UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea in Fisica

Tesi di Laurea

I processi di Hawkes e loro applicazioni in Finanza

Relatore

Prof. Paolo Dai Pra

Laureando

Darko Ivanovski

Anno Accademico 2018/2019

Indice

1 I processi di punto **5**
1.1 I processi di Poisson 5
1.1.1 I processi di Poisson non omogenei 7
1.2 I processi di Hawkes 8
1.3 Simulazione di eventi e stima dei parametri 10

2 Sincronizzazioni e instabilità nei mercati finanziari **13**
2.1 Identificazione di eventi estremi 13
2.2 Dinamica degli eventi negli ultimi anni 15
2.2.1 Dipendenza dalle notizie macroeconomiche 17

3 Modellizzazione del fenomeno **19**
3.1 Processo di Hawkes multivariato 19
3.1.1 Scelta della parametrizzazione 20
3.1.2 Stima dei parametri 22
3.2 Risultati del modello 23
3.3 Conclusioni 25

Introduzione

L'obiettivo di questa tesi è presentare un modello matematico relativamente semplice e accessibile, ma allo stesso tempo potente e molto utilizzato in diversi ambiti. I processi di Hawkes, infatti, sono una particolare tipologia dei più generici processi di punto, e costituiscono un utile modello probabilistico per trattare occorrenze di eventi discreti e interdipendenti. In generale, i processi di punto sono una collezione di eventi casuali che cadono sulla retta temporale. Essi, quindi, possono essere utilizzati per creare modelli in diverse aree di studio, come in geofisica per la modellizzazione dei terremoti di assestamento successivi a un terremoto di grande intensità, o in ecologia per studiare l'osservazione di una determinata specie animale in una data area geografica, o infine per creare modelli che riproducano alcuni comportamenti che emergono nei social network come la viralità di una data notizia. Questi modelli possono, poi, essere utilizzati per cercare di spiegare la natura del processo osservato, per effettuare delle simulazioni e, anche, per predire la probabilità di eventi futuri.

In questa tesi, tuttavia, i processi di Hawkes sono stati applicati alla Finanza per modellizzare i tempi di occorrenza di operazioni su un mercato finanziario. Come si vedrà, tramite un approccio matematico rigoroso ma accessibile, si è riusciti a creare un modello soddisfacente per spiegare delle instabilità interne al mercato finanziario riscontrate negli ultimi anni. Recentemente, infatti, si è assistito a una considerevole crescita del ruolo che la tecnologia occupa anche all'interno della finanza. Al giorno d'oggi, tutti i principali mercati mondiali utilizzano solamente piattaforme elettroniche per effettuare le diverse operazioni. Ne deriva, quindi, una notevole velocità nella propagazione delle informazioni e nell'esecuzione degli ordini di compra-vendita. Tutto ciò ha cambiato la dinamica interna ai mercati finanziari, mostrando degli effetti di sincronizzazione e instabilità prima sconosciuti.

Nel primo capitolo, quindi, vengono presentate le basi matematiche e le definizioni del modello utilizzato. Nel secondo capitolo, invece, vengono presentate le evidenze empiriche del problema di cui si vuole creare un modello, mostrando, per l'appunto, come la dinamica dei mercati sia notevolmente cambiata negli ultimi 15 anni, facendo emergere quei fenomeni di instabilità che la tesi si propone di studiare. Nel terzo capitolo, infine, viene sviluppato concretamente il modello, utilizzando i processi di Hawkes, tramite il quale vengono effettuate delle simulazioni che ben si accordano ai dati empirici rilevati, riuscendo quindi a darne una spiegazione soddisfacente.

Capitolo 1

I processi di punto

Un processo di punto definito sulla semiretta reale non negativa, la quale viene utilizzata solitamente per rappresentare il tempo, consiste in una successione strettamente crescente di tempi aleatori $(T_i)_{i\geq 1}$ priva di punti di accumulazione. Ogni tempo T_i può essere interpretato come il tempo al quale l'evento i si verifica e, di conseguenza, T_i viene chiamato "tempo dell'evento i".

Equivalentemente si può definire un processo di conteggio N_t, dove N_t è una funzione definita per $t \geq 0$ che assume esclusivamente valori interi non negativi. Il suo valore rappresenta il numero di eventi del processo di punto verificatisi prima del tempo t. In definitiva, N_t conta il numero di eventi fino al tempo t ed è univocamente determinata dalla sequenza casuale dei tempi T_i del processo di punto. Si può scrivere:

$$N_t := \sum_{i>0} \mathbb{1}_{[t\geq T_i]} \tag{1.1}$$

dove $\mathbb{1}_{[t\geq T_i]}$ è una funzione che vale 1 se è verificata la condizione $t \geq T_i$ e 0 altrimenti. Si può vedere che, ovviamente, $N_0 = 0$. Inoltre N_t è costante a tratti, tutti i suoi salti hanno ampiezza 1 e si verificano ai vari tempi T_i. È immediato convincersi che la sequenza dei tempi T_i e la funzione N_t sono rappresentazioni equivalenti del processo di punto che si sta descrivendo.

1.1 I processi di Poisson

Il processo di punto più semplice è il processo di Poisson.

Definizione 1.1. *Sia $(\tau_i)_{i\geq 1}$ una sequenza di variabili casuali indipendenti, ognuna con una distribuzione esponenziale di parametro λ, e una successione di tempi $T_n = \sum_{i=1}^{n} \tau_i$. Il processo descritto da questa successione T_n, o equivalentemente dalla funzione N_t, con $t \geq 0$, come definita sopra, è chiamato processo di Poisson con intensità λ.*

Per come è stato definita la successione T_n, si vede che la sequenza delle τ_i rappresenta la distanza temporale tra l'evento i e l'evento $i-1$: il primo evento si verifica a τ_1, il secondo

si verifica a un tempo τ_2 dopo il primo, e così via per ogni evento.
Siccome le τ_i sono definite come variabili casuali, per ogni τ si può scrivere la densità di probabilità e calcolarne il valore atteso.

$$f_\tau(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{1.2}$$

$$\mathbb{E}_\tau[\tau] = \int_{-\infty}^{\infty} t f_\tau(t) dt = \int_0^{\infty} \lambda t e^{-\lambda t} dt = \frac{1}{\lambda} \int_0^{\infty} x e^{-x} dx = \frac{1}{\lambda} \tag{1.3}$$

Questo semplice calcolo indica che il valore atteso del tempo tra due eventi qualsiasi è λ^{-1}, ovvero, intuitivamente, che gli eventi vengono osservati in media λ volte per unità di tempo. Questo giustifica il nome processo di Poisson con *intensità* λ.

Una delle proprietà più importanti del processo di Poisson è la sua *mancanza di memoria*. La mancanza di memoria esprime, intuitivamente, il fatto che la locazione temporale degli eventi futuri è indipendente da quella degli eventi passati. Nel caso in questione del processo di Poisson, la mancanza di memoria indica che la probabilità di osservare un nuovo evento avendone appena osservato uno è la stessa di osservare un nuovo evento avendo già aspettato un arbitrario tempo m dall'evento precedente.
Per vedere ciò, avendo a disposizione la densità di probabilità per la distanza temporale tra due eventi, si calcola la probabilità di osservare un evento tra il tempo 0 e un tempo prefissato t:

$$\mathbb{P}(\tau \leq t) = \int_0^t \lambda e^{-\lambda x} dx = [-e^{\lambda x}]_{x=0}^{x=t} = 1 - e^{-\lambda t} \tag{1.4}$$

Quindi, si può calcolare banalmente anche la probabilità di osservare un evento al tempo $\tau > t$:

$$\mathbb{P}(\tau > t) = e^{-\lambda t} \tag{1.5}$$

A questo punto, supponendo di aver già aspettato un tempo m dall'ultimo evento senza averne osservato nessun altro, si calcola la probabilità di dover aspettare un ulteriore tempo t. Quindi, supponendo che nell'intervallo $[0, m]$ non ci siano eventi, e quindi che $\tau > m$, si ha:

$$\begin{aligned} \mathbb{P}(\tau > (t+m) | \tau > m) &\overset{(a)}{=} \frac{\mathbb{P}(\tau > (t+m), \tau > m)}{\mathbb{P}(\tau > m)} \\ &\overset{(b)}{=} \frac{\mathbb{P}(\tau > (t+m))}{\mathbb{P}(\tau > m)} \overset{(c)}{=} \frac{e^{-\lambda(t+m)}}{e^{-\lambda m}} = e^{-\lambda t} = \mathbb{P}(\tau > t) \end{aligned} \tag{1.6}$$

Nel passaggio (a) si è usata la definizione di probabilità condizionata, mentre nel passaggio (b) si è semplicemente notato che se $\tau > (t+m)$ automaticamente è verificato anche $\tau > m$. Infine, nel passaggio (c) si è utilizzato il risultato ottenuto in (1.5).
Il risultato ottenuto in (1.6) esprime la proprietà di *mancanza di memoria* in forma rigorosa. Qualsiasi sia il tempo m passato dall'osservazione dell'ultimo evento, un nuovo evento ha sempre la probabilità $e^{-\lambda t}$ di essere osservato dopo un aggiuntivo tempo t.

In linea del tutto generale, la *mancanza di memoria* è una proprietà di tutte le variabili casuali con una distribuzione di probabilità esponenziale. Infatti, riscrivendo la (1.6), e denotando la probabilità con $f(x) = \mathbb{P}(\tau > x)$, otteniamo:

$$\frac{f(t+m)}{f(m)} = \frac{\mathbb{P}(\tau > (t+m))}{\mathbb{P}(\tau > m)} = \mathbb{P}(\tau > t) = f(t) \quad \longleftrightarrow \quad f(t+m) = f(t)f(m) \tag{1.7}$$

La funzione f deve quindi soddisfare la proprietà $f(t+m) = f(t)f(m)$, deve cioè essere un omomorfismo tra il gruppo additivo e il gruppo moltiplicativo dei reali, e l'unica classe di funzioni continue che la soddisfa è appunto l'esponenziale.

1.1.1 I processi di Poisson non omogenei

Nei processi di Poisson descritti finora, gli eventi venivano osservati con un'intensità costante λ. Questo modello basilare è sufficiente a descrivere processi semplici, come ad esempio il passaggio delle automobili su una strada poco affollata e in un breve periodo di tempo. Tuttavia, per poter descrivere processi più complessi, quali ad esempio simulare il traffico nelle ore di punta, è necessario poter considerare il caso di un'intensità $\lambda(t)$ che possa dipendere dal tempo e dalla storia del processo fino al tempo t.

Definizione 1.2. *L'intensità di un processo di punto $(N_t)_{t>0}$ è definita come segue, assumendo l'esistenza del limite:*

$$\lambda(t|\mathcal{H}_t) = \lim_{h \to 0} \frac{\mathbb{P}(N_{t+h} - N_t = 1|\mathcal{H}_t)}{h} \tag{1.8}$$

dove $\mathcal{H}_t$ è la storia del processo fino al tempo t, contenente quindi la successione dei tempi T_i . Useremo per brevità la notazione $\lambda(t) := \lambda(t|\mathcal{H}_t)$. Inoltre, la distribuzione di un processo di punto è completamente caratterizzata dalla sua intensità.

La definizione fornisce quindi la caratterizzazione del processo di punto dal punto di vista dell'intensità, equivalente alle altre due caratterizzazioni viste in precedenza tramite il processo di conteggio e la successione dei tempi. Formalmente, si può vedere che $\lambda(t)$ e N_t sono collegate tramite la probabilità di osservare un evento in un piccolo intervallo di tempo h:

$$\begin{aligned} \mathbb{P}(N_{t+h} = n+1|\mathcal{H}_t) &= \lambda(t)h + o(h) \\ \mathbb{P}(N_{t+h} = n+m|\mathcal{H}_t) &= o(h) \qquad \textit{se } m > 1 \\ \mathbb{P}(N_{t+h} = n|\mathcal{H}_t) &= 1 - \lambda(t)h + o(h) \end{aligned} \tag{1.9}$$

Da questa caratterizzazione si vede, dunque, che la probabilità di osservare un nuovo evento durante l'intervallo infinitesimo tra t e $t+h$ è $\lambda(t)h$, mentre la probabilità di osservare più di un evento nello stesso intervallo infinitesimo è trascurabile.

Nei processi di Poisson, e più in generale nei processi di punto con intensità deterministica, i tempi di attesa fra due eventi successivi sono indipendenti. Ciò non avviene quando l'intensità dipende dalla storia del processo, ed è quindi aleatoria.

1.2 I processi di Hawkes

Per riuscire a descrivere situazioni più complicate, e quindi più interessanti, si può introdurre ora una classe di processi di punto nei quali la funzione d'intensità dipende esplicitamente dagli eventi passati, i *processi auto-eccitanti*. In questo tipo di processi, l'osservazione di un evento provoca l'incremento della funzione di intensità del processo stesso.
L'esempio più noto di questa tipologia di processi è il *processo di Hawkes*:

Definizione 1.3. *Sia $(N_t)_{t>0}$ un processo di punto, associato con la sua storia $\mathcal{H}_t$, $t \geq 0$. Come visto in (1.9), il processo di punto è completamente determinato dalla funzione di intensità $\lambda(t)$.*
Il processo di punto si chiama processo di Hawkes se la funzione di intensità $\lambda(t|\mathcal{H}_t)$ assume la forma:

$$\lambda(t|\mathcal{H}_t) = \lambda_0(t) + \sum_{i:t>T_i} \phi(t - T_i) \tag{1.10}$$

dove $\lambda_0(t) : \mathbb{R} \to \mathbb{R}_+$ è una funzione che determina l'intensità di base del processo, indipendente dagli altri eventi, mentre $\phi : \mathbb{R} \to \mathbb{R}_+$ è chiamata memory kernel. Si può vedere, inoltre, che i processi di Hawkes sono un caso particolare dei processi di Poisson non omogenei, nei quali la funzione di intensità è esplicitamente dipendente dagli eventi passati tramite la funzione $\phi(\cdot)$.

Come si vede dalla definizione, tutti gli eventi con un tempo $T_i < t$, cioè tutti gli eventi che sono già stati osservati prima del tempo t, contribuiscono alla funzione di intensità al tempo t.
Più nel nel dettaglio, $\lambda_0(t) > 0$ descrive l'osservazione di eventi innescati da fattori esterni. Questi eventi vengono chiamati *esogeni* e il loro verificarsi è indipendente dai precedenti eventi del processo. La natura di *processo auto-eccitante* del processo di Hawkes deriva dalla sommatoria, dove il kernel $\phi(t - T_i)$ modula il cambiamento che un evento al tempo T_i provoca alla funzione di intensità. Anche se non necessario, tipicamente la funzione $\phi(\cdot)$ è una funzione monotona decrescente, in modo che un evento più recente influenzi maggiormente l'intensità del processo. Tipicamente, le due famiglie di funzioni più utilizzate come kernel nei processi di Hawkes sono: la legge di potenza e la funzione esponenziale.
Per la legge di potenza si ha:

$$\phi(x) = \frac{\alpha}{(x + \delta)^{\eta+1}} \tag{1.11}$$

dove $\alpha \geq 0,\ \delta > 0,\ \eta > 0$ e $\alpha < \eta\delta^{\eta}$.
La più popolare è tuttavia la funzione esponenziale:

$$\phi(x) = \alpha e^{-\delta x} \tag{1.12}$$

con $\alpha \geq 0, \delta > 0$ e $\alpha < \delta$.

Una ulteriore visione equivalente dei processi di Hawkes utilizza il concetto di cluster, tramite il quale si separano gli eventi del processo in due categorie: *esogeni* e *discendenti*. Gli eventi discendenti, come suggerisce il nome, sono innescati da eventi preesistenti del processo, mentre gli eventi esogeni sono innescati da fattori esterni, sono cioè indipendenti

e non hanno una relazione di parentela con altri eventi preesistenti. Tutti gli eventi che discendono, direttamente o indirettamente, da un evento esogeno formano un *cluster* associato all'evento esogeno stesso. L'evento esogeno forma la generazione 0, gli eventi che discendono direttamente da lui formano la generazione 1, gli eventi che discendono da un evento della generazione 1 formano la generazione 2, e così via. Questa visione viene chiamata *struttura ramificata* del processo di Hawkes. Tramite questa struttura si possono definire due importanti quantità del processo: il *fattore di ramificazione*, ovvero il numero di eventi direttamente innescati da un dato evento, e il numero totale di eventi in un dato cluster.

Il fattore di ramificazione n^* è definito come il numero medio di eventi direttamente discendenti da un singolo evento. Tale fattore n^*, intuitivamente, descrive il numero di eventi che appariranno nel processo o, nel contesto dei social network, la *viralità* di un processo. Inoltre, il fattore di ramificazione indica se un cluster avrà un numero finito o infinito di eventi al suo interno. Per $n^* < 1$ il processo viene definito in *regime subcritico* e il numero totale di eventi in un qualsiasi cluster sarà limitato. Gli eventi esogeni continuano a manifestarsi con l'intensità di base $\lambda_0(t)$, ma ognuno di questi genera un cluster con un numero di eventi finito e soprattutto limitato nel tempo. Al contrario, se $n^* > 1$ il processo viene definito in *regime supercritico*, situazione in cui la funzione $\lambda(t)$ continua a crescere e il numero di eventi in ogni cluster è illimitato. Si può calcolare esplicitamente il fattore di ramificazione integrando il kernel $\phi(t)$ sui tempi t:

$$n^* = \int_0^\infty \phi(t)dt \tag{1.13}$$

Imponendo la condizione $n^* < 1$ si ritrovano esattamente le condizioni sulle variabili delle $\phi(x)$ precedentemente presentate.
Focalizzando l'attenzione al caso di $n^* < 1$ si può stimare in maniera più accurata la dimensione di ogni cluster. Sia A_i il numero di eventi attesi nella *generazione* i, con ovviamente $A_0 = 1$. Il numero totale di eventi nel cluster, N_∞, sarà allora la somma di tutte le A_i. Tuttavia, visto che, in media, ogni evento della generazione $i-1$ genera n^* eventi discendenti, si può scrivere:

$$A_i = A_{i-1}n^* = A_{i-2}(n^*)^2 = ... = A_0(n^*)^i = (n^*)^i \tag{1.14}$$

Assumendo quindi $n^* < 1$, si ottiene che il numero di eventi attesi in ogni cluster, N_∞, è la somma di una progressione geometrica convergente:

$$N_\infty = \sum_{i=0}^{\infty} A_i = \sum_{i=0}^{\infty} (n^*)^i = \frac{1}{1-n^*} \tag{1.15}$$

1.3 Simulazione di eventi e stima dei parametri

In questa sezione vengono presentati alcuni cenni su un metodo, il più basilare, per la simulazione di eventi casuali che soddisfano le proprietà dei processi di Hawkes, e un metodo per la stima dei parametri di un processo osservato.

La simulazione di eventi che soddisfano le proprietà dei processi di Hawkes, sostanzialmente, consiste nel simulare una sequenza di τ_i in accordo con una data funzione di intensità $\lambda(t)$. Il metodo che viene presentato è applicabile a un qualsiasi processo di Poisson non omogeneo, e quindi anche al processo di Hawkes. Viene prima presentato il caso di un processo di Poisson omogeneo, e successivamente viene esteso il metodo al caso generale. Come visto, in un processo di Poisson omogeneo le variabili τ_i seguono una distribuzione esponenziale, e le loro funzioni di ripartizione sono quindi $F_\tau(t) = 1 - e^{-\lambda t}$. Siccome sia $F_\tau(t)$ che $F_\tau^{-1}(t)$ sono in forma chiusa, si può utilizzare il *metodo dell'inversione* per simulare i tempi degli eventi del processo. Senza entrare troppo nello specifico, se u è una variabile casuale con una distribuzione uniforme, allora la variabile $\tau^* = F_\tau^{-1}(u)$ è una variabile casuale che ha F_τ come funzione di ripartizione. Considerando che distribuzione di probabilità e funzione di ripartizione sono in relazione biunivoca, la variabile τ^* ha la distribuzione di probabilità che volevamo. Nel caso in esame, si ha $F_\tau^{-1}(u) = -\frac{\ln 1-u}{\lambda}$, con $0 \leq u \leq 1$, ma ovviamente se la variabile u ha una distribuzione uniforme allora anche la variabile $1 - u$ ha la stessa distribuzione. In definitiva simulare processi di Poisson omogenei è semplice: si simula una variabile u con distribuzione uniforme tra 0 e 1, e poi si pone $\tau = -\frac{\ln u}{\lambda}$.

Ora, per simulare eventi nel caso generale bisogna utilizzare una proprietà dei processi di Poisson. Un generico processo di Poisson con intensità λ può essere separato in due processi indipendenti tra di loro, con intensità λ_1 e λ_2, tali che $\lambda = \lambda_1 + \lambda_2$. Ogni evento del processo originale può, quindi, essere visto come un evento di uno dei due nuovi processi indipendenti. Allora, effettuando il ragionamento al contrario, se λ_1 è l'intensità di un processo non omogeneo che si vuole simulare, si può trovare una intensità λ_2 di un un altro processo non omogeneo, di cui però non si è interessati, tale che λ sia costante, ottenendo cioè un processo omogeneo.
In questo modo, è possibile simulare un processo di Poisson non omogeneo: viene simulato, innanzitutto, un processo omogeneo con intensità $\lambda^* \geq \lambda(t), \forall t$, poi, fra tutti gli eventi ottenuti, vengono accettati solo quelli che in ogni istante t fanno in modo che l'intensità del processo simulato sia in accordo con $\lambda(t)$. Si sta, cioè, ottenendo il processo di Poisson non omogeneo diradando il processo omogeneo simulato con un'intensità $\lambda^* \geq \lambda(t)$, scartando gli eventi "in più". Diventa possibile, quindi, costruire un algoritmo che svolga tutte queste operazioni, e quindi simulare un qualsiasi processo non omogeneo, tra cui anche un processo di Hawkes. La complessità temporale di tale algoritmo è $O(N^2)$, dove N è il numero di eventi simulati, ed è dovuta al fatto che per trovare il massimo di $\lambda(t)$ bisogna prima calcolare la sommatoria presente nella formula (1.10), e questo calcolo deve essere ripetuto per ogni evento che si vuole simulare. Ovviamente esistono altri algoritmi più efficienti di questo appena presentato, ma qui non vengono discussi.

L'altro grande problema che si affronta nella modellizzazione di processi auto-eccitanti è quello della stima dei parametri a partire dai dati osservati. Nel caso dei processi di Hawkes, con funzione di kernel esponenziale, si tratta di determinare la funzione $\lambda_0(t)$,

cioè l'intensità di base, e il valore dei parametri α e δ della funzione $\phi(x)$ presentata in (1.12). Uno dei metodi standard utilizzati per stimare questi parametri è il *metodo della massima verosimiglianza.*

Dato un processo di Hawkes con intensità $\lambda(t)$, siano $T_1, ..., T_N$ i tempi dei primi N eventi, si può dimostrare che, condizionatamente a $T_1, ..., T_k$, la densità di $T_{k+1} - T_k$ è:

$$f(t|T_1, ..., T_k) = \lambda(T_k)e^{-\lambda(T_k)t}, \tag{1.16}$$

cioè è esponenziale di parametro $\lambda(T_k)$. Ne segue che la densità di probabilità congiunta di $T_1, T_2 - T_1, ..., T_N - T_{N-1}$ è:

$$\begin{aligned} f(T_1, T_2 - T_1, ..., T_N - T_{N-1}) &= \prod_{k=1}^{N} \lambda(T_k)e^{-\lambda(T_k)(T_{k+1}-T_k)} \\ &= \prod_{k=1}^{N} \lambda(T_k)e^{-\int_0^{T_N} \lambda(t)dt} \end{aligned} \tag{1.17}$$

Da qui, deriva la formula per la funzione di verosimiglianza L, funzione dei diversi parametri da stimare θ, ovvero:

$$L(\theta) = \prod_{i=1}^{N} \lambda(T_i)e^{-\int_o^T \lambda(t)dt} \tag{1.18}$$

Una stima dei parametri può essere quindi trovata cercando il massimo di questa funzione rispetto a θ. Da un punto di vista computazionale, è molto più conveniente massimizzare il logaritmo della funzione di verosimiglianza:

$$l(\theta) = \log L(\theta) = -\int_0^T \lambda(t)dt + \sum_{i=1}^{N(T)} \log \lambda(T_i) \tag{1.19}$$

Visto che il logaritmo è una funzione monotona, massimizzare $l(\theta)$ automaticamente implica massimizzare anche $L(\theta)$. Massimizzare $l(\theta)$ invece che $L(\theta)$ è computazionalmente più conveniente innanzitutto perché effettuare somme è molto più facile che effettuare moltiplicazioni. Inoltre, moltiplicando tanti numeri prossimi a 0 si ottiene una funzione di verosimiglianza molto piccola, con il rischio di finire la precisione a disposizione del calcolatore. Tuttavia, esistono altri possibili problemi computazionali, che qui vengono brevemente accennati.

Come prima cosa si può notare che, nella ricerca dei massimi, si potrebbe incorrere in dei massimi locali, che potrebbero essere anche molto lontani dal massimo globale. Dal punto di vista computazionale, non può esserci la certezza che un massimo trovato sia un massimo globale, tuttavia, si possono implementare diverse idee per riuscire a ottenere con maggiore probabilità un massimo che sia effettivamente globale.

La prima idea è ripetere la stima dei parametri partendo da set di valori iniziali diversi: poiché un algoritmo deve quantomeno inizializzare i parametri θ per calcolare la funzione da massimizzare, cambiando tali valori iniziali si possono ridurre di molto le possibilità di

ottenere massimi locali. Inoltre, si possono utilizzare diversi metodi di ottimizzazione per effettuare i calcoli: se tutti i diversi metodi portano alla stima dello stesso set di parametri, è molto probabile che il massimo trovato sia il massimo globale cercato.

L'altra grossa problematica di questo metodo è la sua complessità computazionale. Nel caso di un processo di Hawkes generico, la complessità è $O(N^2)$, dove N è il numero di eventi, e quindi per un set significativo di dati il calcolo potrebbe essere intrattabile. Notiamo che per massimizzare (1.19) si possono massimizzare separatamente i due termini che compaiono, l'integrale e la sommatoria, poiché non hanno termini in comune. La complessità quadratica nasce dalla doppia sommatoria presente, considerando che anche la funzione di intensità $\lambda(T_i)$ presenta una sommatoria su tutti i diversi eventi. Tuttavia, nel caso di kernel esponenziale, si dimostra che il numero di operazioni richieste può essere diminuito fino a ottenere un algoritmo con complessità $O(N)$, rendendo il problema decisamente più trattabile.

Capitolo 2

Sincronizzazioni e instabilità nei mercati finanziari

Negli ultimi due decenni l'attività di trading nei mercati finanziari ha subito un profondo cambiamento, passando dalle vecchie conversazioni telefoniche con i broker ai nuovissimi algoritmi che sono in grado di operare in modo automatico. Inoltre, la velocità con cui le informazioni vengono processate ha subito un incremento senza precedenti, aprendo la possibilità a grandi fluttuazioni di prezzo di propagarsi molto in fretta. Tutto questo ha portato ad un effetto di sincronizzazione all'interno dei mercati, e la più grande manifestazione di questa sincronizzazione è avvenuta con il crollo del 6 maggio 2010, chiamato anche "Flash Crash". Il crollo è iniziato a causa di un rapido declino dei prezzi nel mercato E-Mini S&P 500 e in pochi minuti l'anomalia è diventata sistemica propagandosi sui vari ETF, sugli indici azionari e sulle singole azioni. Il prezzo del Dow Jones è crollato del 9% in meno di 5 minuti per poi ritornare ai valori precedenti al crollo in soli 15 minuti.
Dopo varie indagini si è scoperto che tale enorme oscillazione è stata causata da un algoritmo automatico che ha eseguito un ordine di vendita per un grande fondo comune. L'effetto è poi stato amplificato dalle operazioni di trading ad alta frequenza. Questo evento ha portato diversi studi a cercare di capire meglio l'influenza degli algoritmi automatici all'interno dei mercati finanziari. Inoltre, il crollo ha mostrato quanto fortemente interconnessi siano diventati i diversi mercati e i diversi strumenti finanziari.

2.1 Identificazione di eventi estremi

Si può, quindi, cercare di studiare come la frequenza di instabilità collettive avvenute in brevi periodi temporali sia cambiata nel corso degli ultimi anni. Riprendendo i risultati di [1], si identificano e si esaminano gli eventi estremi che si sono verificati in un minuto, cioè quegli eventi che hanno mostrato grandi fluttuazioni di prezzo all'interno di un minuto per poi tornare a valori normali poco dopo. L'analisi che viene presentata tiene in considerazione i dati registrati tra il 2001 e il 2013 su un campione di azioni americane ad alta liquidità. Le azioni analizzate sono presenti nell'indice Russell 3000. Per ogni anno vengono prese in considerazione le 140 azioni con la più alta liquidità e viene analizzato il prezzo di chiusura ad ogni minuto durante l'intera sessione del mercato americano. Inoltre,

a questi dati viene applicato un filtro per eliminare i pattern di volatilità infra-giornalieri.

Analizzando i dati nella loro globalità si notano, in effetti, degli evidenti pattern giornalieri. L'ampiezza delle fluttuazioni dei prezzi esibisce delle differenze significative durante i vari momenti della giornata di contrattazione, formando una tipica forma ad U. I movimenti più ampi avvengono all'inizio o alla fine della giornata, mentre durante le ore intermedie le fluttuazioni sono più contenute. Per eliminare questo pattern, e ottenere dei dati che non presentano una struttura periodica, si opera in un modo standard. Si calcola prima la media, su tutti i giorni, delle diverse fluttuazioni di prezzo, riscalate in base alla volatilità giornaliera. Riscalare le fluttuazioni in base alla volatilità fa in modo che grandi fluttuazioni in una giornata a bassa volatilità abbiano un peso maggiore nel calcolo della media. Una volta calcolata questa media, e cioè una volta definito il pattern giornaliero, si filtrano i dati originari dividendo la loro fluttuazione per il pattern ottenuto. In questo modo, una grande variazione di prezzo, in un orario che in media non ne presenta, ha una rilevanza nettamente maggiore rispetto alla stessa variazione di prezzo che tuttavia avviene, per esempio, all'apertura dei mercati in cui è frequente osservare grandi oscillazioni.

Per analizzare in maniera sistematica la successione dei prezzi P_t, si definisce la variabile $r_t = \ln \frac{P_t}{P_{t-1}}$ e la si confronta con una stima della volatilità media. Si possono usare diversi metodi per stimare la volatilità media, anche se il risultato finale è sostanzialmente invariato. Per esempio, si può definire la successione della volatilità al tempo t, σ_t, come una media mobile esponenziale della successione dei prezzi. Più nel dettaglio, ricordando la definizione di r_t di sopra, si può definire la media mobile esponenziale, dipendente da un parametro α, come:

$$\sigma_t = \sqrt{\frac{\pi}{2}}\alpha \sum_{i>0}^{t} (1-\alpha)^{i-1} |r_{t-i}|, \tag{2.1}$$

dove in questo caso è stato fissato $\alpha = \frac{2}{D+1}$, con $D = 60$. Per ulteriori informazioni riguardo alla media mobile esponenziale e a come viene calcolato il valore della volatilità, si faccia riferimento a [2].
Avendo definito sia r_t che σ_t, si dice che è avvenuto un evento estremo al tempo t se

$$\frac{|r_t|}{\sigma_t} > \theta, \tag{2.2}$$

per un certo valore di soglia θ. Nelle analisi qui presentate viene utilizzato quasi sempre $\theta = 4$, tranne in alcuni casi in cui viene esplicitamente specificato l'uso di $\theta = 6,\ 8,\ 10$.

Si dice che un'azione *salta* in un dato minuto t se la sua successione dei prezzi soddisfa la condizione (2.2) per un dato θ. Con *salti coordinati* di molteplicità M, detti anche *cojumps* in inglese, viene invece indicato il simultaneo avvenimento di un salto da parte di M azioni presenti in un sottoinsieme delle azioni analizzate. La molteplicità M dei salti coordinati fornisce, quindi, una misura della natura sistemica dell'evento rilevato, ovvero è una stima di quanto le azioni analizzate siano sincronizzate tra di loro. Un salto con alta molteplicità indica che la sincronizzazione all'interno dell'insieme analizzato è molto elevata per quel particolare evento, mentre se la molteplicità è bassa significa che l'evento si è propagato lievemente.

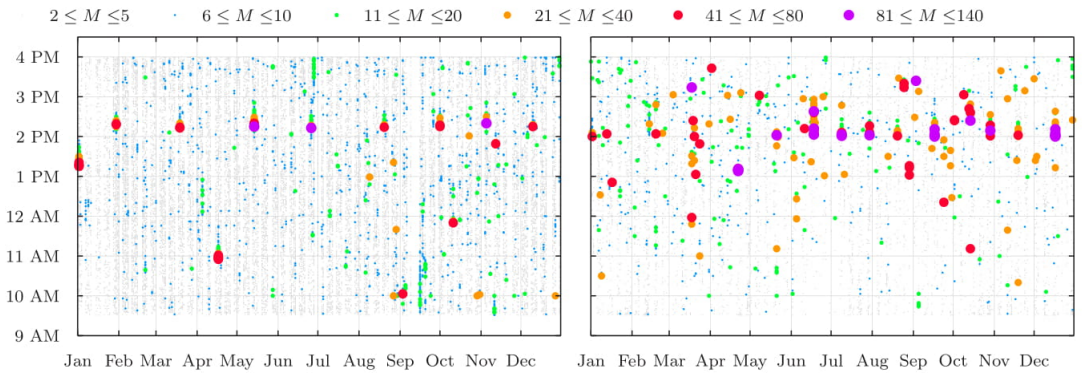

Figura 2.1: Serie temporale dei salti coordinati rilevati in 140 azioni ad alta liquidità nel 2001 (figura a sinistra) e nel 2013 (figura a destra). La dimensione dei cerchi aumenta con la molteplicità dell'evento.

2.2 Dinamica degli eventi negli ultimi anni

Come dettagliatamente discusso in [1], la dinamica dei salti coordinati tra i primi anni 2000 ed oggi è profondamente diversa.
Una rappresentazione grafica di come la instabilità dei mercati finanziari sia cambiata negli ultimi anni è fornita in Figura 2.1, dove viene visualizzata la dinamica dei salti coordinati per $\theta = 4$ nel 2001 e nel 2013. Il grafico riporta in ascissa i diversi giorni dell'anno e in ordinata l'orario del giorno, segnando con dei cerchi colorati i salti con varia molteplicità. Come è evidente, nel 2001 erano presenti molti più salti in termini assoluti rispetto al 2013, e i salti con grande molteplicità erano collocati in uno specifico orario del giorno, corrispondentemente a importanti notizie macroeconomiche quali gli annunci del FOMC. Nel 2013, invece, anche se il numero totale di salti coordinati è diminuito, si osservano in numero nettamente maggiore i salti con una alta molteplicità, i quali sono inoltre distribuiti in maniera più casuale durante tutta la giornata di contrattazione e non più in specifici orari come nel 2001. Questi dati sono una prima indicazione che i moderni mercati finanziari sono diventati più sincronizzati, nel senso che le instabilità si presentano in modo più sistemico, e che queste instabilità sono meno collegate alle notizie macroeconomiche.

Più evidentemente, nel grafico in alto a sinistra di Figura 2.2 è mostrato il numero di minuti in cui è stato rilevato almeno un salto. Si vede in maniera chiara che, qualsiasi sia il valore di θ scelto, il numero totale di salti è leggermente diminuito nel corso degli anni. Tuttavia emerge un comportamento nettamente diverso se si considera la dinamica dei salti coordinati. In alto a destra della Figura 2.2 sono mostrate le frequenze dei salti coordinati per diversi valori della molteplicità (normalizzate ai valori del 2001). Nonostante, quindi, il numero totale di salti sia diminuito nel corso degli anni, è altrettanto evidente come i salti ad alta molteplicità siano diventati anche 10 volte più frequenti. Questo risultato non cambia se viene fissato il numero minimo della molteplicità, se vengono cioè considerati solamente i salti con $M \geq 30$, ma viene cambiato il valore di soglia θ, come mostrato nel grafico in basso a sinistra. Viene osservato chiaramente come il numero di salti con alti

valori di M sia cresciuto negli ultimi anni, indipendentemente dal θ scelto.

Infine, nel grafico in basso a destra della Figura 2.2 è rappresentata in scala logaritmica quella che viene chiamata *funzione di sopravvivenza*, cioè il complemento della funzione di ripartizione, della molteplicità dei salti coordinati. Tale funzione è definita come $\overline{F}_X(M) = \mathbb{P}(X > M) = 1 - F_X(M)$, dove $F_X(M)$ è l'ordinaria funzione di ripartizione della variabile aleatoria X. In sostanza, $\overline{F}_X(M)$ indica la probabilità che un evento, cioè un salto, abbia una molteplicità maggiore al valore M. Banalmente, $\overline{F}_X(0) = 1$, ovvero qualsiasi salto rilevato ha una molteplicità maggiore a 0, la quale per definizione vale al minimo 1.
Quello che emerge è un chiaro comportamento descrivibile con una legge di tipo polinomiale, la quale implica che la probabilità di salti coordinati è ampia anche per eventi altamente sistemici, cioè per salti ad alta molteplicità. Consistentemente con le altre osservazioni, la coda della distribuzione, cioè la regione con un alto M, è diventata più spessa nel corso degli anni, dimostrando nuovamente che il numero di salti ad alta molteplicità è cresciuto. Infine, la flessione della distribuzione in prossimità del valore $M = 140$ è da attribuire al fatto che l'insieme analizzato presentava 140 azioni, e quindi banalmente $\overline{F}_X(140) = 0$.

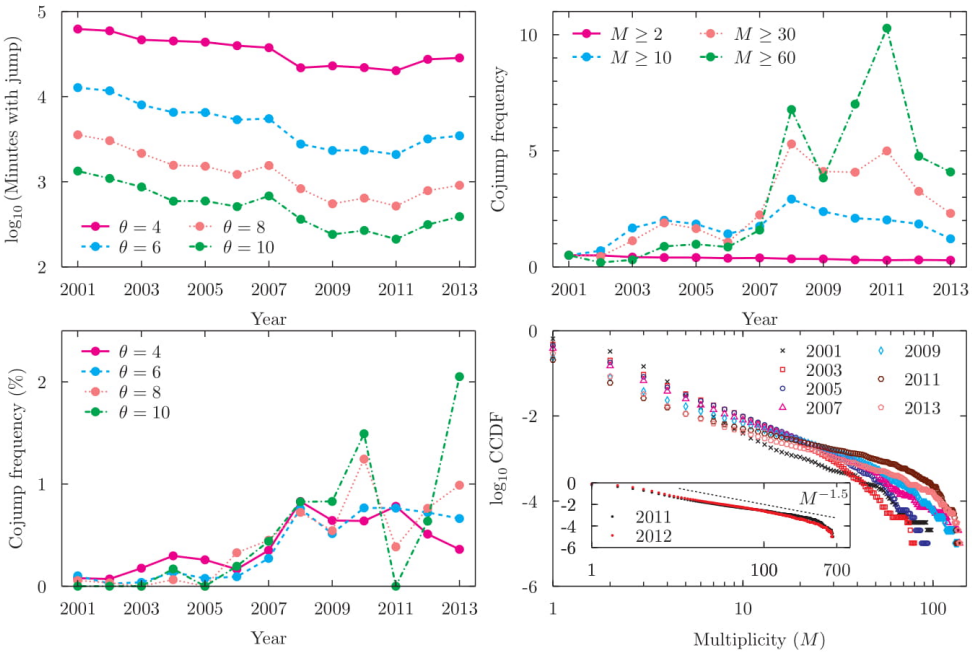

Figura 2.2: In alto a sinistra: numero totale di minuti in cui è stato osservato almeno un salto tra le 140 azioni analizzate, nel corso degli anni, per vari valori di θ.
In alto a destra: per $\theta = 4$, evoluzione nel corso degli anni della frazione di minuti con almeno un salto di molteplicità $M \geq 2, 10, 30, 60$ rispetto ai valori del 2001.
In basso a sinistra: evoluzione nel corso degli anni del rapporto in percentuale tra numeri di salti coordinati con $M \geq 30$ e numero di salti coordinati totali, per diversi valori di θ.
In basso a destra: grafico in scala logaritmica del complemento della funzione di ripartizione della molteplicità dei salti coordinati per diversi anni.

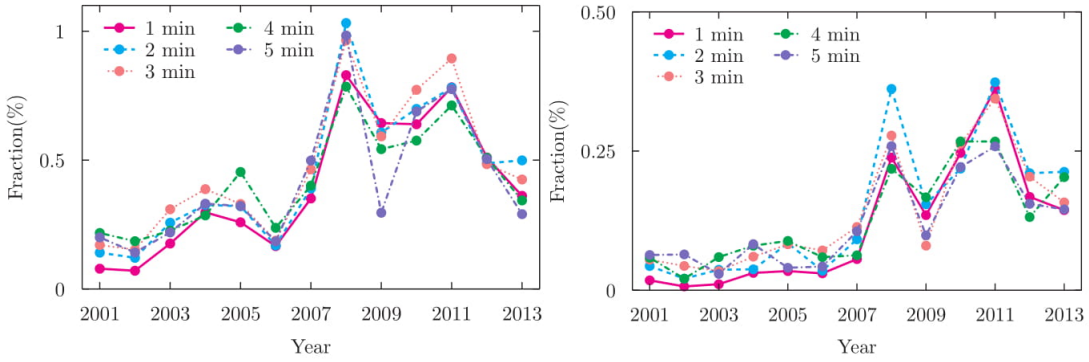

Figura 2.3: Evoluzione nel corso degli anni della frazione di salti coordinati di molteplicità $M \geq 30$ (a sinistra) e $M \geq 60$ (a destra) rispetto al numero totale di salti, per $\theta = 4$, per diversi orizzonti temporali.

Una possibile critica a quanto detto è che un minuto nel 2013 non è equivalente a un minuto nel 2001 in termini di attività del mercato. Quindi, è importante verificare se l'aumento del numero di salti ad alta molteplicità sia semplicemente dovuto al fatto che negli anni passati la sincronizzazione avveniva per tempi superiori al minuto, a causa della più bassa velocità delle operazioni di compra-vendita, oppure è un fenomeno indipendente dalla velocità dei sistemi di comunicazione. Per verificare questa ipotesi, si ripetono le analisi fin qui presentate variando la scala temporale per la rilevazione dei salti da uno fino a cinque minuti. In figura 2.3 viene mostrato come il numero di salti di molteplicità $M \geq 30$ o $M \geq 60$ sia cambiato nel corso degli anni, avendo scelto $\theta = 4$. Si vede che la scelta dell'intervallo temporale è irrilevante per il fenomeno rilevato: a qualsiasi scala temporale, da uno a cinque minuti, è evidente che il numero di salti ad alta molteplicità sia cresciuto nel corso degli anni. Per di più, il numero di salti al alta molteplicità in un minuto nel 2013 è superiore al numero di salti al alta molteplicità in cinque minuti nel 2001. Questo risultato dimostra che l'aumento della velocità nelle transazioni del mercato non ha effetti significativi sull'effetto sistemico dei salti coordinati degli ultimi anni, che appare quindi come un fenomeno endogeno del sistema.

2.2.1 Dipendenza dalle notizie macroeconomiche

Una volta rilevata la dinamica dei salti coordinati negli ultimi anni, si può cercare di capire quale percentuale ha un'origine esterna e quale percentuale possiede invece un'origine endogena. Per analizzare accuratamente la questione, si cerca di studiare quanto frequentemente un salto coordinato sia preceduto da una notizia macroeconomica programmata. Si tengono in considerazione solamente notizie macroeconomiche generali, e non notizie relative a una singola azione, perché è inverosimile che una notizia relativa a una particolare azienda condizioni tutto il mercato.
A questo proposito, vengono considerate 42 delle più importanti categorie di notizie macroeconomiche diffuse, quali ad esempio le conferenze FOMC che, tra le tante cose, decide i tassi di interesse della Fed. Tra queste, vengono tenute in considerazione solamente le 27 categorie che vengono annunciate mentre il mercato è ancora aperto, scartando ovvia-

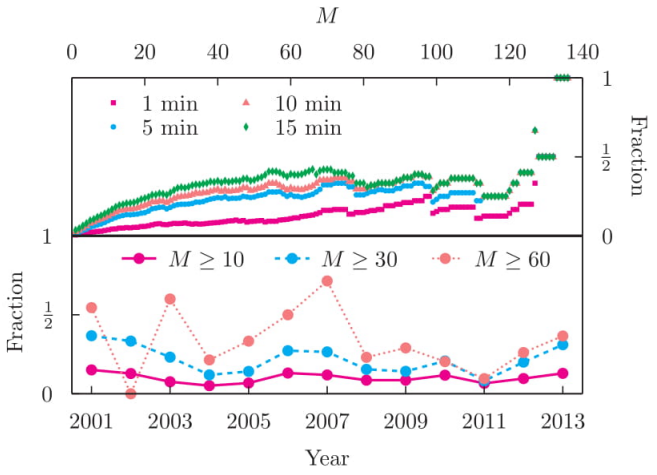

Figura 2.4: In alto: frazione di salti coordinati, nel 2012, con molteplicità maggiore o uguale al valore in ascissa preceduti da una notizia macroeconomica negli ultimi $1, 5, 10, 15$ minuti rispetto al numero totale.
In basso: frazione di salti coordinati con diversa molteplicità per i quali è stata registrata almeno una notizia nei 5 minuti precedenti al salto.

mente le notizie divulgate a mercato chiuso. In questo modo sono state considerate un totale di 2888 notizie tra il 2001 e il 2013.

Nella figura 2.4, in alto, viene analizzata la frazione di salti coordinati, di molteplicità almeno M, che sono preceduti da una notizia macroeconomica negli ultimi $1, 5, 10, 15$ minuti prima della rilevazione del salto stesso. Come si legge dal grafico, meno del 40% dei salti ad alta molteplicità viene preceduto da una importante notizia macroeconomica. Inoltre, si nota che i pattern rilevati per $5, 10, 15$ minuti sono simili tra loro, indicando che se una notizia macroeconomica fa saltare il mercato, questo avviene tipicamente nei primi 5 minuti dalla pubblicazione della stessa.

Per avere una prospettiva storica, nel grafico in basso della figura 2.4, viene mostrato che la frazione di salti coordinati innescati da una notizia macroeconomica è rimasta più o meno costante nel corso degli anni, per diversi valori della molteplicità. L'analisi, quindi, mostra che una grande percentuale dei salti sistemici che vengono rilevati non sono associati alla pubblicazione di una qualche importa notizia macroeconomica. Inoltre, notizie relative a una specifica azienda potrebbero avere un ruolo soltanto per i salti a bassa molteplicità, in quanto è difficile credere che le notizie di una sola azienda possano influenzare tutto il mercato.

Concludiamo quindi che, per salti ad alta molteplicità, devono essere presenti dei meccanismi endogeni all'interno del sistema che giocano un ruolo decisivo nella sincronizzazione delle diverse azioni analizzate.

Capitolo 3

Modellizzazione del fenomeno

Le evidenze empiriche del capitolo precedente suggeriscono che una fetta importante dei salti coordinati è indipendente dalle notizie macroeconomiche. Inoltre, le instabilità dei mercati tendono a propagarsi velocemente ad altri strumenti finanziari. Per questo motivo, per modellizzare accuratamente il fenomeno riscontrato, è importante studiare se e come le instabilità sistemiche inneschino altre instabilità negli istanti successivi, negli stessi o anche in diversi strumenti finanziari. È evidente, infatti, che non si sono mai specificate singole azioni, ma si è sempre parlato di molteplicità in senso generico: non è importante quali azioni abbiano effettuato un salto, quello che interessa è solamente il numero di azioni che lo hanno fatto.

Un primo approccio per costruire un modello è stato effettuato in [2], in cui ogni azione viene rappresentata da un processo di punto, dove ogni salto dell'azione rappresenta un evento del processo. L'intensità di questo processo è data dalla somma di un fattore comune a tutto il mercato e da un termine esogeno. Sia il fattore comune del mercato che il termine esogeno vengono modellizzati da un processo di Hawkes.
Questo approccio è molto efficace nel descrivere i salti con $M = 2$, ma al crescere di M il modello mostra le sue debolezze. Si vede che, ad esempio, nel limite di un infinito numero di azioni N, il modello usato in [2] predice una distribuzione delle molteplicità di tipo gaussiano, in contrasto con l'evidente legge polinomiale osservate in Figura 2.2.

Uno dei problemi concettuali che presenta il modello precedente è che tenta di mantenere l'informazione su quale azione abbia effettuato un salto, modellizzando ogni singola azione con un diverso processo di punto. Un approccio migliore al problema è, invece, modellizzare direttamente un vettore di molteplicità, perdendo l'informazione sull'identità delle azioni che hanno saltato, ma concentrandosi solamente sul loro numero.

3.1 Processo di Hawkes multivariato

Nello specifico, si considera un processo di punto N-dimensionale, caratterizzato dal vettore delle intensità $\boldsymbol{\lambda}_t$. Un evento nella componente i del processo al tempo t indica che, al tempo t, è avvenuto un salto coordinato di molteplicità i. In questo modo, come si voleva, si hanno informazioni solamente sul numero di azioni che hanno effettuato un salto, ma

non è più possibile risalire quali tra le N azioni lo hanno fatto. A questo scopo, si usa un processo di Hawkes multivariato di dimensione N.

Definizione 3.1. *Un processo di Hawkes N-dimensionale è un processo di punto caratterizzato dal vettore delle intensità* $\boldsymbol{\lambda}_t := (\lambda_t^1, ..., \lambda_t^N)^T$*, dove ogni componente* i *soddisfa la relazione*

$$\lambda_t^i = \mu_t^i + \sum_{j=1}^{N} \sum_{T_k^j < t} \nu_j^i (t - T_k^j), \tag{3.1}$$

dove μ_t^i *e* ν_j^i *sono funzioni positive e deterministiche per ogni* $i, j = 1, ..., N$*, mentre la successione* $(T_k^j)_{k>0}$ *corrisponde alla sequenza temporale degli eventi associata alla componente* j *del processo di punto N-dimensionale.*

Ovviamente, se $\mu_t^i = \mu^i$ è una costante e la funzione di kernel ν_j^i è nulla, il processo di Hawkes si riduce a un semplice processo di Poisson N-dimensionale. Altrimenti, se la funzione di kernel è positiva, ogni volta che avviene un evento per una qualsiasi delle componenti del processo multidimensionale, l'intensità λ_t^i cresce di un valore proporzionale al kernel e al tempo trascorso dall'evento.

3.1.1 Scelta della parametrizzazione

Come in ogni problema multidimensionale, la modellizzazione è problematica per via del grosso numero di parametri liberi. Per questo motivo, bisogna fare delle assunzioni per cercare di ridurre al minimo i parametri, senza che il modello perda la sua efficacia.

Innanzitutto, si assume che il vettore $\boldsymbol{\mu} := (\mu_t^1, ..., \mu_t^N)^T$ sia indipendente dal tempo. Inoltre, per la funzione di kernel, si utilizza una delle parametrizzazioni più comuni in termini di funzioni esponenziali:

$$v_j^i(t - T_k^j) := \alpha_{ij} e^{-\beta_{ij}(t - T_k^j)}, \tag{3.2}$$

con $\alpha_{ij} > 0$ e $\beta_{ij} > 0$ per ogni i, j. Il parametro α_{ij} fissa la scala dell'intensità del processo λ^i, determina cioè di quanto un evento di molteplicità j al tempo T_k^j modifichi l'intensità del processo di tipo i. Il parametro β_{ij}, invece, descrive l'inverso del tempo necessario perché il processo i perda la memoria riguardo a un evento avvenuto nel processo j.
Come nel caso unidimensionale presentato precedentemente in (1.12), anche per α_{ij} e β_{ij} esiste una condizione simile a $\alpha/\delta < 1$. Se definiamo la matrice Γ di elementi $\Gamma_{ij} = \frac{\alpha_{ij}}{\beta_{ij}}$, la condizione da imporre è che il suo raggio spettrale (ovvero l'estremo superiore del modulo dei suoi autovalori) sia strettamente minore di 1.
Sempre analogamente al caso unidimensionale, si può dimostrare una formula che è una generalizzazione di (1.15). Nel caso in cui il raggio spettrale di Γ sia minore di 1, si ha, infatti, che il valore atteso delle intensità è:

$$\mathbb{E}[\boldsymbol{\lambda}_t] = (\mathbb{I} - \Gamma)^{-1} \boldsymbol{\mu}, \tag{3.3}$$

dove $\mathbb{I}$ è la matrice identità della dimensione di Γ.

In questo modo, il modello dipende dal vettore $\boldsymbol{\mu}$, di dimensione N, e dalle due matrici α_{ij} e β_{ij}, di dimensione $N \times N$. Per poter ridurre drasticamente la dimensione del problema da $N + 2N^2$ a un numero ragionevole di parametri, vengono effettuate diverse ulteriori assunzioni.

- Si assume che tutte le β_{ij} siano uguali a una costante $\beta > 0$. In questo modo, c'è un'unica scala temporale che caratterizza il decadimento dei vari kernel, ovvero tutte le componenti i del processo hanno la stessa memoria.
- Si impone la condizione $\boldsymbol{\mu} = \eta\mathbb{E}[\boldsymbol{\lambda}_t]$, con $0 < \eta < 1$. Questa condizione è giustificata dal fatto che uno degli obiettivi principali del modello è la capacità di riprodurre la distribuzione del vettore di molteplicità osservata empiricamente. Con questa condizione, infatti, si impone che la distribuzione del processo esogeno, ovvero la distribuzione degli eventi che hanno un'origine esterna, sia proporzionale alla distribuzione dell'intero processo osservato empiricamente. In questo modo, quindi, l'eccitazione reciproca tra le diverse componenti del processo di Hawkes N-dimensionale non deve cambiare la struttura della distribuzione dell'intero processo.

  Si nota che questa assunzione, insieme a (3.3), implica che:

  $$\Gamma\mathbb{E}[\boldsymbol{\lambda}_t] = (1 - \eta)\mathbb{E}[\boldsymbol{\lambda}_t], \tag{3.4}$$

  cioè $\mathbb{E}[\boldsymbol{\lambda}_t]$ (o equivalentemente $\boldsymbol{\mu}$) è un autovettore di Γ di autovalore $1 - \eta$.
  Ora, siccome per definizione tutti i valori della matrice Γ sono strettamente positivi, si può applicare il teorema di Perron-Frobenius. Quindi esiste un unico autovettore con tutte le componenti strettamente positive, e l'autovalore associato a tale autovettore coincide con il raggio spettrale. Siccome $\mathbb{E}[\lambda_t^i] > 0$ per ogni $i = 1, ..., N$ per la definizione di intensità λ_t^i, si può concludere che il raggio spettrale coincide con l'autovalore associato a $\mathbb{E}[\boldsymbol{\lambda}_t]$, cioè $1 - \eta$.
  Per come è definito η, cioè il rapporto tra l'intensità del processo esogeno e l'intensità del processo effettivamente osservato, si può dire che η rappresenta la frazione di eventi spiegabile tramite il processo esogeno, e di conseguenza $1 - \eta$, cioè il raggio spettrale, rappresenta la frazione di eventi spiegabile esclusivamente tramite l'eccitazione reciproca tra le diverse componenti del processo. Quindi, solamente calcolando gli autovalori della matrice Γ, e determinando di conseguenza il suo raggio spettrale, è possibile stimare quale percentuale degli eventi abbia un'origine esogena e quale percentuale, invece, derivi dal meccanismo di eccitazione interno al processo.
- Infine, dopo aver ridotto i parametri relativi alla matrice β_{ij} e al vettore $\boldsymbol{\mu}$, bisogna dare una forma anche alla matrice α_{ij}. Siccome si ha $\alpha_{ij} = \beta\Gamma_{ij}$, basta parametrizzare la matrice Γ che, quindi, descrive l'effetto dell'eccitazione della componente j sulla componente i del processo. Si impone, quindi, che ogni elemento della matrice Γ sia il prodotto di due termini: un termine D_{ii} che dipende solamente dalla componente i su cui si vuole studiare l'effetto del salto avvenuto nella componente j, e di un termine $\sigma(|i - j|)$ che dipende dalla differenza tra le due molteplicità. In questo modo, si può scrivere la matrice Γ come il prodotto di due matrici: $\Gamma = D\Sigma$, dove D è una matrice diagonale di elementi:

  $$D_{ii} := \frac{(1 - \eta)\mu^i}{\sum_{j=1}^{N} \mu^j \sigma(|i - j|)}, \tag{3.5}$$

mentre la matrice Σ è definita come

$$\Sigma_{ij} = \sigma(|i-j|) = (|i-j|+1)^{-\gamma}. \tag{3.6}$$

Scrivendo Γ in questi termini si vede come, in realtà, la matrice dipende dal solo parametro γ, il quale esprime, in un certo senso, la forza con cui avviene l'eccitazione reciproca tra eventi di diversa molteplicità. Ovviamente, il contributo maggiore all'intensità λ^i del processo è dato dal termine di auto-eccitazione, α_{ii}, in cui banalmente $\Sigma_{ii} = 1$. Inoltre, salti coordinati di una data molteplicità influenzano maggiormente salti di molteplicità simile.

In definitiva, il modello presentato è completamente determinato da soli tre parametri, η, β e γ, e da $\mathbb{E}[\boldsymbol{\lambda}_t]$. Tuttavia, è possibile ricavare $\mathbb{E}[\boldsymbol{\lambda}_t]$ a partire dal numero empirico di eventi osservati con una data molteplicità: per il significato di intensità, infatti, il prodotto tra il valor medio dell'intensità e l'intervallo di tempo durante il quale si osserva il processo, deve tendere al numero di eventi osservati. Quindi $\mathbb{E}[\boldsymbol{\lambda}_t]$ è un dato empirico, e in particolare è semplicemente il rapporto tra il numero di eventi osservati e il tempo durante il quale si è osservato il processo, e il modello è parametrizzato da soli tre parametri.

3.1.2 Stima dei parametri

Una stima rigorosa dei tre parametri del modello tramite il metodo della massima verosimiglianza pone diversi problemi computazionali. Si può utilizzare, perciò, un procedimento euristico basato sui momenti. Si considerino a questo proposito le seguenti due quantità:

$$f_\tau^{(1)}(M;J) := \mathbb{P}[\exists t^{'} \in (t, t+\tau] \ t.c. \ M_{t'} \geq J | M_t \geq M], \tag{3.7}$$

$$f_\tau^{(2)}(M) := \mathbb{E}[M_{t'} | M_t \geq M, \exists t^{'} \in (t, t+\tau] \ t.c. \ M_{t'} > 0]. \tag{3.8}$$

La prima, $f_\tau^{(1)}(M;J)$, è la probabilità, condizionata all'osservazione di un evento di molteplicità almeno M al tempo t, di osservare un nuovo evento sistemico di molteplicità almeno J nell'intervallo $(t, t+\tau)$. In altre parole, misura la probabilità che un salto di molteplicità almeno M inneschi un nuovo evento sistemico in un breve intervallo temporale. La seconda quantità, $f_\tau^{(2)}(M)$, invece, è la media della molteplicità dei salti all'interno di un intervallo temporale di lunghezza τ, condizionata al fatto di aver osservato un evento di molteplicità almeno M al tempo t. In altre parole, misura la tipica molteplicità del salto innescato direttamente da un salto di molteplicità almeno M.

Le due quantità appena definite vengono utilizzate, quindi, per stimare i parametri η, γ e β del modello tramite il metodo dei minimi quadrati. Siccome non si è in grado di calcolare analiticamente i momenti di $f_\tau^{(1)}(M;J)$ e $f_\tau^{(2)}(M)$ a partire dal modello, viene effettuata una simulazione Monte Carlo con i tre parametri fissati. A questo punto, data la molteplicità M, si possono calcolare la media, $a_m^{(i)}$ $(i = 1, 2)$, e la deviazione standard, $\delta_m^{(i)}$ $(i = 1, 2)$ sia di (3.7) che di (3.8). Inoltre si hanno a disposizione media $(a_d^{(i)})$ e deviazione standard $(\delta_d^{(i)})$ dei dati empirici raccolti. In questo modo, per ogni $i = 1, 2$, si

può definire una funzione di perdita:

$$\chi^2_{(i)} = \sum_{M \in S} \frac{(a_d - a_m)^2}{\delta_d^2 + \delta_m^2}, \tag{3.9}$$

dove la somma viene effettuata su un insieme S di molteplicità. Infine, si costruisce la funzione di perdita totale, che viene definita come $\chi^2_{(1)} + 0.5\chi^2_{(2)}$, cioè la somma pesata delle funzioni di perdita relative a $f_\tau^{(1)}(M;J)$ e $f_\tau^{(2)}(M)$. A questo punto, si ottiene la stima dei parametri cercati minimizzando la funzione di perdita così definita.

Da un punto di vista pratico, si cerca il minimo della funzione variando i parametri a passi di 0.05. Viene considerato nel dettaglio il caso di $N = 140$ azioni ad alta liquidità dell'indice Russell 3000 nel 2013. Per la calibrazione del modello viene fissato il valore $J = 10$ nell'equazione (3.7), e il valore dell'intervallo temporale di riferimento $\tau = 5$ sia in (3.7) che in (3.8). Come insieme delle molteplicità su cui fare la somma si sceglie $S = \{5, 10, 15, ..., 65, 70\}$. In questo modo, si trova un chiaro minimo corrispondente ai valori $\eta = 0.15$, $\beta = 0.6$, $\gamma = 2.65$. Una volta stimati i tre parametri, si può testare la validità del modello sulla funzione $f_\tau^{(1)}(M;J)$ con $J = 30$ e $J = 60$.

3.2 Risultati del modello

Si possono ora analizzare meglio i valori ottenuti per i parametri. Il valore $\eta = 0.15$ indica che solamente il 15% dei salti coordinati rilevati è spiegabile tramite fattori esterni, e che, quindi, l'85% dei salti è provocato da un meccanismo di eccitazione interno al sistema. Inoltre, la scala temporale tipica della memoria che possiede il processo è $\tau^* = 1/\beta \simeq 1.67$ minuti, e, quindi, si può assumere che dopo un periodo superiore a $3\tau^*$ ($\simeq 5$ minuti) il processo abbia quasi completamento perso la memoria riguardo a un salto precedente.

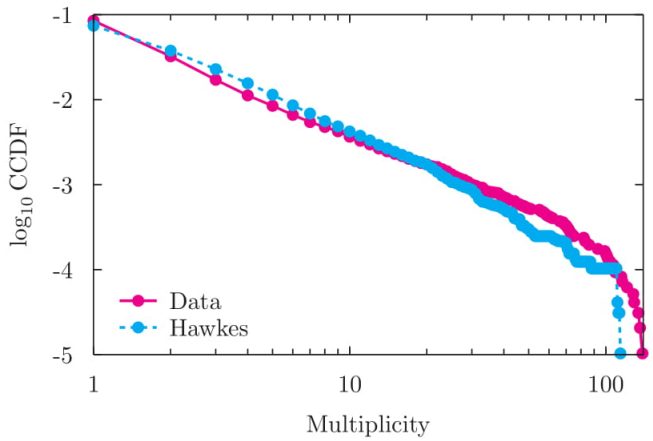

Figura 3.1: Rappresentazione in scala logaritmica del complemento della funzione di ripartizione (o *funzione di sopravvivenza*) della molteplicità dei salti coordinati. La linea piena corrisponde alla distribuzione empirica misurata dal campione di 140 azioni durante il 2013. La linea tratteggiata è la distribuzione ottenuta dalla simulazione del processo di Hawkes multidimensionale.

Infine, il valore relativamente basso di γ indica una forte eccitazione anche tra eventi con un diverso valore della molteplicità.

Analizzando, poi, le simulazioni ottenute utilizzando i parametri stimati, come prima cosa, si vede, in Figura 3.1, che il modello utilizzato riproduce correttamente la funzione di ripartizione della molteplicità dei salti. Ovviamente questo non è sorprendente, in quanto una delle assunzioni che erano state fatte per ridurre il numero di parametri del modello era proprio quella di imporre $\boldsymbol{\mu} = \eta\mathbb{E}[\boldsymbol{\lambda}_t]$, cioè imporre che la distribuzione del processo esogeno coincidesse con la distribuzione osservata. Inoltre, gli eventi del processo che sono generati da correlazioni interne sono regolati dalla matrice Γ, e anch'essa è determinata univocamente da $\mathbb{E}[\boldsymbol{\lambda}_t]$. Quindi, anche la distribuzione degli eventi del processo dovuti alle correlazioni tra le diverse componenti di $\boldsymbol{\lambda}_t$ dipende dai valori empirici. Perciò, visto che sia il fattore esogeno che quello interno dipendono univocamente dai dati empirici utilizzati per la calibrazione, c'è un'ottima accuratezza nel riprodurre la funzione di ripartizione.

La potenza del modello tuttavia appare evidente osservando la Figura 3.2, dove vengono riportate le quantità $f_\tau^{(1)}(M;J)$ e $f_\tau^{(2)}(M)$ precedentemente definite. La linea continua

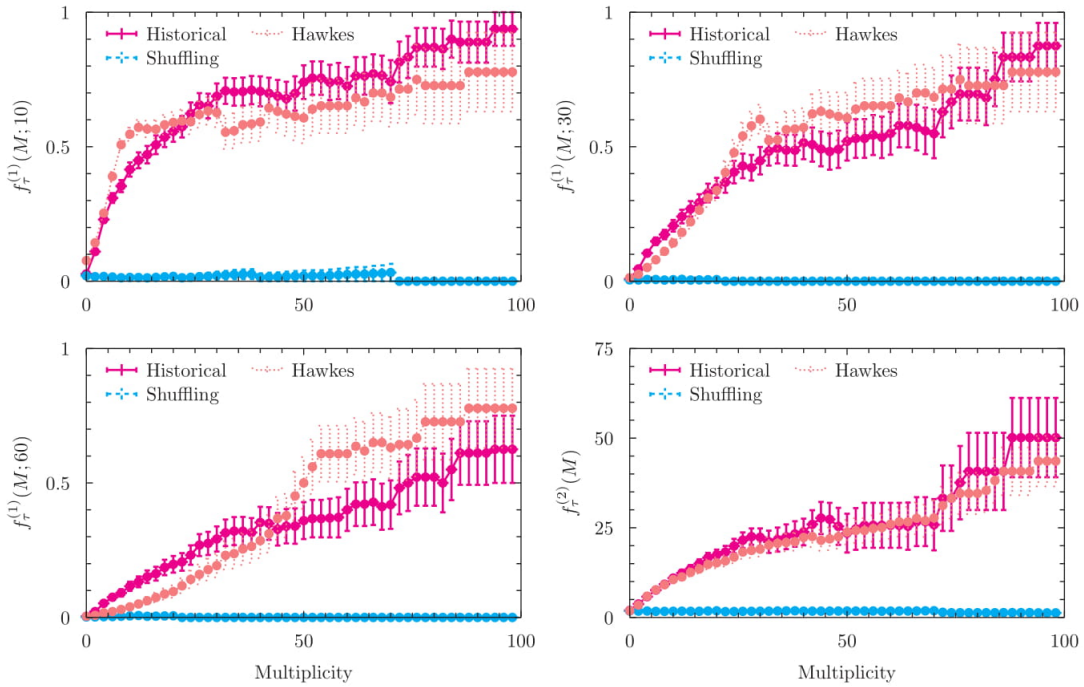

Figura 3.2: In alto a sinistra: probabilità che un salto di molteplicità maggiore o uguale a 10 avvenga in un intervallo di $\tau = 5$ minuti, dopo che al tempo t è stato osservato un salto di molteplicità $M_t \geq M$. Vengono presentati sia i dati storici che i dati simulati. Le barre di errore rappresentano la deviazione standard.
In alto a destra e in basso a sinistra: valore di soglia della molteplicità 10 sostituito con 30 e 60 rispettivamente.
In basso a destra: molteplicità attesa dei salti in un intervallo di $\tau = 5$ minuti seguente un salto di molteplicità $M_t \geq M$.

corrisponde, in tutti i grafici, alle probabilità misurate direttamente dai dati storici, mentre la linea tratteggiata è il risultato della simulazione del modello di Hawkes. Infine, è presente anche un modello di riferimento in cui viene effettuato un rimescolamento delle serie temporali della molteplicità. Viene cioè rappresentata, come riferimento, la simulazione di un modello che non tiene in considerazione la dipendenza temporale tra i diversi salti, rimescolando casualmente i tempi in cui i vari salti sono avvenuti. È quindi evidente come, perdere l'informazione sull'ordine temporale dei salti, ignorando quindi la correlazione tra salti temporalmente vicini tra di loro, porti a una descrizione completamente irrealistica del processo osservato.

Il modello di Hawkes, al contrario, si adatta bene ai dati osservati e, quindi, descrive adeguatamente il meccanismo di eccitazione tra i diversi salti sistemici rilevati. Si nota una leggera discrepanza per il valore $J = 60$, ma la forma generale della curva e i suoi livelli sono comunque ben rappresentati. Il modello di Hawkes, insomma, è un grande miglioramento rispetto al modello di riferimento riportato, che non tiene conto dell'eccitazione reciproca tra un salto e un altro. Inoltre, dai grafici è evidente che, maggiore è la molteplicità di un evento, maggiore è la probabilità che negli istanti successivi si verifichi un ulteriore evento sistemico ad alta molteplicità.

3.3 Conclusioni

Riepilogando, dall'analisi di un insieme di azioni ad alta liquidità quotate sul mercato americano, si è notato che dal 2001 ad oggi il numero di eventi estremi, cioè di salti, è diminuito in modo apprezzabile. Tuttavia, al contrario, è aumentato il numero medio della molteplicità di questi salti, indicando chiaramente che, oggi, quando avviene un salto, questo si propaga velocemente a diversi strumenti finanziari. I mercati, cioè, sono diventati più interconnessi e presentano una forte sincronizzazione interna. Si è visto che solo una frazione (al massimo fino al 40%) di questi salti coordinati tra diverse azioni ha una spiegazione di tipo esogeno, cioè provocata da notizie e fattori macroeconomici. Il rimanente 60% suggerisce che ci sia in atto un meccanismo interno al mercato stesso.

Una delle ipotesi più accreditate per spiegare come mai la sincronizzazione sia aumentata in questo modo negli ultimi anni, è che il numero degli algoritmi di trading che operano in automatico sui mercati è cresciuto notevolmente. Grazie alla innovazione tecnologica, e quindi grazie alla maggior velocità di trasferimento delle informazioni, è oggi possibile una rapida propagazione di ampi movimenti di prezzo tra diversi strumenti finanziari. Inoltre, si è evidenziato come le instabilità sistemiche che si generano in questo modo hanno un duplice effetto: aumentano la probabilità di un ulteriore evento sistemico nell'immediato futuro, e aumentano il numero di strumenti finanziari che risentono della instabilità che si crea.

Considerando anche che la scala temporale in cui il mercato ha, in un certo senso, memoria di un evento appena verificatosi è dell'ordine di qualche minuto, è importante cercare di creare un modello che riesca a studiare accuratamente la dinamiche e gli effetti di questi eventi a breve termine. Il modello di Hawkes qui presentato riesce proprio in questo intento. Tramite un modello che, alla fine, dipende solamente da tre parametri, si è riusciti

a descrivere in modo soddisfacente le dinamiche di breve termine di queste instabilità sistemiche riscontrate empiricamente. In definitiva, il modello fornisce una descrizione realistica del comportamento del mercato in situazioni estreme, il che è importante per diverse prospettive e possibili applicazioni, quali, ad esempio, il trading o il controllo del rischio.

Bibliografia

[1] Lucio Maria Calcagnile et al. ≪Collective synchronization and high frequency systemic instabilities in financial markets≫. In: *Quantitative Finance* 18.2 (dic. 2017), pp. 237–247. ISSN: 1469-7696. DOI: 10.1080/14697688.2017.1403141. URL: http://dx.doi.org/10.1080/14697688.2017.1403141.

[2] Giacomo Bormetti et al. ≪Modelling systemic price cojumps with Hawkes factor models≫. In: *Quantitative Finance* 15.7 (2015), pp. 1137–1156. DOI: 10.1080/14697688.2014.996586. URL: https://doi.org/10.1080/14697688.2014.996586.

[3] Marian-Andrei Rizoiu et al. *A Tutorial on Hawkes Processes for Events in Social Media*. 2017. arXiv: 1708.06401 [stat.ML].