Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Previsione della Volatilità Realizzata tramite analisi del sentiment e riduzione della dimensionalità di un dataset di news finanziarie

Relatore Prof. Massimiliano Caporin
Dipartimento di Scienze Statistiche

Laureando Andrea Coccoli
Matricola N 1096160

Anno Accademico 2016/2017

Indice

Elenco delle tabelle

Introduzione

A partire dallo studio Fama 1970 è stata prodotta una grande quantità di materiale letterario riguardante l'efficienza dei mercati finanziari. Secondo gli studi dell'economista americano tutta l'informazione sui prezzi osservati di un determinato asset è contenuta nella serie storica dei prezzi stessi, quindi non è possibile ottenere rendimenti attesi maggiori di quelli di mercato tramite strategie di trading. Studi successivi hanno invece mostrato che vi sono diverse prove empiriche a sostegno dell'ipotesi opposta: esistono strategie che rendono i mercati non efficienti, nemmeno in forma debole. Nei quasi cinquant'anni che sono trascorsi, molto è cambiato nel mondo della finanza: per esempio, ora è possibile fare trading in tempo reale e praticamente in ogni luogo. Inoltre vi è stata una notevole evoluzione del concetto stesso di *informazione*: se all'inizio si trattava principalmente di notizie economiche fornite dai (relativamente) pochi mezzi di comunicazione specializzati, od ottenute in seguito a consulenze con esperti del settore, lo sviluppo dei mass-media prima e successivamente di internet (in particolare i social network) ha aumentato esponenzialmente il numero di notizie a disposizione di chiunque.

La *sentiment analysis* è una delle metodologie più recenti che sfrutta tutta questa informazione: consiste nell'estrarre informazioni dalle fonti testuali, attraverso l'utilizzo del linguaggio naturale e dell'analisi del testo. Essa permette di assegnare un valore positivo, negativo o neutro ad un tweet, piuttosto che a un articolo pubblicato su internet, o ad una notizia fornita da un provider. Ciò apre le porte a successive analisi statistiche per valutare se questo valore abbia un effetto significativo sull'andamento in borsa del titolo dell'azienda nominata, o sulla sua voltilità. Oltre a questo, va aggiunto che lo sviluppo tecnologico

degli ultimi decenni ha reso più facile ed economicamente poco dispendioso l'accumulo di dati. Se dal punto di vista della facilità di reperimento ciò rappresenta indubbiamente un vantaggio, dall'altro può generare, paradossalmente, un problema: la presenza in un dataset con moltissime variabili e/o osservazioni corrisponde, dal punto di vista prettamente matematico, ad un aumento eccezionale delle dimensioni delle matrici da analizzare e ha spinto, o meglio obbligato, gli studiosi a cercare metodi alternativi a quelli canonici fino ad allora utilizzati, con conseguente rinuncia alle loro proprietà "desiderabili": infatti risulta difficile immaginare una funzione in più di due dimensioni ed è chiaro come i dati si *spargano* velocemente all'aumentare della dimesionalità[1]. Quest'ultimo problema è anche noto col nome di "*curse of dimesionality*". In questo lavoro, per ovviare al problema appena esposto, sono state applicate diverse tecniche di *Dimensionality Reduction* su un dataset creato appositamente per trattare la stima e la previsione della *Volatilità Realizzata (RV)* di un titolo e che fa parte di uno studio non ancora pubblicato. Tale studio ha prodotto, in aggiunta, un algoritmo (che qui è stato validato) per l'identificazione del sentiment contenuto in dati di testo forniti da provider specializzati. Pertanto, è possibile dividere questa tesi in tre parti: la prima contiene una descrizione delle variabili considerate, in particolare i concetti principali riguardanti il *sentiment* e la procedura di validazione dell'algoritmo di *sentiment detection*, verranno anche descritte le caratteristiche della voltilità realizzata e delle tecniche utilizzate per modellarla e successivamente verrà analizzato nel dettaglio il dataset utilizzato; nella seconda si tratteranno le tecniche di riduzione della dimensionalità utilizzate per il dataset, ponendo l'attenzione anche sulle scelte che è stato necessario prendere; infine, l'ultima contiene i risultati delle analisi svolte, ovvero le stime e previsioni un passo in avanti della volatilità realizzata.

[1] Azzalini, Scarpa e Walton 2012, "Data Analysis and Data Mining. An Introduction".

Capitolo 1

Sentiment e Realized Volatility

In questo capitolo verranno descritti i due concetti che stanno alla base del lavoro svolto: da un lato il sentiment riguarda il tentativo, molto ambizioso, di innovare l'analisi di dati finanziari utilizzando il contenuto del testo delle news, mentre dall'altro si trova la variabile d'interesse per le previsioni, rappresentata da una misura di volatilità (ovvero una delle possibili stime della variabilità del mercato).

1.1 Cos'è il sentiment?

Dal punto di vista concettuale, la sentiment analysis rappresenta una sfida difficile per la statistica: lo scopo ultimo è di riuscire a trattare dati soggettivi in maniera oggettiva, assegnando ad una fonte testuale un valore numerico. A partire dai primi anni 2000 sono stati svolti diversi studi, dapprima per cercare una metodologia convincente per "misurare" il contenuto delle notizie e in seguito per scovare una relazione tra questa misura e l'andamento del mercato. Di fatto, lo scopo è sempre stato classificare una news come positiva o negativa (o neutra) in base alle parole che contiene, come mostrato in Dave, Lawrence e Pennock 2003 (definizione di *opinion minig*). Questo "contenuto" può rappresentare una definizione basilare del termine "sentiment", ma non del tutto soddisfacente. Il problema più grande riguarda infatti come effettuare realmente questa misura: inizialmente in Antweiler e Murray 2004 fu sviluppato un algoritmo basato sem-

plicemente sul numero di volte in cui le parole apparivano, in Tetlock 2007 fu introdotta per la prima volta la metodologia basata sulle "liste di parole", poi in Tetlock, Saar-Tsechansky e Macskassy 2008 fu utilizzato il dizionario **H4N** come diretta conseguenza dello studio precedente, mentre in Loughran e McDonald 2011 venne apportata un'ulteriore miglioria.
Lo sviluppo tecnologico massiccio e la diffusione dei computer hanno permesso di ottenere notevoli miglioramenti, ma raramente ciò basta per compiere un'analisi testuale soddisfacente: è noto in letteratura[1] che il semplice utilizzo della "forza bruta" dei calcolatori non porta a sostanziali benefici, perché il solo conteggio delle parole rappresenta un'analisi troppo superficiale.

È chiaro quindi che effettuare tale misura non è sempre banale. In particolare sembra non esserlo in finanza, come è stato dimostrato nello studio già citato di Loughran e McDonald 2011, in cui si è potuto constatare come la classificazione di parole tramite il Dizionario Psicosociologico di Harvard[2] risultasse poco precisa, benché esso venisse utilizzato ampiamente in moltissimi ambiti diversi. Il motivo principale per cui ciò avveniva è che una parte consistente delle parole classificate come negative, in realtà non lo sono in ambito finanziario. Pertanto è stata necessaria una ridefinizione delle liste. In aggiunta alle considerazioni degli autori, durante la scrittura dell'algoritmo utilizzato in questo studio, è stato svolto un lavoro supplementare sulla lista delle negazioni, la quale è stata notevolmente ampliata poiché la presenza di negazioni influisce moltissimo sul significato semantico di un testo o una frase, ma allo stesso tempo non sempre esse vengono espresse in modo chiaro.

1.1.1 Misura del sentiment

Sono stati considerati solo i titoli dell'indice S&P 100 di cui si avevano a disposizione tutte le osservazioni giornaliere nell'intervallo di tempo dal 4 febbraio 2005 al 25 febbraio 2015. Di questi titoli sono state raccolte tutte le *news* fornite da due provider specializzati quali StreetAccount di FactSet e Thomson One di Thomson Reuters, nonché i Trends di Google (basati su Google Search), che mostrano quanto viene ricercato un determinato termine. In seguito, alle news è

[1] Ceron, Curini e Iacus 2014.
[2] Harvard-IV-4 TagNeg (H4N).

stato applicato un algoritmo di *Sentiment Detection* basato sulle liste di parole di Loughran e McDonald 2011[3], modificandone in particolare quella contenente le negazioni: si è passati da solo 6 parole singole, a tre guppi (nominati "parole singole", "sequenza di due parole", "sequenza di tre parole") formati da 28, 24 e 6 elementi, rispettivamente. Per ottenere il sentiment complessivo di una news è stata seguita la seguente procedura:

1. alle parole valutate positive è stato assegnato valore +1, a quelle negative -1 e il valore è stato invertito qualora fosse presente anche una negazione;

2. facendo una media aritmetica del sentiment di tutte le parole valutate, è stata ottenuta una misura chiamata *Sentiment Relativo*, per costruzione compresa tra -1 e +1;

3. al sentiment complessivo è stato assegnato uno dei valori tra 1, -1 e 0, a seconda che quello relativo risultasse maggiore di 0.05, minore di -0.05, o incluso (esterni compresi).

1.1.2 Validazione dell'algoritmo

Con il termine "validazione" si intende fornire delle misure per analizzare il comportamento dell'algoritmo. Nel corso dell'intero lavoro, il *sentiment* complessivo appena descritto è stato trattato come una variabile qualitativa multiclasse, nel senso che ogni osservazione può essere classificata come appartenente a una e una sola delle tre classi non sovrapposte che lo compongono. Chiaramente, queste classi si riferiscono alla positività, negatività o neutralità riscontrate dall'algoritmo e sono state etichettate coi valori +1, -1 e 0, rispettivamente. Trattandosi di più classi, le misure ottenibili dalla *matrice di confusione*[4] forniscono solo informazioni "locali" sulle singole classi, ma per avere un'idea della bontà generale è necessario ottenere dei valori aggregati.

[3] *negative, positive, uncertainty, litigious, strong modal, weak modal.*
[4] Per esempio: accuratezza, richiamo, precisione, tasso d'errore e F-score.

1.1.3 Selezione dei titoli

La validazione è stata fatta selezionando 10 dei titoli considerati all'inizio, senza avere preferenze tra le aziende che venivano inserite[5]. Poi sono state selezionate casualmente 50 news riguardanti ognuno di essi: 25 fornite da StreetAccount e 25 da Thomson One, per un totale di 500. Successivamente, i risultati ottenuti utilizzando l'algoritmo che considera la lista di negazioni di Loughran e McDonald 2011 (chiamato d'ora in poi "no-neg") e quello con la lista ampliata (algoritmo "con-neg") sono stati confrontati con una stima manuale del sentiment. Di seguito, nelle Tabelle 1.1 e 1.2, vengono riportate le matrici di confusione nei due casi.

		Valori previsti		
		-1	0	1
Valori veri	-1	78	13	3
	0	58	123	22
	1	46	73	84

Tabella 1.1: Matrice di confusione dell'algoritmo "no-neg".

		Valori previsti		
		-1	0	1
Valori veri	-1	79	11	4
	0	53	122	28
	1	41	74	88

Tabella 1.2: Matrice di confusione dell'algoritmo "con-neg".

1.1.4 Misure di performance aggregate

Lo studio Sokolova e Lapalme 2009[6] fornisce delle misure semplici, per analizzare, come in questo caso, variabili con più di due categorie: innanzitutto il problema di classificazione viene affrontato come *One-vs-All* e non *One-vs-One*, nel senso che le misure delle singole classi non vengono calcolate rispetto alle

[5] Gli stessi verranno utilizzati anche per tutte le analisi successive: la lista è contenuta nell'Appendice A.1.

[6] In particolare: Table 3, pag.430

altre due considerandole una alla volta, ma come se queste fossero una sola; detto ciò, per ognuna sono stati calcolati i valori dei **falsi** positivi e negativi e dei **veri** positivi e negativi. In seguito, a partire da questi valori, sono state ottenute le seguenti misure aggregate:

1. ***accuratezza media***: media dei rapporti delle somme tra veri positivi e veri negativi con il totale delle osservazioni;
2. ***tasso d'errore***: media dei rapporti delle somme tra falsi positivi e falsi negativi con il totale delle osservazioni;
3. ***micro-precisione***: rapporto tra la somma di tutti i veri positivi e la somma di tutti i veri positivi coi falsi positivi;
4. ***Macro-precisione***: media dei rapporti tra i veri positivi e la somma tra veri positivi e falsi positivi;
5. ***Macro-recupero***: media dei rapporti tra i veri positivi e la somma tra veri positivi e falsi negativi;
6. ***Macro statistica-F*** [7]: una combinazione delle misure "Macro", pari a

$$F_M(\beta) = \frac{(\beta^2+1)Macro-precisione * Macro-recupero}{\beta^2 Macro-precisione + Macro-recupero}. \qquad (1.1)$$

Nella Tabella 1.3 vengono riportati i valori che sono stati calcolati: si può notare da subito come non ci siano i miglioramenti "teorici" che una lista di negazioni più ampia e curata dovrebbe apportare. Infatti, benchè i risultati dell'algoritmo "con-neg" siano numericamente preferibili (tranne la Macro-precisione), dal punto di vista concettuale ci si aspetterebbero probabilmente performance maggiori da questo algoritmo, visto anche quanto lavoro ha richiesto l'affinamento della tecnica di *detection*.

[7] In accordo con quanto viene suggerito in letteratura, si è scelto $\beta = 1$.

Misure	No-neg	Con-neg
Accuratezza media	0.713	0.719
Tasso d'errore	0.287	0.281
Micro precisione	0.570	0.578
Macro precisione	0.596	0.593
Macro recupero	0.616	0.625
Macro F-Statistic	0.606	0.609

Tabella 1.3: Misure di performance multiclasse.

Come già accennato in precedenza, in genere l'analisi testuale non è banale da svolgere e pertanto, trattandosi ancora di una fase quasi "embrionale", è lecito aspettarsi dei risultati non eccelsi. Certo è che un fatto non può essere nascosto: un tasso d'errore di poco inferiore al 30% è comunque un valore abbastanza elevato, che indubbiamente influenzerà in maniera negativa l'interpretabilità e la correttezza delle analisi svolte.

1.2 Cos'è la Volatilità Realizzata?

In finanza il termine volatilità si riferisce a quanto il mercato risulta movimentato, ovvero quanto cambiano i prezzi in un certo intervallo di tempo. Esistono diversi modi per misurare numericamente questo grado di "movimento", ma un fattore da tenere in forte considerazione per la scelta è l'orizzonte temporale al quale si riferisce l'analisi da svolgere. Se per esempio ci si basa sulla teoria della frontiera efficiente di Markowitz, che rappresenta tutt'oggi uno dei capisaldi nell'ambito degli investimenti di mercato, per descrivere la distribuzione condizionata dei rendimenti (ovvero, le variazioni dei prezzi) sono sufficienti la media e la varianza campionarie (o la matrice di covarianza, nel caso di un portafoglio di strumenti finanziari), perché l'orizzonte temporale a cui ci si riferisce è tipicamente quello mensile e nelle serie storiche mensili, la varianza è un buon indicatore di variabilità, specialmente se in aggiunta viene trattata con modelli che ammettono la presenza di eteroschedasticità condizionata (come i modelli della famiglia GARCH). Se però le operazioni vengono svolte ad intervalli più brevi come i giorni, o addirittura ne vengono fatte diverse all'interno dello stes-

so giorno, la varianza risulta essere un indicatore che in genere sovrastima la volatilità effettivamente osservata[8].

Un tentativo importante di risolvere questo problema è stato fatto in Andersen e Bollerslev 1998a, nel quale gli autori suggeriscono di dividere il giorno in molti sotto-intervalli (per esempio, di 5 minuti) e calcolare come misura di volatilità per quel giorno, la somma dei quadrati dei rendimenti intra-giornalieri, definendola appunto *Volatilità Realizzata*. Ciò è possibile e sensato, perché considerando le dinamiche dei processi di Itô che descrivono i modelli dei prezzi a tempo continuo, quando si ragiona in termini di rendimenti si può scrivere:

$$r_t = \int_{t-1}^{t} \mu_s ds + \int_{t-1}^{t} \sigma_s dW_s$$

dove r_s indica il rendimento al tempo s, μ_s il drift in s, σ_s la volatilità istantanea in s e W_s il moto Browniano. In particolare, il termine che interessa stimare è il secondo (che prende il nome di *Volatilità Integrata*), anche perché il primo è spesso nullo o trascurabile: nell'articolo citato, viene dimostrato che la varianza condizionata dei modelli GARCH corrisponde alla Volatilità Integrata e una stima discretizzata di essa è rappresentata proprio dalla Volatilità Realizzata. Tale stima risulta più recisa, aumentando il numero di periodi in cui si divide l'intervallo di tempo[9].

1.2.1 Come viene modellata la Volatilità Realizzata

Moltissimi studi riguardanti la modellazione sono stati svolti, specialmente in seguito alla massiccia diffusione dei dati finanziari a frequenza elevata e ultra-elevata. Vale la pena nominarne almeno due: si tratta di Corsi 2009 e Barndorff-Nielsen e Shephard 2004. Il primo è stato fondamentale per l'introduzione di modelli a memoria lunga per descrivere il processo della volatilità realizzata, mentre il secondo per l'utilizzo di somme e prodotti dei valori assoluti dei rendimenti[10] come stima della volatilità e dei *jumps* come approssimazione della sua componente discontinua. Studi empirici come Andersen, Bollerslev e Diebold 2007 hanno mostrato che buona parte della variabilità che può essere cattura-

[8] Ad esempio in Andersen e Bollerslev 1998b, in particolare nella Figura 2, pag.228.

[9] In realtà, al diminuire della grandezza dei periodi, cresce anche l'effetto del rumore dovuto alla microstruttura del mercato.

[10] *Power & Bipower Variation.*

ta deriva dalla *persistenza* della volatilità, ovvero la dipendenza dinamica dai ritardi. Il modello definitivo viene chiamato HAR-TCJ e assume la seguente forma:

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 C_{t-5} + \beta_3 C_{t-22} + \beta_4 J_{t-1} + \epsilon_t. \tag{1.2}$$

I β sono i coefficienti delle variabili, i termini RV indicano la Volatilità Realizzata ai tempi t e t-1, J_{t-1} è il jump al tempo t-1, mentre le componenti "C" indicano le medie dei precedenti 5 e 22 valori della volatilità[11], utilizzate come approssimazioni della "memoria" del processo.

1.2.2 Variabili aggiuntive

Il contributo di questo studio all'argomento discusso in precedenza vuole essere valutare se l'aggiunta delle variabili relative alle news apporta dei miglioramenti sostanziali, oppure se esse rappresentano solamente una fonte di "disturbo" per il modello (1.2). Pertanto, quest'ultimo è stato modificato come segue:

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 C_{t-5} + \beta_3 C_{t-22} + \beta_4 J_{t-1} + \beta_{News}^T News_{t-1} + \epsilon_t. \tag{1.3}$$

Questa volta β_{News}^T rappresenta un vettore (trasposto) di coefficienti e $News_{t-1}$ una matrice con tante colonne quante le variabili selezionate per l'analisi. Si rimanda ai prossimi capitoli la descrizione di tutte le variabili potenzialmente utilizzabili e di quelle effettivamente incluse nell'analisi.

[11] Si riferiscono quindi alla settimana e al mese lavorativi precedenti.

Capitolo 2

Descrizione del dataset

La scelta dei titoli è ricaduta sull'indice S&P 100 perché include aziende molto conosciute e con una elevata capitalizzazione. Inoltre, il dataset è stato creato considerando 10 variabili macroeconomiche relative agli Stati Uniti d'America[1], i cui valori sono stati forniti da Thomson Reuters. Il quadro complessivo delle variabili già a disposizione si completa con i dati su annunci e previsioni degli *Earnings per share* (EPS), forniti da FactSet.

2.1 Definizione delle nuove variabili

Tutto sommato, la procedura di *detection* descritta nel capitolo precedente risulta essere abbastanza standard nell'ambito dell'analisi testuale, in quanto ha portato semplicemente ad assegnare un valore numerico all'informazione contenuta. Tuttavia, come già accennato, è stato possibile fare un passo oltre le misure solitamente utilizzate attraverso la definizione di nuove. Esse si basano su uno schema di nove "concetti". Ognuno di essi risulta specifico nella reazione che può causare nel mercato e si riferisce, oltre che al periodo di riferimento, anche a quelli precedenti di lunghezza pari o maggiore (se il riferimento è il giorno, allora le variabili verranno costruite in base all'informazione rilasciata

[1] Fiducia dei Consumatori, Indice dei Prezzi al Consumo, Tasso d'Interesse del FOMC, PIL, Produzione Industriale, Bilancia dei Pagamenti, Richieste di Indennizzo di Disoccupazione, Variazione dei Nuovi Occupati nel mese precedente (esclusi operai e statali), Indice dei Prezzi alla Produzione e Vendite al dettaglio.

quel giorno e nei vari giorno, settimana e mese precedenti).
Scendendo nel dettaglio, i concetti sono:

- **misure standard**:

  1. numero di news;
  2. numero di parole;
  3. sentiment.

  Queste sono state incluse perché approssimano la quantità e la qualità delle informazioni;

- **quantità anomala**: ovvero una quantità di news oltre una certa soglia, che può sorprendere il mercato e quindi influenzarlo;

- **incertezza**: occorrenza di news con sentiment opposti, all'interno del periodo di riferimento;

- **indice di rilascio delle notizie**: una misura della quantità di informazioni rilasciate durante il periodo di riferimento, che tiene anche in considerazione se un rilascio improvviso in un breve intervallo abbia un effetto diverso rispetto ad un rilascio graduale su un intervallo più lungo. Per ogni intervallo di tempo, l'indice è definito come sommatoria della *k-esima* potenza (con $k \geq 1$) del numero di news o parole rilasciate nei sotto-intervalli che lo compongono (per esempio, il giorno viene diviso in intervalli di 5, 10 o 15 minuti, mentre la settimana o il mese in serie da 1, 2 o 5 giorni);

- **variazione della quantità**: variazione, tra i periodi, della quantità di news o parole, per valutare se il mercato venga influenzato non solo dal rilascio di informazioni, ma anche dalla loro numerosità;

- **persistenza/interazione delle news**: quando una **quantità anomala** si manifesta in due periodi consecutivi;

- **inversione del sentiment**: quando il sentiment di un periodo è l'opposto di quello dei periodi precedenti;

- **variazione della quantità condizionata al sentiment**: variazione positiva condizionata al sentiment del periodo di riferimento e negativa rispetto a quello dei periodi precedenti, per tener conto del sentiment del periodo con più informazioni (quando se ne confrontano diversi);

- **sentiment condizionato alla quantità**: sentiment del periodo di riferimento, condizionato alla quantità di informazioni rilasciate durante lo stesso e nei periodi più lunghi, per valutare se gli investitori si basano sul primo, tenendo conto della seconda e se lo fanno in maniera più o meno evidente.

2.1.1 EPS e variabili Macro

Dato che per gli EPS e le variabili macroeconomiche sono stati forniti sia i valori reali che quelli previsti, sono state create due misure standardizzate che accorpassero di fatto l'informazione di entrambe le singole serie, in conformità con quanto viene spesso fatto in letteratura: si tratta dei rapporti tra gli errori di previsione e le loro deviazioni standard e per gli EPS prendono il nome di punteggi *SUE* (Standardized Unexpected Earnings)[2].

2.1.2 Google Trends

Per analizzare i *Trends* di Google è stato costruito un indice, chiamato appunto Google Search Index (GSI). Poiché si avevano a disposizione solo le serie giornaliere dei singoli mesi (relative al valore massimo del rispettivo mese, che è pari a 100) e quella dei dati aggregati mensilmente (relativi al massimo su tutto il periodo di osservazione), si è deciso di creare dapprima la serie giornaliera relativa. Ogni osservazione di essa è il risultato del rapporto tra il valore della serie giornaliera in ogni giorno e la somma di tutti i valori di quel mese. Per ottenere la serie giornaliera sull'intero periodo di osservazione, è bastato moltiplicare tra di loro quella relativa appena descritta e quella mensile. Infine, si è ottenuto il GSI, dividendo i singoli valori della serie giornaliera complessiva moltiplicati per cento, per i valori massimi registrati nell'intero periodo.

[2]Non hanno invece un nome specifico per le variabili macroeconomiche.

2.1.3 Variabili relative agli orizzonti temporali e trasformazioni

Oltre che a considerare le informazioni del mercato sull'andamento dell'ultima settimana, dell'ultimo mese e nelle sole fasi di apertura, è sembrato opportuno tenere conto anche di un fenomeno ben noto in letteratura: l'accumulo d'informazione che si verifica durante la chiusura del mercato, nella fase detta *overnight*.

Infine, sono state create delle trasformazioni delle variabili precedenti come *radice con segno, logaritmo* (del valore più 1) *con segno* e *potenza quadrata con segno*, per catturare eventuali relazioni non-lineari tra attività di mercato e indicatori. In più il quadro è completato, con le variabili indicatrici del fatto che ogni variabile precedentemente descritta fosse positiva, negativa o nulla (a patto che avesse effettivamente senso considerare le varie possibilità[3]).

2.2 Riassumendo

Per concludere la descrizione è opportuno riassumere le variabili complessivamente considerate per ogni titolo nella seguente maniera:

- le variabili riferite ai **giorni** risultano essere 868;
- 496 quelle riferite alla fase **overnight**;
- 589 sia quelle **settimanali** che **mensili**;
- le variabili **multi-periodali** sono in tutto 2232;
- le misure sugli **EPS** sono 32, quelle sulle variabili **macro** 320;
- infine, per i **Google Trends** se ne contano 33.

Il totale complessivo ammonta a 5159, a cui vanno aggiunte le variabili che si riferiscono alla persistenza della volatilità, mentre il numero di osservazioni è di 2531. Pertanto risulta evidente l'impossibilità di analizzare questa mole di dati con i metodi tradizionali e la necessità di ricorrere a metodi di dimensionality reduction.

[3] Per esemepio, per le variabili che possono essere solo positive non aveva senso considerare la possibilità di osservarne valori negativi.

Capitolo 3

Metodi per la riduzione della dimensionalità

Una parte consistente del lavoro ha riguardato la definizione delle variabili su cui sono state realmente svolte le analisi: la dimensionalità proibitiva del dataset ha reso questo passo fondamentale e dovuto. Esistono moltissime tecniche più o meno complesse che potevano essere applicate, ma tutte in generale ben definite in letteratura e comunque, ampiamente utilizzate. Pertanto, è stato necessario fare delle scelte[1]: dopo una scrematura iniziale, atta ad eliminare tutte le variabili che risultavano non adatte ad essere analizzate[2], le rimanenti sono state divise in qualitative e quantitative (queste ultime sempre in maggioranza) e trattate in maniera opportuna attraverso tre metodi differenti:

- *Correspondence Analysis* per le variabili qualitative;
- *Principal Component Analysis* e *Indipendent Component Analsysis* per le variabili quantitative.

[1] Chiaramente, sarà molto importante l'apporto di lavori futuri, che utilizzeranno altre tecniche o avranno a disposizione aggiornamenti dell'algoritmo di detection.

[2] Da un lato si sono riscontrati numerosissimi casi di variabili che erano costanti o combinazioni lineari di altre e pertanto andavano forzatamente eliminate; dall'altro si è scelto di rinunciare "volutamente" ad alcune variabili, perché ritenute avere un contenuto informativo ridotto, o comunque non sufficiente per giustificarne l'inclusione nel dataset: si tratta delle molte variabili categoriali in cui le osservazioni risultavano appartenere in grande maggioranza a una sola delle classi e di altre numeriche che assumevano una ristretta gamma di valori. È stato scelto di utilizzare una soglia, del tutto arbitraria e opinabile, imponendo che fossero eliminate le variabili in cui almeno l'85% delle osservazioni assumeva lo stesso valore.

La procedura di analisi ha previsto due fasi: nella prima sono state considerate le iniziali 1508 osservazioni e in base questo dataset è stato definito, a seconda dei vari approcci, quante componenti o dimensioni considerare; nella seconda parte è stato utilizzato il resto del dati per eseguire delle stime step-rolling di un'osservazione (ripetute per ogni mese), con previsione un passo in avanti della RV, attraverso dei modelli lineari con standard error robusti. Le previsioni sono state poi confrontate attraverso test di Diebold-Mariano. Maggiori dettagli vengono lasciati alle sezioni e ai capitoli successivi.

3.1 Variabili Qualitative

La Correspondence Analysis (CA) è una tecnica sviluppata nella seconda metà degli anni '80 ed ha come scopo principale la rappresentazione grafica dei punti di un dataset di variabili categoriali, in una maniera alternativa ai classici grafici a dispersione. In presenza di due sole variabili si utilizza spesso la tabella di contingenza dei valori, per ottenere le frequenze (assolute o relative) di ogni possibile combinazione, permettendo di confrontarle tra di loro e fornendo un'analisi più compatta dei dati. All'aumentare del numero di variabili e/o di categorie per ogni variabile, risulta però ben presto difficile utilizzare questa strategia, a meno di un adeguato[3] aumento della dimensionalità, per il motivo già accennato che i dati si "spargono". A questo punto il grafico a dispersione non è più uno strumento valido da utilizzare.

3.1.1 *"Profili e Spazio dei Profili"*

Nel caso di due sole variabili categoriali, è possibile riassumere comodamente le osservazioni in una tabella a doppia entrata con tante colonne quante le categorie di una delle variabili e tante righe quante le categorie dell'altra. In ogni entrata vengono inserite le frequenze relative, $f_{(i,j)}$, delle osservazioni appartenenti alla categoria *i-esima* della variabile le cui categorie formano le righe della tabella e alla *j-esima* di quella le cui categorie formano le colonne. Inoltre, si può aggiungere una colonna che conterrà le somme per ogni riga delle frequenze, $n_{i,+} = \sum_j f_{(i,j)}$ e allo stesso modo una riga contenente le somme per ogni

[3] E spesso questa *adeguatezza* molto difficilmente è raggiungibile.

colonna, $n_{+,j} = \sum_i f_{(i,j)}$. La Tabella 3.1 mostra un esempio di due variabili con $i = 4$ e $j = 3$.

	Categoria1	Categoria2	Categoria3	$Totale_r$
CategoriaA	$f_{(1,1)}$	$f_{(1,2)}$	$f_{(1,3)}$	$n_{1,+}$
CategoriaB	$f_{(2,1)}$	$f_{(2,2)}$	$f_{(2,3)}$	$n_{2,+}$
CategoriaC	$f_{(3,1)}$	$f_{(3,2)}$	$f_{(3,3)}$	$n_{3,+}$
CategoriaD	$f_{(4,1)}$	$f_{(4,2)}$	$f_{(4,3)}$	$n_{4,+}$
$Totale_c$	$n_{+,1}$	$n_{+,2}$	$n_{+,3}$	$n_{+,+}$

Tabella 3.1: Esempio di tabella a doppia entrata.

Chiaramente, vale: $n_{+,+} = \sum_j n_{+,j} = \sum_i n_{i,+}$.

Utilizzando la nomenclatura di Greenacre 2016, si ha che i vettori contenenti i valori ottenuti dividendo $f_{(i,j)}$ per $n_{i,+}$ siano detti *Profili Riga*, mentre quelli ottenuti dividendo $f_{(i,j)}$ per $n_{+,j}$ siano detti *Profili Colonna*. Per costruzione, ogni profilo gode della proprietà di sommare a 1 o, equivalentemente, al 100%. Inoltre, viene confermato che: *«A seconda che l'interesse dello studio sia incentrato su una variabile piuttosto che l'altra...»*, *«... si può decidere se concentrarsi su un "Profilo" piuttosto che l'altro, con la certezza che i risultati della CA saranno invarianti rispetto a questa scelta.»*[4]

Con riferimento ad una tabella che raggruppa tutti i profili riga, è possibile considerare una riga aggiuntiva contenente i valori ottenuti come rapporto tra i vari $n_{+,j}$ e $n_{+,+}$: questo nuovo vettore viene chiamato profilo *riga Medio*. Facendo lo stesso coi profili colonna e considerando il vettore contenente i rapporti tra $n_{i,+}$ e $n_{+,+}$, si ha invece il profilo *colonna Medio*.

3.1.2 Rappresentazione grafica dei Profili

Dal punto di vista grafico un profilo gode di una proprietà interessante: dato che esso contiene m punti che sommano a uno, può essere rappresentato in uno spazio *(m-1)-dimensionale* detto Simplesso e compreso all'interno del contorno ottenuto unendo tutte le coppie degli m vettori unitari, sugli m assi perpendicolari: in pratica, gli elementi dei profili possono essere visti come le coordinate che li identificano nel piano cartesiano in cui le categorie rappresentano le etichette

[4] Cap.2, pag.11.

dei nuovi assi. Il nuovo piano è detto *Sistema a Coordinate Baricentriche*, dove il riferimento al "baricentro" è legato al concetto di "media pesata". I vertici del contorno delimitato dai vettori unitari è detto *Spazio dei Profili*.

3.1.3 Alcuni concetti: Centroide, Massa, Distanza e *Inertia*

Dato che la CA permette una trattazione simmetrica di righe e colonne della tabella di contingenza, è stato scelto di riferirsi d'ora in avanti al caso delle variabili in riga e quindi ai corrispondenti Profili Riga.
I Profili Medi svolgono un ruolo fondamentale, perché rappresentano una media pesata (o *Centroide*) dei profili stessi, in cui i pesi sono i singoli valori, anche chiamati *Masse*, che vengono assegnati ai vertici. Perciò un profilo tenderà a posizionarsi più vicino al vertice al quale è assegnato un valore più elevato. Tali pesi sono proporzionali alle somme per riga della tabella iniziale, ovvero non sono altro che i Profili Colonna Medi. Sempre per le proprietà "simmetriche" della CA, concentrandosi invece sui profili colonna, si avrebbe che le masse sono corrispondenti ai Profili Riga Medi.

A questo punto ci si può chiedere se vi sia una differenza significativa tra le varie classi. Nel caso ciò non si verificasse, si direbbe che i Profili sono "simili" e le classi *omogenee* rispetto a quella variabile. Se si utilizza un test statistico, l'ipotesi di omogeneità non può essere rifiutata ogni volta che le frequenze osservate non sono *uguali* a quelle previste, ma solo quando sono *molto* diverse; ovvero serve definire una misura di *Distanza* delle classi dall'omogeneità. In letteratura, viene spesso utilizzata la statistica χ^2, definita come somma dei quadrati delle differenze tra valori previsti e osservati delle classi, divisi per i valori previsti:

$$\chi^2 = \sum \frac{(valori\ osservati - valori\ previsti)^2}{valori\ previsti}. \tag{3.1}$$

Dato che la categoria *i-esima* contiene in totale $n_{i,+}$ elementi, ci si aspetta che la percentuale di appartenenti alla categoria *j-esima* sia pari al corrispondente valore del Profilo Riga Medio. Per cui, con riferimento all'Equazione (3.1)

e alla Tabella 3.1:

$$valore\ osservato_{(i,j)} = f_{(i,j)} \tag{3.2}$$

$$valore\ previsto_{(i,j)} = n_{i,+}\frac{n_{+,j}}{n_{+,+}}. \tag{3.3}$$

La statistica complessiva è la somma delle statistiche riferite alle singole categorie e il valore che assume viene confrontato con il quantile di una distribuzione χ^2, con numero di gradi di libertà pari al prodotto tra i numeri delle categorie delle due variabili, diminuiti di uno.

Benché questa misura di omogeneità sia del tutto consistente sia in ambito teorico che pratico, è altresì possibile definire una statistica che indichi invece il grado di discrepanza: dividendo tutti i numeratori e denominatori di (3.1) per i rispettivi $n_{i,+}$ al quadrato, si ottengono al numeratore i profili (riga) osservati e previsti, mentre il denominatore può essere riscritto come $\frac{n_{i,+}}{profilo\ previsto_i}$. In pratica, l'equazione (3.1) può essere riscritta come:

$$\chi^2 = \sum_i n_{i,+}\frac{(profilo\ osservato_i - profilo\ previsto_i)^2}{profilo\ previsto_i}. \tag{3.4}$$

Infine, dividendo entrambi i termini in (3.4) per $n_{+,+}$, si ottiene una quantità nota come *Inertia Totale* e indicata spesso con ϕ^2. Essa misura quanta varianza è presente in una tabella e assume la forma seguente:

$$\phi^2 = \frac{\chi^2}{n_{+,+}} = \sum_i \frac{n_{i,+}}{n_{+,+}}\frac{(profilo\ osservato_i - profilo\ previsto_i)^2}{profilo\ previsto_i}. \tag{3.5}$$

Per rappresentare graficamente la distanza χ^2 è necessario "trasformarla" in maniera tale da ricondursi all'usuale distanza euclidea. Se si considerano i *k-esimi* elementi di due profili, x_k e y_k, come coordinate cartesiane, allora la distanza euclidea dei profili sarà pari alla radice quadrata della somma dei quadrati delle differenze tra gli elementi: $\sqrt{\sum_k (x_k - y_k)^2}$; mentre nella distanza χ^2, ogni termine sarà diviso per il corrispettivo valore c_k del profilo medio: $\sqrt{\sum_k \frac{(x_k-y_k)^2}{c_k}}$. La soluzione che consente di passare dall'una all'altra è in realtà molto semplice: dato che si può scrivere $\frac{(x_k-y_k)^2}{c_k}$ come $(\frac{x_k}{\sqrt{c_k}} - \frac{y_k}{\sqrt{c_k}})^2$, è sufficiente considerare $\frac{x_k}{\sqrt{c_k}}$ e $\frac{y_k}{\sqrt{c_k}}$ come nuove coordinate "trasformate".

Sempre in Greenacre 2016, si legge: *«L'essenza della CA è di individuare un sottospazio di poche dimensioni, che contenga approssimativamente i profili, ovvero identifica le dimensioni lungo le quali c'è una bassa dispersione di questi punti ed elimina le direzioni a bassa variazione, riducendo così la dimensionalità»*[5]. Tali dimensioni vengono chiamate *Assi Principali* e la loro quota di Inertia rispetto a quella Totale è detta *Inertia Principale*: più questa è vicina al 100%, meno la riduzione di dimensionalità provoca perdita d'informazione rilevante. Le nuove coordinate con cui i dati vengono ricodificati sono dette quindi *Coordinate Principali*.

3.1.4 Multiple Correspondence Analysis

Finora si è parlato solamente della CA classica, ovvero con sole due variabili implicate nell'analisi. Tuttavia, è possibile applicare delle tecniche che derivano direttamente dalla CA, anche su interi set di variabili (omogenee)[6]. Il primo passo consiste nel generare la *Matrice Indicatrice*, la quale ha tante righe quante osservazioni e tante colonne quanto il numero complessivo di classi di tutte le variabili considerate (codificate tramite variabili indicatrici). Un esempio viene fornito nella seguente Tabella 3.2:

			Variabile1				Variabile2			Variabile3		
Variabili			Categorie				Categorie			Categorie		
1	2	3	1	2	3	4	1	2	3	1	2	3
4	1	1	0	0	0	1	1	0	0	1	0	0
3	2	2	0	0	1	0	0	1	0	0	1	0
2	2	1	0	1	0	0	0	1	0	1	0	0
1	2	3	1	0	0	0	0	1	0	0	0	1
4	3	3	0	0	0	1	0	0	1	0	0	1
⋮	⋮	⋮			⋮				⋮			⋮

Tabella 3.2: Esempio di Matrice Indicatrice.

La Multiple Correspondence Analsysis (MCA) può essere definita semplicemente come una CA applicata alla Matrice Indicatrice, anche perché l'Inertia

[5] Cap.6, pag.43.
[6] Abdi e Valentin 2007; Camiz e Gomes 2013.

di tale matrice assume una forma molto comoda per l'analisi: utilizzando la nomenclatura presente in Greenacre 2016[7], si assuma di avere Q variabili e che la *q-esima* abbia J_q categorie, con $J = \sum_q J_q$ numero totale di categorie; si definisca la matrice indicatrice Z, con J colonne, come composta dalle Z_q sotto-matrici con J_q colonne; allora, l'Inertia totale di Z è pari alla media delle Inertiae delle sotto-matrici Z_q. Dato che ogni riga delle Z_q ha un solo 1 e tutti zeri, i profili riga delle Z_q giacciono sui vertici e le Inertiae sono pari a uno in ogni asse principale della sotto-matrice mentre quella totale è pari alla sua dimensionalità, che è $J_q - 1$. Pertanto, l'Inertia totale di Z è $\phi^2(Z) = \frac{1}{Q}\sum_q \phi^2(Z_q) = \frac{1}{Q}\sum_q (J_q - 1) = \frac{J-Q}{Q}$. Siccome J-Q è la dimensionalità di Z, si ha che $\frac{1}{Q}$ è l'Inertia media per ogni dimensione e può essere considerata come una sorta di soglia di decisione per valutare se inserire o meno la corrispondente dimensione nella MCA.

In alternativa, un metodo ugualmente efficace e indissolubilmente legato al precedente si basa sulla *Matrice di Burt*, simmetrica e definita come il prodotto tra la trasposta della Matrice Indicatrice e la matrice stessa: $B = Z^T Z$. Per come è costruita, risulta quindi che le Inertiae principali di B siano i quadrati di quelle di Z; ecco perché le coordinate standard sono le stesse in entrambi i metodi, mentre le coordinate principali differiscono: utilizzando la matrice di Burt, queste sono pari alle coordinate standard moltiplicate per le radici quadrate delle Inertiae principali. L'Inertia della Matrice di Burt è pari alla media delle Inertiae di tutte le sotto-matrici: le matrici sulla diagonale principale rappresentano le cross-tabulazioni di ogni variabile con sé stessa e dato che le loro Inertiae sono definite come nella Matrice Indicatrice, vengono anche calcolate allo stesso modo (sono esattamente pari al numero di variabili meno uno) e risultano più elevate delle Inertiae delle matrici fuori dalla diagonale. Di conseguenza, la stima dell'Inertia complessiva di B viene artificialmente "gonfiata", mentre quelle principali vengono ridotte, introducendo un termine di errore.

3.1.5 Joint Correspondence Analysis

Una soluzione è fornita dalla Joint Correspondence Analysis (JCA): si tratta di un algoritmo iterativo che permette di massimizzare l'Inertia spiegata, imponendo a priori il numero di dimensioni da considerare. Per "liberarsi" delle matrici

[7] Cap. 18, pag. 140

sulla diagonale di B, esse vengono inizialmente trattate come valori mancanti per massimizzare l'adattamento alle Inertiae fuori dalla diagonale: partendo dalla soluzione MCA, i valori delle sotto-matrici vengono sostituiti da delle stime ottenute tramite imputazione dei valori mancanti e si ottiene così una matrice *Modificata* di Burt. A questa viene nuovamente applicata la CA per avere una nuova soluzione, da cui ottenere un'altra matrice modificata. Il processo viene ripetuto fino a convergenza e ad ogni iterazione l'adattamento migliora.

La letteratura resta ancora molto divisa nella scelta del metodo migliore da utilizzare: oramai è appurato che la MCA "semplice" produce una forte distorsione nella stima dell'Inertia e pertanto, non è consigliato procedere con essa. D'altro canto però, sono stati proposti diversi miglioramenti dell'algoritmo, ognuno con pro e contro. In questo studio si è deciso di utilizzare, per scopi diversi, i due principali: quella che potremmo definire *Adjusted* MCA è stata utilizzata sui primi 1508 dati (per ogni titolo) in modo da definire una volta per tutte il numero di dimensioni da utilizzare nelle previsioni; per le stime rolling è stata invece utilizzata la metodologia Joint CA. Nella Adjusted MCA lo scopo è ridefinire l'Inertia totale della matrice di Burt, solo per le sotto-matrici fuori dalla diagonale ed "aggiustare" le Inertiae principali della soluzione MCA standard: per come vengono calcolate le Inertiae e per come è costruita B, l'Inertia media delle matrici fuori dalla diagonale è pari a $\frac{Q}{Q-1}(\phi^2(B) - \frac{J-Q}{Q})$. Supponendo che la generica Inertia principale (autovalore) della MCA su B sia λ_k, quella "aggiustata" viene calcolata come: $\lambda_k^{Adj} = (\frac{Q}{Q-1})^2 \times (\sqrt{\lambda_k} - \frac{1}{Q})^2$.
Il motivo per il quale sono stati utilizzati entrambi è presto spiegato: dato che è difficile stabilire a priori il numero di dimensioni ottimale, nella prima fase è stato utilizzato il metodo che «... è ritenuto essere la migliore scelta di default»[8] per questo scopo, poiché valuta tutte le dimensioni possibili e considera come significative solo quelle a cui è associato un autovalore maggiore del reciproco del numero di variabili[9]; nella seconda parte, è stato utilizzato il metodo JCA perché massimizza la quota d'Inertia, per un determinato numero di dimensioni.

[8] Nenadic e Greenacre 2007.
[9] Le decisioni sono state prese seguendo lo studio Ben Ammou e Saporta 2003.

3.2 Variabili Quantitative

Le variabili numeriche sono state trattate attraverso tecniche che hanno lo scopo di identificare un certo numero di variabili latenti e che producono pertanto una riduzione della dimensionalità. A differenza del caso delle variabili categoriali, i due metodi utilizzati sono diversi, ma legati l'uno all'altro: il primo è il metodo delle Componenti Principali (PCA), ampiamente noto in letteratura e facente parte, come la CA, della branca della statistica multivariata nota col nome di Analisi Fattoriale, il secondo è quello delle Componenti Indipendenti (ICA).

3.2.1 Analisi delle Componenti Principali

Nella PCA le nuove variabili sono ottenute attraverso una combinazione lineare di quelle iniziali e vengono proiettate su diversi assi del nuovo sistema cartesiano, in ordine decrescente di varianza (dalla più esplicativa, alla meno). Tutto ciò viene fatto dopo aver standardizzato le variabili iniziali (rendendole a media nulla e varianza unitaria) e permette di ottenere come risultato i cosiddetti *Scores*. Essi hanno varianza pari al rispettivo autovalore della matrice di covarianza delle variabili iniziali e sono incorrelati tra di loro. Nell'Appendice B viene riportata l'analisi svolta sui *Loadings*, ovvero sui coefficienti che permettono di passare dalle variabili originali agli scores.

3.2.2 Analisi delle Componenti Indipendenti

Dato che l'incorrelazione non implica sempre indipendenza, si è deciso di sfruttare gli Scores per valutare se ciò abbia anche un effetto sulla bontà delle previsioni: la ICA ha una forma del tutto simile alla precedente e ai modelli fattoriali, ma in questo caso le componenti utilizzate (chiamate *Sorgenti* o *Segnali*) sono tra di loro indipendenti, come suggerisce il nome. In genere è richiesto uno sbiancamento dei dati, attraverso centramento rispetto alla media, ma trattandosi di Scores ottenuti con PCA, essi hanno già media nulla e pertanto si può procedere immediatamente con la stima delle matrice di *Mixing*, che renda le componendi indipendenti. Un'importante assunzione della ICA è la non-gaussianità delle Sorgenti, poiché da due variabili gaussiane incorrelate non si otterrebbe alcuna informazione sulle direzioni delle colonne della matrice di mixing, che viene

valutata tramite la negentropia nell'algoritmo *fastICA*. Esso venne sviluppato nello studio Hyvärinen e Oja 2000 e risulta tutt'oggi uno degli algoritmi più utilizzati, pertanto se n'è fatto uso anche in questo lavoro.

3.3 Risultati delle analisi

Con riferimento alle prime 1508 osservazioni delle variabili quantitative, è stato scelto per ogni titolo il numero di componenti dei due metodi in base alla quota cumulata di varianza spiegata: nel caso della PCA si è deciso di scegliere il 50% e utilizzare il numero che permetteva di avvicinarsi più a tale soglia (il minimo è stato 8, mentre il massimo 13, come si può vedere dal grafico in Figura 3.1).

Per quanto riguarda l'ICA, per ogni titolo sono state valutate diverse opportunità: la scelta è stata fatta graficamente, caso per caso, optando per il numero minimo di componenti tra diverse possibilità, in base alla necessità di utilizzare più o meno segnali, condizionata all'andamento della varianza cumulata. Il titolo Apple è mostrato in Figura 3.2, gli altri vengono riportati nell'Appendice C.

Passando alle variabili qualitative, in Figura 3.3 sono mostrate le Inertiae delle prime cinque dimensioni delle prime 1508 osservazioni di tutti i titoli: non tutte sono state utilizzate per l'analisi successiva, perché mai si è verificato che più di tre fossero significative; in particolare, nel grafico in Figura 3.4 vengono indicati esattamente, per ogni titolo, quante dimensioni sono state considerate. Inoltre, si può notare come la prima dimensione sia quella che spiega la maggior parte dell'Inertia totale in quasi tutti i casi e come ciò vada di pari passo con la grandezza del corrispettivo autovalore.

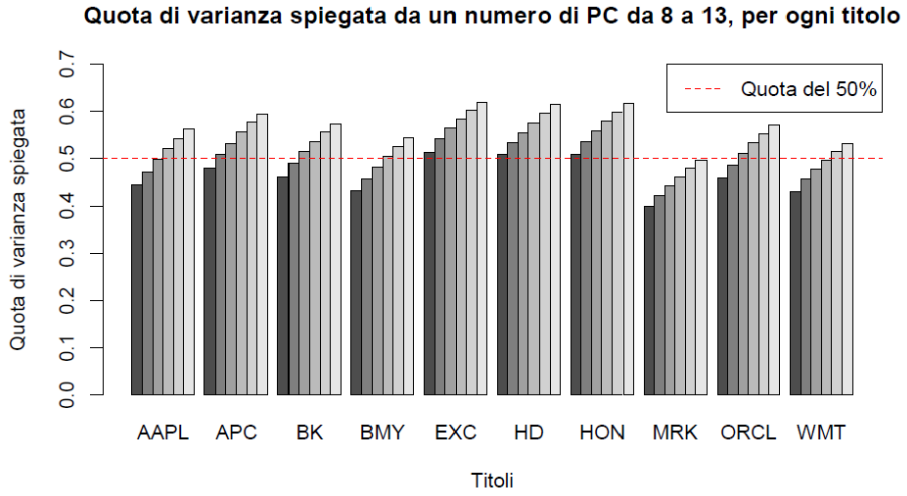

Figura 3.1: Quote della varianza spiegata dalle PC.

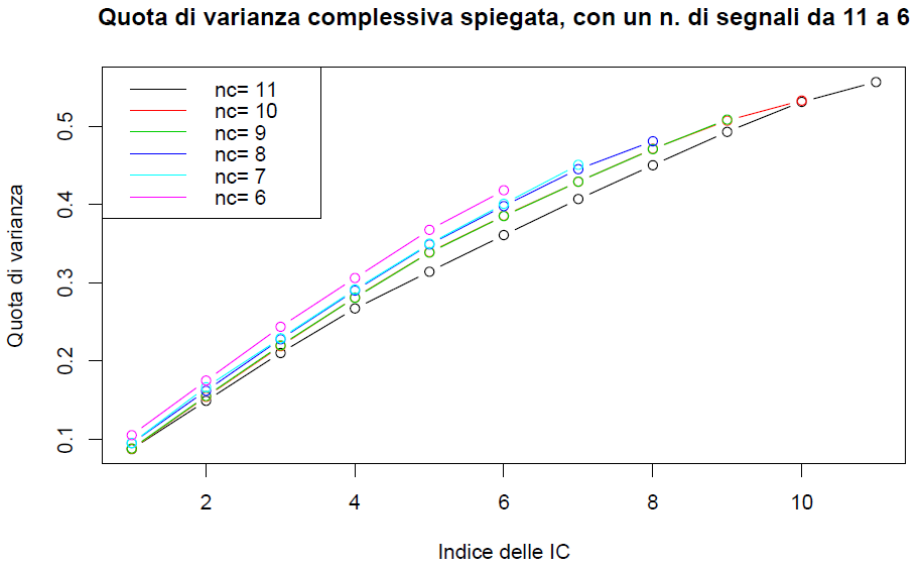

Figura 3.2: Quote della varianza spiegata dalle IC.

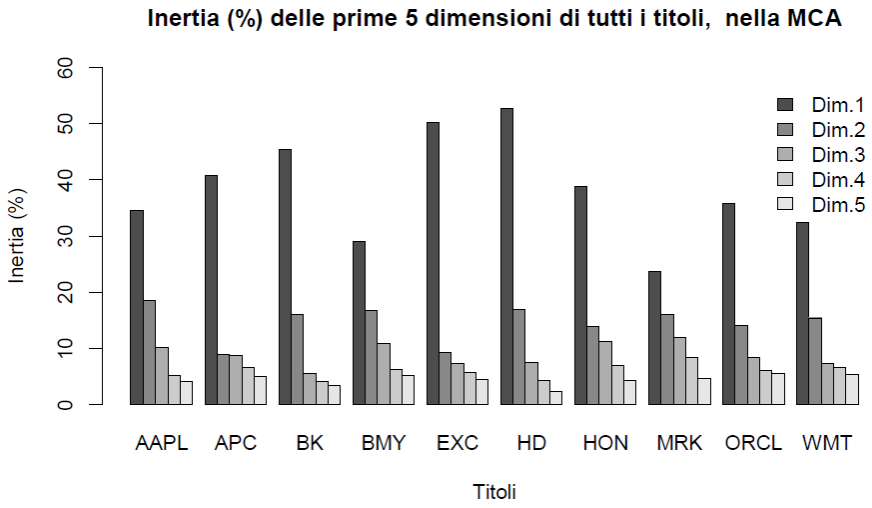

Figura 3.3: Inertia MCA.

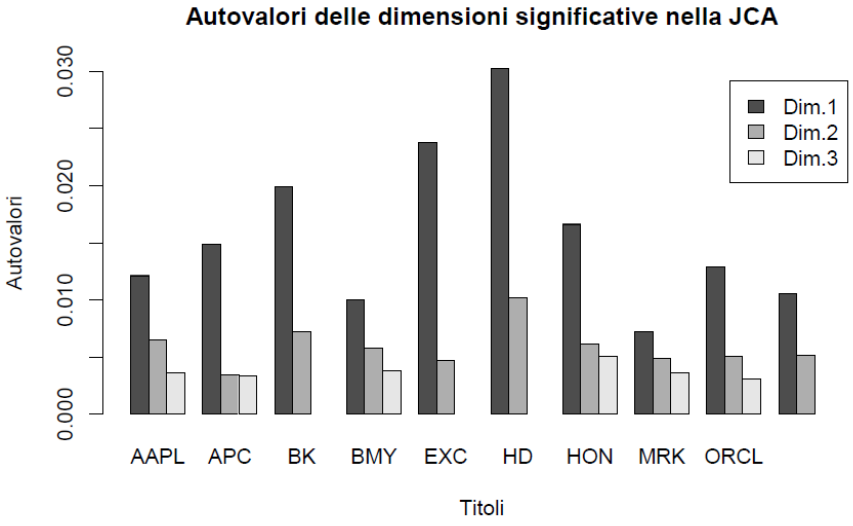

Figura 3.4: Autovalori JCA.

Capitolo 4

Analisi delle previsioni

Una volta definiti per ogni titolo il numero di componenti principali e indipendenti, nonché quello di dimensioni, sono state eseguite le stime step-rolling e le previsioni un passo in avanti. Considerando solo le ultime 1023 osservazioni, sono state individuate le date corrispondenti all'ultimo giorno di ogni mese e i valori da inizio periodo fino a questi punti utilizzati per definire scores, segnali e coordinate principali con cui fare le previsioni di tutto il mese successivo: i dati fino all'ultimo giorno del mese precedente hanno formato l'insieme di stima per tre diversi modelli lineari e quelli del giorno successivo sono stati usati per fare la previsione. Nel passo seguente, la stima veniva fatta fino al giorno in cui in precedenza si era fatta la previsione, mentre la previsione con i dati del giorno dopo. L'operazione è stata ripetuta fino a fine mese (e in quel punto sono stati ristimati scores, segnali e sorgenti) e poi per tutti i mesi fino alla fine del periodo di studio. Ciò ha prodotto 1005 valori previsti della volatilità realizzzata[1], poi confrontati con quelli realmente osservati.

4.1 I modelli

Nello studio si è fatto uso di modelli lineari con standard error robusti, in quanto le stime sono state ottenute minimizzando la funzione *Bisquare* di Tukey che

[1] Il primo mese delle ultime 1023 osservazioni è stato considerato parte dell'insieme di stima per la prima previsione.

garantisce più robustezza rispetto alla funzione di perdita di Huber[2]. Come riferimento è stato utilizzato il modello con i soli ritardi della volatilità (e il jump) tra le variabili esplicative; gli altri due hanno in comune la presenza delle coordinate principali, ottenute dalle variabili categoriali e si differenziano per la presenza degli scores o delle sorgenti:

1. **Modello coi soli ritardi**:

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 C_{t-5} + \beta_3 C_{t-22} + \beta_4 J_{t-1} + \epsilon_t;$$

2. **Modello con PC**:

$$RV_t = \delta_0 + \delta_1 RV_{t-1} + \delta_2 C_{t-5} + \delta_3 C_{t-22} + \delta_4 J_{t-1} + \delta_{pca}^T S_{pca,t-1} + \delta_{ca}^T COORD_{t-1} + \xi_t;$$

3. **Modello con IC**:

$$RV_t = \theta_0 + \theta_1 RV_{t-1} + \theta_2 C_{t-5} + \theta_3 C_{t-22} + \theta_4 J_{t-1} + \theta_{ica}^T S_{ica,t-1} + \theta_{ca}^T COORD_{t-1} + \zeta_t.$$

$S_{pca,t-1}$ indica la matrice degli Scores, $S_{ica,t-1}$ quella delle Sorgenti, $COORD_{t-1}$ quella delle Coordinate Principali e i vari β, δ e θ i rispettivi coefficienti associati.

4.2 Le previsioni

Un primo confronto tra valori previsti e osservati è stato fatto graficamente: singolarmente per ogni titolo, ma congiuntamente per i tre modelli, sono stati prodotti dei grafici a dispersione e ai vari punti è stato assegnato un colore in base al fatto che la previsione sovra- o sotto-stimasse il valore osservato sul mercato. L'esempio del titolo Apple è riportato in Figura 4.1, mentre i restanti grafici sono riportati in Appendice D.

Si nota subito che è caratteristica comune a tutti e tre i modelli il fatto che le previsioni sottostimino, in prevalenza, i valori reali indipendentemente dal loro valore assoluto: infatti, sia nel caso di valori elevati di volatilità, sia in caso di valori più piccoli, il valore previsto risulta minore dell'osservazione con frequenza

[2]Come mostrato in Stuart 2011.

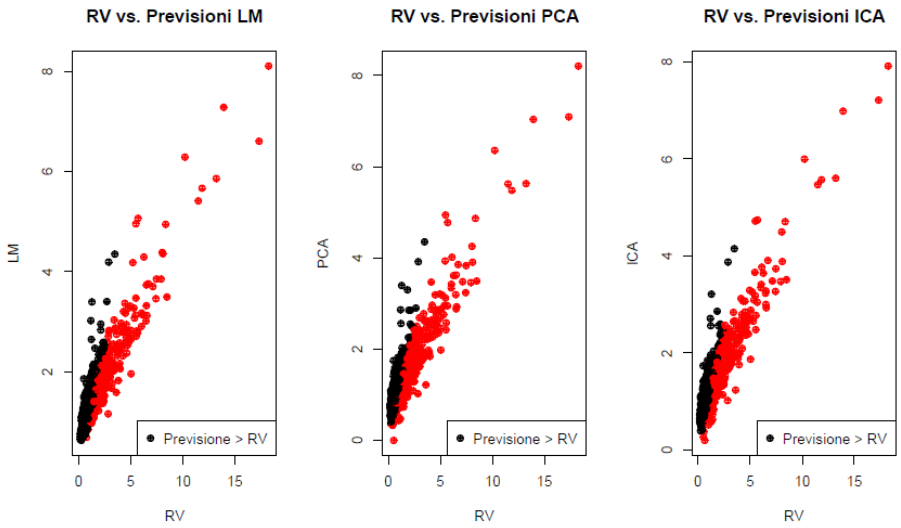

Figura 4.1: Confronto tra valori previsti e osservati del titolo Apple.

maggiore. Ciò si può vedere anche osservando nel dettaglio i valori "estremi": in alcuni casi, le previsioni hanno valori che sono la metà, un terzo, o anche meno di quelli effettivi. Insomma, quando si tratta di periodi a volatilità più elevata, raramente le previsioni sono effettivamente confrontabili con quanto avviene nella realtà. Questo è un problema che affligge spesso i modelli lineari, poiché lo scopo principale per cui vengono utilizzati consiste nel descrivere l'andamento generale del fenomeno e non adattarsi a valori che possono essere considerati anomali[3].

4.3 I coefficienti

Oltre ad un'analisi delle previsioni, ne è stata svolta una anche sui coefficienti dei modelli stimati, con lo scopo di valutare se le tecniche di dimensionality reduction avessero o meno effetto nel modificarne i valori. Sono state valutate due "proprietà": la *Persistenza* e la *Significatività*.

[3] Una precisazione sul titolo Anadarko Petroleum: dalle previsioni è stato scartato il valore nella posizione 781, in quanto sarebbe stato confrontato con uno osservato durante un giorno particolare. Si tratta di inizio aprile 2014, quando la società dovette pagare una penale di 5 miliardi di USD, in seguito a una condanna in tribunale per "Danno Ambientale".

4.3.1 Rappresentazione della persistenza

La prima si rifà al concetto espresso nei capitoli precedenti, in cui è stato spiegato che una buona parte della variabilità della RV può essere catturata considerandone i valori ritardati (o loro trasformazioni) e la componente discontinua: per tutto il periodo di previsione, è stato valutato graficamente l'andamento del valore di un coefficiente "cumulato" che contenesse l'effetto di tutte queste variabili. Tale coefficiente è stato inteso come somma dell'intercetta con i coefficienti relativi al valore precedente, alla media della settimana e del mese precedenti e al jump. Questo è stato fatto per ogni titolo e per ognuno dei metodi utilizzati: in ogni grafico è presente una serie storica del coefficiente di persistenza ottenuto dalle stime dei coefficienti delle rispettive variabili nel modello lineare semplice, in quello in cui si è tenuto conto degli scores e in quello in cui sono state considerate le sorgenti.
In pratica, con riferimento ai modelli della Sezione 4.1 si ha:

$$\beta_{persistenzaLM} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4; \tag{4.1}$$

$$\beta_{persistenzaPCA} = \delta_0 + \delta_1 + \delta_2 + \delta_3 + \delta_4; \tag{4.2}$$

$$\beta_{persistenzaICA} = \theta_0 + \theta_1 + \theta_2 + \theta_3 + \theta_4. \tag{4.3}$$

4.3.2 Rappresentazione della significatività

Il concetto di significatività è stato rappresentato graficamente in due maniere: alle serie storiche dei coefficienti descritte in precedenza sono state aggiunte altre due, in cui si è tenuto conto anche dei coefficienti di scores o sorgenti (a seconda del caso) e dimensioni, risultati significativi al 95% o 99% (ecco perché nei grafici relativi a ogni metodologia, le serie storiche ulteriori sono due). Seguendo lo stesso ragionamento fatto nella Sezione precedente e nella 4.1, le equazioni (4.2) e (4.3) possono essere riscritte con delle aggiunte:

$$\beta^{0.95}_{signif.PCA} = \beta_{persistenzaPCA} + {}_{0.95}\bar{\mathbf{I}}_{(1\times s)}\delta_{pca} + {}_{0.95}\bar{\mathbf{I}}_{(1\times nd)}\delta_{ca}; \quad (4.4)$$

$$\beta^{0.99}_{signif.PCA} = \beta_{persistenzaPCA} + {}_{0.99}\bar{\mathbf{I}}_{(1\times s)}\delta_{pca} + {}_{0.99}\bar{\mathbf{I}}_{(1\times nd)}\delta_{ca}; \quad (4.5)$$

$$\beta^{0.95}_{signif.ICA} = \beta_{persistenzaICA} + {}_{0.95}\bar{\mathbf{I}}_{(1\times nc)}\theta_{ica} + {}_{0.95}\bar{\mathbf{I}}_{(1\times nd)}\theta_{ca}; \quad (4.6)$$

$$\beta^{0.99}_{signif.ICA} = \beta_{persistenzaICA} + {}_{0.99}\bar{\mathbf{I}}_{(1\times nc)}\theta_{ica} + {}_{0.99}\bar{\mathbf{I}}_{(1\times nd)}\theta_{ca}. \quad (4.7)$$

Ogni $\mathbf{I}$ è un vettore riga di variabili indicatrici del fatto che il vettore colonna (post-moltiplicato) dei coefficienti contenga o meno un valore significativo al livello ricercato in ogni posizione: ${}_{Livello\ ricercato}\bar{\mathbf{I}}_{(1\times Numero\ di\ variabili\ considerate)}$.[4]

4.4 Persistenza dei coefficienti

Come già fatto in precedenza, viene mostrato in seguito solo il grafico di un titolo[5] per ogni metodologia, mentre gli altri vengono inseriti nell'Appendice E. Ogni figura è costituita da due grafici: quello superiore contiene le serie appena descritte, in quello inferiore vi sono rappresentate le serie ottenute contando il numero di coefficienti significativi ai due livelli, per ogni osservazione.

[4] nd è il numero di dimensione scelto con la correspondence analysis, s quello di scores o segnali ottenuto con le rispettive metodologie.

[5] Apple.

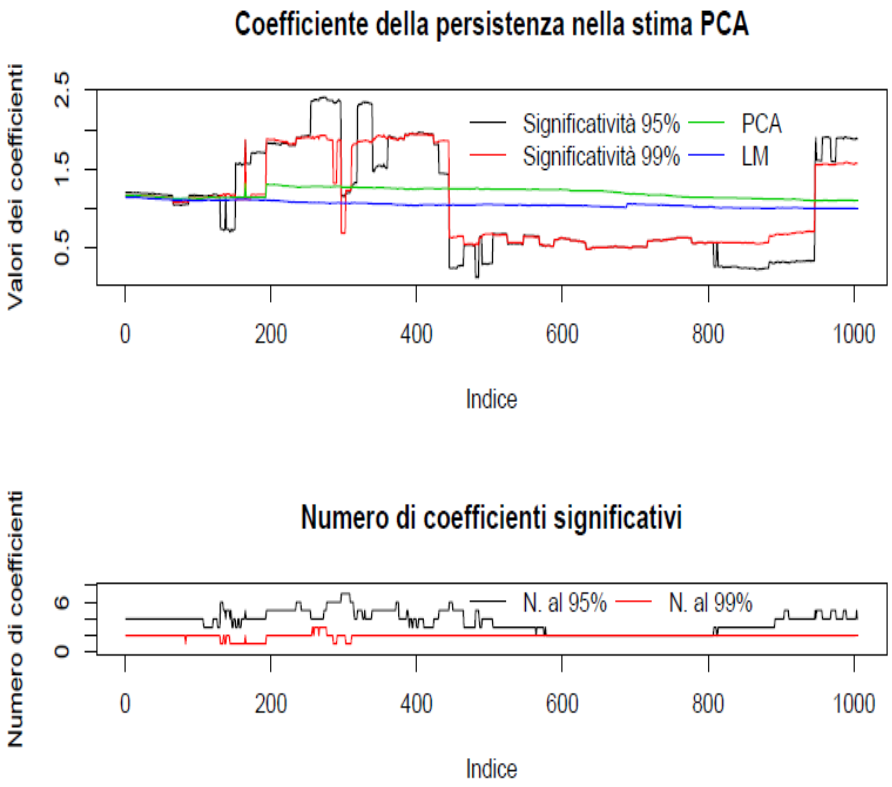

Figura 4.2: Grafico dei coefficienti delle PC.

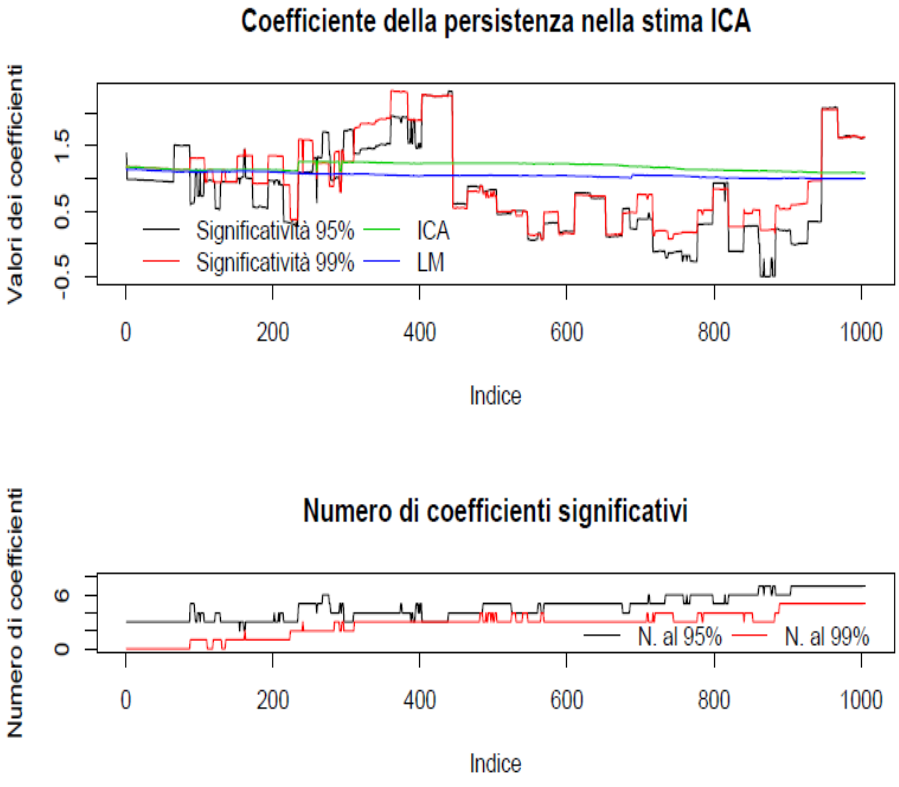

Figura 4.3: Grafico dei coefficienti delle IC.

L'aggiunta di ulteriori variabili ha avuto un effetto sull'andamento dei coefficienti che era prevedibile all'inizio: la stima "LM" risulta infatti molto più liscia e regolare e anche le serie delle sole stime "PCA" e "ICA" lo sono abbastanza. A questo punto però, non è ancora chiaro se l'andamento più oscillatorio e irregolare delle stime comprendenti anche i coefficienti significativi di PCA e ICA corrisponda a variazioni significative dal nostro *benchmark*, rappresentato dalle stime LM, né se effettivamente ciò abbia permesso di migliorare le previsioni. Questo verrà valutato nella sezione successiva.

Vale la pena spendere qualche commento anche sui grafici inferiori, riguardanti il numero di componenti significative: una caratteristica comune più o meno a tutti i casi è il fatto che siano sempre "poche" le variabili aggiuntive, ad avere un impatto significativo, anche solo al 95%. Infatti, si può notare come nei casi in cui la strategia abbia avuto più successo, sia stata selezionata solo la metà circa delle variabili a disposizione. D'altro canto, per diversi titoli non si è registrato alcun coefficiente significativo per buona parte parte del periodo (se non per tutto) e specialmente al 99%. A questo punto, l'apporto fornito dalle variabili news potrebbe essere interpretato più come un disturbo che qualcosa d'informativo sul fenomeno: purtroppo non è facile dire se ciò dipenda dalla struttura del dataset, dalle metodologie per le stime, o se semplicemente le variabili prese in considerazione non abbiano veramente alcun effetto sulla volatilità del mercato. Per capirlo, non resta che analizzare le performance dei modelli in quanto ad errori nelle previsioni.

4.5 Analisi degli errori

L'Errore Quadratico Medio (Mean Squared Error, MSE) è una delle statistiche che vengono usate solitamente per valutare la correttezza delle stime e confermare o smentire l'adeguatezza del modello scelto. La maggior parte di esse si basa sulla stessa quantità: gli errori di previsione. Essi rappresentano le differenze tra valori osservati e valori predetti, le quali vengono poi sommate, eventualmente considerandole elevate ad una certa potenza (la seconda, nel caso del MSE), in valore assoluto, in rapporto ai valori reali, o altro. In definitiva, viene considerata una funzione di perdita attesa. Dal punto di vista puramente economico, la

scelta dei "parametri" che definiscono tale funzione è fondamentale, perché i risultati devono essere interpretati al fine di fornire delle linee guida per prendere decisioni[6].

Lo stesso discorso può essere applicato nel caso in cui si desideri confrontare più modelli in base alla loro qualità predittiva: considerando un test statistico con ipotesi nulla l'equivalenza tra i modelli, è possibile definire una misura di discrepanza tra di essi come valore assoluto della differenza delle funzioni di perdita scelte e verificare che sia in media uguale a zero. Nello studio Diebold e Mariano 1995 viene fornito un test, molto utilizzato in econometria, che può essere applicato ad una vasta gamma di casi e che tiene conto anche dell'eventuale presenza di autocorrelazione nella serie degli errori. L'ipotesi nulla è di uguaglianza delle previsioni e come alternativa possono essere esplorate tutte le possibilità: sia bidirezionale, per concludere semplicemente che i due modelli non sono statisticamente uguali nel prevedere, sia unidirezionale, per valutare se uno dei due sia migliore dell'altro. In questo lavoro, dato che già in precedenza il modello lineare semplice è stato considerato il "benchmark", si è deciso di confrontarlo coi metodi contenenti le componenti principali e indipendenti, imponendo come ipotesi alternativa il fatto che questi fossero "meno accurati" del primo. I risultati dei test sono riportati nella Tabella 4.1, con in grassetto i p-value nei casi in cui l'ipotesi nulla è stata accettata al livello del 5%:

Senza contare il titolo Bank of New York, che di fatto è l'unico di un azienda propriamente "finanziaria" (quindi può aver risentito maggiormente della crisi iniziata nel 2007) e in cui la statistica test ha avuto valore positivo, H_0 è stata accettata solo in due occasioni: quando ciò accade, si può concludere che i due metodi hanno avuto la stessa capacità predittiva, mentre quando la si rifiuta, è possibile avere ulteriori informazioni sulla preferenza in base al segno della statistica. Siccome essa è costruita come differenza tra la funzione di perdita del modello lineare e quella del modello con PC/IC, se il suo segno è negativo, vuol dire che la seconda funzione di perdita è maggiore e quindi il primo modello è preferibile. Purtroppo, tali considerazioni evidenziano che i risultati non sono molto positivi: la superiorità delle prestazioni del modello lineare in quasi tutte le prove è indice del fatto che utilizzando le news in questo modo, esse forniscano

[6]Per esempio, si possono considerare anche funzioni di perdita asimmetriche, per trattare gli errori di stima positivi in maniera diversa rispetto a quelli negativi.

	PCA		ICA	
	Valore	p-value	Valore	p-value
AAPL	-3,31	$< 0,001$	-2,91	0,002
APC	-1,81	0,035	-2,42	0,008
BK	3,57	~ 1	3,68	~ 1
BMY	-4,50	$< 0,001$	-4,12	$< 0,001$
HD	-1,60	**0,055**	-2,35	0,01
HON	-2,24	0,01	-2,36	0,01
MRK	-3,31	$< 0,001$	-1,71	0,04
ORCL	-4,90	$< 0,001$	-4,99	$< 0,001$
WMT	-1,85	0,03	-1,27	**0,10**

Tabella 4.1: DM-test: valori delle statistiche e p-value.

un apporto minimo nell'interpretazione del fenomeno. Si può dire che rappresentino addirittura una sorta di *rumore*, il quale provoca un peggioramento nelle previsioni invece di un miglioramento.

4.5.1 Analisi grafica

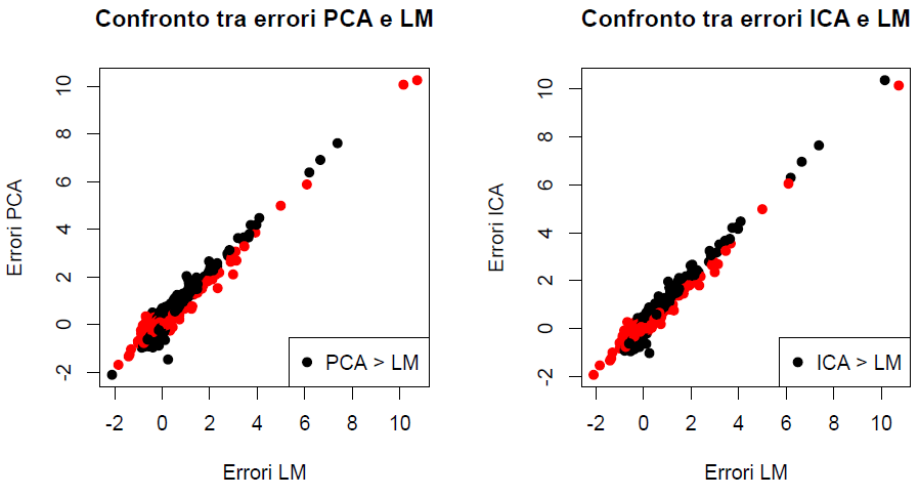

Figura 4.4: Confronto tra gli errori di previsione, per il titolo Apple.

L'analisi grafica svolta sugli errori e riassunta nella Figura 4.4 (i restanti grafici sono riportati nell'Appendice F) non fornisce informazioni aggiuntive:

i risultati dei test precedenti sono molto più informativi; anche perché all'apparenza si può notare una certa similarità tra dati dei vari modelli, la quale nasconde parte di quella che invece può essere considerata la conclusione dello studio: il modello lineare risulta migliore, nella sua semplicità.

Capitolo 5

Conclusioni

Lo scopo di questo studio era di valutare se esiste un'alternativa più sofisticata alla stima ai minimi quadrati standard, in ambito finanziario. In particolare, per prevedere l'andamento della Volatilità Realizzata. Essa è una delle misure di volatilità del mercato più utilizzate in letteratura, specialmente da quando è nata la concezione di Dati ad Alta ed Altissima Frequenza (*HF-* e *UHF-data*) e le strategie di trading hanno cominciato a svilupparsi con orizzonti temporali intragiornalieri. Un fatto noto grazie a diversi studi è la *persistenza* di questa variabile, pertanto sono sempre stati considerati alcuni ritardi di essa, nonché la componente discontinua denominata "Jump". Un primo miglioramento è stato apportato considerando un modello semplice, ma con standard error stimati in maniera robusta rispetto ai valori estremi, ai quali le stime classiche sono solitamente molto sensibili. Data l'enorme quantità d'informazioni testuali fornita (quasi) gratuitamente agli analisti finanziari, in seguito allo sviluppo di internet, dei social media e delle metodologie adatte per trattare dati del genere, si è pensato di poter apportare un ulteriore sviluppo attraverso l'utilizzo di un algoritmo per l'analisi del sentimento di news finanziarie e tecniche di dimensionality reduction applicate a grandi dataset contenenti anche tali variabili, per cercare di "concentrare" l'informazione. Ciò è stato fatto considerando i dati di 10 titoli dell'indice S&P 100 e suddividendo le variabili in qualitative e quantitative, in modo da poterle trattare in maniera adeguata alle loro caratteristiche. Due sono stati i modelli utilizzati come possibili alternative: nel primo le variabili quan-

titative sono state trattate tramite Analisi delle Componenti Principali (PCA), il cui numero è stato scelto caso per caso in base a quante fossero necessarie per avvicinarsi o andare oltre alla quota del 50% di varianza spiegata; nel secondo è stata utilizzata l'Analisi delle Componenti Indipendenti (ICA), applicata a tutte le nuove variabili prodotte dalla PCA, scegliendo quante considerarne in base a una valutazione grafica di quale fosse il numero che spiegava una buona parte della variabilità, ferma restando la necessità di avere un modello parsimonioso. In entrambi i casi, le variabili qualitative sono state inserite in seguito ai risultati ottenuti con l'Analisi delle Corrispondenze (CA): il numero di coordinate è stato definito in base a quanti autovalori della matrice di Burt fossero significativi, ottenendo così una certa quota di *Inertia* spiegata. Per ogni metodo, sono state fatte stime step-rolling a un passo e successive previsioni un passo in avanti. Per valutare quale modello fosse migliore, sono state fatte valutazioni sugli errori di previsioni tramite test di Diebold-Mariano, considerando come H_0 che i modelli con PC e IC avessero le stesse capacità predittive di quello lineare. I risultati portano a conclusioni abbastanza chiare: le variabili news producono quasi unicamente un *rumore* aggiuntivo alle stime invece di ridurlo.

I motivi per cui ciò accade possono essere molteplici, a cominciare dall'algoritmo stesso: esso è stato costruito ex-novo e sicuramente potrà e dovrà essere migliorato in futuro; magari partendo dalla ridefinizione delle liste di parole, visto che (come mostrato dalle misure di performance multiclasse utilizzate nell'analisi) le modifiche a quella di parole negative non hanno apportato miglioramenti considerevoli. Inoltre, va tenuto conto che le variabili news possono essere viste come dei "disturbi" per l'andamento dei titoli: esse rappresentano un flusso di informazioni fornito all'improvviso e in grado di incrinare o rafforzare le convinzioni degli investitori, perciò una trattazione *Event Study* potrebbe risultare più appropriata.

Per concludere, può essere che una o più delle scelte che, inevitabilmente, è stato necessario fare, abbiano distorto i risultati e annullato i benefici informativi contenuti nelle news: questo può accadere quando si ha a che fare con la riduzione della dimensionalità, nel senso che l'eccessiva parsimoniosità ricercata può essere parallelamente causa di una riduzione del contenuto informativo, fornito dalle variabile aggiuntive. Il riferimento è diretto principalmente ai metodi

per le variabili quantitative: la scelta di una soglia per la varianza spiegata può aver causato i problemi sopra citati, perché troppo bassa per risultare utile a spiegare la variabilità del dataset. Per quanto riguarda le variabili categoriali, forse il problema è di minore entità: nella scelta delle coordinate ci si è basati sui risultati teorici disponibili per la CA, i quali sono comunque molto incerti su alcune questioni; tuttavia, le quote di Inertiae spiegate sono risultate molto buone, nonostante una drastica riduzione della dimensionalità, a testimonianza che il metodo è consistente. Perciò una possibilità sarebbe considerare un numero maggiore di coordinate anche se i corrispondenti autovalori non sono significativi, o lo sono ad un livello inferiore: si è visto infatti, che già solo 5 coordinate fornivano, per tutti i titoli, quote ancora più elevate di Inertia spiegata. Così facendo, è vero che si andrebbe "contro" le nozioni teoriche su cui ci si è basati per tutto il lavoro, ma probabilmente si avrebbero risultati migliori. Per le variabili numeriche una soluzione potrebbe essere fornita dai Metodi di Regolarizzazione come il Lasso, o in generale le Reti Elastiche, oltre che a quelli non parametrici nell'ambito del Machine Learning: dopo aver definito le componenti principali (e poi quelle indipendenti nella seconda delle trattazioni) e degli opportuni parametri di regolarizzazione, si potrebbe applicare ad esse uno dei metodi citati, in modo da eliminare quelle non significative e valutare i nuovi risultati sia dal punto di vista grafico, che tramite test statistico.

Bibliografia

Abdi, H. e D. Valentin (2007). «Multiple Correspondence Analysis». In: Program in Cognition and Neurosciences, The University of Texas at Dallas.

Andersen Torben, G. e T. Bollerslev (1998a). «Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts». In: *International Economic Review* 39.4, pp. 885–905.

— (1998b). «Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements and Longer Run Dependencies». In: *The Journal of Finance* 3.1, pp. 219–265.

Andersen Torben, G., T. Bollerslev e X. Diebold Francis (2007). «Roughing it up: Including Jump Components in the Measurement, Modelling and Forecasting of Return Volatility». In: *Review of Economics and Statistics* 89.4, pp. 701–720.

Antweiler, W. e Frank Z. Murray (2004). «Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards». In: 59.3, pp. 1259–1294.

Azzalini, A., B. Scarpa e G. Walton (2012). *Data Analysis and Data Mining. An Introduction*. Oxford University Press, USA.

Barndorff-Nielsen Ole, E. e N. Shephard (2004). «Power and Bipower Variation with Stochastic Volatility and Jumps». In: *The Journal of Financial Econometrics* 2.1, pp. 1–37.

Ben Ammou, S. e G. Saporta (2003). «On the connection between the distribution of eigenvalues in Multiple Correspondence Analysis and Log-Linear Models». In: *CARME2003*. Correspondence Analysis and Related Methods, Barcelone, 29 juin- 2 juillet 2003. X, France.

Camiz, Sergio e Gastão Coelho Gomes (2013). «Joint Correspondence Analysis Versus Multiple Correspondence Analysis: A Solution to an Undetected Problem». In: *Classification and Data Mining*. A cura di Antonio Giusti, Gunter Ritter e Maurizio Vichi. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 11–18.

Ceron, A., L. Curini e S.M. Iacus (2014). *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. SxI - Springer for Innovation. Springer Milan.

Corsi, F. (2009). «A Simple Approximate Long-Memory Model of Realized Volatility». In: *The Journal of Financial Econometrics* 7.2, pp. 174–196.

Dave, Kushal, Steve Lawrence e David M. Pennock (2003). «Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews». In: *Proceedings of the 12th International Conference on World Wide Web*. WWW '03. New York, NY, USA: ACM, pp. 519–528.

Diebold Francis, X. e S. Mariano Roberto (1995). «Comparing Predictive Accuracy». In: *The Journal of Business and Economic Statistics* 13.3, pp. 253–263.

Fama, Eugene F. (1970). «Efficient Capital Markets: A Review of the Theory and Empirical Work». In: 25.2, pp. 383–417.

Greenacre, M. (2016). *Correspondence Analysis in Practice*. Chapman & Hall/-CRC Interdisciplinary Statistics. Taylor & Francis Group.

Hyvärinen, A. e E. Oja (2000). «Independent component analysis: algorithms and applications». In: 13.4 and 13.5, pp. 411–430.

Loughran, T. e B. McDonald (2011). «When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks». In: *The Journal of Finance* 66.1, pp. 35–65.

Nenadic, Oleg e Michael Greenacre (2007). «Correspondence analysis in R, with two-and three-dimensional graphics: The ca package». In: *Journal of Statistical Software* 20.3.

Sokolova, Marina e Guy Lapalme (2009). «A systematic analysis of performance measures for classification tasks». In: *Information Processing & Management* 45.4, pp. 427–437.

Stuart, C. (2011). «Robust regression». In: *Department of Mathematical Sciences, Durham University* 169.

Tetlock, Paul C. (2007). «Giving Content to Investor Sentiment: The Role of Media in the Stock Market». In: 62.3, pp. 1139–1168.

Tetlock, Paul C., M. Saar-Tsechansky e S. Macskassy (2008). «More than Words: Quantifying Language to Measure Firms' Fundamentals». In: 63.3, pp. 1437–1467.

Appendice A

Titoli selezionati

Titoli	**Simboli**
Apple	AAPL
Anadarko Petroleum Corporation	APC
The Bank of New York Mellon Corporation	BK
Bristol-Myers Squibb Company	BMY
Exelon Corporation	EXC
The Home Depot, Inc.	HD
Honeywell International Inc.	HON
Merck & Co. Inc.	MRK
Oracle Corporation	ORCL
Wal-Mart Stores Inc.	WMT

Tabella A.1: Titoli selezionati.

Appendice B

Loadings della PCA

Un fatto interessante da notare è che i coefficenti delle prime componenti tendono, in quasi ogni caso, ad avere tutti un segno concorde, mentre quelli delle altre sono più centrati intorno allo zero. Ciò dimostra l'importanza delle prime componenti nello spiegare la varianza totale del fenomeno.

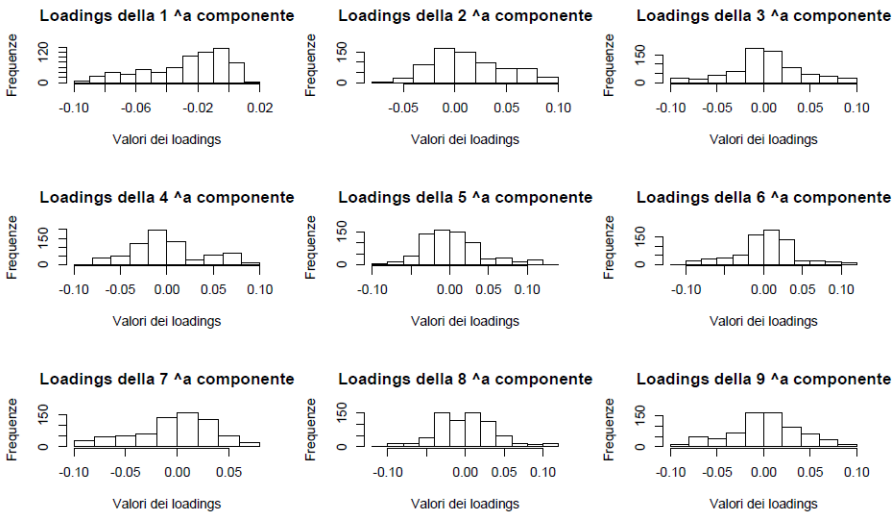

Figura B.1: Apple.

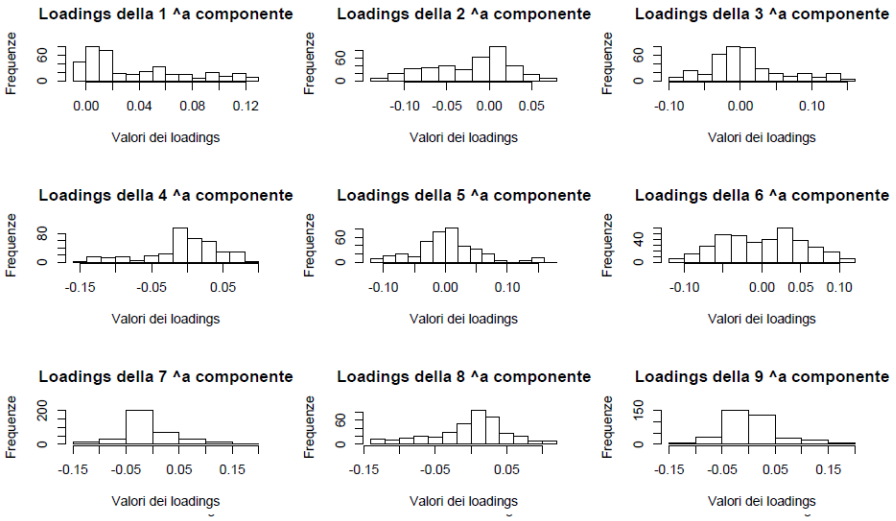

Figura B.2: Anadarko.

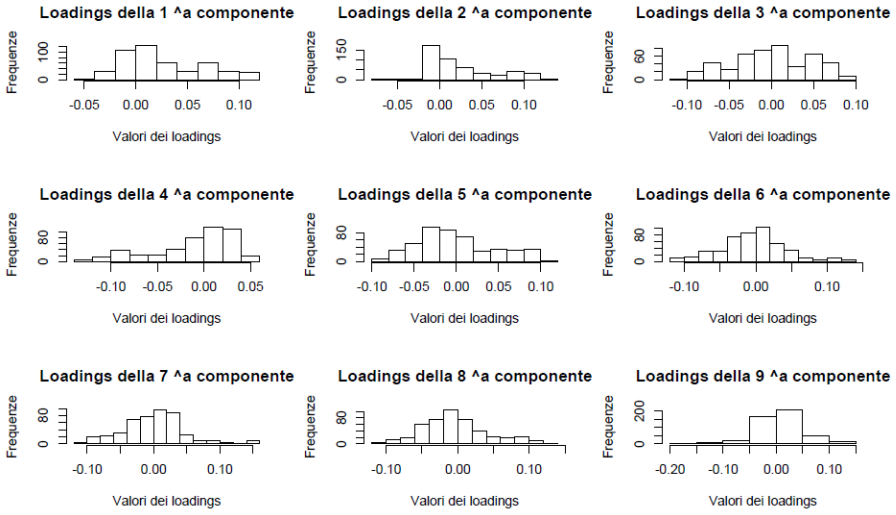

Figura B.3: Bank of NY.

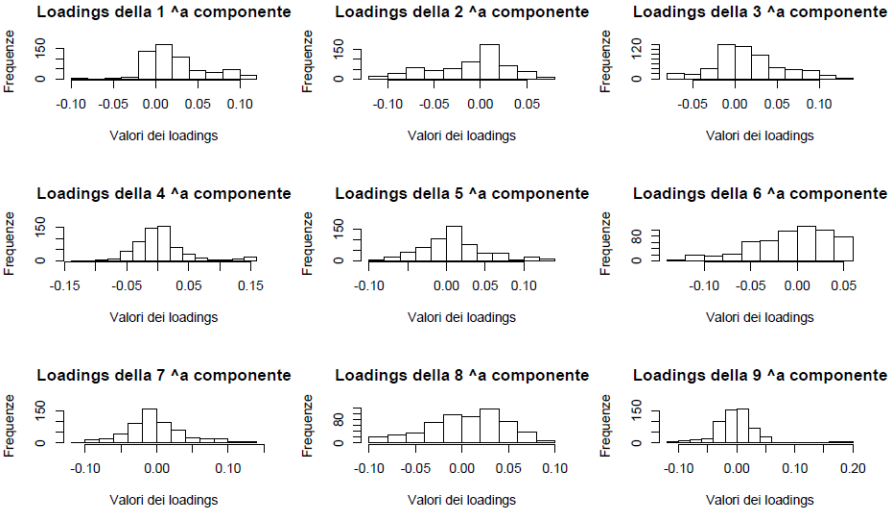

Figura B.4: Bristol-Meyers.

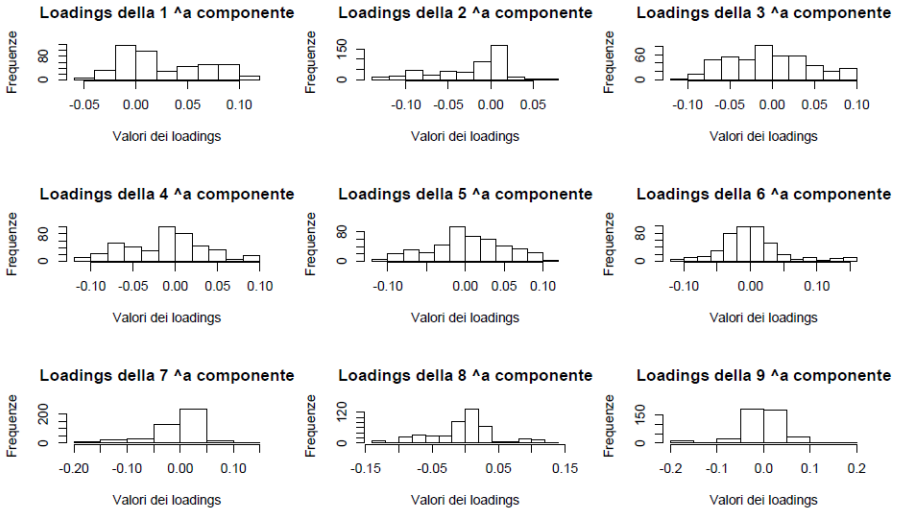

Figura B.5: Exelon.

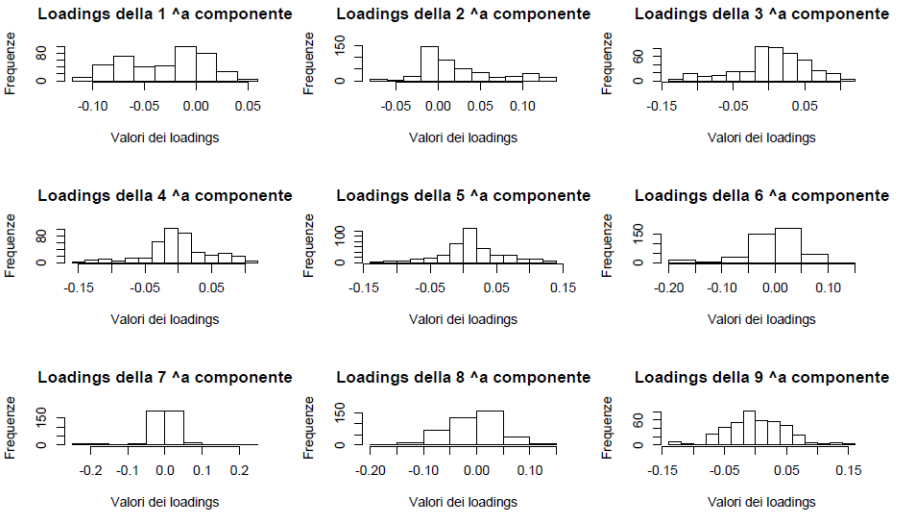

Figura B.6: Home Depot.

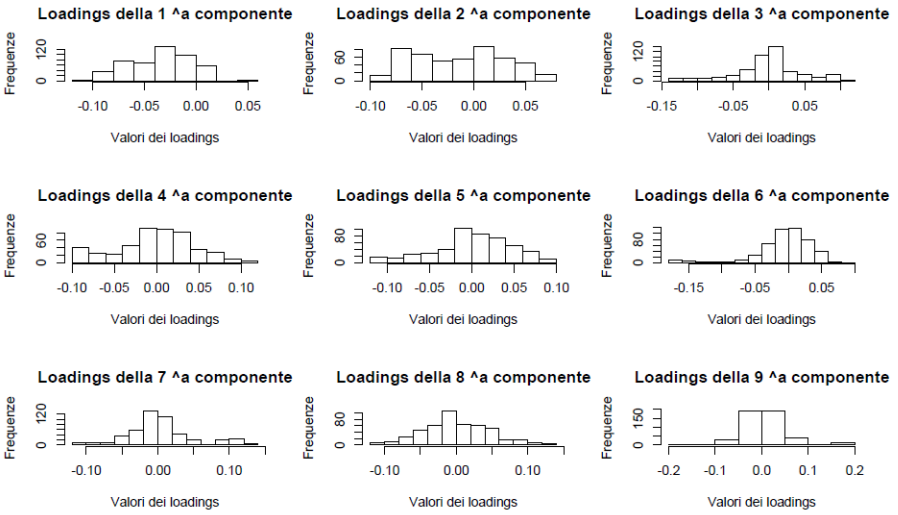

Figura B.7: Honeywell.

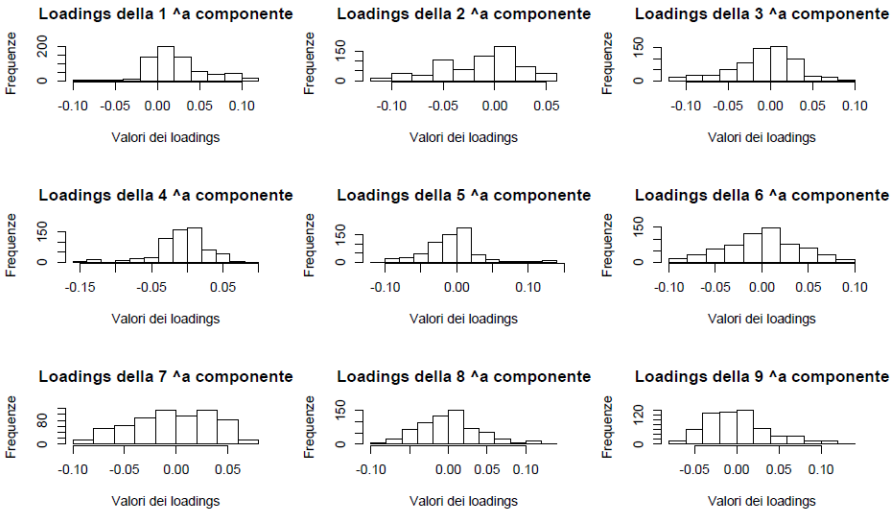

Figura B.8: Merk.

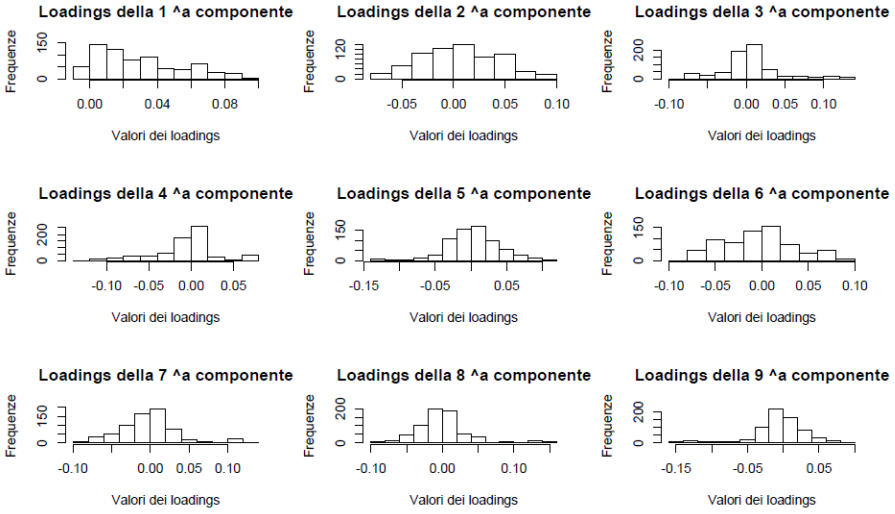

Figura B.9: Oracle.

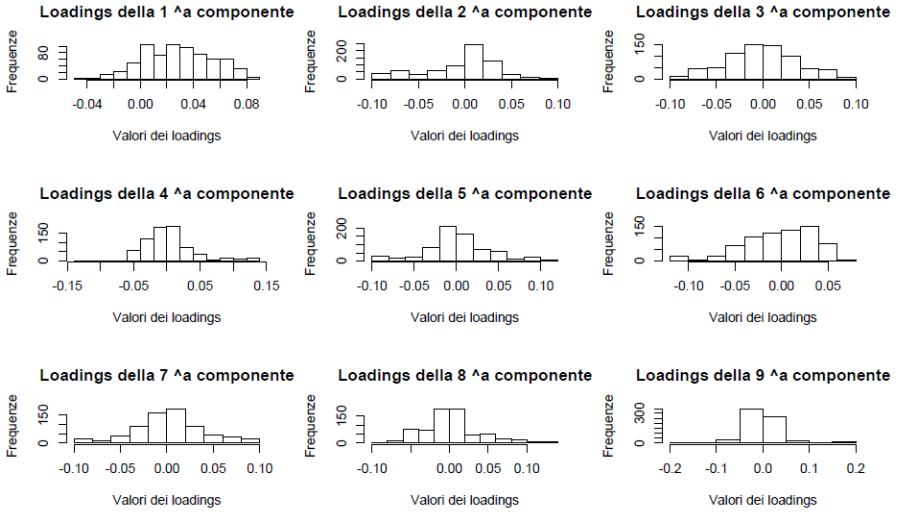

Figura B.10: Wal-Mart.

Appendice C

Varianza spiegata dalle IC

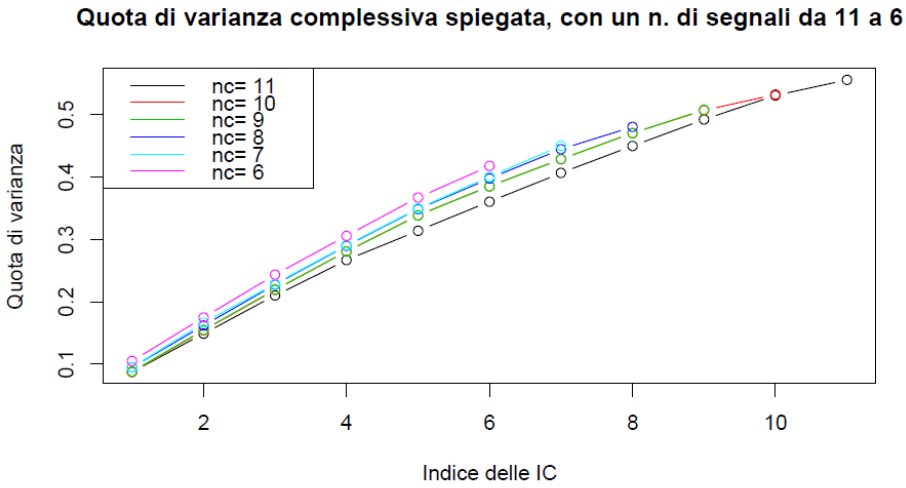

Figura C.1: Anadarko.

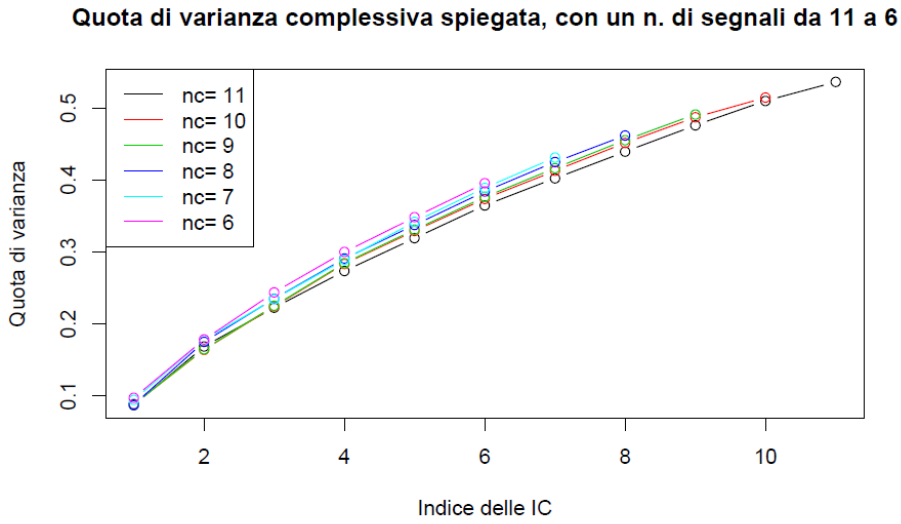

Figura C.2: Bank of NY.

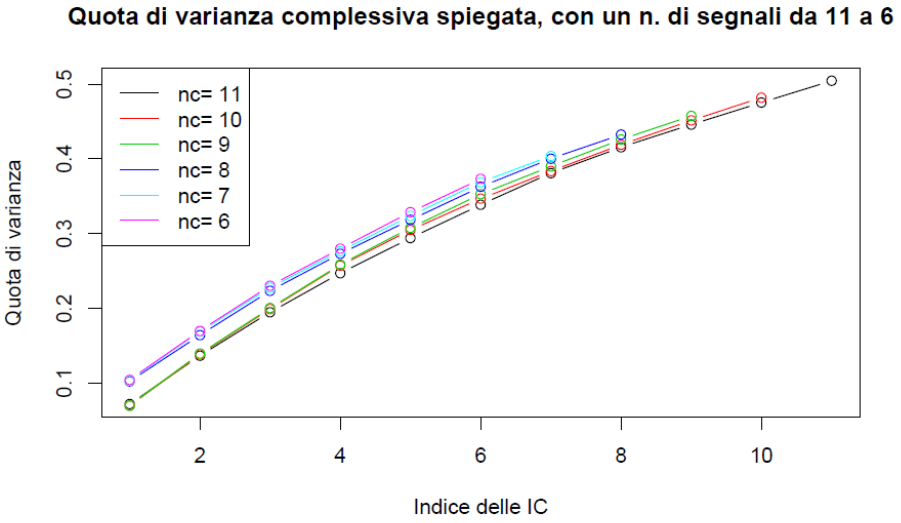

Figura C.3: Bristol-Meyers.

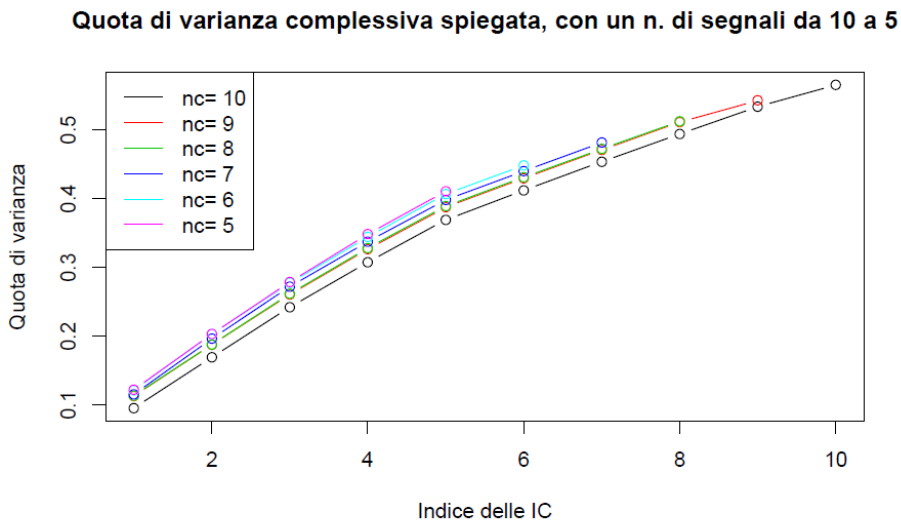

Figura C.4: Exelon.

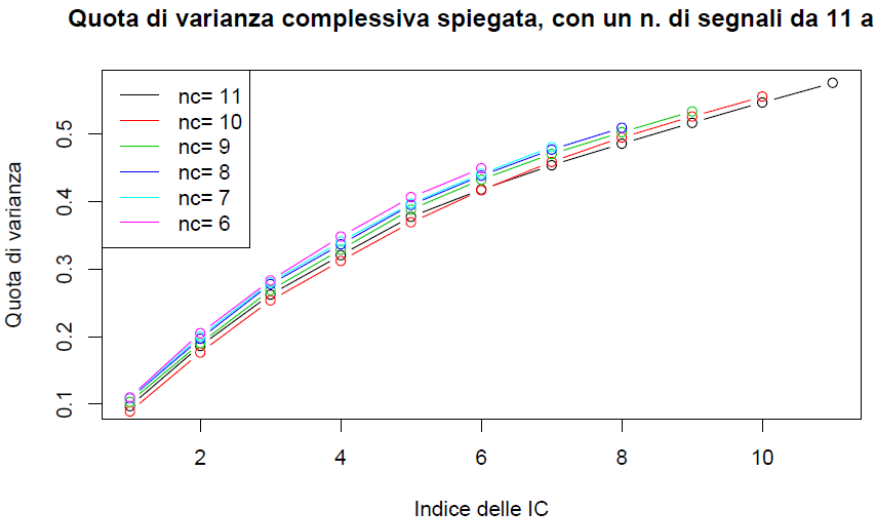

Figura C.5: Home Depot.

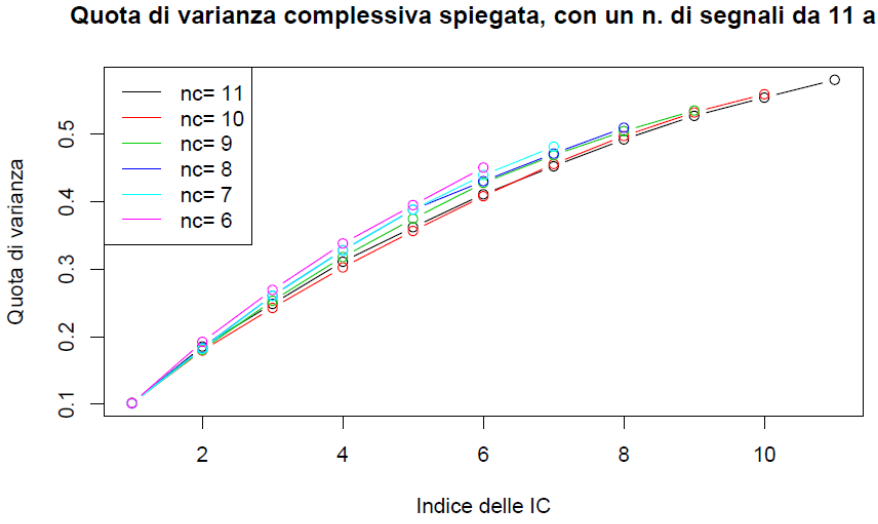

Figura C.6: Honeywell.

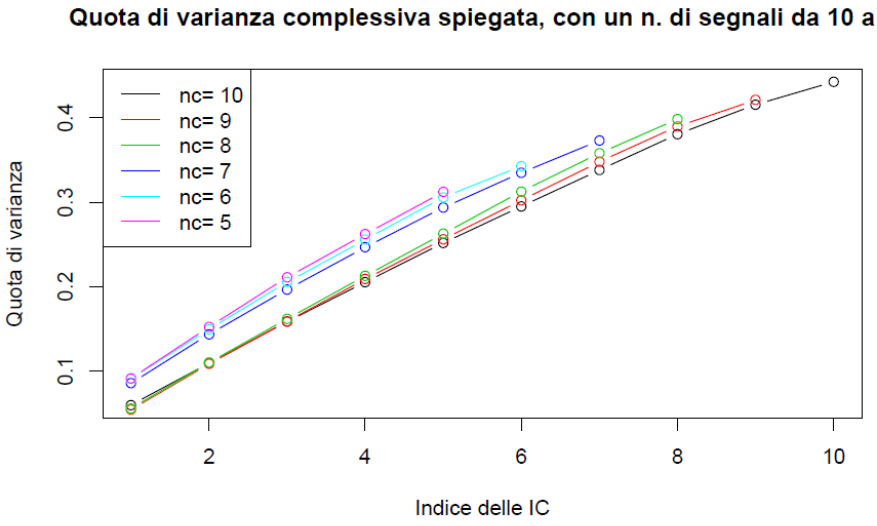

Figura C.7: Merk.

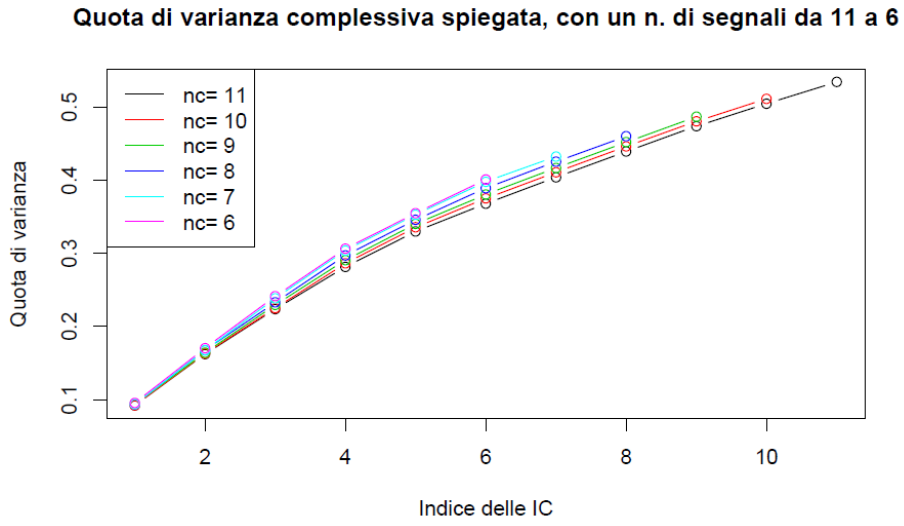

Figura C.8: Oracle.

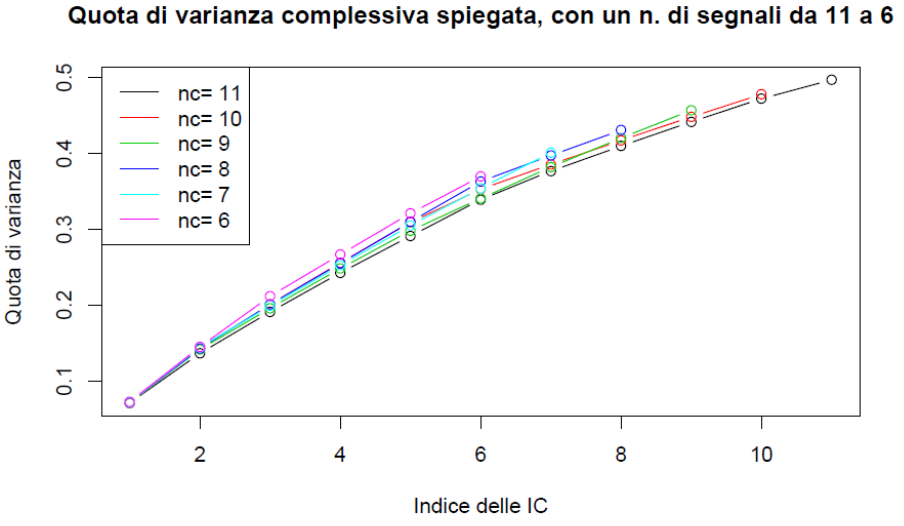

Figura C.9: Wal-Mart.

Appendice D

Confronto tra previsioni e valori osservati

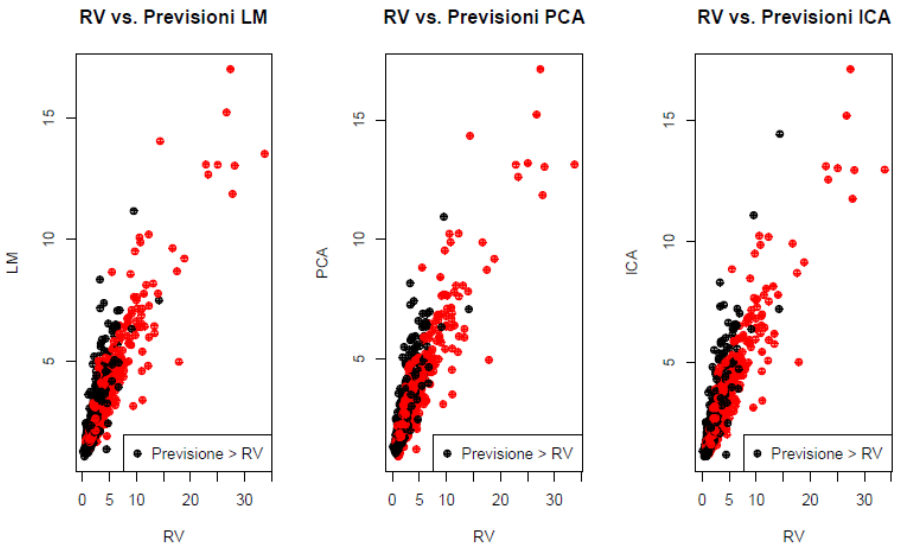

Figura D.1: Anadarko.

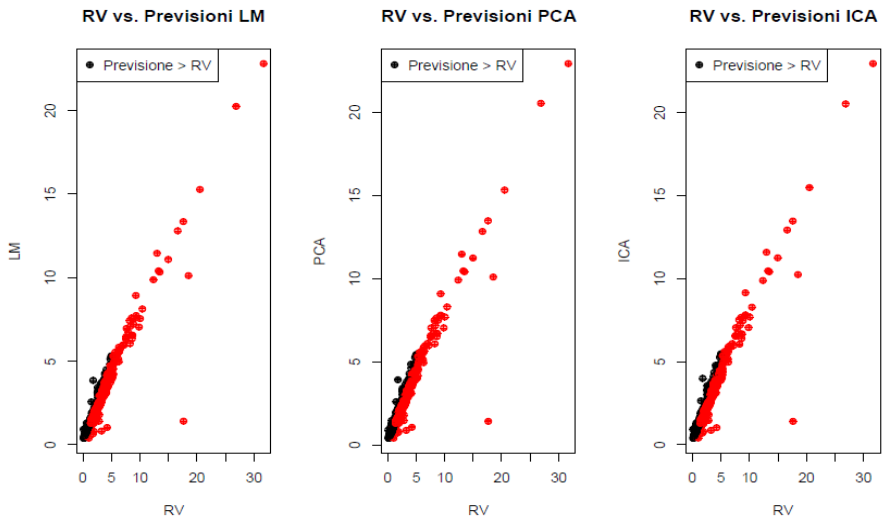

Figura D.2: Bank of NY.

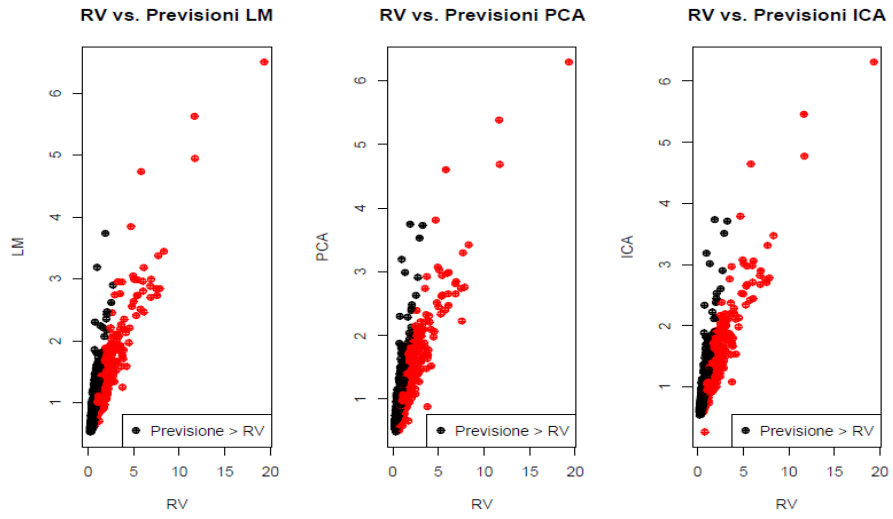

Figura D.3: Bristol-Myers.

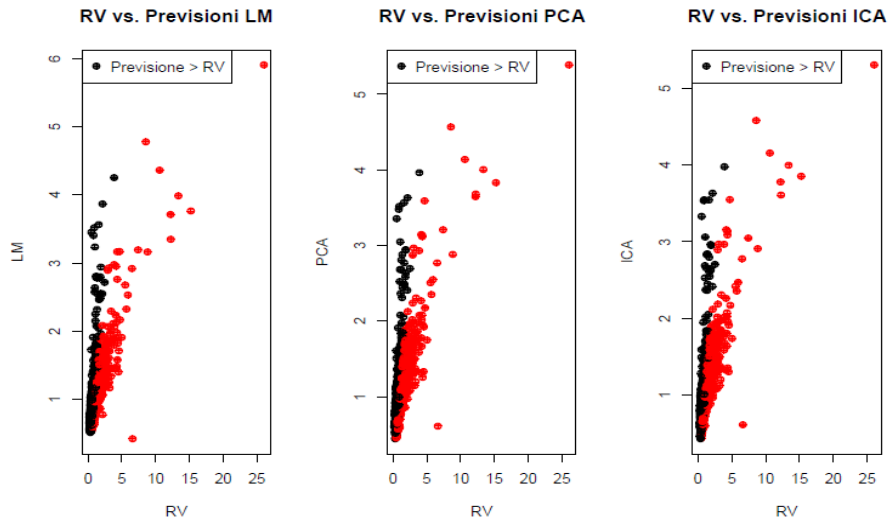

Figura D.4: Exelon.

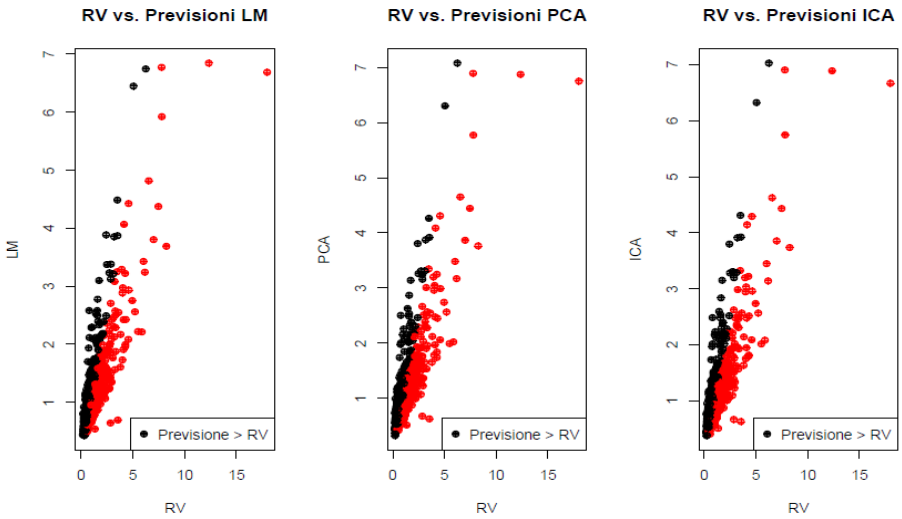

Figura D.5: Home Depot.

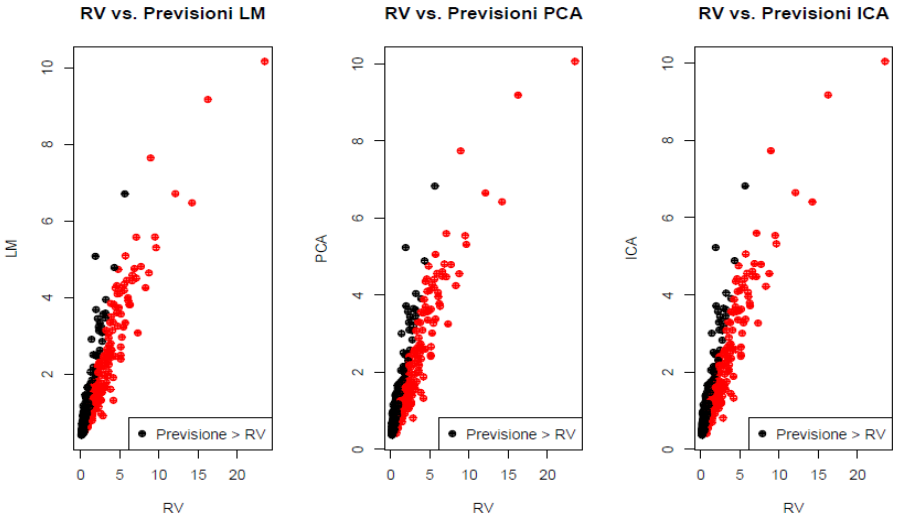

Figura D.6: Honeywell.

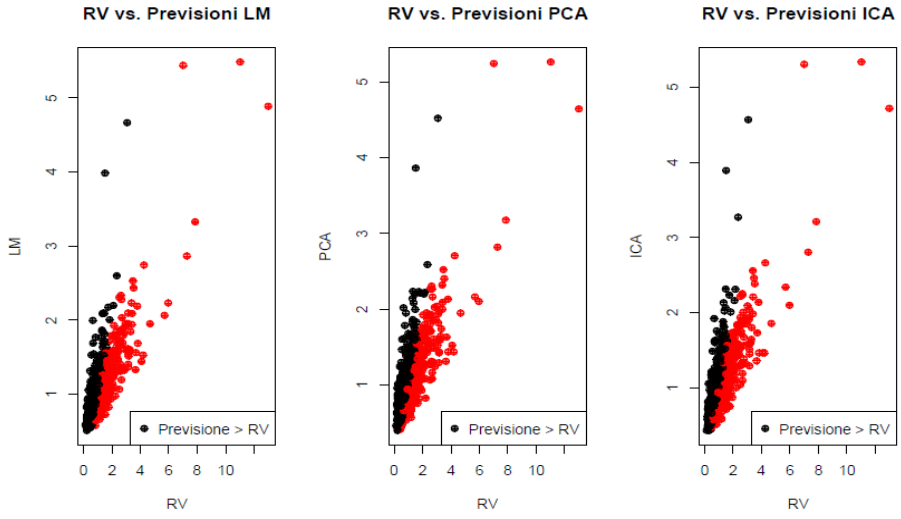

Figura D.7: Merk.

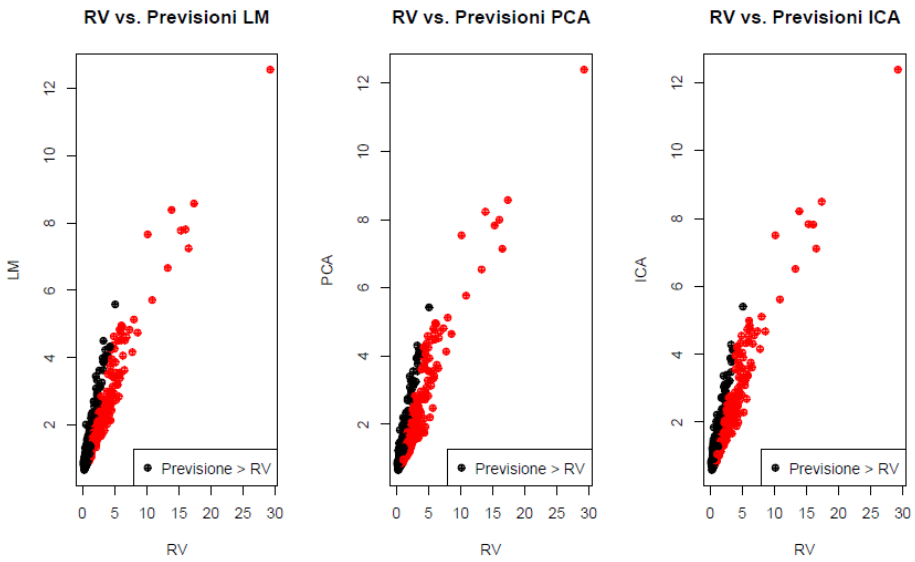

Figura D.8: Oracle.

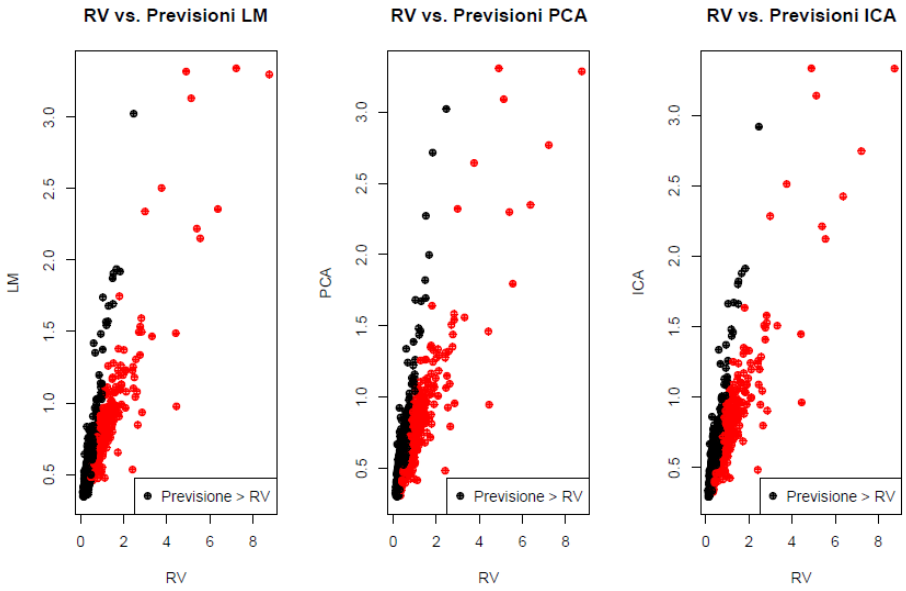

Figura D.9: Wal-Mart.

Appendice E

Coefficienti di persistenza

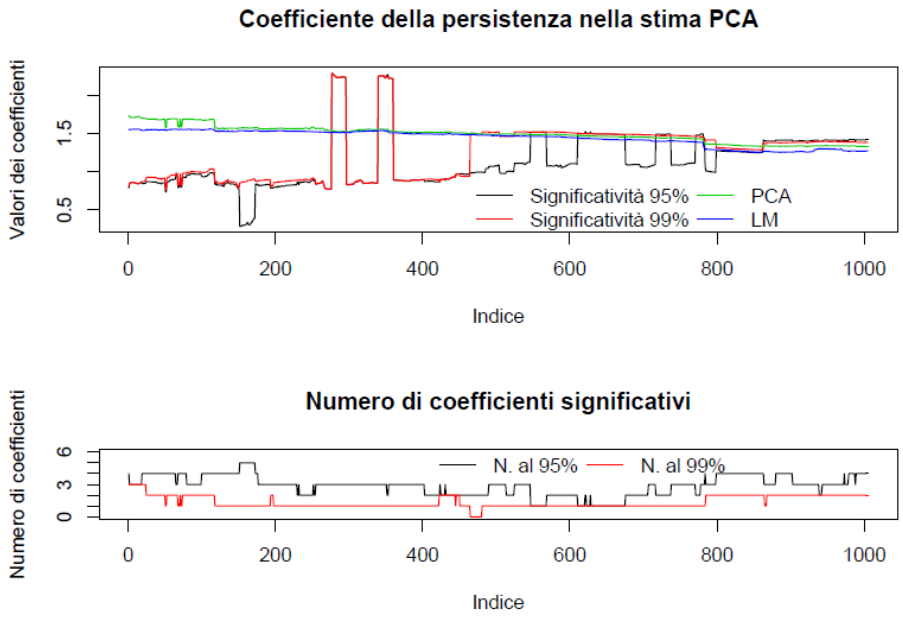

Figura E.1: Grafico dei coefficienti delle PC, titolo Anadarko.

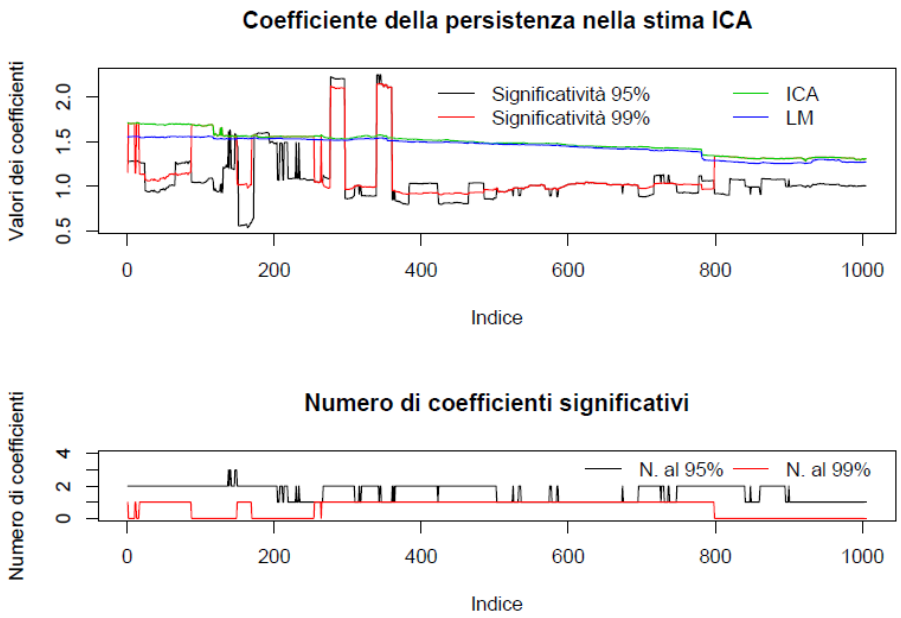

Figura E.2: Grafico dei coefficienti delle IC, titolo Anadarko.

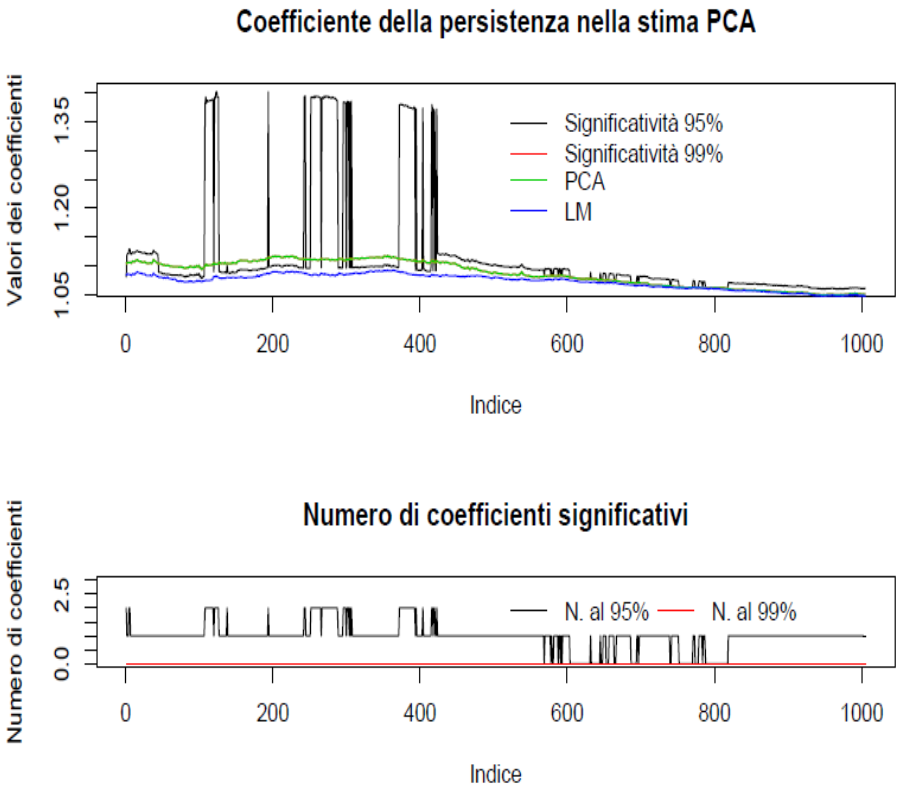

Figura E.3: Grafico dei coefficienti delle PC, titolo Bank of NY: nessuna variabile significativa al 99%.

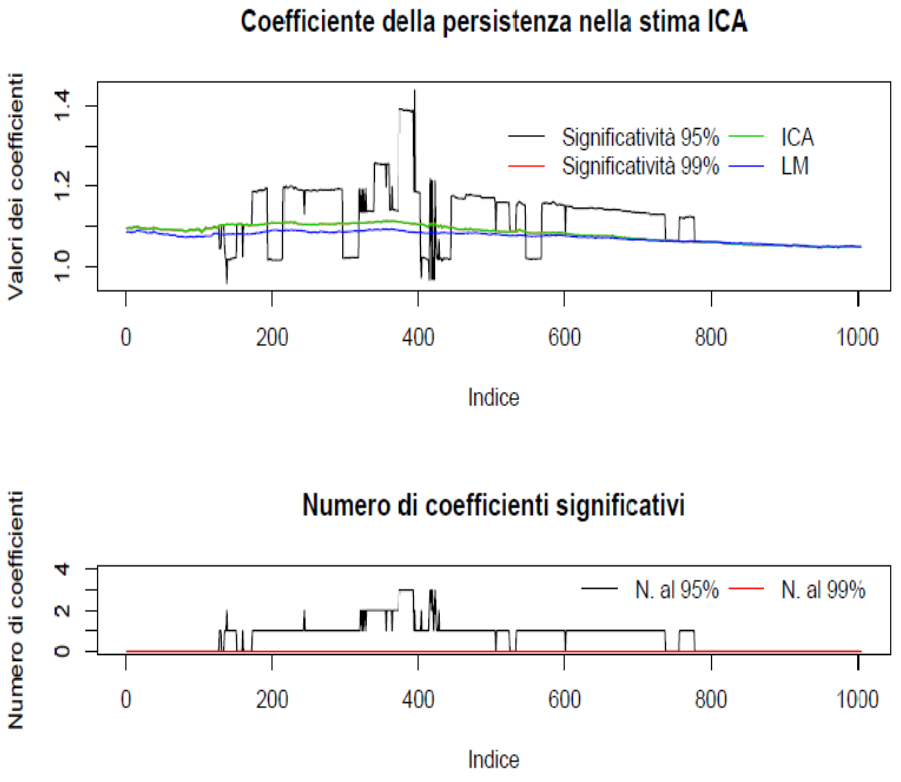

Figura E.4: Grafico dei coefficienti delle IC, titolo Bank of NY: nessuna variabile significativa al 99%.

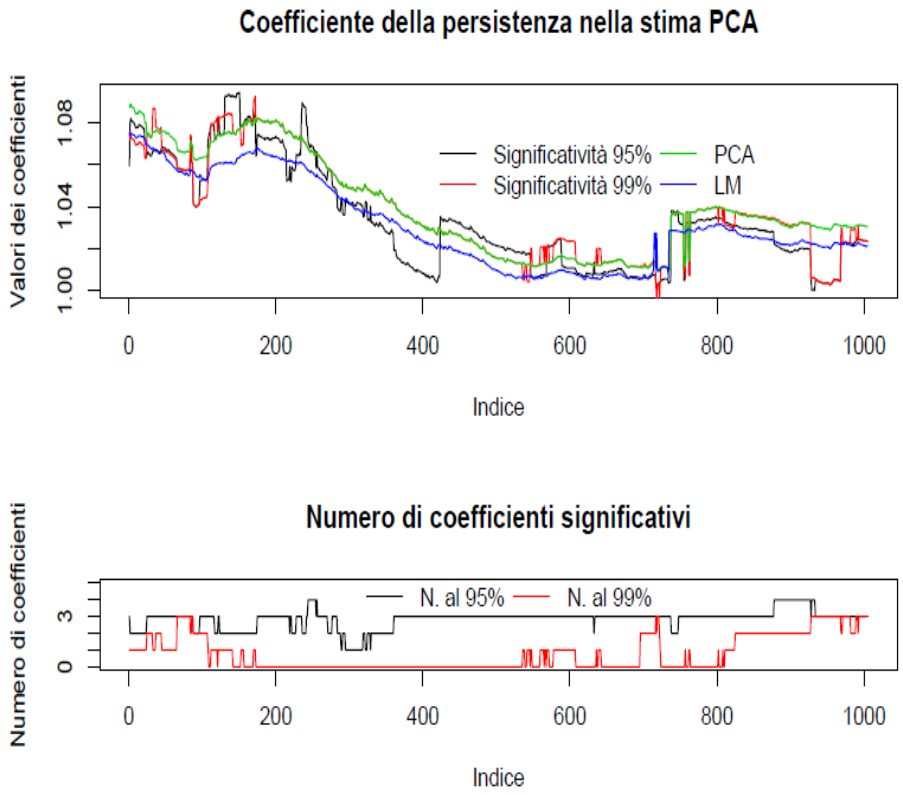

Figura E.5: Grafico dei coefficienti delle PC, titolo Bristol-Myers.

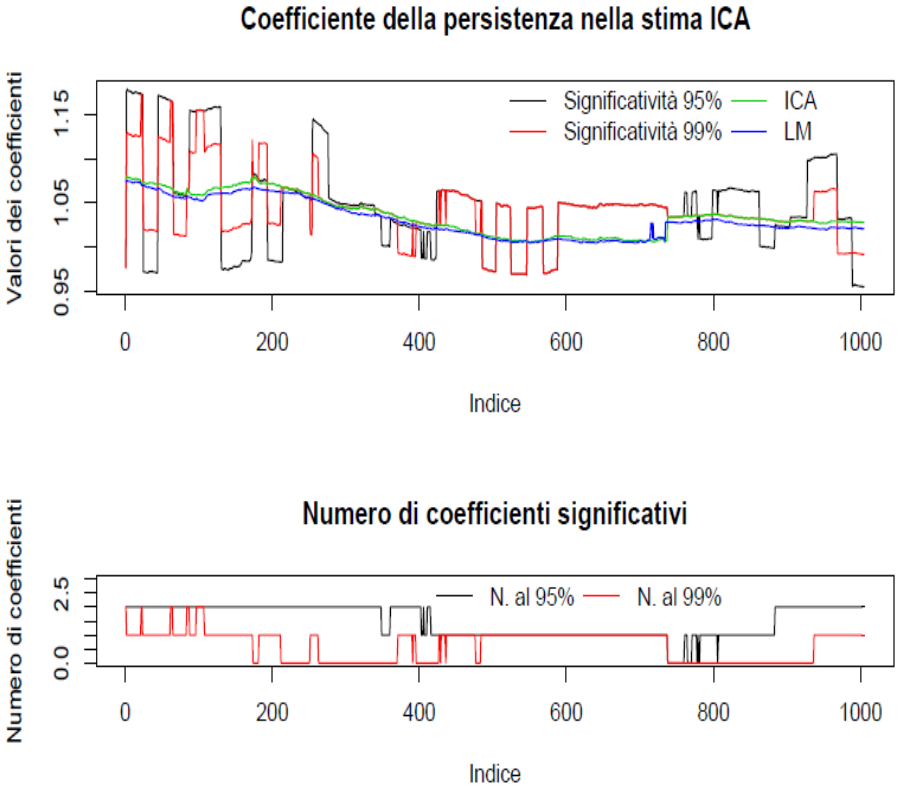

Figura E.6: Grafico dei coefficienti delle IC, titolo Bristol-Myers.

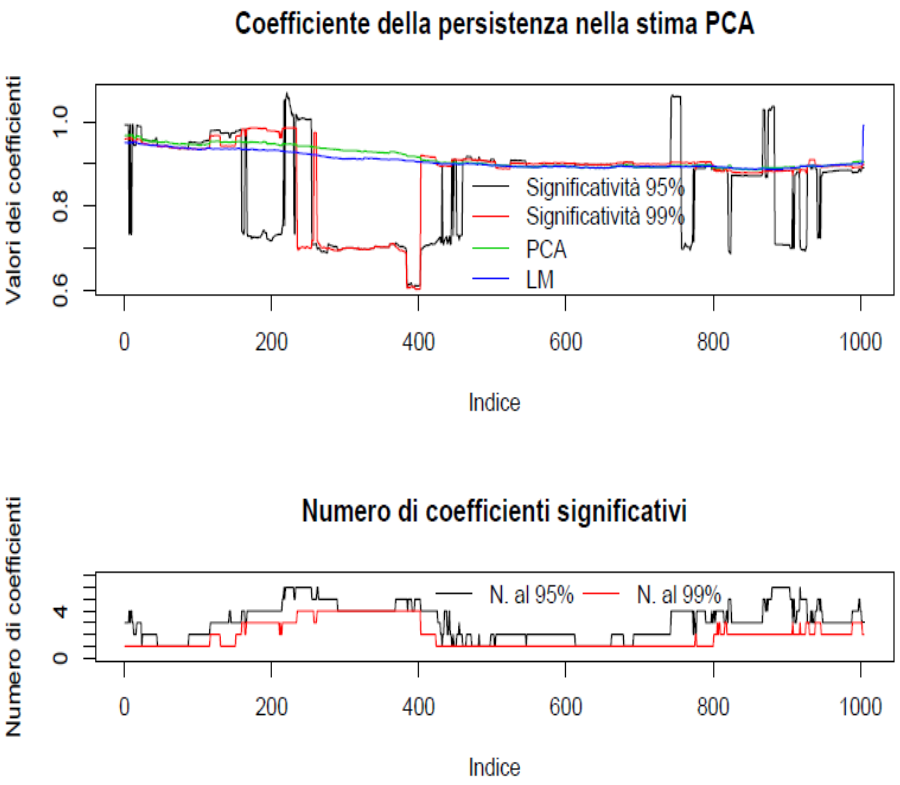

Figura E.7: Grafico dei coefficienti delle PC, titolo Exelon.

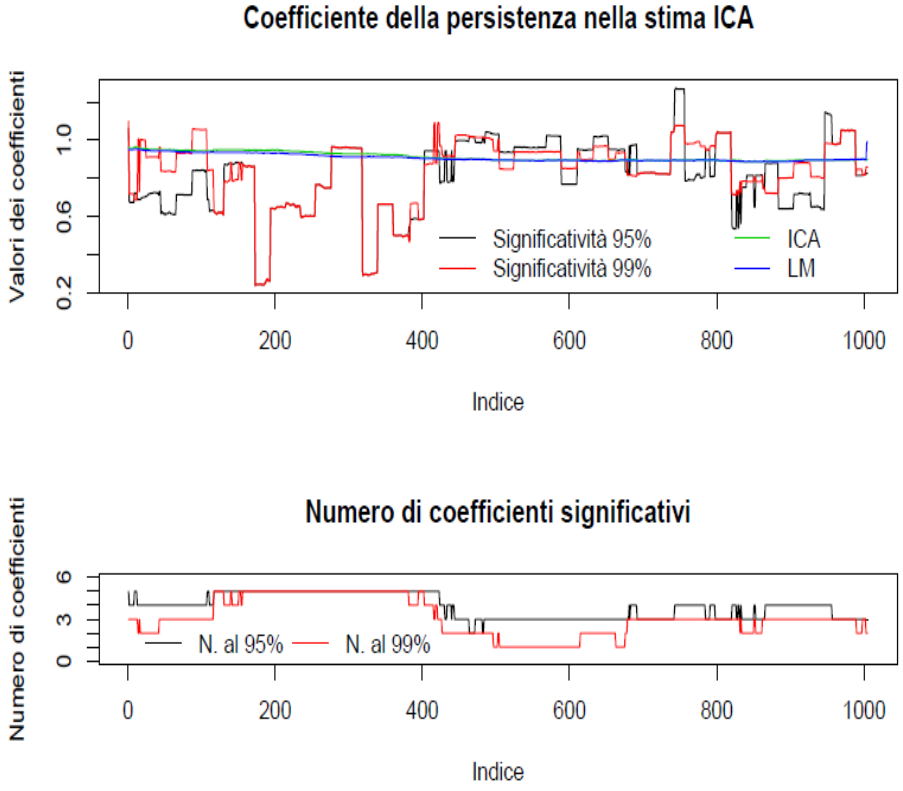

Figura E.8: Grafico dei coefficienti delle IC, titolo Exelon.

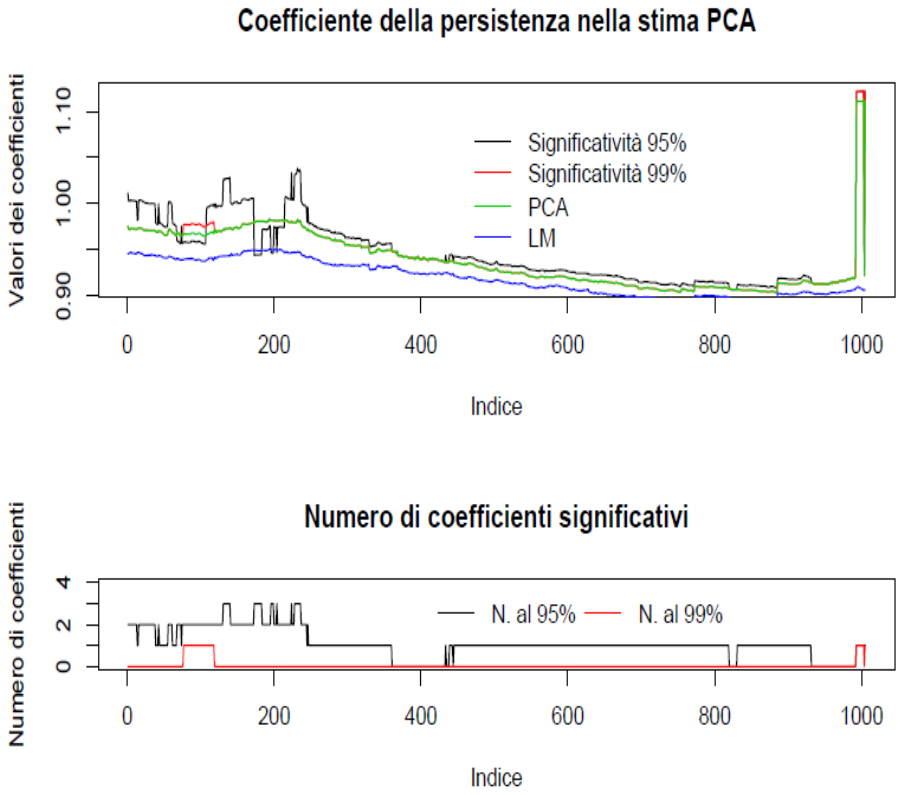

Figura E.9: Grafico dei coefficienti delle PC, titolo Home Depot.

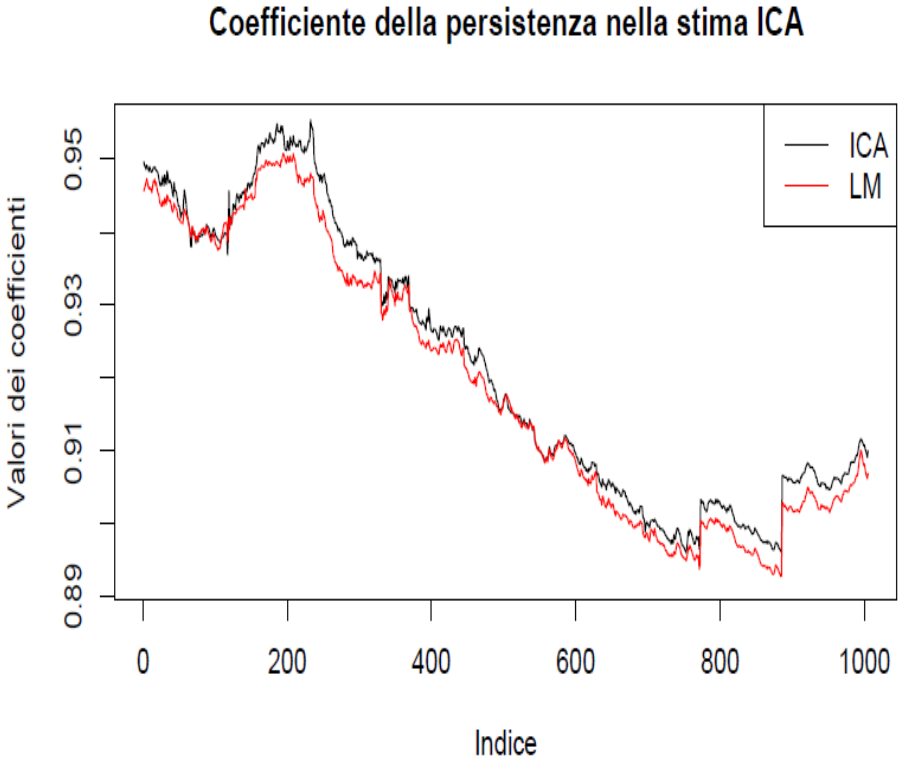

Figura E.10: Grafico dei coefficienti delle IC, titolo Home Depot: nessuna variabile è risultata significativa.

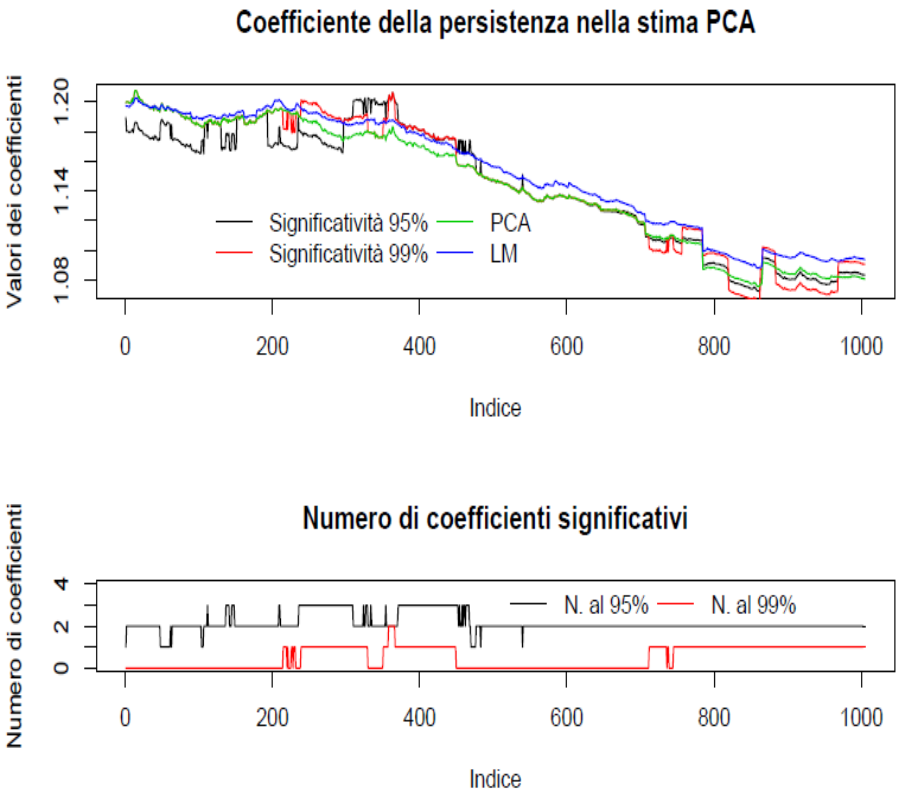

Figura E.11: Grafico dei coefficienti delle PC, titolo Honeywell.

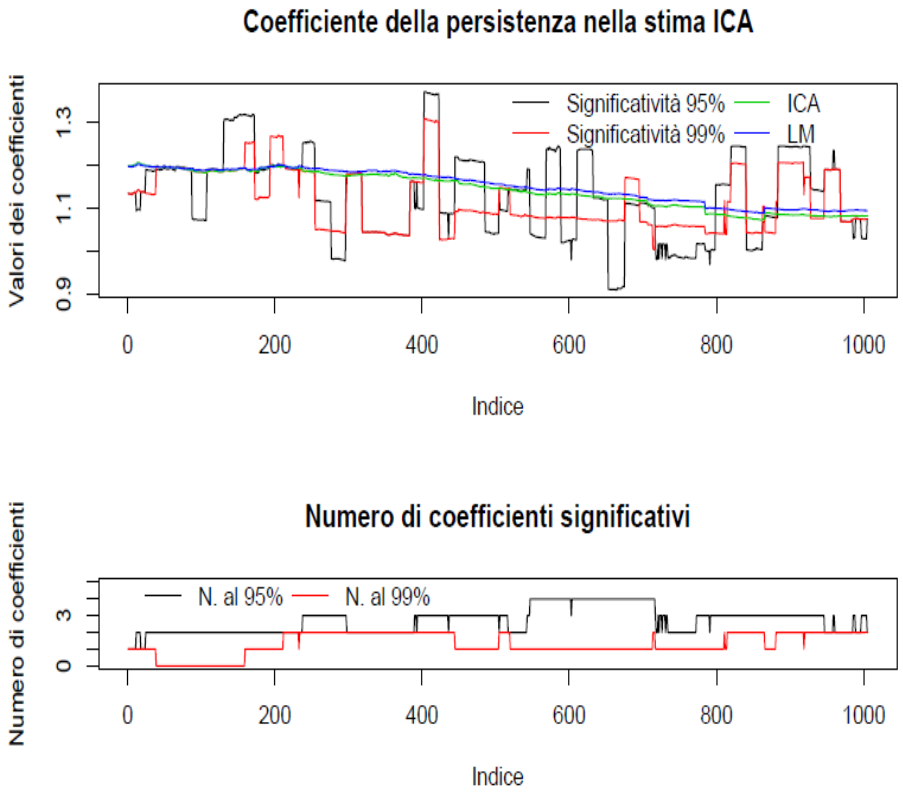

Figura E.12: Grafico dei coefficienti delle IC, titolo Honeywell.

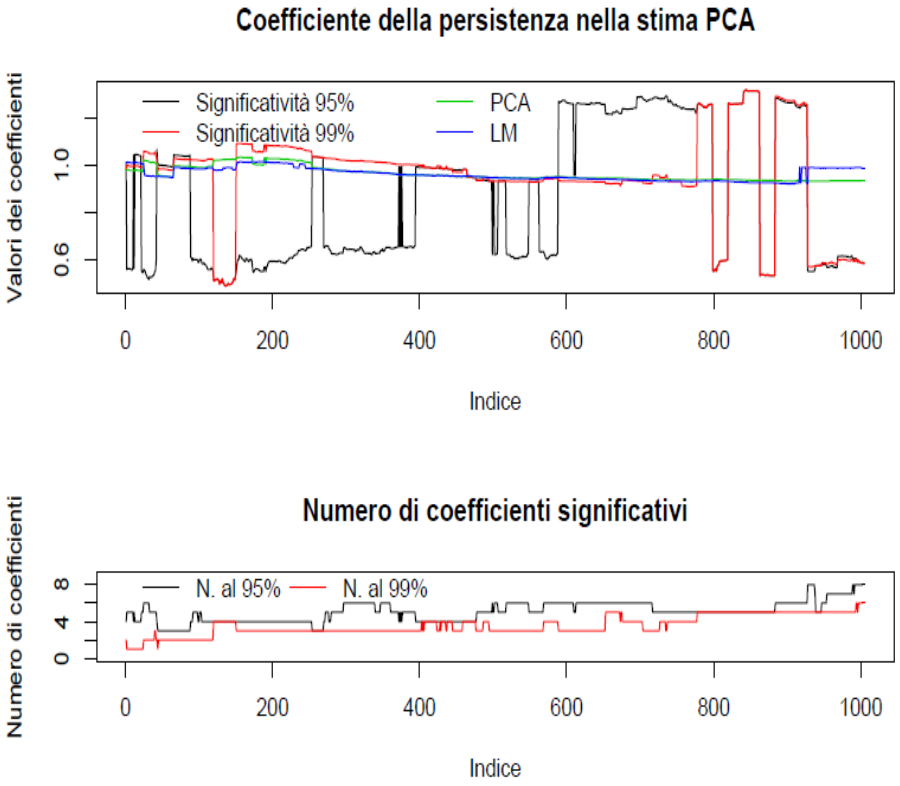

Figura E.13: Grafico dei coefficienti delle PC, titolo Merck.

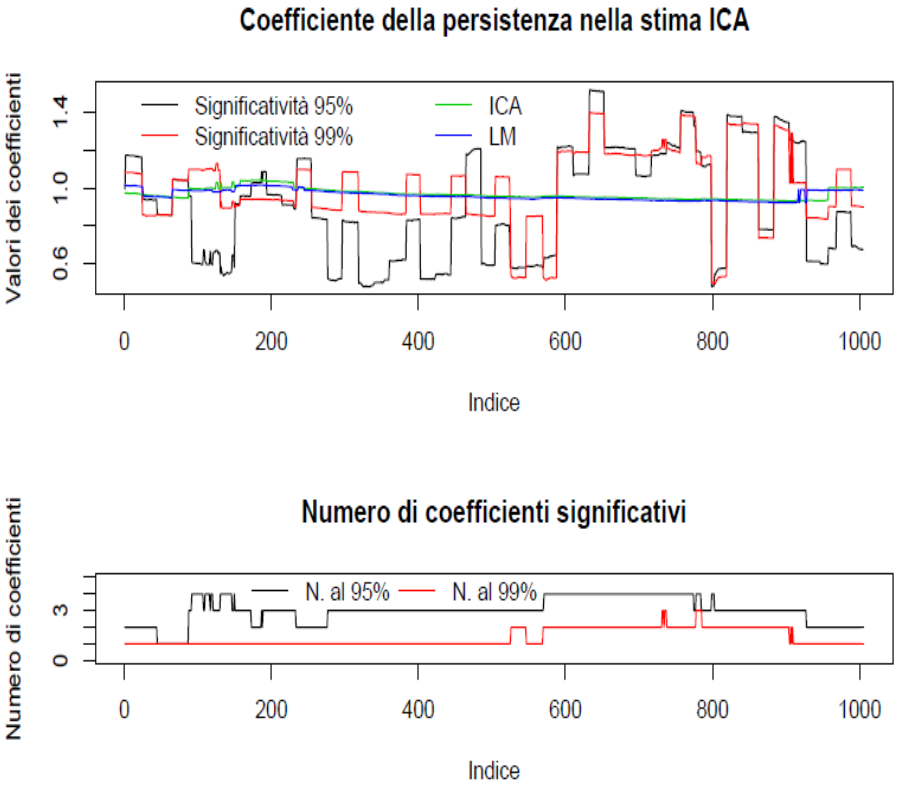

Figura E.14: Grafico dei coefficienti delle IC, titolo Merck.

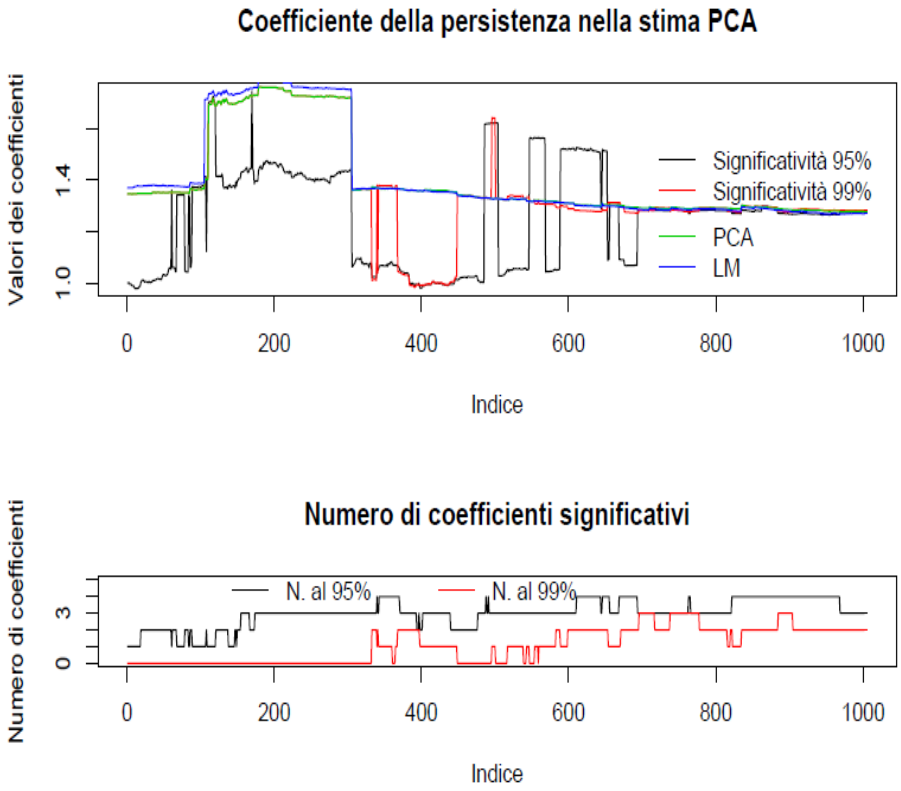

Figura E.15: Grafico dei coefficienti delle PC, titolo Oracle.

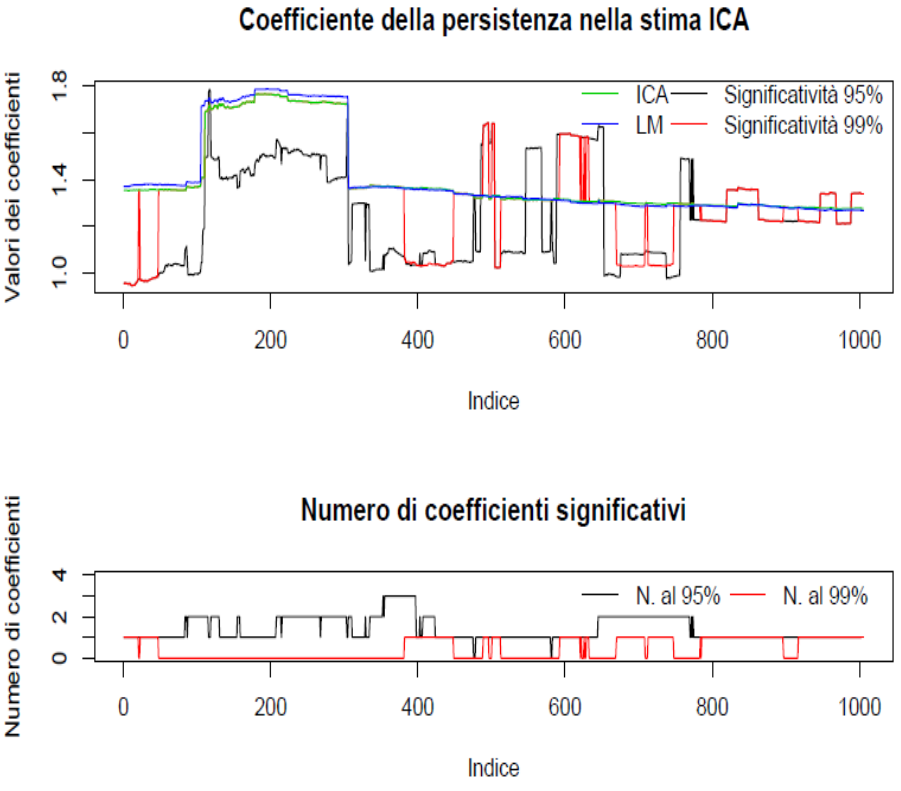

Figura E.16: Grafico dei coefficienti delle IC, titolo Oracle.

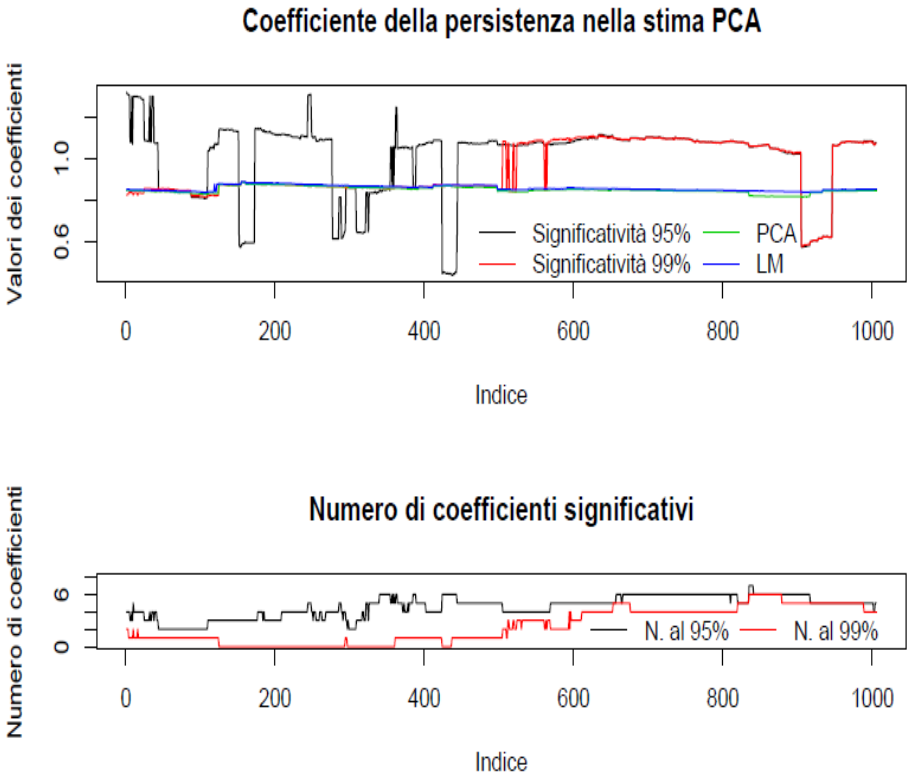

Figura E.17: Grafico dei coefficienti delle PC, titolo Wal-Mart.

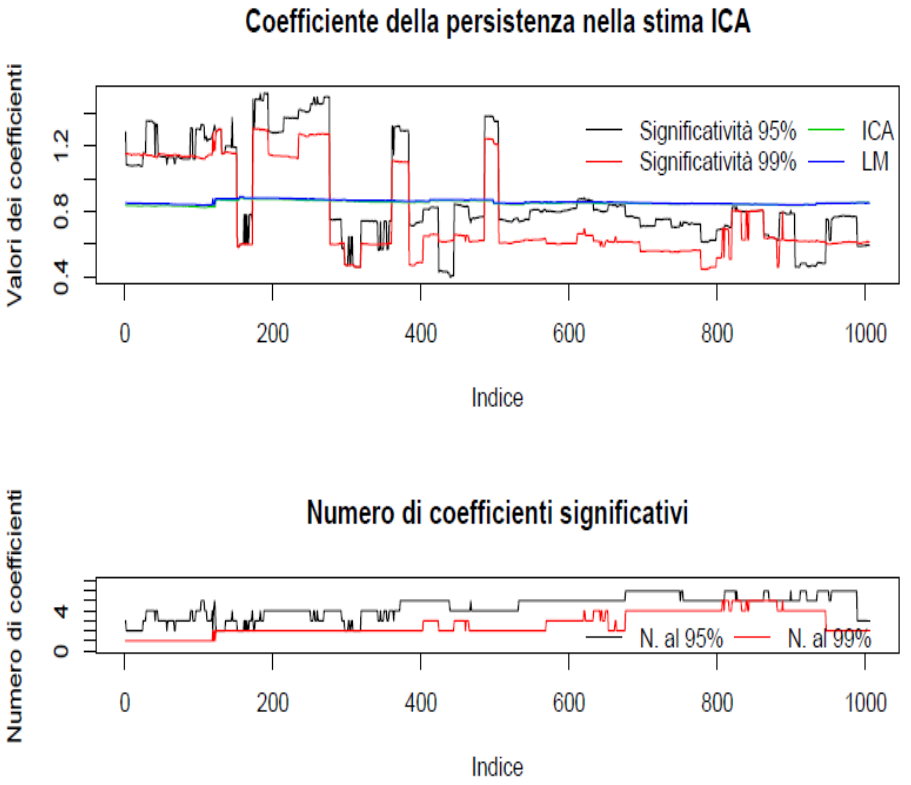

Figura E.18: Grafico dei coefficienti delle IC, titolo Wal-Mart.

Appendice F

Errori di previsione

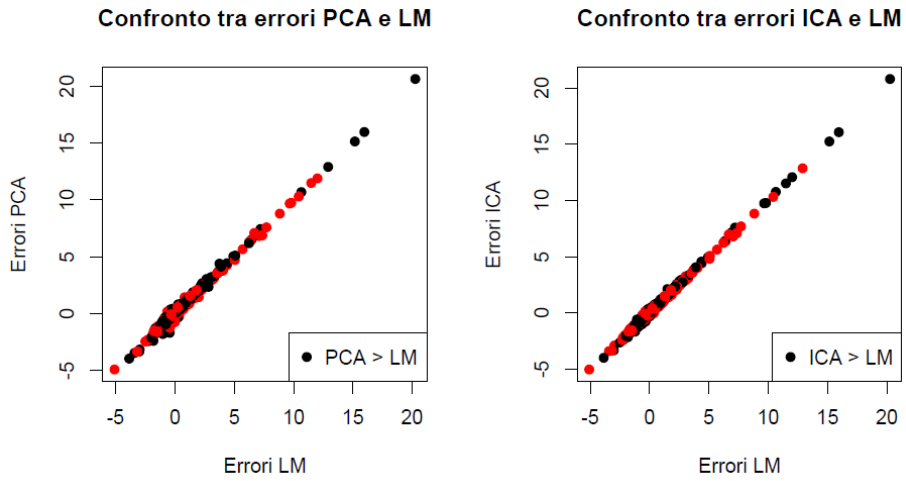

Figura F.1: Confronto tra gli errori di previsione, per il titolo Anadarko.

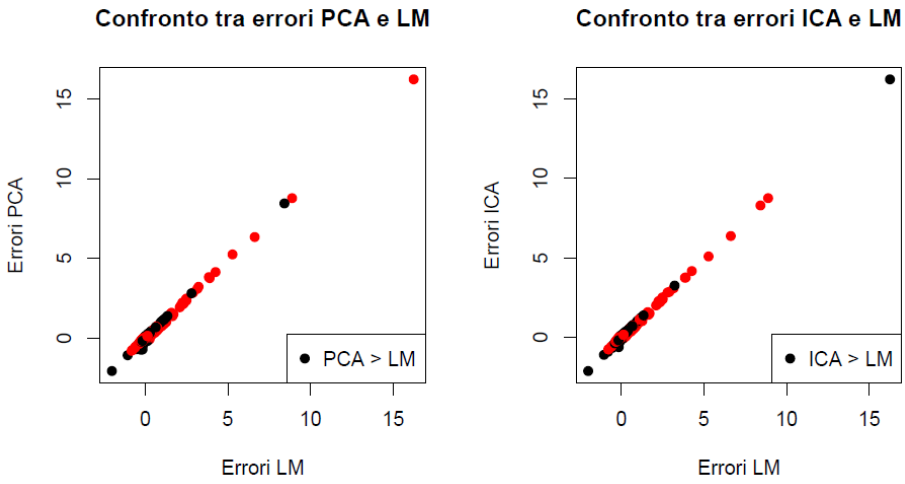

Figura F.2: Confronto tra gli errori di previsione, per il titolo Bank of NY.

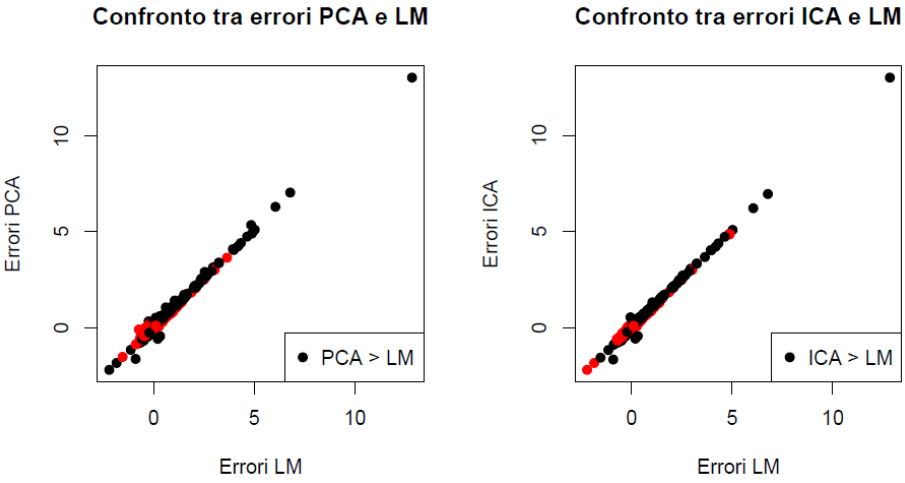

Figura F.3: Confronto tra gli errori di previsione, per il titolo Bristol-Myers.

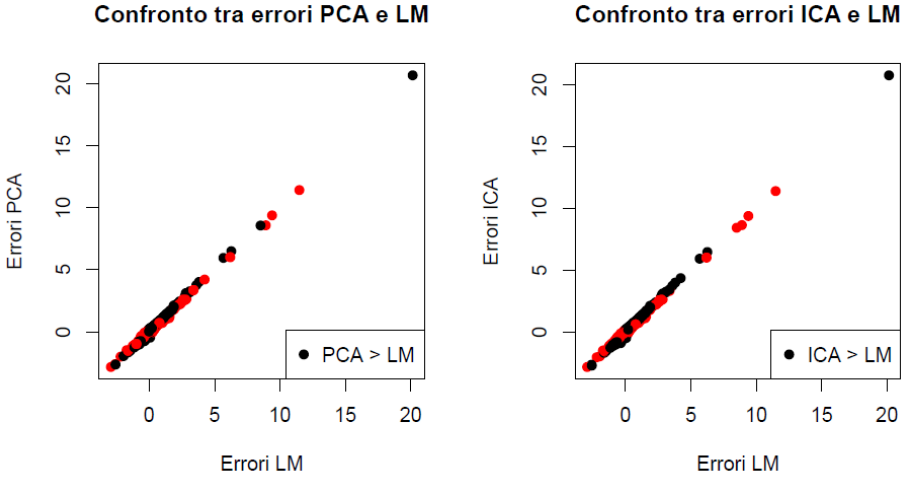

Figura F.4: Confronto tra gli errori di previsione, per il titolo Exelon.

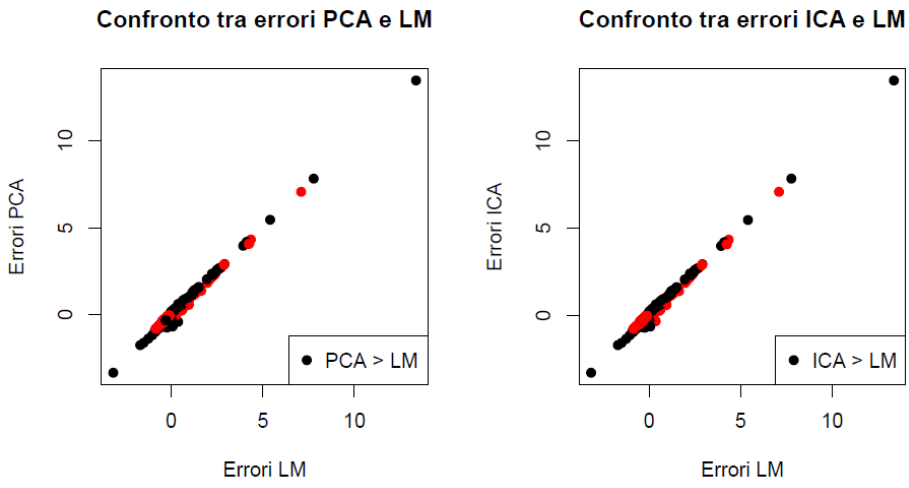

Figura F.5: Confronto tra gli errori di previsione, per il titolo Home Depot.

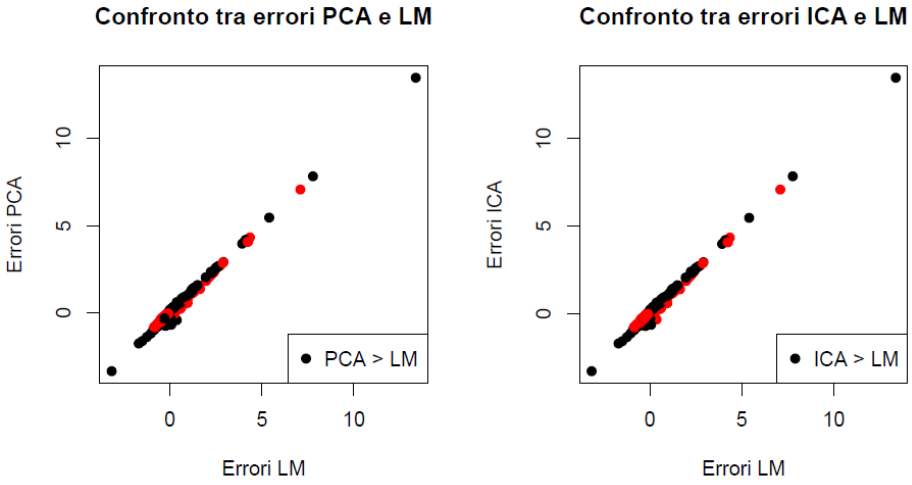

Figura F.6: Confronto tra gli errori di previsione, per il titolo Honeywell.

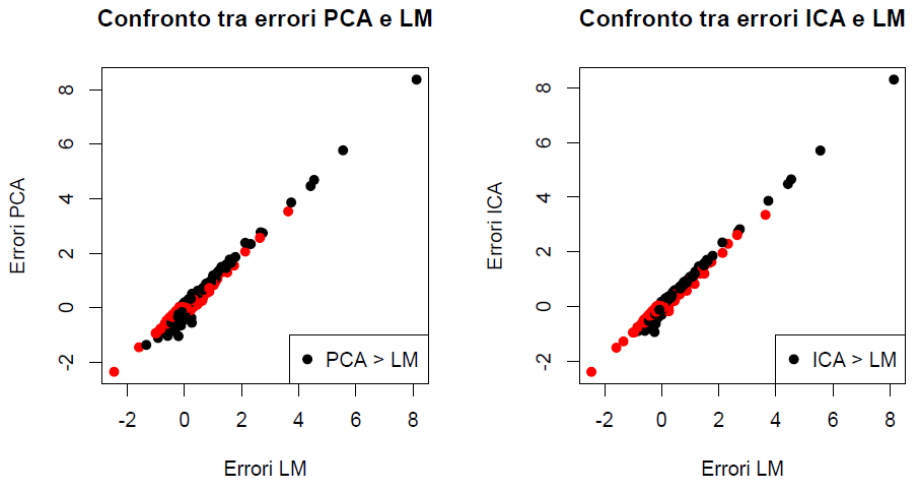

Figura F.7: Confronto tra gli errori di previsione, per il titolo Merck.

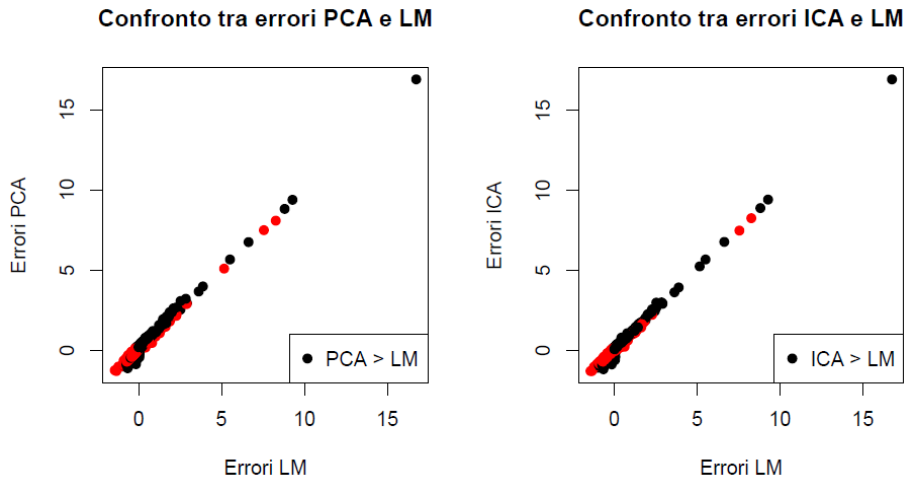

Figura F.8: Confronto tra gli errori di previsione, per il titolo Oracle.

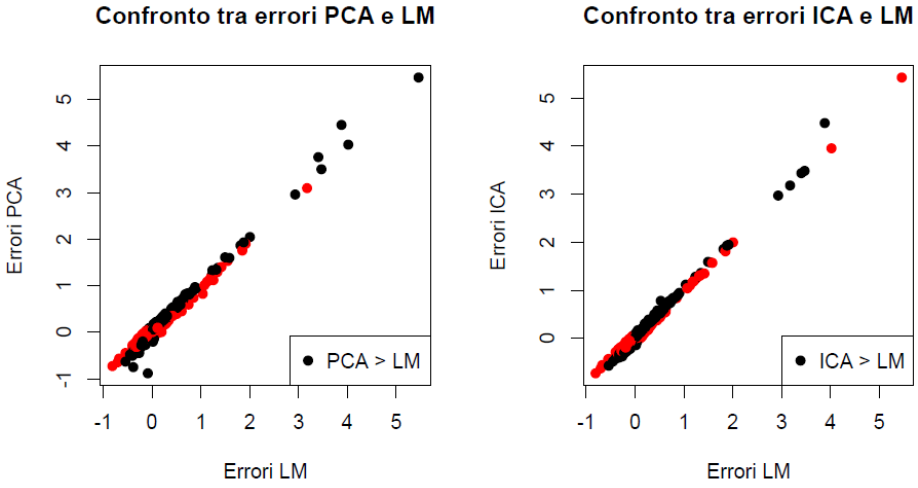

Figura F.9: Confronto tra gli errori di previsione, per il titolo Wal-Mart.