura, quali *batch size*, *epochs* e *dropout*, sono stati invece mantenuti fissi, così da garantire stabilità, confrontabilità dei risultati e sostenibilità computazionale. Infine, il criterio di ottimizzazione adottato è stato l'errore quadratico medio (MSE), in coerenza con quanto impiegato nel framework XGBoost. L'adozione di una metrica comune ha garantito omogeneità di confronto tra i due approcci e ha permesso di valutare le performance in termini comparabili. Al termine della fase di *tuning* e validazione, il modello LSTM ottimale è stato riaddestrato congiuntamente su training e validation set e successivamente applicato al test set per la generazione delle previsioni finali.

4.6: Risultati dell'analisi

L'analisi empirica è stata condotta su tre indici azionari rappresentativi di aree geografiche distinte: lo Standard and Poor's 500 (SPX) per gli Stati Uniti, lo Stoxx 600 Europe (STOXX) per l'Europa e l'MSCI AC Asia Pacific (MXAP) per l'Asia. Per evitare disallineamenti temporali, è stata costruita l'intersezione delle date valide fra i tre indici, scegliendo solo quelle in cui tutti e tre gli indici presentano un prezzo di chiusura. Il risultato è un dataset di 6270 rendimenti logaritmici giornalieri complessivi, che costituisce la base del lavoro. Per ciascun indice sono stati considerati tre dataset parzialmente sovrapposti: il primo (01/1998-06/2025) relativo all'intero periodo, mentre il secondo (01/2007-06/2025) e il terzo (01/2017-06/2025), riferiti a finestre temporali più recenti, contando rispettivamente 4299 e 2082 osservazioni. Le verifiche preliminari hanno confermato la stazionarietà dei rendimenti (ADF Test) e la non normalità della loro distribuzione (Jarque–Bera Test), in linea con i noti fatti stilizzati delle serie finanziarie. Per ciascun esperimento, i dati sono stati suddivisi in training set, validation set e test set, mantenendo l'ordine temporale e seguendo una ripartizione orientativa 60/20/20, verificando nuovamente le ipotesi di stazionarietà e non normalità su ogni sottocampione. Dopo aver aggregato training e validation set per ogni esperimento (4931 osservazioni per il primo esperimento, 3432 per il secondo e 1713 per il terzo) abbiamo stimato i parametri dei modelli econometrici tradizionali attraverso il metodo della massima verosomiglianza. Mentre per l'EWMA si è dovuta mantenere necessariamente l'ipotesi di normalità, per il GARCH(1,1) e l'EGARCH(1,1,1) la leptocurtosi dei rendimenti è stata incorporata ipotizzando una distribuzione *t* di Student, stimando i gradi di libertà ν congiuntamente agli altri parametri via massima verosimiglianza. Grazie a questa assunzione i due modelli adottano ora la denominazione di GARCH-t(1,1) ed EGARCH-t(1,1,1). Di seguito, abbiamo scelto la varianza realizzata giornaliera (RV) come variabile target, calcolata come media aritmetica dei rendimenti al quadrato su finestre mobili di 10, 21 e 63 giorni lavorativi. Successivamente abbiamo addestrato i modelli di machine learning, XGBoost e LSTM su di essa nei training set di ogni combinazione tra esperimento e indice. Nonostante le differenze strutturali di base tra questi due modelli, si è cercato nell'analisi di mantenere una coerenza metodologica: come metrica di selezione è stato utilizzato il *mean squared error* (MSE) per entrambi i modelli, mentre nella scelta degli iperparametri sono stati ottimizzati quelli con maggiore

potere predittivo attraverso una *grid search* compatta, lasciando fissi gli iperparametri consolidati in letteratura per mantenere i modelli stabili. Dopo aver effettuato l'*early stopping* sul validation set e riaddestrato i modelli di machine learning su training e validation set, i cinque modelli stimati (EWMA, GARCH-t(1,1), EGARCH-t(1,1,1), XGBoost e LSTM) sono stati applicati su un test set disgiunto, valutandone le performance mediante due metriche di errore, valutando le previsioni effettuate dai modelli attraverso due misure di errore:

• RMSE (*root mean squared error*), calcolato sulla volatilità realizzata, che penalizza gli errori estremi maggiormente rispetto a quelli più ridotti. Valori più vicini allo zero, indicano stime più accurate.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\sigma_i - \sqrt{RV_i})^2}$$

(4.17)

• QLIKE Loss (*quasi-likelihood loss*), calcolata sulla varianza realizzata, misura che tende a penalizzare maggiormente le sottostime della varianza rispetto alle sovrastime, aspetto rilevante in un'ottica di risk management. Valori più negativi corrispondono a maggiore accuratezza.

$$QLIKE = \frac{1}{n}\sum_{i=1}^{n}[\ln(\sigma_i^2) + \frac{RV_i}{\sigma_i^2}]$$

(4.18)

Di seguito possiamo osservare i grafici delle previsioni della volatilità giornaliera dei modelli trattati per i tre indici e nei tre esperimenti. In ogni grafico ci sono tre pannelli a seconda della definizione di volatilità realizzata giornaliera adottata.

• **Primo esperimento**

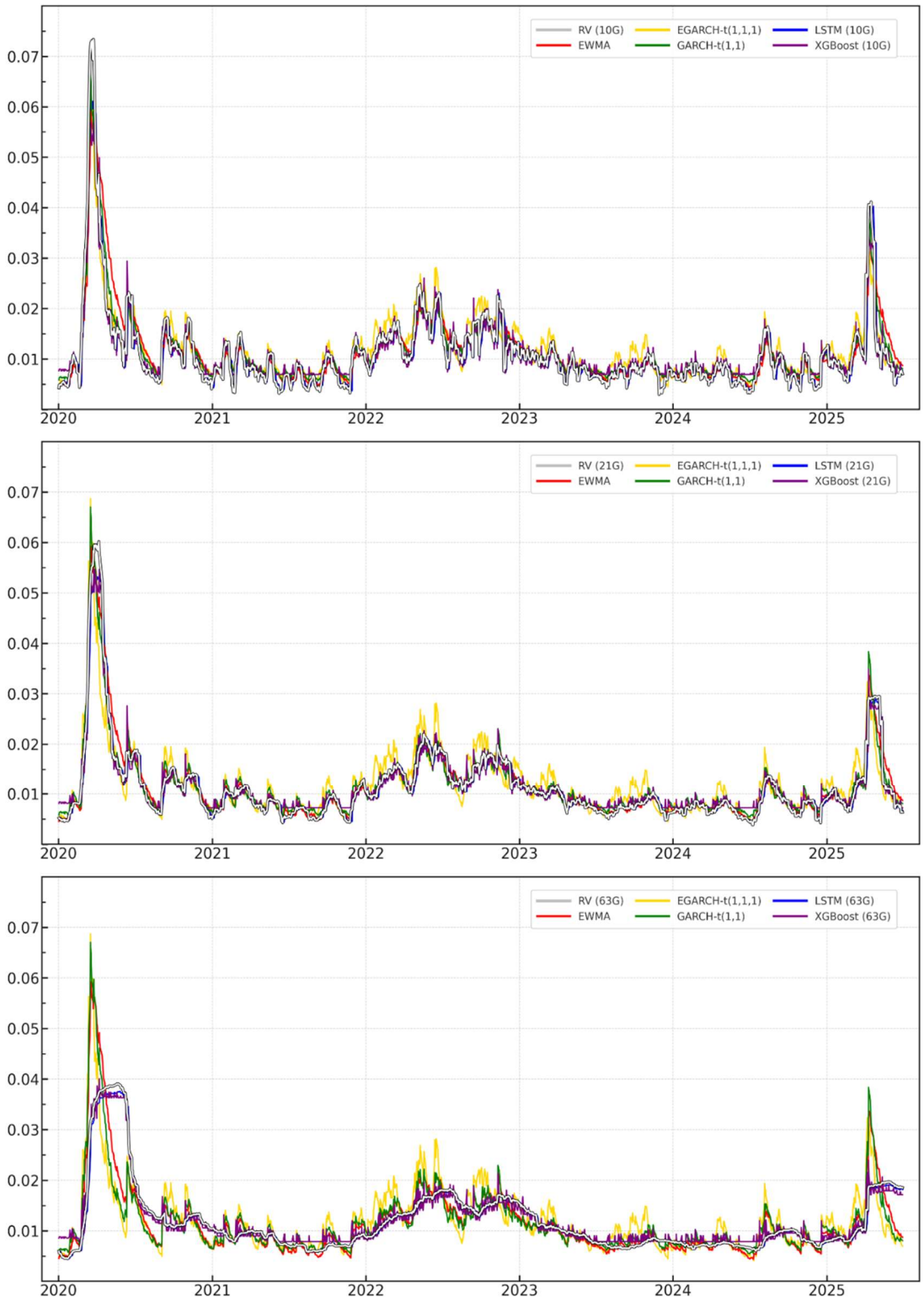

Figura 4.7: Previsioni della volatilità giornaliera, primo esperimento, SPX

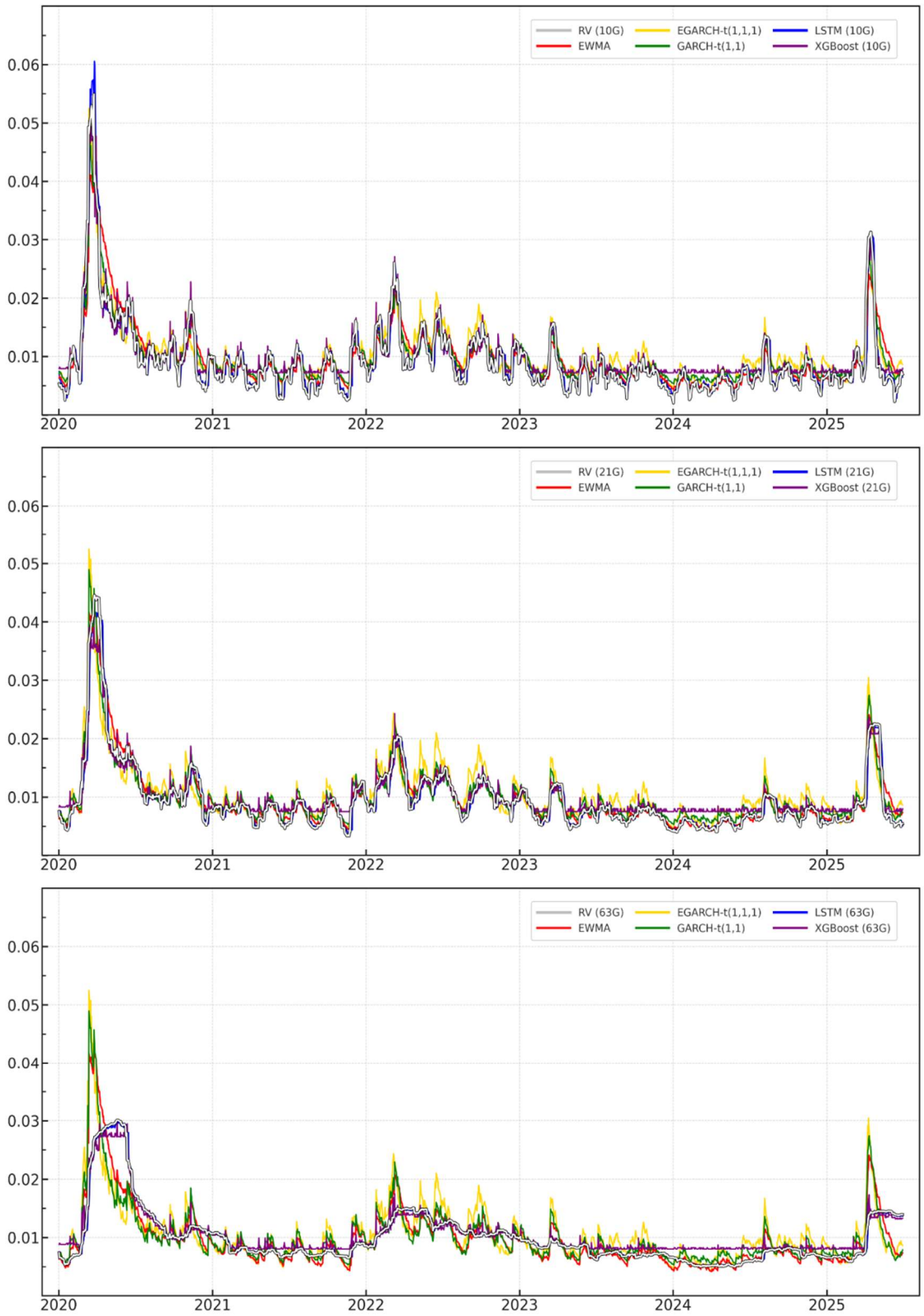

Figura 4.8: Previsioni della volatilità giornaliera, primo esperimento, STOXX

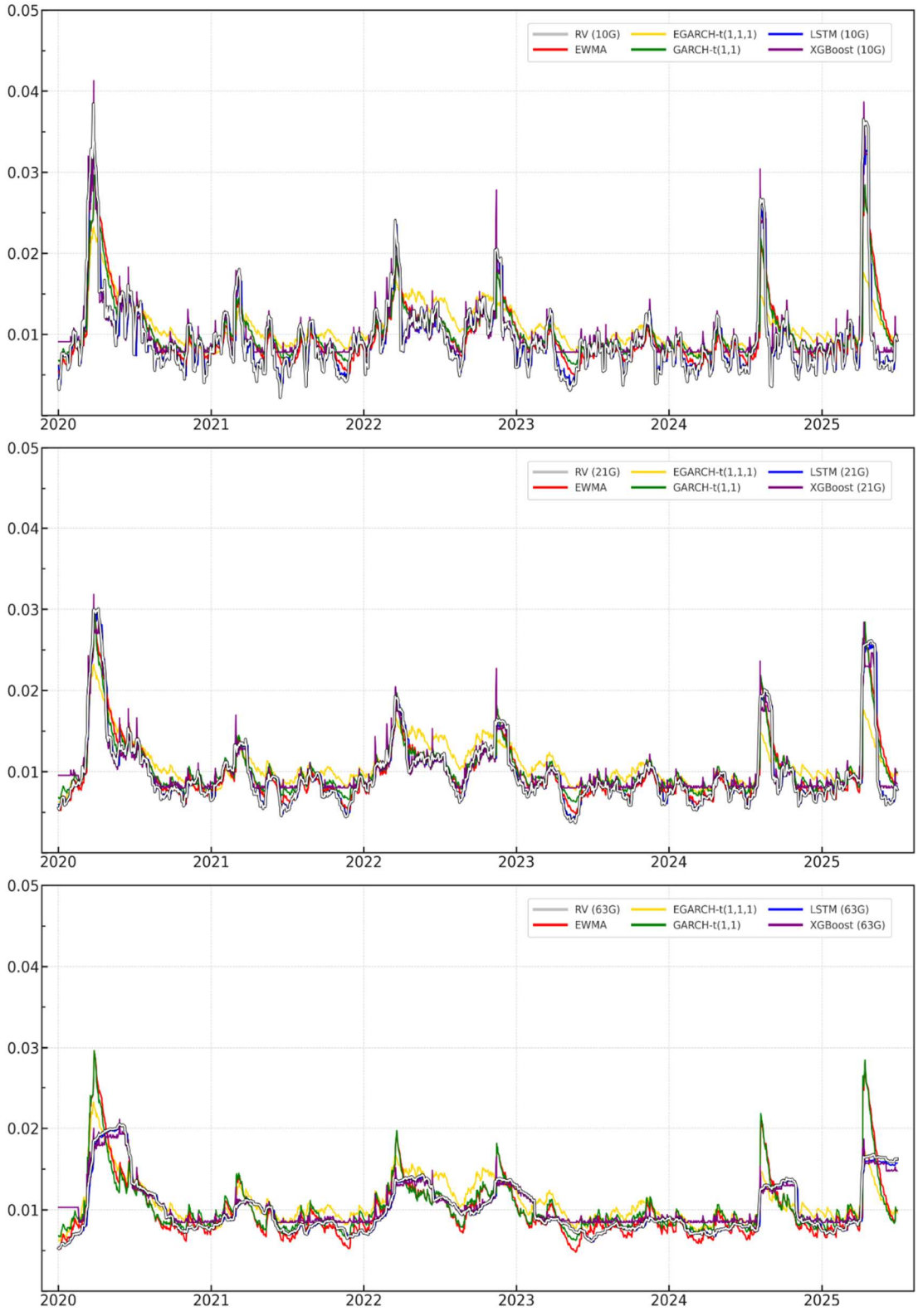

Figura 4.9: Previsioni della volatilità giornaliera, primo esperimento, MXAP

• **Secondo esperimento**

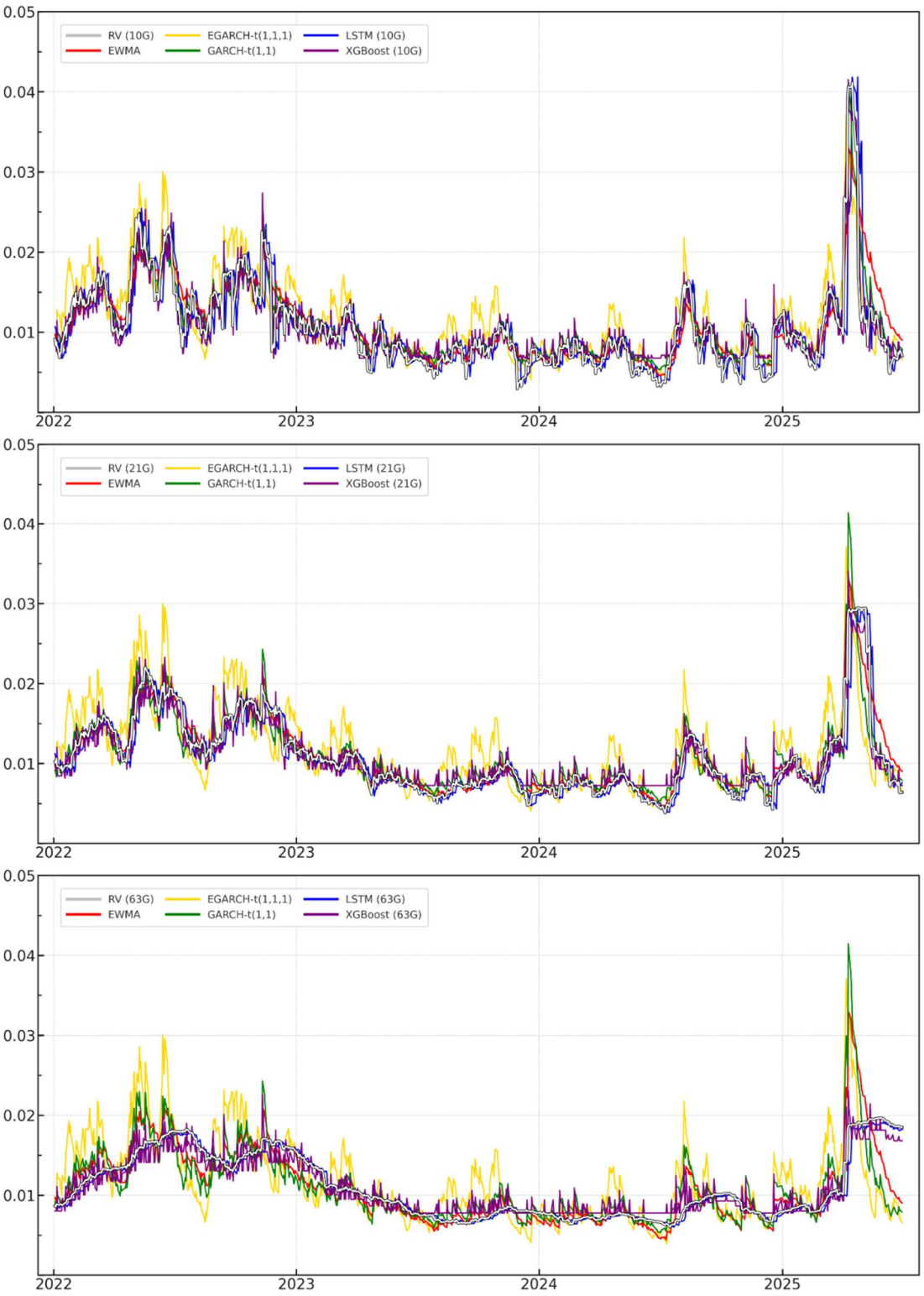

Figura 4.10: Previsioni della volatilità giornaliera, secondo esperimento, SPX

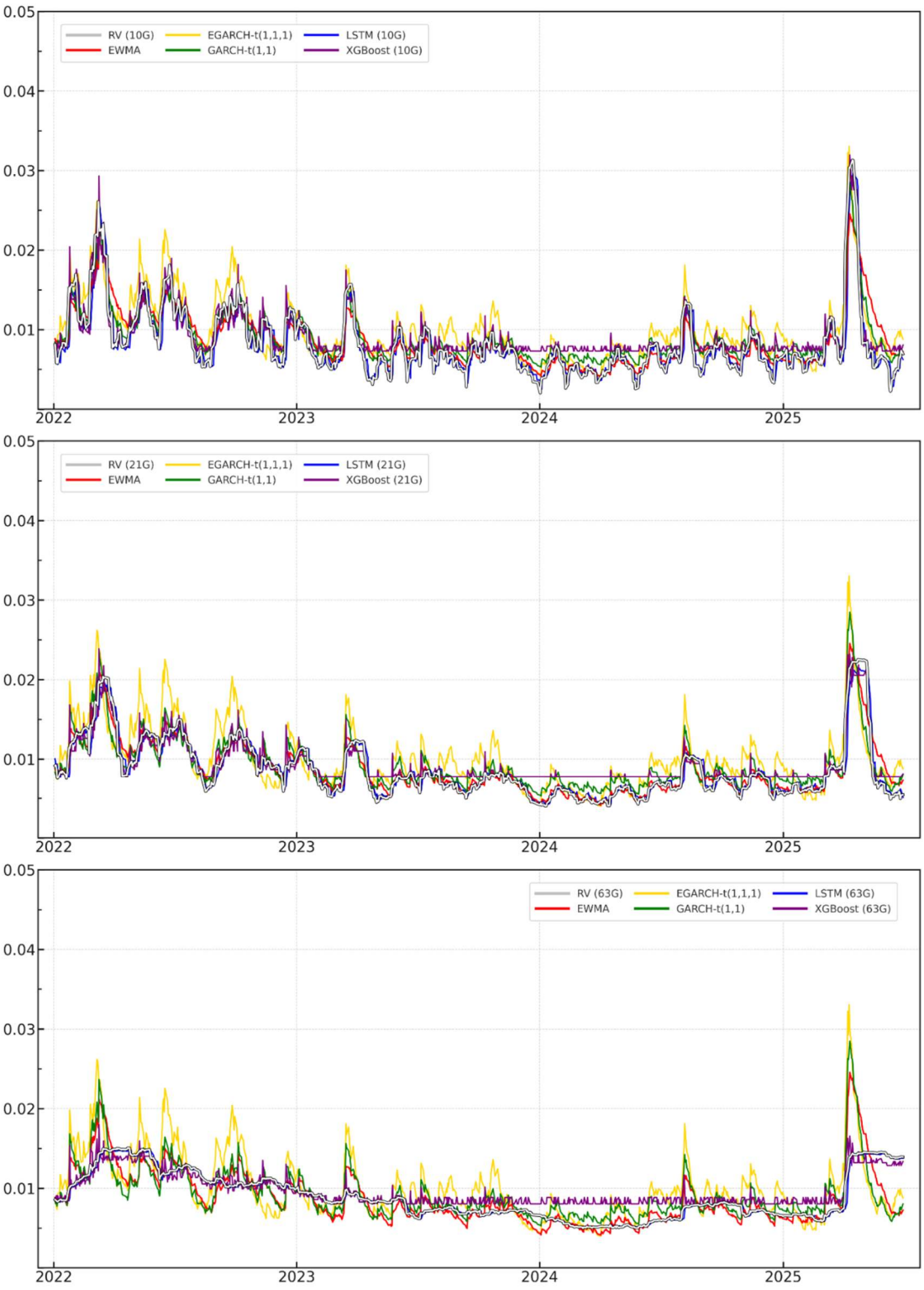

Figura 4.11: Previsioni della volatilità giornaliera, secondo esperimento, STOXX

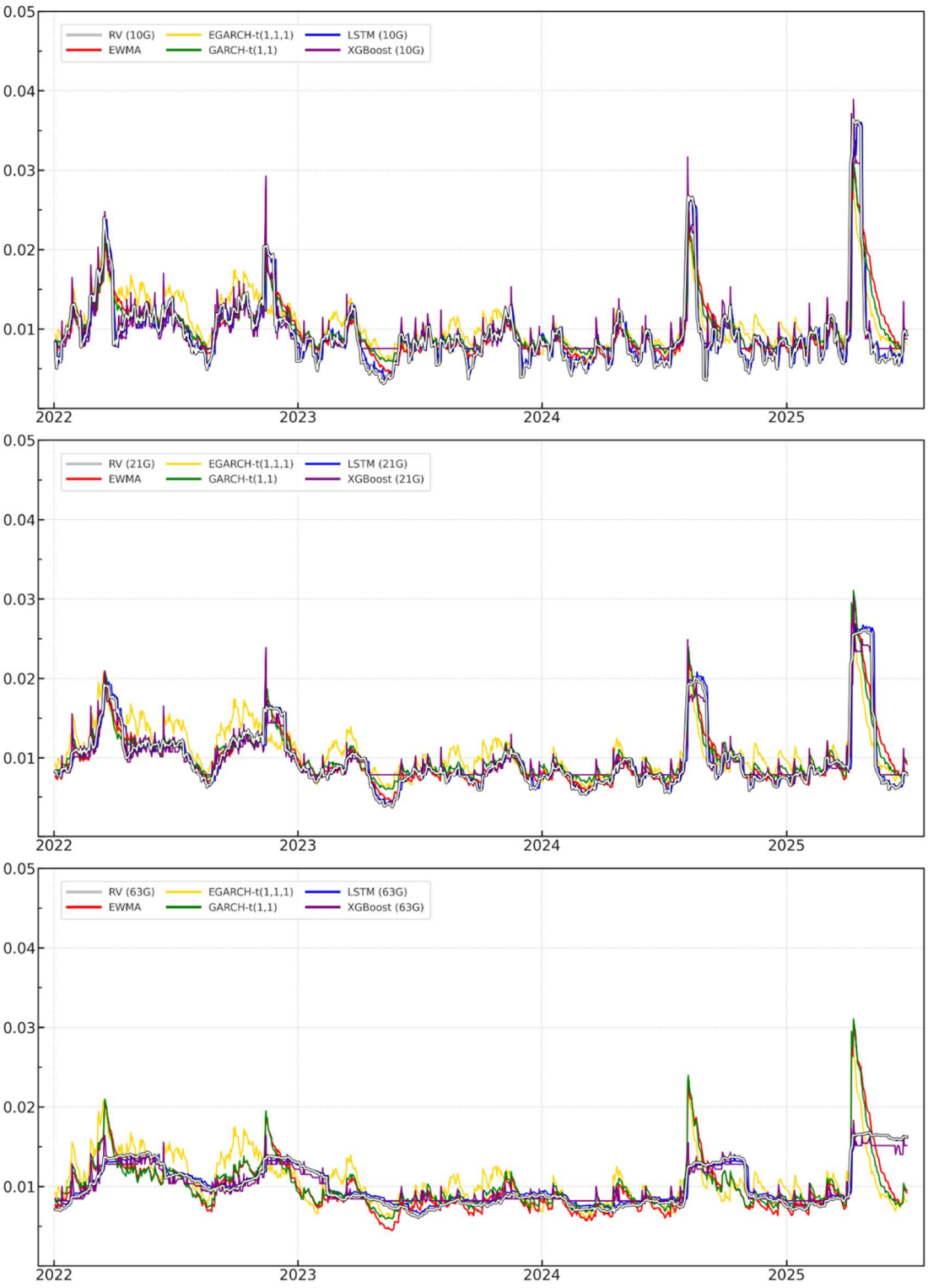

Figura 4.12: Previsioni della volatilità giornaliera, secondo esperimento, MXAP

• **Terzo esperimento**

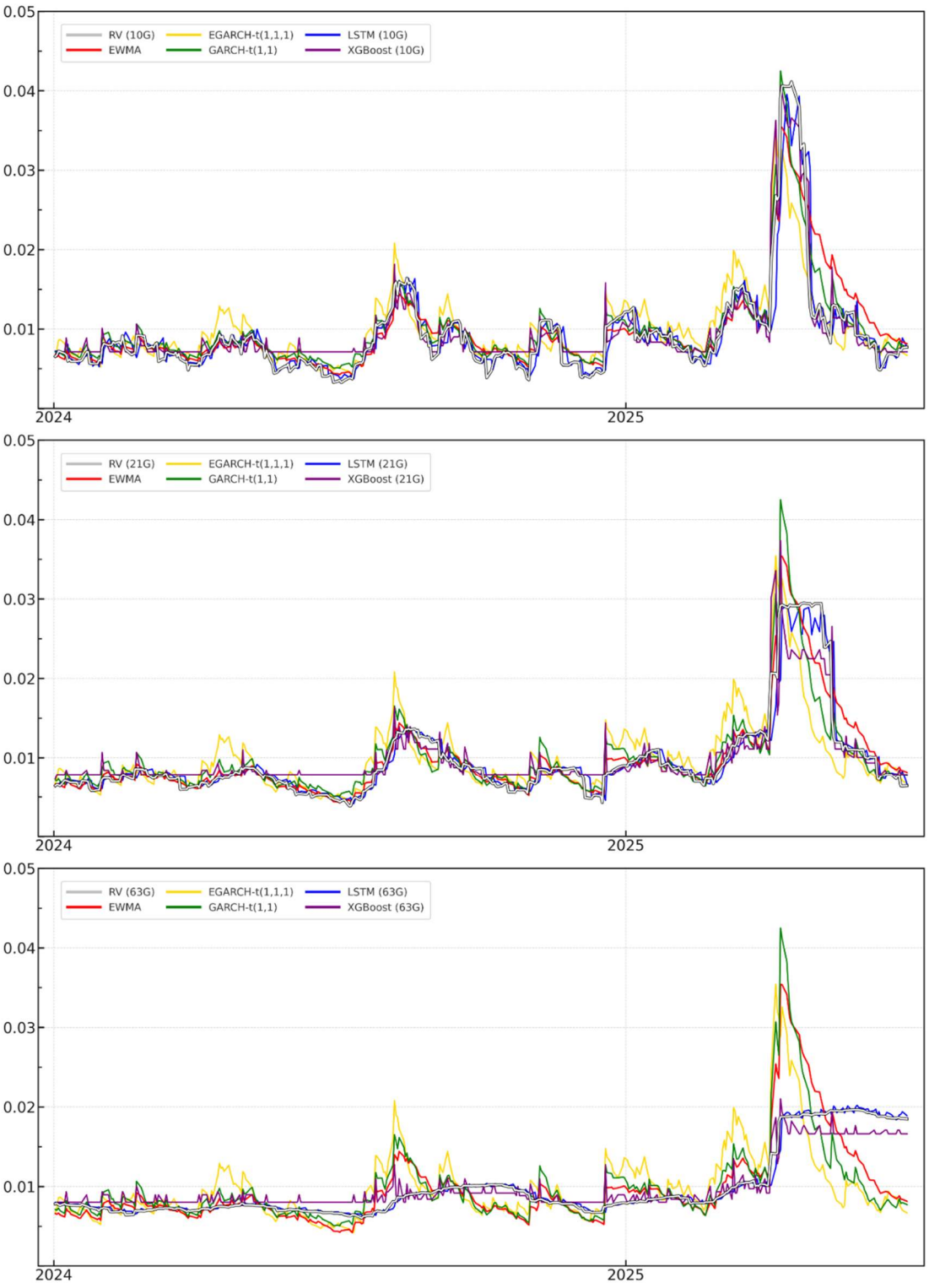

Figura 4.13: Previsioni della volatilità giornaliera, terzo esperimento, SPX

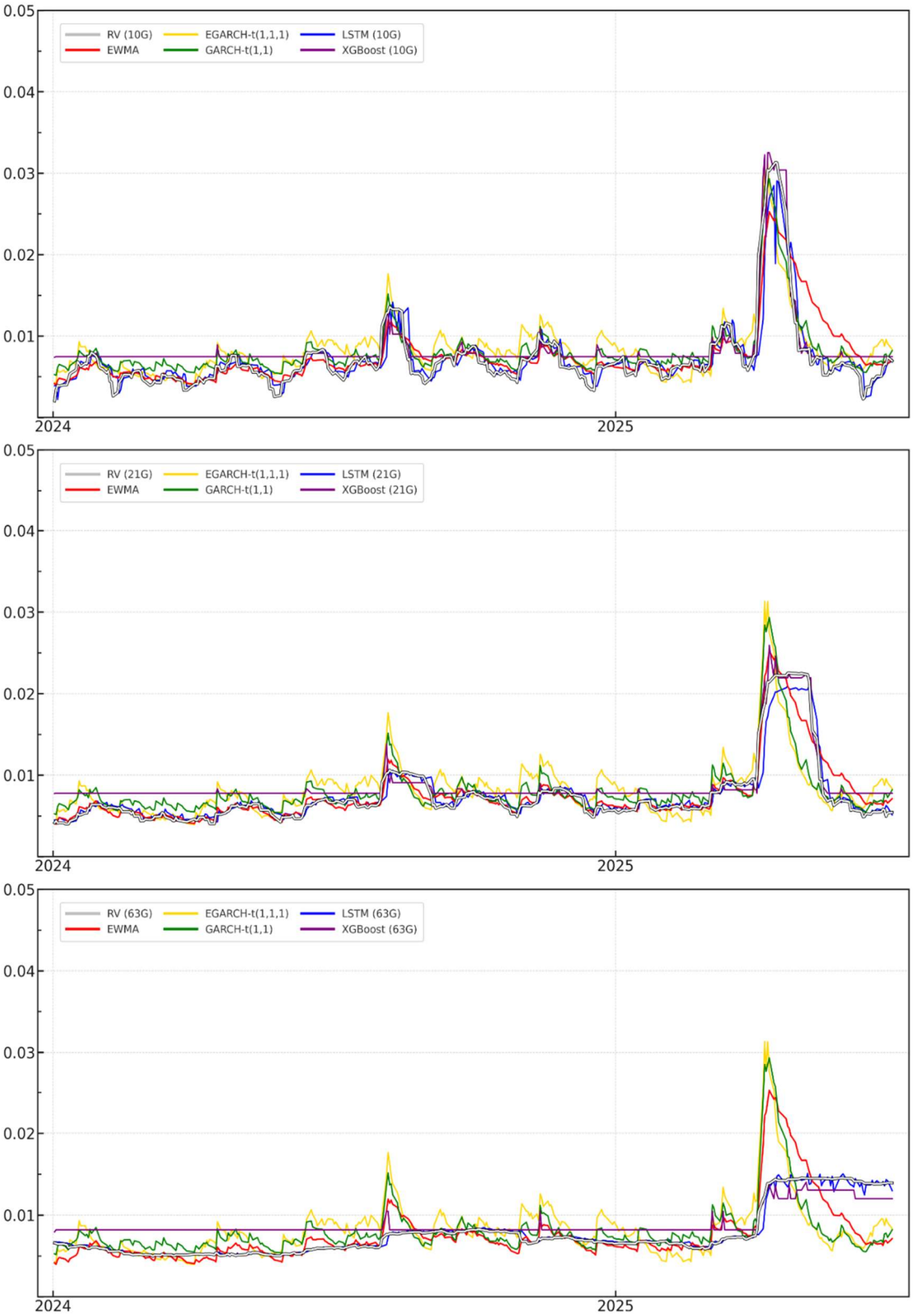

Figura 4.14: Previsioni della volatilità giornaliera, terzo esperimento, STOXX

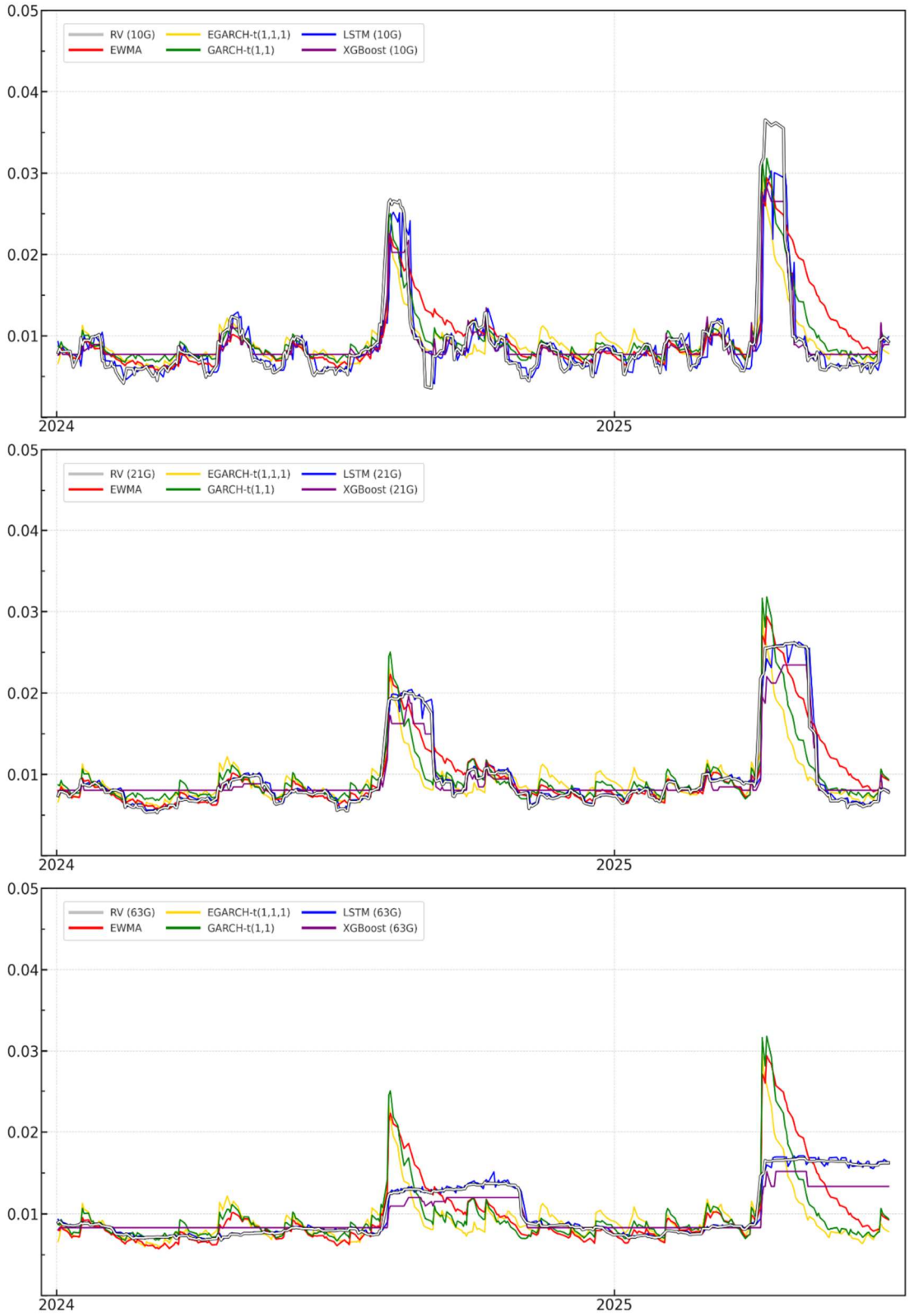

Figura 4.15: Previsioni della volatilità giornaliera, terzo esperimento, MXAP

I grafici riportano le stime della volatilità giornaliera ottenute dai cinque modelli, per i tre indici e i tre esperimenti considerati.

Prima di discutere i risultati, si fornisce una sintesi degli eventi esogeni (geopolitici, macroeconomici e sanitari) che hanno determinato i principali rialzi di volatilità, suddivisi nelle finestre di training, validation e test per ciascun esperimento:

• Primo esperimento: nel training set ricadono gli anni di instabilità successivi alla bolla *dot-com* e all'11 settembre 2001, la crisi del 2008 e quella dei debiti sovrani europei. Nel validation set compaiono, fra gli episodi più rilevanti, il *panic* globale 2015–2016 (crollo del petrolio e timori su Cina/yuan) e la Brexit; il test si apre nel 2020 con lo scoppio della pandemia di COVID-19 seguito dall'invasione del Donbass da parte della Russia, mentre per i mercati asiatici si rileva il sell-off di inizio 2024, causato da uno stress persistente del settore immobiliare cinese. A ridosso dei giorni nostri osserviamo un picco rilevante di volatilità causato dagli annunci sull'introduzione di dazi commerciali da parte del presidente degli Stati Uniti d'America Donald Trump.

• Secondo esperimento: il training set comprende la grande crisi del 2008 e dei debiti sovrani, il validation set copre la Brexit e il COVID, mentre il test set include lo scoppio della guerra Russia-Ucraina e le recenti notizie di politica economica statunitense.

• Terzo esperimento: include nel training set la Brexit e il COVID, nel validation set la guerra russo-ucraina, mentre il test set contiene per l'Asia (MXAP), il sell-off di inizio 2024 sui listini cinesi, oltre che l'annuncio dei dazi statunitensi.

Dall'analisi grafica dei test set emerge una chiara mappa di sensibilità agli shock per i tre indici. Lo SPX reagisce soprattutto al COVID-19 e alle recenti notizie di politica economica statunitense; lo STOXX mostra l'impatto più marcato della guerra russo-ucraina; MXAP è dominato dal sell-off di inizio 2024 sui listini asiatici.

In ciascuna figura sono mostrati tre pannelli, ciascuno per ogni base di calcolo della volatilità realizzata giornaliera, computata come media aritmetica dei rendimenti giornalieri elevati al quadrato, utilizzando una finestra mobile di 10, 21 e 63 giorni.

La scelta di adottare come *benchmark* una media mobile semplice nasce dall'esigenza di non incorporare le ipotesi specifiche dei modelli introdotti, così da non favorire alcuna classe di metodi, mantenendo la neutralità del confronto e garantendo la replicabilità dei

risultati. La scelta della base di calcolo della volatilità realizzata giornaliera è un elemento cruciale in questo tipo di analisi, poiché è la grandezza sulla quale sono calcolate le misure di errore che saranno presentate a breve.

Dal confronto grafico emerge con immediatezza che sia la volatilità realizzata giornaliera costruita su finestra mensile (21G) e, ancor più, trimestrale (63G) evidenzia un marcato effetto eco, fenomeno problematico che abbiamo affrontato nel secondo capitolo e che riguarda le stime della volatilità ottenute attraverso medie semplici. Tale problema risulta particolarmente evidente nel terzo esperimento, dove la finestra di test è più contenuta. L'effetto eco è presente, seppure in misura minore, anche nella definizione bisettimanale (10G): ad esempio, nel pannello MXAP del terzo esperimento si osserva una persistenza residua dopo gli *spike*[82]. Tuttavia, la sua intensità è nettamente inferiore rispetto ai casi mensile e trimestrale, perché il livello della RV, una volta esaurito il picco, rientra tipicamente entro l'orizzonte della stessa finestra. In sintesi, la durata dell'effetto eco è pressoché pari all'ampiezza della finestra utilizzata come base di calcolo della volatilità realizzata.

Per quanto riguarda i modelli econometrici tradizionali EWMA, GARCH-t(1,1) ed EGARCH-t(1,1,1), si può osservare che le loro stime risultano indipendenti dalla definizione di volatilità giornaliera realizzata adottata (10G/21G/63G) e, per questo, appaiono quasi sovrapponibili nei tre pannelli di ciascuna figura. In media i tre modelli generano stime simili della volatilità giornaliera; tuttavia, in corrispondenza degli *spike* i modelli GARCH-t tendono a sovrastimare il picco e a rientrare più rapidamente rispetto a EWMA. Tale comportamento, in ottica prudenziale, è coerente con l'ipotesi di innovazioni distribuite secondo una *t* di Student che cattura la leptocurtosi dei rendimenti. Osservando i picchi più pronunciati, per SPX e STOXX la dominanza si alterna tra GARCH-t(1,1) ed EGARCH-t(1,1,1); per MXAP, invece, prevale più spesso il GARCH-t(1,1).

Un'ulteriore regolarità è che, dopo uno shock, EGARCH-t(1,1,1) è in genere il primo a decrescere, seguito da GARCH-t(1,1) e quindi da EWMA. La maggiore rapidità di EGARCH è coerente con la dinamica logaritmica della varianza e con il *leverage effect*:

[82] Per *spike* si intende un picco improvviso e di breve durata in una serie temporale.

i rendimenti negativi accrescono la volatilità più dei positivi, favorendo un rientro più rapido una volta esaurito l'impulso.

I modelli di machine learning invece, sono addestrati per ogni esperimento sulle tre definizioni di RV, producendo stime diverse per ogni pannello.

Dall'osservazione dei grafici emerge che XGBoost mostra un comportamento eccessivamente conservativo nelle fasi di rientro: la previsione non scende al di sotto di una soglia pressoché costante pari a 0,078%. Quando la volatilità realizzata giornaliera è inferiore a 0,078%, la stima di XGBoost si blocca su tale livello, o presenta rapide oscillazioni in prossimità del *floor*), risultando quindi poco adatta a descrivere regimi di bassa volatilità. L'effetto è particolarmente evidente nel terzo esperimento dello STOXX, in cui la volatilità realizzata giornaliera rimane per diversi mesi sotto il livello menzionato. In prossimità degli *spike* XGBoost tende a sovrastimare l'altezza del picco, sebbene per un intervallo di tempo molto ridotto, pari a un giorno. Tale fenomeno appare maggiormente visibile nell'indice asiatico per una base di calcolo della volatilità realizzata giornaliera settimanale e bisettimanale. Nel complesso, la presenza del pavimento a 0,078% limita la reattività del modello nei regimi tranquilli e ne degrada l'accuratezza quando la volatilità effettiva è persistentemente inferiore a tale soglia.

Il modello LSTM, al contrario, mostra un'ottima capacità di *tracking* della volatilità realizzata, riproducendo con fedeltà sia la fase ascendente sia il successivo rientro. Il comportamento tende a migliorare al crescere dell'orizzonte della finestra di calcolo (21G/63G), dove la serie è meno rumorosa. Negli *spike*, soprattutto su finestre settimanali e bisettimanali (10G/21G), tende a comparire una lieve sottostima della volatilità realizzata giornaliera nella parte alta del picco.

Di seguito sono riportate le misure di errore computate a partire dai test set dei modelli esaminati sulle tre definizioni di volatilità giornaliera realizzata (10G/21G/63G): l'RMSE, calcolato sulla volatilità giornaliera realizzata, e la QLIKE Loss, riferita alla varianza realizzata giornaliera. Si ricorda che le graduatorie tra modelli non sono univoche e oggettive, ma dipendono intrinsecamente dalla definizione di volatilità realizzata adottata, oltre che naturalmente dalla metrica di errore impiegata.

Tabella 4.20: RMSE sulla volatilità realizzata giornaliera su base bisettimanale (10G)

Index Exp	EGARCH	EGARCH Rank	EWMA	EWMA Rank	GARCH	GARCH Rank	LSTM	LSTM Rank	XGB	XGB Rank
SPX Exp1	0.003017	4	0.003281	5	0.002404	1	0.002409	2	0.002471	3
STOXX Exp1	0.002470	4	0.002735	5	0.002173	3	0.002060	1	0.002081	2
MXAP Exp1	0.003599	5	0.002953	4	0.002831	3	0.001949	2	0.001916	1
SPX Exp2	0.003115	5	0.002515	3	0.001855	1	0.002965	4	0.001938	2
STOXX Exp2	0.002603	5	0.002143	4	0.001773	2	0.001606	1	0.001968	3
MXAP Exp2	0.002893	5	0.002555	4	0.002346	3	0.002102	2	0.001959	1
SPX Exp3	0.003051	5	0.002483	4	0.001877	1	0.002155	3	0.001945	2
STOXX Exp3	0.002279	5	0.002248	4	0.001728	1	0.001756	2	0.002144	3
MXAP Exp3	0.003102	4	0.003430	5	0.002593	2	0.002762	3	0.002463	1

Tabella 4.21: QLIKE Loss sulla volatilità realizzata giornaliera su base bisettimanale (10G)

Index Exp	EGARCH	EGARCH Rank	EWMA	EWMA Rank	GARCH	GARCH Rank	LSTM	LSTM Rank	XGB	XGB Rank
SPX Exp1	-8.2150	5	-8.2347	4	-8.2495	1	-8.2480	2	-8.2435	3
STOXX Exp1	-8.5144	5	-8.5348	3	-8.5414	2	-8.5695	1	-8.5237	4
MXAP Exp1	-8.2955	5	-8.3620	3	-8.3598	4	-8.4211	1	-8.4060	2
SPX Exp2	-8.2071	4	-8.2483	3	-8.2683	1	-8.2068	5	-8.2540	2
STOXX Exp2	-8.6284	5	-8.6829	3	-8.6829	2	-8.7095	1	-8.6512	4
MXAP Exp2	-8.3616	5	-8.4007	4	-8.4008	3	-8.4244	1	-8.4172	2
SPX Exp3	-8.5180	5	-8.5611	3	-8.5733	1	-8.5702	2	-8.5415	4
STOXX Exp3	-8.9493	4	-8.9890	2	-8.9880	3	-9.0067	1	-8.9205	5
MXAP Exp3	-8.4882	4	-8.4613	5	-8.5082	2	-8.4929	3	-8.5142	1

Tabella 4.22: RMSE sulla volatilità realizzata giornaliera su base mensile (21G)

Index Exp	EGARCH	EGARCH Rank	EWMA	EWMA Rank	GARCH	GARCH Rank	LSTM	LSTM Rank	XGB	XGB Rank
SPX Exp1	0.003561	5	0.001726	3	0.002117	4	0.001464	1	0.001661	2
STOXX Exp1	0.002853	5	0.001443	2	0.001893	4	0.001191	1	0.001805	3
MXAP Exp1	0.002714	5	0.001579	3	0.001671	4	0.001112	1	0.001419	2
SPX Exp2	0.003632	5	0.001369	1	0.002030	4	0.001660	3	0.001610	2
STOXX Exp2	0.003031	5	0.001184	2	0.001806	4	0.000927	1	0.001699	3
MXAP Exp2	0.002613	5	0.001491	3	0.001720	4	0.001237	1	0.001449	2
SPX Exp3	0.003655	5	0.001760	2	0.002656	4	0.001421	1	0.002391	3
STOXX Exp3	0.003002	5	0.001333	2	0.002446	4	0.001022	1	0.001937	3
MXAP Exp3	0.003283	5	0.002027	3	0.002707	4	0.001159	1	0.001721	2

Tabella 4.23: QLIKE Loss sulla volatilità realizzata giornaliera su base mensile (21G)

Index Exp	EGARCH	EGARCH Rank	EWMA	EWMA Rank	GARCH	GARCH Rank	LSTM	LSTM Rank	XGB	XGB Rank
SPX Exp1	-8.1296	5	-8.2008	2	-8.1853	3	-8.2079	1	-8.1829	4
STOXX Exp1	-8.4222	5	-8.4964	2	-8.4709	3	-8.5063	1	-8.4503	4
MXAP Exp1	-8.2836	5	-8.3582	2	-8.3463	4	-8.3781	1	-8.3509	3
SPX Exp2	-8.0988	5	-8.2194	1	-8.1945	4	-8.2041	2	-8.1965	3
STOXX Exp2	-8.5261	5	-8.6398	2	-8.6021	3	-8.6501	1	-8.5798	4
MXAP Exp2	-8.3001	5	-8.3666	2	-8.3522	4	-8.3761	1	-8.3554	3
SPX Exp3	-8.3611	5	-8.4843	2	-8.4460	3	-8.4907	1	-8.4365	4
STOXX Exp3	-8.8026	5	-8.9479	2	-8.8633	3	-8.9612	1	-8.8490	4
MXAP Exp3	-8.3170	5	-8.4189	3	-8.3826	4	-8.4562	1	-8.4207	2

Tabella 4.24: RMSE sulla volatilità realizzata giornaliera su base trimestrale (63G)

Index Exp	EGARCH	EGARCH Rank	EWMA	EWMA Rank	GARCH	GARCH Rank	LSTM	LSTM Rank	XGB	XGB Rank
SPX Exp1	0.005911	5	0.004283	3	0.005208	4	0.000688	1	0.001382	2
STOXX Exp1	0.004215	5	0.002932	3	0.003687	4	0.000491	1	0.001381	2
MXAP Exp1	0.002122	3	0.002259	4	0.002417	5	0.000462	1	0.001071	2
SPX Exp2	0.004608	5	0.002421	3	0.003378	4	0.000615	1	0.001420	2
STOXX Exp2	0.003635	5	0.002133	3	0.002739	4	0.000322	1	0.001562	2
MXAP Exp2	0.003155	5	0.002558	3	0.002728	4	0.000587	1	0.000931	2
SPX Exp3	0.004852	5	0.003603	3	0.004457	4	0.000478	1	0.001440	2
STOXX Exp3	0.003666	5	0.002608	3	0.003397	4	0.000409	1	0.001998	2
MXAP Exp3	0.003766	4	0.003174	3	0.003798	5	0.000439	1	0.001373	2

Tabella 4.25: QLIKE Loss sulla volatilità realizzata giornaliera su base trimestrale (63G)

Index Exp	EGARCH	EGARCH Rank	EWMA	EWMA Rank	GARCH	GARCH Rank	LSTM	LSTM Rank	XGB	XGB Rank
SPX Exp1	-7.7795	5	-7.9841	3	-7.9086	4	-8.1042	1	-8.0740	2
STOXX Exp1	-8.2244	5	-8.3054	3	-8.2577	4	-8.3928	1	-8.3469	2
MXAP Exp1	-8.2183	4	-8.2214	3	-8.2146	5	-8.2856	1	-8.2637	2
SPX Exp2	-7.8653	5	-8.0800	3	-8.0064	4	-8.1493	1	-8.1240	2
STOXX Exp2	-8.3480	5	-8.4487	3	-8.4006	4	-8.5415	1	-8.4744	2
MXAP Exp2	-8.1215	5	-8.1776	3	-8.1737	4	-8.2747	1	-8.2652	2
SPX Exp3	-8.0110	5	-8.2210	3	-8.1275	4	-8.4000	1	-8.3672	2
STOXX Exp3	-8.5867	5	-8.7216	3	-8.6146	4	-8.8645	1	-8.7439	2
MXAP Exp3	-7.9928	5	-8.1694	3	-8.0658	4	-8.2929	1	-8.2663	2

La finestra temporale trimestrale (63G) mostra il dominio incontrastato del modello LSTM, seguito da XGBoost per entrambi i criteri di errore (RMSE su volatilità e QLIKE su varianza). Questo è un esito atteso in quanto una finestra lunga appiattisce il rumore della serie della volatilità realizzata e fornisce un target più regolare, favorendo i modelli di machine learning, addestrati direttamente su questa serie, beneficiano della maggiore stabilità del segnale.

Come affermato in precedenza, questa finestra risulta la meno interessante delle tre, in quanto ciò che cerchiamo è una stima della volatilità che sia reattiva, senza tuttavia perdere troppo contenuto informativo.

Sulla finestra mensile (21G) continua a dominare il modello LSTM secondo entrambe le misure di errore, ad eccezione del secondo esperimento dello SPX, per cui vince l'EWMA, che migliora il suo posizionamento medio, ponendosi sempre almeno secondo in base al criterio della QLIKE.

La finestra bisettimanale (10G) è la più rilevante, in quanto presenta stime tra i modelli più simili rispetto agli altri casi. Questo è il caso in cui la stima della volatilità realizzata giornaliera risulta più rumorosa, e di conseguenza anche i modelli di machine learning forniranno stime più volatili, ben più simili a quelle generate dai modelli econometrici tradizionali. A differenza delle finestre temporali più lunghe, in questo caso non ci sono modelli che dominano nettamente rispetto ad altri, ma le graduatorie dipendono a seconda dell'indice. In prima battuta i risultati migliori sono raggiunti dal GARCH-t(1,1,1) per lo S&P 500, da LSTM per lo STOXX e da XGBoost per MXAP, anche se ci sono alcune eccezioni a seconda dell'indice e dell'esperimento considerato. In particolare secondo l'RMSE, il GARCH-t(1,1,1) domina nel terzo esperimento dello STOXX, mentre LSTM sul primo e secondo esperimento del MXAP, secondo la QLIKE.

Un chiarimento fondamentale riguarda l'EGARCH-t(1,1,1), che risulta nella quasi totalità dei casi il modello peggiore, secondo entrambe le metriche di errore. La causa è da imputarsi al fatto che nel nostro impianto sperimentale il target di riferimento è la volatilità realizzata costruita come media semplice dei rendimenti al quadrato. Tale *proxy* è, per definizione, simmetrica: non distingue tra *shock* positivi e negativi e tende a smorzare meccanicamente i picchi con un ritardo proporzionale all'ampiezza della finestra. L'EGARCH (1,1,1), al contrario, è un modello che incorpora esplicitamente il

leverage effect, ovvero l'asimmetria della risposta della volatilità agli *shock* (i ribassi fanno crescere la volatilità più dei rialzi di pari entità). Questa differenza concettuale spiega perché, nel confronto contro un target simmetrico, l'EGARCH risulti penalizzato dalle metriche di errore (RMSE su volatilità e QLIKE su varianza, proprio a causa del fatto di cogliere informazione economica aggiuntiva (l'asimmetria) che il target non contiene.

Considerando nuovamente i primi pannelli dei grafici (RV 10G) si può osservare che mentre in corrispondenza degli *spike*, tutti i modelli forniscono stime simili, in caso di moderati rialzi di volatilità l'EGARCH-t(1,1,1) tende a sovrastimare sistematicamente la volatilità rispetto agli altri modelli; si tratta di un atteggiamento prudente, e apprezzato in un'ottica di risk management. Inoltre l'EGARCH-t(1,1,1) è molto reattivo nelle discese a seguito di uno *spike* e riesce a modellare bene eventuali ulteriori rialzi improvvisi della volatilità. Questo è un pregio dei modelli che incorporano leptocurtosi dei rendimenti e asimmetria della volatilità tra le ipotesi, a differenza di modelli che ipotizzano distribuzioni normali, come l'EWMA.

Una soluzione potrebbe essere quella di leggere simultaneamente le stime di una LSTM, addestrata sulla volatilità realizzata giornaliera calcolata come media aritmetica (su una base di calcolo compresa tra le due settimane e il mese lavorativo), e di un modello tradizionale che incorpori l'effetto leva della volatilità, oltre che la leptocurtosi dei rendimenti, come ad esempio l'EGARCH-t(1,1,1).

La letteratura ha prodotto una sconfinata varietà di modelli GARCH, i quali presentano diverse sfumature tra di loro e che possono essere scelti per stimare la volatilità di una posizione detenuta in portafoglio a seconda delle caratteristiche del mercato considerato.[83] In alternativa la LSTM potrebbe essere allenata sulla volatilità realizzata definita da un modello EWMA, con il parametro λ stimato a seconda del set di dati prescelto. Tuttavia, in caso di rapida discesa dopo uno *spike*, sarebbe preferibile dare priorità alle stime dell'EGARCH-t(1,1,1), in quanto come abbiamo già affermato, l'EWMA presenta dei cali di volatilità più lenti, che rischiano di non cogliere le condizioni attuali di mercato.

[83] Per una visione introduttiva completa dei modelli GARCH si veda: T. Bollerslev, Glossary to ARCH (GARCH), *Oxford Scholarship Online*, 2010, pp. 137–163

Conclusioni

Questo lavoro ha analizzato la capacità di differenti approcci, sia econometrici (EWMA, GARCH, EGARCH) sia di machine learning (XGBoost, LSTM), di stimare la volatilità realizzata di tre indici azionari ampi e diversificati, rappresentativi dei mercati di tre distinte aree geografiche: Stati Uniti, Europa e Asia/Oceania.

La fase preliminare ha confermato le consuete regolarità empiriche tipiche dei rendimenti azionari: oltre alla stazionarietà, ipotesi necessaria per poter sviluppare i modelli econometrici e verificata attraverso l'Augmented Dickey-Fuller Test (ADF), è stata riscontrata la presenza di una lieve asimmetria negativa e di una forte leptocurtosi, verificate dal Jarque Bera Test (JB), che ha rigettato l'adozione di una distribuzione normale per descrivere i rendimenti.

In un'ottica di risk management utilizzare questa distribuzione può causare una sottostima sistematica della frequenza con la quale si possono osservare rendimenti estremi. Infatti, la probabilità che si verifichino variazioni di prezzo dei fattori di mercato distanti dal valore medio è dunque più elevata di quella implicita in una distribuzione normale. Dunque la probabilità di conseguire perdite superiori al VaR parametrico calcolato, per esempio, con livello di confidenza del 99% è in realtà superiore all'1%.

Per riflettere più adeguatamente la probabilità associata a movimenti estremi dei fattori di mercato è stata scelta una distribuzione *t* di Student, caratterizzata da un parametro ν, denominato "gradi di libertà", che ne controlla il grado di leptocurtosi. Al crescere di ν la distribuzione empirica tenderà a quella normale, mentre valori più bassi indicano un maggiore livello di leptocurtosi.

Queste considerazioni, a differenza del modello EWMA che deve ammettere necessariamente l'ipotesi di distribuzione normale, sono state incorporate nelle ipotesi dei modelli GARCH, che adottano la denominazione di GARCH-t(1,1) ed EGARCH-t(1,1,1). In queste ultime due casistiche la stima dei gradi di libertà avviene contestualmente agli altri parametri attraverso il metodo della massima verosomiglianza.

I modelli di machine learning invece non richiedono specifiche ipotesi sulla distribuzione dei rendimenti dei fattori di mercato, in quanto apprendono direttamente dai dati la

relazione tra input e variabile target, senza imporre a priori una certa forma funzionale o una distribuzione dei rendimenti.

La natura dei due modelli di machine learning affrontati è ben diversa: mentre XGBoost è un metodo di *gradient boosting* su alberi decisionali, LSTM al contrario è una rete neurale ricorrente, che sfrutta la propria architettura con porte interne e una memoria di stato per apprendere dinamiche sequenziali latenti nei dati.

I modelli di machine learning sono stati sulla varianza realizzata giornaliera nei training set di ogni combinazione tra esperimento e indice, utilizzando come metrica di selezione il *mean squared error* (MSE) per entrambi i modelli, al fine di garantire comparabilità.

La varianza realizzata giornaliera è stata calcolata come media aritmetica dei rendimenti al quadrato su finestre mobili di 10, 21 e 63 giorni lavorativi, cioè rispettivamente su base bisettimanale, mensile e trimestrale, in modo da non favorire nessuna classe di modelli, mantenendo il confronto neutrale.

Nonostante le differenze strutturali di base tra questi due modelli, si è cercato nell'analisi di mantenere una coerenza metodologica, oltre che nella scelta della misura di errore per l'addestramento, nella scelta degli iperparametri, ottimizzando quelli con maggiore potere predittivo attraverso una *grid search* compatta, e lasciando fissi gli iperparametri consolidati in letteratura per mantenere i modelli stabili.

Dopo aver effettuato l'*early stopping* sul validation set e riaddestrato i modelli di machine learning su training e validation set, i cinque modelli stimati (EWMA, GARCH-t(1,1), EGARCH-t(1,1,1), XGBoost e LSTM) sono stati applicati su un test set disgiunto, valutandone le performance mediante due metriche di errore, l'RMSE calcolato sulle stime della volatilità e la QLIKE Loss calcolata sulle stime della varianza.

Mentre la prima è una misura simmetrica, che penalizza allo stesso modo errori positivi e negativi, dando tuttavia maggior peso a quelli più estremi, la seconda tende a penalizzare maggiormente le sottostime della varianza rispetto alle sovrastime, risultando in un'ottica di risk management.

I risultati dell'analisi dipendono strettamente dalla definizione di volatilità giornaliera adottata, in quanto la volatilità è una variabile latente e dunque non presenta una definizione univoca. In particolare ciò vale solo per i modelli di machine learning, in

quanto sono direttamente addestrati su di essa, mentre i modelli econometrici tradizionali forniscono stime uguali per ogni definizione di volatilità realizzata adottata. Al crescere dell'ampiezza della finestra utilizzata per definire la volatilità realizzata, questa risulta meno rumorosa e di conseguenza anche i modelli di machine learning forniranno stime meno volatili, più distanti da quelle fornite dai modelli econometrici.

Tuttavia una finestra trimestrale (63G) risulta poco interessante, in quanto il benchmark viene aggiornato molto lentamente ed è caratterizzato da un marcato effetto eco, che si trasmette nelle stime dei modelli di machine learning.

La finestra bisettimanale (10G) risulta sicuramente la più rilevante, in quanto è quella che presenta la definizione di volatilità realizzata giornaliera più rumorosa. Di conseguenza anche i modelli di machine learning forniranno stime più volatili, meno divergenti da quelle generate dai modelli econometrici tradizionali.

Per queste ragioni, mentre su finestre lunghe il confronto tra i modelli è dominato dai modelli di machine learning, al ridursi dell'ampiezza del campione utilizzato per definire la volatilità realizzata giornaliera i risultati diventano meno netti.

In conclusione possiamo affermare che il modello LSTM fornisce le stime più accurate, mentre l'EGARCH-t(1,1,1) risulta generalmente quello peggio classificato, in quanto è l'unico modello affrontato che include tra le ipotesi il *leverage effect* della volatilità. In particolare il benchmark di riferimento, cioè la volatilità realizzata giornaliera costruita come media semplice dei rendimenti al quadrato è per definizione simmetrica, non distinguendo tra innovazioni positive e negative.

Tuttavia questa evidenza non deve portare a scartare l'EGARCH-t(1,1,1), ma anzi a considerarlo nelle diverse soluzioni proposte di seguito, in quanto è l'unico modello tra quelli esaminati che aggiunge un ulteriore fatto stilizzato rispetto alle ipotesi comuni.

Nella nostra analisi è stata scelta una versione basilare di LSTM, per non sfavorire eccessivamente i modelli tradizionali, il potenziale delle reti neurali non si limita a queste impostazioni. Questi modelli sono infatti estendibili per costruzione: ad esempio è possibile inserire ulteriori *feature*, come volatilità implicite ricavate dai pezzi delle opzioni o rendimenti rilevati a frequenze infragiornaliere, oppure si possono aggiungere ulteriori strati nascosti tra input e output. Tale flessibilità è teoricamente illimitata e

consente di adattare il modello al regime di mercato e alle specificità dell'asset. Tuttavia, a questa potenza corrisponde un costo: oltre ai tempi di predisposizione e di calcolo della rete, la crescita di complessità introduce il rischio di ottenere stime *black-box*, cioè senza collegamenti trasparenti tra input e output. Questo ha ovvie implicazioni per l'uso nella gestione del rischio: un sistema a scatola nera può costituire una vera e propria fonte di rischio autonoma (*model risk*) in quanto è difficile da essere supervisionato e potrebbe causare problemi di conformità normativa, specialmente per quanto riguarda la dimostrazione della validità del modello.

Al contrario, i modelli econometrici tradizionali mantengono i vantaggi competitivi di interpretabilità, rapidità di calcolo e replicabilità. Proprio per queste ragioni risultano efficienti in molte applicazioni operative e spesso preferibili qualora si ricerchi trasparenza, solidità e tempi di risposta contenuti.

Una soluzione conclusiva potrebbe essere, piuttosto che scegliere tra econometria e intelligenza artificiale, far dialogare le due metodologie, avendo sempre ben chiaro il contesto e le finalità per cui si opera. In particolare si potrebbe addestrare direttamente una LSTM sulle stime giornaliere fornite da un modello econometrico che incorpori sia la leptocurtosi dei rendimenti sia l'effetto leva della volatilità, come ad esempio l'EGARCH-t(1,1,1), ipotizzando che la volatilità realizzata giornaliera evolva nel tempo in base ai parametri del modello prescelto. Dai grafici possiamo osservare che le stime dei modelli tradizionali non si discostano molto tra gli esperimenti, dunque appare preferibile stimarli su un campione di dati sufficientemente ampio, così da poter mantenere un elevato contenuto informativo. Un modello ibrido di questo tipo permetterebbe di ottenere i vantaggi di entrambi i modelli, producendo stime di volatilità affidabili e sostenibili nel tempo, grazie al compromesso fra trasparenza dei modelli GARCH e potenza predittiva delle LSTM.

Bibliografia

[1] Acerbi, C., & Tasche, D. (2002). Expected Shortfall: A natural coherent alternative to Value at Risk. *Economic Notes* 31(2): 379–388.

[2] Alexander, C. (1996). *The Handbook of Risk Management and Analysis*. Chichester: Wiley.

[3] Andersen, T. G., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review* 39(4): 885–905.

[4] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Vega, C. (2007). Real-Time Price Discovery in Stock, Bond and Foreign Exchange Markets. *Journal of International Economics* 73(2): 251–277.

[5] Arago, F. (1806). *Mémoire sur la dispersion des erreurs dans les mesures géodésiques*.

[6] Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance* 9(3): 203–228.

[7] Barone, E., & Barone, G. (2022). *Unscrambling Codes: From Hieroglyphs to Market News*. Milano: Egea.

[8] Bekaert, G., & Wu, G. (2000). Asymmetric Volatility and Risk in Equity Markets. *Review of Financial Studies* 13(1): 1–42.

[9] Berk, J., & DeMarzo, P. (2018). *Finanza aziendale I*. Milano: Pearson Italia, cap. 10, 11, 12.

[10] Bessel, F. W. (1818). Über die Wahrscheinlichkeit der Beobachtungsfehler. *Astronomische Nachrichten* 1: 255–267.

[11] Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3): 637–654.

[12] Blattberg, R., & Gonedes, N. (1974). A Comparison of the Stable and Student Distributions as Statistical Models for Stock Prices. *Journal of Business* 47(2): 244–280.

[13] Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31(3): 307–327.

[14] Bollerslev, T. (2010). Glossary to ARCH (GARCH). *Oxford Scholarship Online*: 137–163.

[15] Boston Consulting Group (2015). *Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries*.
[16] Brenner, M., & Galai, D. (1989). New Financial Instruments for Hedging Changes in Volatility. *Financial Analysts Journal* 45(4): 61–65.
[17] Cao, J., Chen, J., & Hull, J. (2020). A neural network approach to understanding implied volatility movements. *Quantitative Finance* 20(9): 1405–1413.
[18] Cboe Global Indices (2012). *Cboe VVIX℠ White Paper: The Cboe VIX of VIX Index*. Chicago Board Options Exchange.
[19] Cboe Global Indices (2022). *Cboe Volatility Index® Methodology*. Chicago Board Options Exchange.
[20] Chebyshev, P. L. (1867). Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées* 12: 177–184.
[21] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16 Proceedings*: 785–794.
[22] Chew, L. K. (1994). *Fundamentals of Probability Theory for Finance*. Singapore: World Scientific: 63–70.
[23] Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *EMNLP*: 1724–1734.
[24] Christie, A. A. (1982). The Stochastic Behavior of Common Stock Variances: Value, Leverage and Interest Rate Effects. *Journal of Financial Economics* 10(4): 407–432.
[25] Christoffersen, P. F., & Diebold, F. X. (2000). Financial asset returns, direction-of-change forecasts, and volatility clustering. *International Economic Review* 41(4): 915–941.
[26] Christoffersen, P., & Jacobs, K. (2004). Which GARCH model for option valuation? *Management Science* 50(9): 1204–1221.
[27] Collard, F., Dellas, H., Diba, B., & Loisel, O. (2018). Optimal Monetary and Fiscal Policy in Large Crises. *Journal of Monetary Economics* 97: 42–63.
[28] Cornish, E. A., & Fisher, R. A. (1937). The Percentile Points of Distributions Having Known Cumulants. *Proceedings of the Cambridge Philosophical Society* 34(2): 335–341.

[29] Culkin, R., & Das, S. R. (2017). Machine learning in finance: The case of deep learning for option pricing. *Journal of Investment Management* 15(4): 92–100.
[30] de Finetti, B. (1935). *Sulle operazioni finanziarie: Note di matematica finanziaria*. Roma: Istituto Italiano degli Attuari.
[31] de Finetti, B. (1940). Il problema dei pieni. *Giornale dell'Istituto Italiano degli Attuari* 11: 1–88.
[32] de Finetti, F., & Nicotra, L. (2008). *Bruno de Finetti: un matematico scomodo*. Livorno: Salomone Belforte & C.
[33] Derman, E. (2004). *My Life as a Quant: Reflections on Physics and Finance*.
[34] Derman, E. (2011). *Models. Behaving. Badly: Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life*. Free Press.
[35] Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1(1): 83–106.
[36] Dowd, K. (2002). *Measuring Market Risk*. Chichester: John Wiley & Sons.
[37] Duffie, D., & Pan, J. (1997). An Overview of Value at Risk. *Journal of Derivatives* 4(3): 7–49.
[38] Engle, R. F. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50(4): 987–1007.
[39] Engle, R. F., & Mezrich, J. (1996). GARCH for Groups. *Risk*: 36–40.
[40] Engle, R. F., & Patton, A. J. (2007). What good is a volatility model? In *Forecasting Volatility in the Financial Markets*. Oxford: Butterworth-Heinemann: 47–63.
[41] Fama, E. F. (1965). The Behavior of Stock-Market Prices. *Journal of Business* 38(1): 34–105.
[42] Ferguson, R., & Green, A. (2018). Deeply learning derivatives. *SSRN Working Paper* (October).
[43] Figlewski, S. (1994). Forecasting Volatility Using Historical Data. *Financial Markets, Institutions & Instruments* 3(2): 1–88.
[44] French, K. R. (1980). Stock Returns and the Weekend Effect. *Journal of Financial Economics* 8(1): 55–69.
[45] French, K. R., & Roll, R. (1986). Stock Return Variances: The Arrival of Information and the Reaction of Traders. *Journal of Financial Economics* 17(1): 5–26.

[46] FTSE Russell. *FTSE Italia All-Share Index Series Ground Rules*. Ultima versione.
[47] FTSE Russell. *FTSE MIB Index Series Ground Rules*. Ultima versione.
[48] Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburg: Perthes et Besser.
[49] Gately, E. (1996). *Neural Networks for Financial Forecasting*. New York: John Wiley & Sons.
[50] Girsanov, I. V. (1960). On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability and Its Applications* 5(3): 285–301.
[51] Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48(5): 1779–1801.
[52] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*.
[53] Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* 20(7): 873–889.
[54] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.
[55] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning in finance. *arXiv:1602.06561*.
[56] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation* 9(8): 1735–1780.
[57] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4(2): 251–257.
[58] Hull, J. C. (2021). *Opzioni, futures e altri derivati* (11ª ed., a cura di E. Barone). Milano: Pearson Italia, cap. 15, 19, 22, 23.
[59] Hull, J. C. (2023). *Risk Management and Financial Institutions* (6th ed.). Hoboken: Wiley, cap. 1, 7, 10, 12, 15, 16, 25.
[60] Hull, J. C. (2021). *Machine Learning in Business, un'introduzione alla scienza dei dati* (3ª ed.).

[61] Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance* 49(3): 851–889.
[62] IASB (2014). *International Financial Reporting Standard 9 – Financial Instruments (IFRS 9)*. London: IFRS Foundation.
[63] ISO/IEC (2023). *ISO/IEC 42001:2023 – Information technology – Artificial intelligence – Management system*. International Organization for Standardization.
[64] J. P. Morgan (1995). *RiskMetrics Monitor*, Fourth Quarter 1995. New York: J.P. Morgan.
[65] J. P. Morgan/Reuters (1996). *RiskMetrics — Technical Document*. New York: J.P. Morgan.
[66] Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* 30(1): 175–193.
[67] Lorente de Nó, R. (1933). Vestibulo-Ocular Reflex Arc. *Archives of Neurology and Psychiatry* 30(2): 245–291.
[68] Ma, Y., Han, R., & Wang, W. (2021). Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications* 165: 113973.
[69] Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7(1): 77–91.
[70] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5: 115–133.
[71] Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4(1): 141–183.
[72] Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.
[73] MSCI. *MSCI AC Asia Pacific Index Methodology*. MSCI Barra, ultima versione.
[74] Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59(2): 347–370.
[75] Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML Proceedings*: 1310–1318.
[76] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1): 246–256.

[77] Pearson, K. (1894). *Contributions to the Mathematical Theory of Evolution*. London: Dulau & Co.

[78] Poon, S.-H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature* 41(2): 478–539.

[79] Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

[80] Puke, M., & Schweikert, K. (2025). Coherent forecasting of realized volatility. Working paper, University of Hohenheim.

[81] Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. Paris: Bachelier.

[82] Ramòn y Cajal, S. (1909). *Histologie du système nerveux de l'homme & des vertébrés*. Vol. II. Paris: A. Maloine: 149.

[83] Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine learning for stock selection. *Financial Analysts Journal* 75(3): 70–88.

[84] Roll, R. (1984). Orange Juice and Weather. *American Economic Review* 74(5): 861–880.

[85] Rogalski, R. J., & Vinso, D. E. (1978). Empirical Properties of Foreign Exchange Rates. *Journal of International Economics* 8(3): 379–396.

[86] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6): 386–408.

[87] Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13(3): 341–360.

[88] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323(6088): 533–536.

[89] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3(3): 210–229.

[90] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11): 2673–2681.

[91] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19(3): 425–442.

[92] Sheppard, K. (2021). *Financial Econometrics Notes*. University of Oxford.

[93] Simon, H. A. (1984). On the behavioral and rational foundations of economic dynamics. *Journal of Economic Behavior and Organization* 5(1): 35–55.
[94] Sironi, A., & Resti, A. (2021). *Rischio e valore nelle banche. Misura, regolamentazione, gestione*. Milano: Egea, ristampa aggiornata, cap. 6, 7, 10.
[95] Standard & Poor's. *S&P 500 Index Methodology*. S&P Dow Jones Indices, ultima versione.
[96] STOXX Ltd. *STOXX Europe 600 Index Guide*. Qontigo/Deutsche Börse Group, ultima versione.
[97] STOXX Ltd. *Guide to the EURO STOXX 50 Volatility Index (VSTOXX)*. Deutsche Börse Group, ultima edizione.
[98] Student [Gosset, W. S.] (1908). The probable error of a mean. *Biometrika* 6(1): 1–25.
[99] Veronesi, P. (1999). Stock Market Overreaction to Bad News in Good Times: A Rational Expectations Equilibrium Model. *Review of Financial Studies* 12(5): 975–1007.
[100] Wilson, T. (1993). Value at Risk. *Risk Magazine* 6(10): 111–117.
[101] Zakoïan, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control* 18(5): 931–955.
[102] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2024). *Dive into Deep Learning*. Cambridge: Cambridge University Press.

Normativa e vigilanza

[104] Comitato di Basilea per la Vigilanza Bancaria. 1993. *The Supervisory Treatment of Market Risks*. Basilea: Banca dei Regolamenti Internazionali.
[105] Comitato di Basilea per la Vigilanza Bancaria. 1996. *Amendment to the Capital Accord to Incorporate Market Risks*. Basilea: BIS.
[106] Comitato di Basilea per la Vigilanza Bancaria. 2004. *International Convergence of Capital Measurement and Capital Standards: A Revised Framework (Basilea II)*. Basilea: BIS.
[107] Comitato di Basilea per la Vigilanza Bancaria. 2009. *Revisions to the Basel II Market Risk Framework (Basilea 2.5)*. Basilea: BIS.
[108] Comitato di Basilea per la Vigilanza Bancaria. 2010 (rev. 2011). *Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems*. Basilea: BIS.
[109] Comitato di Basilea per la Vigilanza Bancaria. 2016 (rev. 2019). *Minimum Capital Requirements for Market Risk (Fundamental Review of the Trading Book – FRTB)*. Basilea: BIS.
[110] Consiglio delle Comunità Europee. 1993. *Direttiva 93/6/CEE del Consiglio, del 15 marzo 1993, sull'adeguatezza patrimoniale delle imprese di investimento e degli enti creditizi*.