# Effets de rétroaction en finance : applications à l'exécution optimaleet aux modèles de volatilité

Pierre Blanc

UNIVERSITÉ PARIS-EST

École Doctorale : Mathématiques et Sciences et Technologies de l'Information et de la Communication

Thèse de Doctorat
Spécialité : Mathématiques appliquées

*présentée par*

**Pierre BLANC**

---

# Effets de rétroaction en finance : applications à l'exécution optimale et aux modèles de volatilité

---

**Thèse dirigée par Aurélien Alfonsi**
**au CERMICS, École des Ponts ParisTech**

**Thèse soutenue le 9 Octobre 2015 devant le jury composé de :**

| | |
|---|---|
| **Aurélien Alfonsi** | ***Directeur de thèse*** |
| **Jean-Philippe Bouchaud** | ***Examinateur*** |
| **Michel Crouhy** | ***Examinateur*** |
| **Jim Gatheral** | ***Rapporteur*** |
| **Olivier Guéant** | ***Examinateur*** |
| **Bernard Lapeyre** | ***Examinateur*** |
| **Mathieu Rosenbaum** | ***Rapporteur*** |

**Version du 9 Octobre 2015**

# Résumé

Dans cette thèse, nous considérons deux types d'applications des effets de rétroaction en finance. Ces effets entrent en jeu quand des participants de marché exécutent des séquences de transactions ou prennent part à des réactions en chaîne, ce qui engendre des pics d'activité.
La première partie présente un modèle d'exécution optimale dynamique en présence d'un flux stochastique et exogène d'ordres de marché. Nous partons du modèle de référence d'Obizheva et Wang [93], qui définit un cadre d'exécution optimale avec un impact de prix mixte. Nous y ajoutons un flux d'ordres modélisé à l'aide de processus de Hawkes, qui sont des processus à sauts présentant une propriété d'auto-excitation. À l'aide de la théorie du contrôle stochastique, nous déterminons la stratégie optimale de manière analytique. Puis nous déterminons les conditions d'existence de Stratégies de Manipulation de Prix, telles qu'introduites par Huberman et Stanzl [78]. Ces stratégies peuvent être exclues si l'auto-excitation du flux d'ordres se compense exactement avec la résilience du prix. Dans un deuxième temps, nous proposons une méthode de calibration du modèle, que nous appliquons sur des données financières à haute fréquence issues de cours d'actions du CAC40. Sur ces données, nous trouvons que le modèle explique une partie non-négligeable de la variance des prix. Une évaluation de la stratégie optimale en *backtest* montre que celle-ci est profitable en moyenne, mais que des coûts de transaction réalistes suffisent à empêcher les manipulations de prix.
Ensuite, dans la deuxième partie de la thèse, nous nous intéressons à la modélisation de la volatilité intra-journalière. Dans la littérature, la plupart des modèles de volatilité rétroactive se concentrent sur l'échelle de temps journalière, c'est-à-dire aux variations de prix d'un jour sur l'autre. L'objectif est ici d'étendre ce type d'approche à des échelles de temps plus courtes. Nous présentons d'abord un modèle de type ARCH ayant la particularité de prendre en compte séparément les contributions des rendements passés intra-journaliers et nocturnes. Une méthode de calibration de ce modèle est étudiée, ainsi qu'une interprétation qualitative des résultats sur des rendements d'actions américaines et européennes. Dans le chapitre suivant, nous réduisons encore l'échelle de temps considérée. Nous étudions un modèle de volatilité à haute fréquence, dont l'idée est de généraliser le cadre des processus Hawkes pour mieux reproduire certaines caractéristiques empiriques des marchés. Notamment, en introduisant des effets de rétroaction quadratiques inspirés du modèle à temps discret QARCH [102], nous obtenons une distribution en loi puissance pour la volatilité ainsi que de l'asymétrie temporelle.

**Mots clés :** Calibration, Backtest, Modèle d'Impact, Exécution Optimale, Processus de Hawkes, Microstructure de Marché, Trading Haute-Fréquence, Manipulations de Prix, Modèles de Volatilité, Volatilité Rétroactive, Modèles ARCH, Volatilité à Haute Fréquence, Symétrie par Renversement du Temps.

# Abstract

In this thesis we study feedback effects in finance and we focus on two of their applications. These effects stem from the fact that traders split meta-orders sequentially, and also from feedback loops. Therefore, one can observe clusters of activity and periods of relative calm.
The first part introduces an dynamic optimal execution framework with an exogenous stochastic flow of market orders. Our starting point is the well-known model of Obizheva and Wang [93] which defines an execution framework with both permanent and transient price impacts. We modify the price model by adding an order flow based on Hawkes processes, which are self-exciting jump processes. The theory of stochastic control allows us to derive the optimal strategy as a closed formula. Also, we discuss the existence of Price Manipulations Strategies in the sense of Huberman and Stanzl [78], which can be excluded from the model if the self-exciting property of the order flow exactly compensates the resilience of the price. The next chapter studies a calibration protocol for the model, which we apply to tick-by-tick data from CAC40 stocks. On this dataset, the model is found to explain a significant part of the variance of prices. We then evaluate the optimal strategy with a series of backtests, which show that it is profitable on average, although realistic transaction costs can prevent manipulation strategies.
In the second part of the thesis, we turn to intra-day volatility modeling. Previous works from the volatility feedback literature mainly focus on the daily time scale, i.e. on close-to-close returns. Our goal is to use a similar approach on shorter time scales. We first present an ARCH-type model which accounts for the contributions of past intra-day and overnight returns separately. A calibration method for the model is considered, that we use on US and European stocks, and we provide some qualitative insights on the results. The last chapter of the thesis is dedicated to a high-frequency volatility model. We introduce a continuous-time analogue of the QARCH [102] framework, which is also a generalization of Hawkes processes. This new model reproduces several important stylized facts, in particular it generates a time-asymmetric and fat-tailed volatility process.

**Keywords :** Calibration, Backtest, Market Impact Model, Optimal Execution, Hawkes Processes, Market Microstructure, High-frequency Trading, Price Manipulations, Volatility Modeling, Volatility Feedback, ARCH Models, High-Frequency Volatility, Time Reversal Invariance.

# Remerciements

Je voudrais tout d'abord exprimer ma gratitude envers mon directeur de thèse Aurélien Alfonsi. J'admire son efficacité, sa rigueur et son enthousiasme, qui ont été les moteurs de nos séances de travail. Je le remercie pour sa disponibilité et son ouverture d'esprit, ainsi que pour les nombreuses opportunités qu'il m'a données de présenter mes résultats.

Je tiens aussi à remercier Jean-Philippe Bouchaud, auprès de qui j'ai tellement appris depuis mon stage de master à CFM. J'espère avoir su m'inspirer de certaines de ses nombreuses qualités scientifiques, ainsi que de sa bienveillance et de son implication.

Bien sûr, je suis reconnaissant envers la Fondation Natixis pour avoir participé au financement de ma thèse. En particulier, je voudrais remercier Michel Crouhy pour ses encouragements, et pour avoir accepté de faire partie de mon jury. Merci aussi à Julien Puvilland et Marc Souaille de la banque Natixis pour leur aide dans mes travaux empiriques.

Les enseignements de Bernard Lapeyre, Mathieu Rosenbaum et Olivier Guéant constituent les piliers de mon apprentissage des mathématiques financières, c'est pourquoi je suis très heureux qu'ils participent tous les trois à mon jury. Je remercie tout particulièrement Mathieu Rosenbaum et Jim Gatheral d'avoir accepté d'être rapporteurs pour ma thèse.

Mes années au CERMICS ont été une expérience professionnelle aussi profitable qu'agréable, et je voudrais remercier tous mes collègues. Notamment, je suis reconnaissant envers Jean-François Delmas et Benjamin Jourdain qui m'ont confié un rôle d'enseignement dans le cadre de leurs cours de probabilités. Merci à Isabelle Simunic pour son aide toujours efficace et sympathique. Je salue également tous les doctorants et stagiaires, que je remercie pour nos discussions animées et ressourçantes.

Je remercie le « CFM crew », Ahcène, James, Gilles, Camille, Stephen et Iacopo, et tous les postdocs, doctorants et stagiaires que j'ai eu la chance de côtoyer à CFM. Merci à Jonathan pour notre collaboration fructueuse, ainsi qu'à mon mentor Rémy qui m'a appris à dompter les données financières et le langage R.

J'ai aussi une pensée particulière pour tous les participants des Doctoriales Paris-Est de Juin 2014, pour notre semaine riche en innovations et en émotions, mais aussi pour les soirées et pique-niques incroyables que nous avons faits depuis.

Ces remerciements seraient incomplets si je ne mentionnais pas mes soutiens les plus anciens. Merci à Marc, pour nos discussions politiques et « philosophiques » endiablées ; à Thibaut, pour nos nombreuses années d'amitié ; à Pierre-Antoine et Valentin, les gratteurs de Saint-Michel ; à Alexandre, mon camarade bandit stochastique ; à Simon et Charles, mes amis londoniens. Merci à David pour

son humour rafraichissant, et pour ses conseils qui m'ont beaucoup aidé. Un massif merci aux Skipponts, Paul, Rémi, Madi, Amarou et Samir, avec qui nous constituons aujourd'hui bien plus qu'une liste BDS underground. Évidemment, je n'oublie pas le Skippont exceptionnel Casimir, que je remercie pour nos séances de sport matinales, nos délires et nos soirées, et que je serai heureux de fréquenter encore plus l'année prochaine.

Mes pensées les plus affectueuses vont à mes parents et à mon grand-père, qui n'ont pas manqué un seul épisode de mon épopée étudiante, ainsi qu'à mon frère, ma belle-soeur et tout le reste de ma famille. Merci à Christian et Françoise pour leur si grande bienveillance à mon égard, et pour tous mes séjours reposants à Cognac.

Enfin, je veux écrire quelques mots pour celle qui est à mes côtés depuis déjà plusieurs années, et à qui je dois tellement. Je veux lui dédier cette thèse, puisque son soutien quotidien a été crucial dans sa réalisation. Merci à toi, Charlotte, merci infiniment.

# Table des matières

# Introduction

Les travaux présentés dans cette thèse se décomposent en deux parties. Bien que distinctes dans la méthodologie et la finalité, elles se rejoignent sur l'utilisation des processus stochastiques dits auto-excitants. Ces processus permettent de modéliser les séquences d'ordres ainsi que les réactions en chaîne, qui sont des effets observés en pratique sur les marchés financiers, par lesquels l'activité semble former des « regroupements » (*clustering* en anglais). Autrement dit, des périodes d'agitation intense se succèdent avec des moments de calme plat, et l'activité (ou *volatilité*) n'est pas répartie de manière uniforme dans le temps. Ce phénomène influence de manière considérable les mesures de risques, les stratégies de liquidation d'actifs ainsi que le prix de nombreux produits dérivés. En conséquence, les modèles mathématiques qui en tiennent compte présentent un grand intérêt pratique.

La première partie de la thèse traite une problématique d'exécution optimale. Il s'agit d'une approche mathématique pour liquider la position d'un portefeuille pour un actif donné. Nous partons d'un modèle de référence dans ce domaine et nous y ajoutons d'autres acteurs, dont les ordres présentent une propriété d'auto-excitation. Quant à la deuxième partie, il s'agit d'une étude à la fois quantitative et empirique de la volatilité intra-journalière, c'est-à-dire sur des échelles de temps assez courtes, entre l'ouverture et la fermeture des marchés pour une même journée. Notamment, un modèle intra-journalier peut permettre de comprendre au niveau « microscopique » les éléments qui forment la volatilité à basse fréquence (au niveau journalier, hebdomadaire, mensuel ou annuel).

Sans s'attarder sur les détails techniques qui seront abordés dans le corps de la thèse, cette introduction présente les principaux résultats obtenus, ainsi que leur motivation.

## Introduction à l'exécution optimale

En finance, un enjeu naturel est d'acheter ou vendre une certaine quantité d'actifs (c'est-à-dire d'actions, de monnaie, de contrats optionnels...) au meilleur prix possible, en un temps donné. Par exemple, un investisseur peut vouloir acheter un certain nombre d'actions d'une entreprise quotée en bourse, car son analyse le mène à penser qu'il s'agit d'un investissement rentable, ou que le prix auquel elle est actuellement traitée est bas. Dans ce cas, il voudra se procurer ces actions en un certain laps de temps (une heure, une journée, une semaine), avant que cette opportunité ne disparaisse. Considérons un autre exemple : une banque d'affaire, après avoir vendu une option d'achat à un client, se couvre du risque de marché en achetant l'action sous-jacente. Si elle ne possède pas suffisamment d'actions de ce type au moment de la vente du contrat, elle doit acheter la quantité manquante pour un prix acceptable. Elle devra le faire rapidement pour former sa couverture dès que possible. Enfin, une entreprise européenne peut être en possession, du fait de ses exportations, d'une quantité importante de dollars. Dans ce cas, si elle ne veut pas les conserver, il faudra les convertir en euros à un taux satisfaisant. En général, elle devra le faire dans un temps limité pour éviter de s'exposer à un mouvement adverse de ce taux.

Dans tous ces cas, la complication à prendre en compte est que lorsqu'on achète massivement un actif en un court laps de temps, on pousse progressivement le prix vers le haut. Ceci est une conséquence logique de la loi de l'offre et de la demande, qui régit les marchés financiers. On agit donc contre son propre intérêt, puisque l'on rend le prix à l'achat de moins en moins intéressant. De même, quand on vend une grande quantité d'actifs, le prix est poussé vers le bas. Ce mécanisme est appelé « impact de marché » (*market impact* ou *price impact* en anglais). Il s'agit donc de trouver un compromis

entre la vitesse d'exécution de la transaction et la réduction de son impact sur le prix. C'est ce que l'on appelle « exécution optimale ».

D'autre part, l'électronisation des marchés financiers a pris une ampleur considérable depuis une vingtaine d'années. Les « échanges », plate-formes d'interaction entre négociants où les actifs sont achetés et vendus, ne sont plus un endroit de rencontre physique. Il s'agit de hangars abritant de colossaux serveurs informatiques, qui permettent aux transactions d'être faites de manière automatisée n'importe où dans le monde. Parmi les conséquences de cette évolution, on compte l'accélération du rythme des interactions sur les marchés ainsi qu'une informatisation indispensable des méthodes d'échange. Cette informatisation a popularisé les méthodes de *trading* « algorithmique », c'est-à-dire où les transactions ne sont pas décidées directement par des humains, mais par des ordinateurs qui suivent des critères logiques programmés au préalable. Cela permet d'agir de manière plus rapide et de gagner en réactivité. Toutefois, pour mettre au point les algorithmes qui régissent les décisions automatisées, les différentes stratégies doivent être formulées de façon mathématique. C'est pourquoi, en particulier, les problématiques d'exécution optimale gagnent plus que jamais à être traitées par une approche quantitative.

Nous introduisons maintenant les bases de l'exécution optimale mathématique. Ce domaine est en expansion depuis le début des années 2000, et doit ses premiers pas aux travaux de Bertsimas et Lo [24] et Almgren et Chriss [9]. On considère un *trader* particulier, qui veut liquider sa position, c'est-à-dire vendre ou acheter des actifs, selon la situation initiale. Dans ces premières approches, l'impact de marché des transactions est modélisé comme étant linéaire en fonction de leur volume (c'est-à-dire la quantité d'actifs échangés lors de chaque transaction). Le coût de la stratégie du *trader* à chaque instant est donc une fonction quadratique de sa vitesse d'exécution, plutôt que linéaire dans le cas où l'impact est ignoré. Cela permet de pénaliser une exécution trop rapide, tout en imposant que celle-ci soit achevée avant une échéance fixée à l'avance. Les mouvements de prix dûs aux autres participants de marché traitant le même actif sont modélisés par une martingale, c'est-à-dire par un processus sans tendance, pour lequel la meilleure estimation de la valeur future est la valeur actuelle. La formulation mathématique simplifiée du problème d'exécution permet de recourir à la théorie du contrôle stochastique pour déterminer la stratégie optimale sous forme d'une expression analytique. Dans le cas du modèle d'Almgren et Chriss [9], cette stratégie consiste simplement à exécuter la transaction à vitesse constante, c'est-à-dire à diviser la quantité d'actifs à acheter (ou à vendre s'il s'agit d'un programme de vente) de manière uniforme sur tout le temps imparti. Bien qu'elle puisse sembler naive à premier abord, cette méthode de liquidation prend mieux en compte l'impact de marché qu'une stratégie où l'on liquiderait toute la position d'un coup, sans attendre. Une telle stratégie pourrait siphonner la liquidité présente sur le marché, et avoir un impact et un coût considérables.

Il faut noter que dans ces premiers travaux, la structure temporelle de l'impact de marché a une forme particulière. Elle comporte une partie « immédiate », qui augmente le coût instantané de la transaction sans affecter les cours à venir, et une partie permanente, qui modifie définitivement le prix mais n'affecte pas la stratégie de liquidation. Une extension naturelle est d'introduire un impact « transient » ou « temporaire », c'est-à-dire de supposer que l'impact d'une transaction sur le marché décroît avec le temps. Cela permet de modéliser une certaine élasticité des prix, qui absorbent l'impact petit à petit et oscillent autour d'une moyenne mobile. Pour formaliser cette décomposition de l'impact, on introduit la notion de « propagateur », qui est une fonction $G(t)$ qui décrit la manière dont l'impact d'une transaction effectuée au temps 0 évolue avec le temps $t$.

Considérons un participant de marché voulant liquider sa position sur l'intervalle de temps $[0,T]$. On note $x(t)$ sa position au temps $t \in [0,T]$, c'est-à-dire la quantité d'actif qu'il possède dans son portefeuille. On suppose que $x(0) = x_0 \in \mathbb{R}$ est connu, avec $x_0 > 0$ pour un programme de vente et $x_0 < 0$ pour un programme d'achat (puisque liquider un certain nombre d'actifs en notre possession revient à les vendre, et liquider une position à découvert revient à acheter les actifs). Puisque la liquidation doit être achevée au temps $T$, on impose $x(T) = 0$. Au temps $t \in [0,T]$, la transaction effectuée par le participant de marché peut s'écrire comme la variation de sa position, c'est-à-dire $\dot{x}(t)\mathrm{d}t$, où $\dot{x}$ est la dérivée temporelle de $x$. Plus la vitesse de liquidation du participant est élevée, plus il va impacter le prix : on suppose donc qu'à l'instant $t$, l'impact instantané est proportionnel à $f(\dot{x}(t))\ \mathrm{d}t$, où $f$ est une fonction croissante. Dans le modèle de propagateur, l'impact global de la stratégie sur le prix $P_t$ entre 0 et $t \in [0,T]$ est donné par l'équation

$$P_t = P_t^0 + \int_0^t G(t-s) f(\dot{x}(s))\ \mathrm{d}s, \tag{1}$$

où $P_t^0$ est le prix non-impacté. Autrement dit, en modifiant sa position de la quantité $\dot{x}(t)\mathrm{d}t$ sur l'intervalle $[t, t+\mathrm{d}t]$, le *trader* transforme le prix $P_t$ en $P_t + G(0) f(\dot{x}(t))\ \mathrm{d}t$, puis cet impact est propagé dans le temps par la fonction $G$. De tels modèles à propagateur ont par exemple été étudiés par Bouchaud et al. [31], Gatheral [63], Gatheral et al. [64], Alfonsi et al. [7] et Obizheava et Wang [93]. Avec cette formalisation, on peut distinguer

- L'impact immédiat $G(0) - G(0^+)$ (où on note $G(0^+) = \lim_{t \to 0^+} G(t)$), qui est non nul si et seulement si le propagateur n'est pas continu en zéro. Il implique un surcoût d'exécution, sans impact visible sur les prix.
- L'impact permanent $G(+\infty)$ (où on note $G(+\infty) = \lim_{t \to +\infty} G(t)$), qui se traduit comme l'impact des transactions sur les prix à basse fréquence.
- L'impact transient $G(0^+) - G(+\infty)$, qui est la partie qui est progressivement absorbée par le marché, et qui n'influence les prix qu'à des fréquences moyennes ou hautes.

Obizheava et Wang [93] ont été les premiers à résoudre explicitement le problème d'exécution optimale en présence d'impact transient. Ils choisissent un propagateur $G$ de forme exponentielle, une fonction d'impact $f$ linéaire, et autorisent la stratégie à faire des transactions en bloc, c'est-à-dire que la position $x(t)$ du *trader* est remplacée par un processus $X_t$ pouvant faire des sauts. L'équation (1) devient

$$P_t = P_t^0 + \frac{1}{q} \int_0^t \left[\nu + \lambda \exp(-\rho(t-s))\right]\ \mathrm{d}X_s,$$

où $\nu \in [0,1]$ est l'impact permanent, $\lambda = 1 - \nu$ est l'impact transient, $\rho > 0$ est la vitesse de résilience de l'impact et $f(x) = x/q$ avec $q > 0$ une mesure de liquidité. La limite $\rho \to 0^+$ donne un impact purement permanent, tandis que $\rho \to +\infty$ correspond au modèle à impact immédiat d'Almgren et Chriss [9], ce qui est démontré rigoureusement par Kallsen et Muhle-Karbe [84]. Bien que relativement simple dans le choix de la forme du propagateur, le modèle d'Obizheava et Wang [93] généralise donc la plupart de ses prédécesseurs, tout en permettant une forme d'impact plus réaliste. Dans un tel modèle où la fonction d'impact $f$ est linéaire, le coût d'exécution d'une stratégie $X$ est donné par

$$C(X) = \int_0^T P_t\ \mathrm{d}X_t + \frac{1}{2q} \sum_{0 \le \tau \le T} (\Delta X_\tau)^2,$$

où le deuxième terme pénalise les sauts $\Delta X_\tau = X_{\tau^+} - X_\tau$, qui apparaissant en quantité dénombrable dans la stratégie. Le surcoût quadratique des sauts découle directement de l'impact linéaire : si on

exécute en bloc une quantité $y$ au temps $t$, le coût $\pi_t(y)$ est donné par l'équation

$$\pi_t(y) = \int_0^y \left(P_t + \frac{x}{q}\right) \, \mathrm{d}x = P_t \; y + \frac{y^2}{2q},$$

car lorsqu'on a déjà exécuté $x \in [0, y)$, on a déplacé le prix de la quantité $x/q$ et on paye la quantité $\mathrm{d}x$ à ce nouveau prix.

L'idée est ensuite de trouver la stratégie optimale $X^*$, c'est-à-dire celle qui minimise le coût moyen $\mathbb{E}[C(X^*)]$. Si certaines conditions techniques sont satisfaites, en particulier si le processus $(X_t, P_t)$ peut être représenté de manière Markovienne (ou « sans mémoire »), la théorie du contrôle optimal stochastique peut permettre de déterminer $X^*$ de manière explicite. C'est le cas du modèle d'Obizheava et Wang [93]. Rappelons que la position initiale du *trader* est $X_0 = x_0$ et que $X_{T^+} = 0$ est imposé. La stratégie optimale $X^*$ est donnée par deux sauts de même taille au début et à la fin de la période

$$\Delta X_0^* = \Delta X_T^* = -\frac{x_0}{2 + \rho T}, \tag{2}$$

et un taux de *trading* constant sur $(0, T)$,

$$\mathrm{d}X_t^* = -\frac{\rho x_0}{2 + \rho T} \, \mathrm{d}t. \tag{3}$$

Cette stratégie est dite « en seau » (*bucket-shaped* en anglais). Une proportion $2/(2+\rho T)$ du volume est exécutée par des transactions en bloc en $t = 0$ et $t = T$, et la proportion restante $\rho T/(2+\rho T)$ de manière continue sur l'intervalle ouvert $(0, T)$. Il est facile de voir sur ces équations qu'en prenant $\rho = 0$, la partie de *trading* continu disparait, et la stratégie exécute la moitié de la transaction au début de la période et l'autre moitié à la fin. En revanche, pour $\rho \to +\infty$, seule la partie continue subsiste et le taux d'exécution devient $x_0/T$, ce qui revient à la stratégie uniforme d'Almgren et Chriss [9]. Entre ces deux extrêmes, plus la vitesse de résilience $\rho$ de l'impact est grande, plus on peut faire des transactions intermédiaires en sachant que leur impact va disparaître au fur et à mesure. Si la résilience est très lente, il est plus rentable d'exécuter une grande partie de l'ordre immédiatement, puis de laisser autant de temps que possible au marché pour absorber l'impact, avant de compléter la transaction.

Une notion importante dans les modèles d'exécution est celle des stratégies de manipulation de prix (*Price Manipulation Strategies* ou PMS en anglais). La formalisation mathématique de ce concept est dûe à Huberman et Stanzl [78] : une PMS est la donnée d'un horizon $T > 0$ et d'une stratégie $(X_t)_{t\in[0,T]}$ tels que

$$X_0 = X_{T^+} = 0, \quad \mathbb{E}[C(X)] < 0,$$

où $C(X)$ est le coût d'exécution et $\mathbb{E}[C(X)]$ est le coût moyen. Il s'agit donc d'une stratégie à somme nulle (un *round trip* en anglais), c'est-à-dire dont les positions initiale et finale sont identiques, et dont l'exécution rapporte de l'argent au lieu d'en coûter (en moyenne). Aussi appelé « arbitrage dynamique » dans la littérature, ce concept étend la notion d'arbitrage, classique en finance, au cadre de l'exéctution optimale. Toutefois, contrairement aux arbitrages standards, une PMS est une stratégie qui n'est pas nécessairement toujours gagnante, mais qui l'est juste en moyenne. On considère qu'une manière d'affirmer si un modèle d'exécution est « bon » est de vérifier que les PMS y sont impossibles, car leur existence contredit le bon fonctionnement du marché. Dans l'article d'Obizheava et Wang [93], les auteurs montrent que le coût moyen de la stratégie optimale est

positivement proportionnel au carré de la quantité $x_0$ à liquider. Il est donc toujours positif, pourtant c'est le coût minimal par définition de la stratégie optimale, ce qui implique que les PMS sont impossibles dans le modèle.

L'article d'Obizheava et Wang [93] a inspiré de nombreux travaux dans le domaine de l'exécution optimale. Notamment, la littérature comporte beaucoup d'extensions de leur modèle, dans différentes directions. Citons-en quelques-unes :

- Gatheral [63] étudie les propriétés de la stratégie d'exécution optimale ainsi que les possibilités d'arbitrage dynamique pour différentes formes de fonction d'impact $f$ et de propagateur $G$ (cf l'équation (1)). Notamment, cette étude permet de réconcilier le cadre des modèles à propagateur avec certaines observations empiriques.
- Alfonsi et al. [5] reprennent l'étude mise en oeuvre par Obizheava et Wang [93] pour un impact $f$ non linéaire, et déterminent la stratégie d'exécution optimale sous forme d'une équation implicite. Ces résultats ont un grand intérêt pratique puisque la linéarité de $f$ n'est pas considérée comme réaliste.
- Fruth et al. [60] considèrent le cas où la vitesse de résilience $\rho$ et la mesure de liquidité $q$ ne sont plus des constantes, mais des fonctions déterministes (c'est-à-dire non-aléatoires) du temps, $\rho(t)$ et $q(t)$. Cela permet de prendre en compte les saisonnalités intra-journalières prévisibles de l'activité financière. Ils déterminent la stratégie optimale et les conditions d'absence de PMS dans ce cas. Alfonsi and Infante [2] généralisent ces résultats à des fonctions d'impact $f$ non linéaires.

Toutes les approches citées précédemment ont deux points communs : elles se limitent à une modélisation par propagateur, et elles sont statiques, c'est-à-dire que la stratégie optimale est déterministe. On peut déterminer la stratégie à l'avance et s'y tenir jusqu'à ce que la liquidation soit achevée, et ce indépendamment de l'évolution des cours et du comportement des autres participants de marché. Cela repose sur la modélisation du prix non affecté $P_t^0$ par une martingale, dont l'observation ne donne aucune information utile pour l'exécution. Pour aller plus loin dans les modèles mathématiques d'exécution, deux possibilités apparaissent :

- Remplacer le propagateur par un objet plus riche, permettant une modélisation plus flexible et plus réaliste. Cela peut s'avérer compliqué, car un des avantages du propagateur est qu'il donne un cadre favorable aux calculs. Toutefois, Donier et al. [48] proposent un modèle de carnet d'ordres latent inspiré d'arguments de réaction-diffusion, qui permet de reproduire de nombreuses observations empiriques. Ils obtiennent une équation d'impact très générale, dont le propagateur n'est qu'un cas particulier. Dans ce modèle, il est démontré que les stratégies de manipulation de prix sont impossibles, mais la stratégie optimale n'est pas calculable explicitement dans le cas général. Malgré cela, cette étude constitue une nouvelle approche prometteuse.
- Tout en s'en tenant à un modèle à propagateur, dans lequel les calculs sont plus simples, il est possible de s'intéresser à l'aspect dynamique de l'exécution optimale. Pour cela, on ajoute d'autres participants de marché, et on détermine comment réagir à leurs actions de manière optimale, en temps réel. Schied et Zhang [101] modélisent le cas où deux *traders* veulent liquider le même actif simultanément. Par une approche d'équilibre de Nash, les auteurs montrent que la stratégie optimale commune consiste à vendre dès que l'autre achète et réciproquement, ce qui est un scénario peu souhaitable pour le marché dans son ensemble. En ajoutant un certain niveau de coûts de transaction quadratiques, ce comportement disparaît, ce qui suggère que ces coûts de transaction peuvent en fait diminuer le coût d'exécution global et rendre le marché plus efficient.

C'est dans le deuxième point que s'inscrit l'étude [3] présentée dans le chapitre 1 de cette thèse. Contrairement à Schied et Zhang [101], nous ne modélisons pas les différents participants de marché

de manière symmétrique. Au lieu de cela, nous considérons un *trader* de référence, comme le font Obizheava et Wang [93], qui veut liquider sa position de manière optimale. Nous modélisons les autres acteurs comme un flux stochastique de transactions, auxquelles le *trader* de référence peut réagir de manière instantanée pour s'adapter aux sauts de prix. Nous présentons ce modèle dans la partie suivante de l'introduction.

## Exécution optimale dynamique en présence d'un flux stochastique d'ordres de marché

Dans les modèles « classiques » de finance, l'évolution du prix est modélisée comme un processus stochastique, souvent une diffusion, de manière exogène. Cela signifie que le mouvement des cours est dû à d'autres participants de marché qui « n'observent pas » les actions de l'utilisateur du modèle. Cette hypothèse améliore grandement la tractabilité mathématique des problèmes considérés, en se centrant sur le point de vue de l'utilisateur. Les articles présentés dans la partie I de cette thèse transposent cette approche à un modèle d'exécution optimale dynamique.

Nous partons du modèle statique d'exécution optimale d'Obizheava et Wang [93], auquel nous ajoutons un flux d'ordres stochastique et exogène, c'est-à-dire que ce flux est aléatoire mais ne dépend pas des transactions effectuées par le *trader* de référence. Comme les ordres arrivent sur le marché de manière discrète en temps, modéliser ce flux à l'aide d'un processus à sauts est une approche naturelle. Parmi ces processus, le plus utilisé en termes de modélisation est celui de Poisson. Nous présentons donc le modèle de prix

$$P_t = P_t^0 + \frac{1}{q}\int_0^t [\nu + \lambda \exp(-\rho(t-s))]\ (\mathrm{d}X_s + \mathrm{d}N_s^+ - \mathrm{d}N_s^-), \tag{4}$$

où $P^0$ est une martingale quelconque et $N^+, N^-$ sont deux processus de Poissons indépendants entre eux et indépendants de $P^0$, et de même intensité $\kappa_0 > 0$. Les processus $N^+$ et $N^-$ représentent respectivement un flux d'ordres d'achat (qui impactent positivement le prix), et un flux d'ordres de vente (qui impactent négativement le prix). Par souci de simplicité, les amplitudes des sauts de $N^\pm$ sont supposées être des variables aléatoires indépendantes, imprévisibles et distribuées selon une même loi $\mu$ sur $\mathbb{R}^+$. Un des intérêts de ce modèle est que le profil d'impact des transactions est le même pour le *trader* de référence (modélisé par $X$) que pour les autres (modélisés par $N^\pm$), ce qui lui donne une certaine cohérence.

La problématique est identique à celle de l'étude d'Obizheava et Wang [93] : le *trader* de référence veut liquider une position $x_0$ sur l'intervalle de temps $[0, T]$, et ce en minimisant son coût moyen. La théorie du contrôle optimal stochastique permet ici de déterminer la stratégie optimale à l'aide d'un argument de vérification. Cette méthode consiste à « deviner » la forme fonctionnelle du coût moyen minimal, aussi appelé fonction valeur, par rapport aux variables d'état du problème, puis à l'injecter dans la dynamique du modèle pour obtenir un certain nombres d'équations. Quand, comme ici, la résolution de ces équations se fait de manière analytique, on obtient explicitement la stratégie optimale et la fonction valeur. Nous trouvons que la stratégie optimale $X^*$ se décompose sous la forme

$$X^* = X^{\mathrm{OW}} + X^{\mathrm{trend}} + X^{\mathrm{dyn}},$$

où

- $X^{\mathrm{OW}}$ est la stratégie optimale du modèle d'Obizhaeva et Wang [93], donnée par les équations (2) et (3). Ce terme est proportionnel à la position initiale $x_0$, et il correspond à ce qu'on obtient si les processus $N^+$ et $N^-$ sont remplacés par zéro (ce qui est logique puisque l'on revient alors au modèle statique de départ).
- $X^{\mathrm{trend}}$ est la stratégie de « tendance » (*trend* en anglais). Elle est nulle si le marché est à l'équilibre à l'instant initial, et indique comment dégager un profit dans le cas contraire.
- $X^{\mathrm{dyn}}$ est la stratégie « dynamique », proportionnelle aux processus $N^+$ et $N^-$. C'est ce terme qui nous intéresse le plus, puisqu'il donne la réaction optimale aux transactions observées, et rend la stratégie dynamique. Notamment, si un ordre de marché est posté au temps $\tau \in (0,T)$, alors $N^+$ ou $N^-$ saute, et la stratégie réagit en sautant immédiatement après, selon l'équation

$$\Delta X^{\mathrm{dyn}}_\tau \;=\; -\,\frac{1+\rho(T-\tau)}{2+\rho(T-\tau)}\,(\Delta N^+_\tau - \Delta N^-_\tau), \tag{5}$$

  où $\rho$ est la vitesse de résilience. Autrement dit, en réponse à la transaction observée, la stratégie effectue une transaction dans le sens contraire (une vente si on observe un achat, et réciproquement), pour une proportion $(1+\rho(T-\tau))/(2+\rho(T-\tau)) \in [1/2,1]$ du volume observé. La proportion est d'autant plus grande que $\rho(T-\tau)$ est grand, cette quantité mesurant la capacité du prix à revenir à la moyenne avant l'échéance $T$.

On peut interpréter la stratégie optimale comme suit : supposons que le marché est à l'équilibre à l'instant initial (ce qui n'est pas une hypothèse très réductrice) donc $X^{\mathrm{trend}} \equiv 0$. La stratégie se décompose alors de manière additive comme la stratégie de liquidation de la position initiale $x_0$, et une stratégie qui réagit de manière dynamique aux sauts observés. Comme les sauts de $N^\pm$ sont régis par un processus de Poisson, leurs temps d'arrivée sont répartis de manière totalement imprévisible et sans mémoire sur l'intervalle $[0,T]$. Donc, juste après un saut de $N^+$ par exemple, nous n'avons aucune information sur les sauts suivants de $N^+$ et $N^-$. En revanche, à cause de la partie transiente $\lambda\exp(-\rho(t-s))$ de l'impact, on sait qu'une partie $\lambda \in [0,1]$ du saut de prix positif qu'on vient d'observer va revenir à zéro, donc que le prix va probablement baisser dans un futur proche. Il est alors intéressant de vendre immédiatement une certaine quantité d'actif pour la racheter plus tard.

Ceci est en fait une stratégie de manipulation de prix (PMS), comme on le vérifie en calculant le coût moyen. Ceci n'est pas surprenant car dans le modèle, chaque saut de $N^\pm$ donne une information sur la tendance de prix à venir. De plus, nous montrons dans le chapitre 1 que la connaissance de $\rho$ n'est pas requise pour dégager du profit, car on obtient encore une PMS en remplaçant systématiquement la proportion $(1+\rho(T-\tau))/(2+\rho(T-\tau))$ par $1/2$. Ce modèle n'est donc pas compatible avec un bon fonctionnement du marché, ce qui nous incite à l'enrichir.

En fait, il est clair que le problème du modèle de Poisson est qu'en l'absence du *trader* de référence, le prix non-affecté

$$P^{(X\equiv 0)}_t = P^0_t + \frac{1}{q}\int_0^t [\nu + \lambda\exp(-\rho(t-s))]\;(\mathrm{d}N^+_s - \mathrm{d}N^-_s) \tag{6}$$

ne peut pas être une martingale, donc son évolution est partiellement prévisible. Nous montrons en effet dans le chapitre 1 que l'absence de PMS dans ce modèle est équivalente au fait que $P^{(X\equiv 0)}$ soit une martingale. Or, pour obtenir cette propriété de martingale avec des processus à sauts $N^+$ et $N^-$ plus généraux, il faut nécessairement introduire une structure d'auto-corrélation dans les sauts. Autrement dit, pour compenser la force de rappel du prix dûe à l'impact transient, il faut que juste

après un saut de $N^+$ faisant monter le prix, il soit plus probable d'observer un nouveau saut de $N^+$ que d'observer un saut de $N^-$. Ce phénomène est observé en pratique dans les marchés financiers, et est provoqué par le découpage des ordres (*splitting* en anglais), ainsi qu'aux spéculateurs qui utilisent des stratégies de poursuite de tendances (*trend-following* en anglais). Nous renvoyons à l'article de Toth et al. [105] pour plus de détails sur ce point. Le *splitting* est le fait qu'un *trader* qui arrive sur le marché avec une grande quantité d'actif à acheter ou à vendre découpe cette quantité en plusieurs ordres de petite taille pour réduire son impact. C'est ce que nous avons vu dans notre introduction à l'exécution optimale. Par exemple, quand on observe un ordre d'achat sur le marché, la probabilité qu'il fasse partie d'une séquence d'ordres d'achat postés par un même *trader* est non négligeable. Cela crée une structure d'auto-corrélation positive dans le flux d'ordres.

Pour modéliser cette structure d'auto-corrélation, tout en conservant une bonne tractabilité mathématique, une certaine classe de processus stochastiques à sauts est tout à fait adaptée : les processus de Hawkes. Introduits par Alan G. Hawkes en 1971 [72], les processus de Hawkes sont des processus à sauts à intensité stochastique, c'est-à-dire que l'intensité de saut est aléatoire et variable dans le temps. De manière générale, si $J$ est un processus à sauts, le fait qu'il soit d'intensité stochastique $\kappa_t$ signifie que presque sûrement,

$$\forall t \geq 0, \quad \frac{1}{h}\,\mathbb{E}[J_{t+h} - J_t | \mathcal{F}_t] \underset{h\to 0^+}{\longrightarrow} \kappa_t, \tag{7}$$

où $\mathcal{F}_t = \sigma((J_s, \kappa_s)_{s\leq t})$ est la filtration naturelle du processus. Autrement dit, à tout instant $t$, sachant le passé du processus, le nombre moyen de sauts entre $t$ et $t + \mathrm{d}t$ est $\kappa_t\ \mathrm{d}t$. Les processus de Hawkes des processus à intensité stochastique où $\kappa_t$ est simplement donnée par une équation auto-régressive linéaire :

$$\forall t \geq 0, \quad \kappa_t = \kappa_\infty + \int_{-\infty}^{t} \phi(t-s)\ \mathrm{d}J_s, \tag{8}$$

avec $\kappa_\infty > 0$ une constante symbolisant l'intensité « de base », et $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ une fonction mesurable et intégrable appelée le noyau de Hawkes (*Hawkes kernel* en anglais). L'équation (8) signifie que l'intensité $\kappa_t$ est toujours supérieure à l'intensité de base $\kappa_\infty$, et qu'elle s'en écarte d'autant plus que le processus $J$ a sauté dans le passé proche. En effet, à chaque fois qu'un saut de $J$ se produit, l'intensité $\kappa_t$ saute de $\phi(0) \geq 0$, puis est progressivement rappelée vers $\kappa_\infty$ de manière continue et déterministe à travers le noyau $\phi$. C'est pourquoi on parle de dynamique auto-excitante (*self-exciting* en anglais), car le processus saute plus quand l'intensité est élevée, et les sauts du processus augmentent eux-même l'intensité. Contrairement à un processus de Poisson où les sauts sont répartis de manière uniforme sur l'intervalle de temps, un processus de Hawkes forme des périodes de calme (où $\kappa_t$ est proche de $\kappa_\infty$ et le processus $J$ saute peu) et des périodes d'agitation (des *clusters* en anglais, où $\kappa_t$ s'éloigne de $\kappa_\infty$ et le processus $J$ saute beaucoup). Notons toutefois que les processus de Poisson sont un cas particulier des processus de Hawkes, puisque si l'on prend le noyau $\phi$ identiquement nul, l'intensité $\kappa_t$ reste constante et égale à $\kappa_\infty$. Si l'on injecte l'équation (7) dans l'équation (8), on obtient l'égalité en moyenne

$$\forall t \geq 0, \quad \mathbb{E}[\kappa_t] = \kappa_\infty + \int_{-\infty}^{t} \phi(t-s)\ \mathbb{E}[\kappa_s]\mathrm{d}s. \tag{9}$$

Notons $||\phi|| = \int_0^\infty \phi(u)\ \mathrm{d}u \in [0, +\infty)$. En considérant l'équation (9), on voit que trois régimes sont possibles :

- Le régime sur-critique $||\phi|| > 1$, où $\kappa_t \underset{t\to+\infty}{\longrightarrow} +\infty$ avec probabilité non nulle si $\kappa_0 > 0$, presque sûrement si de plus $\kappa_\infty > 0$. Dans ce régime, les sauts deviennent infiniment fréquents, ce qui a peu de sens en termes de modélisation.
- Le régime critique $||\phi|| = 1$, qui est similaire au régime sur-critique si $\kappa_0 > 0$ et $\kappa_\infty > 0$, mais peut avoir plus de sens si $\kappa_\infty = 0$ (voir Brémaud et Massoulié [34]). Dans ce deuxième cas, il s'agit d'un processus où l'auto-excitation est très forte, mais qui peut tout de même rester stable car l'intensité de base est nulle.
- Le régime sous-critique $||\phi|| < 1$, qui est en général le plus intéressant pour modéliser des phénomènes réels. Dans ce cas, le processus converge vers un état stationnaire quand $t \to +\infty$ (voir Hawkes et Oakes [74]), c'est-à-dire qu'il tend vers un équilibre où sa loi de probabilité ne varie plus. Cela implique que la moyenne de l'intensité $\kappa_t$ converge vers une constante, dont la valeur découle de l'équation (9) :

$$\mathbb{E}[\kappa_t] \underset{t\to+\infty}{\longrightarrow} \overline{\kappa} = \frac{\kappa_\infty}{1 - ||\phi||}. \tag{10}$$

  Cette formule a une interprétation claire puisqu'on retrouve $\overline{\kappa} = \kappa_\infty$ pour un processus de Poisson ($||\phi|| = 0$), et $\overline{\kappa}$ diverge quand $||\phi||$ tend vers 1.

La valeur de $||\phi||$ est donc un paramètre très important pour un processus de Hawkes. On l'appelle le ratio de branchement (*branching ratio* en anglais). Il peut être interprété comme le nombre moyen de sauts « engendrés » par chaque saut. En faisant le parallèle avec une dynamique de population, on comprend alors pourquoi $||\phi|| = 1$ est la valeur critique au-delà de laquelle les sauts deviennent infiniment fréquents.

L'utilisation des processus de Hawkes a connu un grand essor en finance quantitative ces dernières années. Sans essayer d'être exhaustif, énumérons certains de ces travaux. Bacry et al. [11], [12], [10], [16] présentent plusieurs versions d'un modèle de prix à haute fréquence utilisant ces processus, et étudient leurs propriétés mathématiques et la façon d'estimer leurs paramètres en pratique. Filimonov et Sornette [58], suivis de Hardiman et al. [68], [69] mènent des études empiriques de l'activité des marchés financiers en modélisant les changements de prix par un processus de Hawkes, et discutent de la valeur du *branching ratio* qui mesure la réflexivité de l'activité financière.

La partie I de cette thèse rend compte de nos résultats en exécution optimale dynamique. L'article [3] présenté dans le chapitre 1 généralise le modèle de prix (4) en remplaçant les processus de Poisson indépendants $N^+$ et $N^-$ par un processus de Hawkes $(N^+, N^-)$ de dimension 2, où $N^+$ (resp. $N^-$) est d'intensité stochastique $\kappa_t^+$ (resp. $\kappa_t^-$). La stratégie optimale est encore calculée explicitement, et comme nous l'espérions, elle diffère de la stratégie (5) du modèle de Poisson par l'ajout d'un terme de signe opposé. Un certain jeu de paramètres permet à ces deux termes de se compenser, ce qui lisse la stratégie optimale et la rend plus réaliste. En fait, ce même jeu de paramètres permet également de faire en sorte que le prix non affecté (6) soit une martingale, ce qui empêche les stratégies de manipulation de prix. Cela fournit un cadre cohérent d'exécution optimale dynamique pour lequel un équilibre de marché est possible, et où la stratégie optimale est connue si cet équilibre n'est pas tout à fait respecté.

Ensuite, le chapitre 2 étudie une généralisation du modèle précédent, ainsi que sa calibration sur un jeu de données financières fourni par la banque d'investissement Natixis. En permettant au propagateur et au noyau de Hawkes d'avoir des formes plus générales que l'exponentielle, de nouveaux résultats théoriques sont obtenus, et permettent d'appliquer le modèle de manière plus réaliste. Une méthode de calibration du modèle est ensuite présentée, puis appliquée sur données simulées et sur

données réelles (extraites d'historiques de prix *tick-by-tick* d'actions du CAC40). Nous proposons une interprétation qualitative de ces résultats empiriques, et menons une évaluation en *back-test* de la stratégie optimale sur notre jeu de données.

## Introduction aux modèles auto-régressifs de volatilité

Nous introduisons maintenant la deuxième partie de cette thèse, concernant la modélisation de la volatilité des marchés financiers. La volatilité est une mesure du niveau d'agitation du prix d'un actif sur une période donnée. On qualifie un actif de volatil quand son prix a des mouvements importants et/ou fréquents, quelle que soit leur direction. Il ne s'agit donc pas de détecter des tendances de prix à la hausse ou à la baisse, mais de mesurer des variations en valeur absolue. La notion de volatilité est étroitement liée à celle de risque, que l'on peut définir comme la probabilité d'événements défavorables (en un sens à préciser selon le contexte). En finance, on considère souvent le risque que le prix d'un actif monte beaucoup (si, pour une raison quelconque, on doit l'acheter), ou baisse beaucoup (si on le possède déjà dans son portefeuille). Ces risques sont d'autant plus importants que la volatilité de l'actif est élevée.

La modélisation de la volatilité est directement motivée par un trait psychologique humain appelé l'aversion au risque : nous n'aimons pas en général prendre des risques non nécessaires. Dans une certaine mesure, nous sommes plus soucieux de nous protéger contre les événements très défavorables, que de maximiser notre gain moyen (financier ou autre). Ceci est bien illustré par l'exemple suivant : on considère un jeu où l'on a une probabilité de 60% de doubler le montant de son compte en banque, et 40% de tout perdre. A priori, une grande majorité des personnes à qui on proposerait ce jeu refuseraient d'y participer, bien que le joueur soit gagnant en moyenne. Nous avons donc naturellement tendance à « pénaliser » le risque dans notre processus de décision. La théorie de la sélection de portefeuilles est une application typique de ce concept en finance. Lorsqu'un investisseur construit son portefeuille, l'approche standard consiste à sélectionner des actifs dont le rendement moyen est aussi bon que possible pour un niveau de volatilité maximal donné. Réciproquement, on peut aussi minimiser la volatilité de son portefeuille pour un niveau de rendement minimal imposé. C'est la théorie de Markowitz [89], développée dès 1952 par Harry Markowitz. Pour mener à bien une telle démarche, il est essentiel de pouvoir évaluer la volatilité des actifs avec précision.

Considérons un modèle à temps discret, c'est-à-dire où le temps évolue sur une grille d'entiers : dans ce cas, l'instant suivant immédiatement le temps $t$ est le temps $t+1$. Soit un actif dont le prix au temps $t$ est noté $P_t$. On appelle rendement de l'actif au temps $t$ la quantité

$$R_t = \frac{P_{t+1} - P_t}{P_t},$$

qui est l'incrément relatif du prix. Pour simplifier le cadre mathématique, on remplace les rendements par les log-rendements

$$r_t = \log\left(\frac{P_{t+1}}{P_t}\right) = \log(1 + R_t),$$

ce qui est équivalent pour $R_t$ petit. Les rendements sont alors additifs, c'est-à-dire que le log-rendement entre $t$ et $t+2$ est la somme des deux log-rendements : $\log(P_{t+2}/P_t) = \log(P_{t+1}/P_t) +$

$\log(P_{t+2}/P_{t+1}) = r_t + r_{t+1}$. On définit ensuite la volatilité $\sigma$ de l'actif comme l'écart-type « conditionnel » de ses rendements (ou de ses log-rendements). Plus précisément, un modèle de volatilité prend la forme

$$r_t = \sigma_t \, \xi_t, \tag{11}$$

où $\sigma_t$ est la volatilité au temps $t$, et $\xi_t$ est une variable aléatoire indépendante de $\sigma_t$ et de variance unitaire, appelée résidu. En général, le résidu est pris de moyenne nulle, et la suite $(\xi_t)$ est formée de variables indépendantes et identiquement distribuées. Différents modèles de volatilité spécifient ensuite différentes dynamiques pour le processus $\sigma_t$.

Le plus classique d'entre eux est la marche aléatoire, qui est la version en temps discret du célèbre mouvement Brownien. Ce modèle suppose que la volatilité $\sigma$ est constante, ce qui implique que les log-rendements $r_t = \sigma\xi_t$ sont eux-mêmes des variables indépendantes et identiquement distribuées. Dans ce modèle très simple, on peut estimer la volatilité constante par l'écart-type empirique des rendements

$$\hat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_t - \overline{r})^2},$$

où la moyenne empirique $\overline{r} = \frac{1}{T}\sum_{t=1}^{T} r_t$ est censée être proche de zéro. La tractabilité mathématique de ce modèle le rend attractif, et il est souvent utilisé en pratique. Toutefois, il ne permet pas de reproduire un certain nombre d'observations empiriques. Une des plus connues d'entre elles est le fait que la volatilité forme des « regroupements » (ce qu'on appelle *volatility clustering* en anglais), c'est-à-dire que l'on observe une succession de périodes de volatilité élevée et de périodes de volatilité faible, au lieu d'avoir une répartition uniforme. En effet, Benoit Mandelbrot [88] écrit en 1963 au sujet des marchés financiers : *... large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes.* Cela se traduit mathématiquement par une auto-corrélation positive des rendements absolus $|r_t|$ et des rendements au carré $r_t^2$. C'est ce phénomène que cherchent à capturer les modèles de type ARCH (*Auto-Regressive Conditional Heteroskedasticity*).

Le premier modèle ARCH a été introduit par Engle [52] en 1982. Dans le cadre de l'équation (11), il propose la dynamique suivante pour le processus de volatilité :

$$\sigma_t^2 = s^2 + g \, r_{t-1}^2 = s^2 + g \, \sigma_{t-1}^2 \xi_{t-1}^2, \tag{12}$$

où $s^2 > 0$ est une constante qui représente le niveau « de base » de la volatilité, et $g > 0$ est le paramètre de rétroaction (*feedback* en anglais). Dans ce modèle, la volatilité au carré est un processus auto-régressif de portée 1, avec un bruit multiplicatif $\xi$. Malgré sa simplicité, cette dynamique permet de rendre compte du phénomène de *volatility clustering*, puisque une haute volatilité $\sigma_{t-1}$ au temps $t-1$ va impacter à la hausse la volatilité suivante $\sigma_t$. Ce mécanisme de *feedback* est d'autant plus fort que le paramètre $g$ est grand. Notamment, si l'on passe à la valeur moyenne dans l'équation (12), on obtient

$$\mathbb{E}[\sigma_t^2] = s^2 + g \, \mathbb{E}[\sigma_{t-1}^2],$$

puisque $\xi_{t-1}$ est idépendant de $\sigma_{t-1}$ par construction, centré et de variance unitaire. Il est clair sur cette nouvelle équation que $\mathbb{E}[\sigma_t^2]$ ne peut converger vers une constante quand $t \to +\infty$ que si $g < 1$ : sans cela, il est impossible d'atteindre un régime stationnaire. Si cette condition est vérifiée,

on obtient alors

$$\mathbb{E}[\sigma_t^2] \underset{t\to+\infty}{\longrightarrow} \overline{\sigma^2} = \frac{s^2}{1-g}. \tag{13}$$

Ce raisonnement et cette formule ne sont pas sans rappeler leurs analogues pour les processus de Hawkes, voir l'équation (10) dans la partie précédente de cette introduction. Cela est naturel puisqu'ils sont caractéristiques des processus auto-régressifs (ou auto-excitants), qu'ils soient exprimés en temps continu comme les processus de Hawkes ou en temps discret comme la volatilité ARCH. Pour un niveau de base $s^2 > 0$ fixé, plus le *feedback* est fort, plus la valeur moyenne du processus est élevée, et la convergence vers un équilibre devient impossible au-delà d'un certain niveau critique de rétroaction.

La principale limitation du modèle précédent est qu'il ne permet qu'un *feedback* de $\sigma_{t-1}$ sur $\sigma_t$. Pourtant, on peut considérer que $\sigma_{t-\tau}$ devrait aussi impacter directement $\sigma_t$ pour $\tau \geq 2$. C'est ce qui est observé empiriquement : la volatilité est positivement auto-corrélée, et cet effet est en fait à mémoire longue. Cela appelle l'extension naturelle qu'est le modèle ARCH($q$) :

$$\sigma_t^2 = s^2 + \sum_{\tau=1}^{q} k(\tau)\, r_{t-\tau}^2, \tag{14}$$

où $k(\tau)$ est le noyau de *feedback* et $q \geq 1$ est sa portée. C'est alors la forme du noyau $k$ qui décrit la mémoire de l'effet de *feedback* et détermine la structure d'auto-corrélation de $\sigma_t$. On obtient dans ce cadre une condition de stationnarité similaire au modèle ARCH(1), où $g$ est remplacé dans l'équation (13) par la somme $\sum_{\tau=1}^{q} k(\tau)$ des coefficients du noyau.

De nombreuses extensions du modèle ARCH ont été considérées dans la littérature. Citons en particulier les modèles GARCH (*Generalized* ARCH) et FIGARCH (*Fractionally Integrated* GARCH), introduits par Bollerslev et al. [28], [29], qui ajoutent les valeurs passées de la volatilité $\sigma^2$ aux variables explicatives. Toutefois, la plupart de ces études ont un point de vue plus économétrique qu'empirique, et les paramètres des modèles en question sont difficiles à interpréter. Nous considérons donc une extension du modèle ARCH dans une autre direction, où le *feedback* a une forme quadratique générale

$$\sigma_t^2 = s^2 + \sum_{\tau=1}^{q} L(\tau)\, r_{t-\tau} + \sum_{1\leq\tau,\tau'\leq q} K(\tau,\tau')\, r_{t-\tau} r_{t-\tau'}.$$

Ce modèle appelé QARCH (*Quadratic* ARCH) a été introduit par Sentana [102], puis étudié par Zumbach [111] et Borland et Bouchaud [30]. Le noyau linéaire $L$, souvent appelé noyau de levier, permet de modéliser le fait que la volatilité augmente davantage quand les cours chutent que quand ils montent. C'est pourquoi l'estimation de $L$ donne en général des coefficients négatifs. Mais c'est le noyau quadratique $K$ (qui peut être écrit comme une matrice avec $q$ lignes et $q$ colonnes) qui décrit la structure fine des effets de rétroaction. Si la matrice $K$ est diagonale, on retrouve le modèle ARCH (14). En revanche, une partie hors-diagonale non nulle indique que les corrélations réalisées $r_{t-\tau} r_{t-\tau'}$ entre les différents rendements passés entrent en jeu dans le *feedback*. Cela fournit un cadre assez général qui regroupe plusieurs modèles antérieurs.

Chicheportiche et Bouchaud [39] proposent une étude détaillée du modèle QARCH, ainsi que sa calibration sur des rendements d'actions américaines. Ils trouvent que la partie diagonale du noyau $K$

constitue l'effet dominant du *feedback* et décroit lentement, comme une loi puissance $K(\tau,\tau) \sim g\tau^{-\alpha}$ avec $g > 0$ et $\alpha \in (1, 3/2)$. Les coefficients hors-diagonaux sont statistiquement différents de zéro mais n'exhibent pas une structure facilement interprétable. Ces résultats montrent que la rétroaction de la volatilité est un phénomène complexe où interagissent de nombreuses échelles de temps. Toutefois, la plupart des modèles que nous avons présentés jusqu'à présent se situent à des échelles de temps supérieures ou égales à la journée. Pour aller plus loin dans la description du processus de volatilité, il semble nécessaire de se pencher sur des échelles plus courtes. En effet, l'analyse microscopique peut permettre, comme dans d'autres contextes, de mieux comprendre ce qui est observé au niveau macroscopique. La dernière partie de cette introduction présente cette nouvelle problématique.

## Modèles auto-régressifs de volatilité intra-journalière

Sur la plupart des marchés financiers, une journée de *trading* est organisée comme suit : des enchères sont organisées le matin, puis le marché ouvre (ce moment est appelé l'*open* en anglais), puis des transactions sont effectuées en continu jusqu'à la fermeture en fin d'après-midi (le *close* en anglais). Un rendement journalier (*daily return* en anglais) est l'incrément relatif (ou logarithmique) entre le prix $C_j$ d'un actif au *close* d'un jour $j$ et le prix $C_{j+1}$ au *close* du jour $j+1$ : $r^{\text{daily}} = \ln(C_{j+1}/C_j)$. Entre ces deux instants, 24 heures s'écoulent. Pendant les 17 premières heures environ, les cours n'évoluent pas puisque le marché est fermé. Pour autant, des événements ou des annonces peuvent se produire pendant la nuit, et engendrent des intentions d'achat et de vente qui ne peuvent pas être exécutées avant le lendemain. Quand arrivent les enchères au matin du jour $j+1$, ces intentions « latentes » sont résolues, et le prix saute avant même que le marché n'ouvre. Cela explique pourquoi le prix d'ouverture $O_{j+1}$ du jour $j+1$ est en général différent du prix de fermeture $C_j$ du jour $j$, ce qui produit un rendement nocturne (*overnight return* en anglais) : $r^{\text{N}} = \ln(O_{j+1}/C_j)$. Après l'*open*, le marché reprend son fonctionnement normal et les transactions sont exécutées en temps réel jusqu'au *close* du jour $j+1$, et on définit naturellement le rendement intra-journalier (*intra-day return* en anglais) par $r^{\text{D}} = \ln(C_{j+1}/O_{j+1})$. Nous obtenons donc la décomposition additive

$$r^{\text{daily}} = r^{\text{N}} + r^{\text{D}},$$

qui met en évidence les deux composantes du rendement journalier. Nous identifions deux subtilités dans cette décomposition :

- Les deux types de rendement $r^{\text{N}}$ et $r^{\text{D}}$ sont de natures complètement différentes. Le rendement nocturne transcrit de manière instantanée les informations accumulées pendant la nuit, tandis que le rendement intra-journalier correspond à une évolution progressive des cours à laquelle les participants de marché peuvent réagir en temps réel. Il semble donc intéressant de comprendre comment interagissent les deux volatilités correspondantes.
- Puisqu'il s'agit d'un saut de prix unique, le rendement nocturne ne peut pas être décomposé à son tour. En revanche, les prix sont cotés en continu quand le marché est ouvert, ce qui permet d'écrire le rendement intra-journalier comme une somme de rendements à plus haute fréquence (30 minutes, 5 minutes, 10 secondes...). On peut ainsi définir un processus de volatilité à l'intérieur d'une même journée de *trading*, et la description de ce processus peut expliquer les propriétés du rendement intra-journalier dans son ensemble.

Le premier point a été considéré par exemple par Gallo [61] et Tsiakas [106]. C'est également l'objet du chapitre 3 de cette thèse, issu de l'article [27], où nous construisons une structure ARCH de

dimension 2 qui modélise les volatilités nocturne et intra-journalière de manière jointe. Nous calibrons ce modèle sur des rendements d'actions européennes et américaines, et trouvons que les différentes rétroactions (jour sur jour, nuit sur jour, etc.) ont des formes différentes. De plus, la volatilité nocturne est sensiblement plus endogène, c'est-à-dire que la partie expliquée par les effets de *feedback* est plus grande.

En ce qui concerne le deuxième point, deux approches sont possibles. La première consiste à découper la journée de *trading* en « paniers » (*bins* en anglais) de longueur $\Delta t$, et à appliquer un modèle de volatilité en temps discret aux temps $0,\ \Delta t,\ 2\Delta t,\ 3\Delta t,\ldots$ Dans ce cas, la connaissance du prix au début de chaque *bin* $[k\Delta t, (k+1)\Delta t]$ est suffisante pour définir des rendements à l'intérieur de la journée. Engle et Sokalska [55] suivent une démarche de ce type. La deuxième approche est de modéliser le prix à haute fréquence par un processus à sauts en temps continu : il s'agit de l'approche « microstructurelle ». Cela correspond à la réalité du marché puisque pendant la journée, les prix sont cotés à tout instant mais ne sont mis à jour que de façon discrète en temps. En ayant recours à des arguments de convergence de processus, on peut alors déterminer la volatilité à basse fréquence qui est engendrée par le modèle microstructurel. La littérature récente comporte plusieurs études de ce genre, telles que celles de Cont et De Larrard [40] et Abergel et Jedidi [1] pour des processus de Poisson, ou encore Bacry et al. [18] et Jaisson et Rosenbaum [80], [81] pour les processus de Hawkes. Le dernier chapitre de cette thèse regroupe les deux approches présentées dans ce paragraphe. Nous calibrons un modèle QARCH sur des *bins* de 5 minutes, et nous montrons que pour des fréquences plus élevées, nous pouvons considérer son analogue en temps continu. Ce nouveau modèle est une généralisation des processus de Hawkes où l'on ajoute au *feedback* des effets quadratiques « hors-diagonaux ». De la sorte, nous parvenons à reproduire plusieurs caractéristiques empiriques importantes du processus de volatilité, telle que sa distribution en loi puissance et son asymétrie temporelle.

# Première partie

# Exécution optimale dynamique en présence d'un flux stochastique d'ordres de marché

# Chapitre 1

# Exécution optimale dynamique dans un modèle de prix basé sur les processus de Hawkes

Ce chapitre est un article écrit avec Aurélien Alfonsi [3] et accepté pour publication dans la revue *Finance and Stochastics*.

**Abstract.** We study a linear price impact model including other liquidity takers, whose flow of orders follows a Hawkes process. The optimal execution problem is solved explicitly in this context, and the closed-formula optimal strategy describes in particular how one should react to the orders of other traders. This result enables us to discuss the viability of the market. It is shown that Poissonian arrivals of orders lead to quite robust Price Manipulation Strategies in the sense of Huberman and Stanzl [78]. Instead, a particular set of conditions on the Hawkes model balances the self-excitation of the order flow with the resilience of the price, excludes Price Manipulation Strategies and gives some market stability.

## 1.1 Introduction

When modeling the price of an asset, we typically distinguish at least three different time scales. At the low-frequency level, the price can often be well approximated by a diffusive process. At the other end, when dealing with very high frequencies, some key features of the Limit Order Book (LOB) dynamics have to be modeled. In between, price impact models consider an intra-day mesoscopic time scale, somewhere between seconds and hours. They usually ignore most of the LOB events (limit orders, cancellations, market orders, etc.) and focus on describing the price impact of the transactions. Their goal is to be more tractable than high-frequency models and to bring quantitative

results on practical issues such as optimal execution strategies. The usual setup is well-described in Gatheral [63], who defines the price process $S$ as

$$S_t \;=\; S_0 + \int_0^t f(\dot{x}_s) G(t-s) \mathrm{d}s + \int_0^t \sigma dZ_s,$$

where $\dot{x}_s$ is the rate of trading of the liquidating agent at time $s < t$, $f(v)$ represents the instantaneous price impact of an agent trading at speed $v$, $G$ is called a « decay kernel » and $Z$ is a noise process. The quantity $f(v)G(+\infty)$ is usually called the « permanent impact », $f(v)G(0)$ the « immediate impact » and $f(v)[G(0^+)-G(+\infty)]$ the « transient impact ». The pioneering price impact models of Bertsimas and Lo [24] and Almgren and Chriss [9] consider a linear impact, with an immediate and a permanent part (which corresponds to $f(v) = \alpha v,\ G(0) > 0,\ G(0^+) = G(+\infty) > 0$ with the previous notations). These models ignore the transient part of the impact which is due to the resilience of the market and cannot be neglected when trading frequently. For that purpose, Obizhaeva and Wang [93] have considered a model that includes in addition a linear transient impact that decays exponentially (i.e. $f(v) = \alpha v,\ G(u) = \lambda + (1-\lambda)\exp(-\rho u),\ 0 \leq \lambda \leq 1,\ \rho > 0$). However, empirical evidence on market data shows that the price impact is not linear but rather concave, see e.g. Potters and Bouchaud [95], Eisler et al. [50], Mastromatteo, Tóth and Bouchaud [90], Donier [47] and more recently, Farmer, Gerig, Lillo and Waelbroeck [57]. Extensions or alternatives to the Obizhaeva and Wang model that include non-linear price impact have been proposed by Alfonsi, Fruth and Schied [5], Predoiu, Shaikhet and Shreve [97], Gatheral [63] and Guéant [66] to mention a few. Similarly, the exponential decay of the transient impact is not truly observed on market data, and one should consider more general decay kernels. Alfonsi, Schied and Slynko [6] and Gatheral, Schied and Slynko [64] consider the extension of the Obizhaeva and Wang model when the transient impact has a general decay kernel. Another simplification made by these models is that they generally assume that when the liquidating trader is passive, the price moves according to a continuous martingale, that sums up the impact of all the orders issued by other participants. However, if one wants to use these models at a higher frequency, they would naturally wonder how these orders (at least the largest ones) can be taken into account in the strategy, and if the martingale hypothesis for the price can be relaxed. This is one of the contributions of the present paper.

On the other hand, high-frequency price models aim at reproducing some statistical observations made on market data such as the autocorrelation in the signs of trades, the volatility clustering effect, the high-frequency resilience of the price, etc., and to obtain low-frequency asymptotics that are consistent with continuous diffusions. At very high frequencies, one then has to describe LOB dynamics, or a part of it. Such models have been proposed by Abergel and Jedidi [1], Huang, Lehalle and Rosenbaum [77], Cont and de Larrard [40], Garèche et al. [62], among others. However, as stressed in [40], LOB events are much more frequent than price moves. Thus, it may be relevant to model the price at the slightly lower frequency of midpoint price changes. For example, Robert and Rosenbaum [99] have proposed a model based on a diffusion with uncertainty zones that trigger the price changes. Recently, Bacry et al. [11] presented a tick-by-tick price model based on Hawkes processes, that reproduces well some empirical facts of market data. This model has then been enriched by Bacry and Muzy [10] to describe jointly the order flow and the price moves. In fact, there is a very recent and active literature that focuses on the use of mutually exciting Hawkes processes in high-frequency price models. Without being exhaustive, we mention here the works of Da Fonseca and Zaatour [44], Zheng, Roueff and Abergel [109], Filimonov and Sornette [58] and Hardiman, Bercot and Bouchaud [68]. Asymptotic and low-frequency behaviour of such models has

been investigated recently by Bacry et al. [12] and Jaisson and Rosenbaum [79].

The present paper is a contribution to this also mutually exciting literature. Its main goal is to make a bridge between high-frequency price models and optimal execution frameworks. On the one hand, Hawkes processes seem to be rich enough to describe satisfactorily the flow of market orders. On the other hand, price impact models are tractable and well-designed to calculate trading costs. The aim of our model is to grasp these two features. Thus, we consider an Obizhaeva and Wang framework where market buy and sell orders issued by other traders are modeled through Hawkes processes. This enables us to make quantitative calculations and to solve the optimal execution problem explicitly. We obtain a necessary and sufficient condition on the parameters of the Hawkes model to rule out Price Manipulation Strategies that can be seen as high-frequency arbitrages. Interestingly, we also show that modeling the order flow with a Poisson process necessarily leads to those arbitrages.

The paper is organized as follows. In Section 1.2, we set up the model and present a general criterion to exclude Price Manipulation Strategies. Section 1.3 summarizes our main results. Section 2.2.3 gives the solution of the optimal execution problem along with several comments and insights on the optimal strategy. Eventually, we analyze the existence of Price Manipulation Strategies in our model in Section 1.5 and give the conditions under which they are impossible. Cumbersome explicit formulas and technical proofs are gathered in the Appendix.

## 1.2 Model setup and the optimal execution problem

### 1.2.1 General price model

We start by describing the price model itself, without considering the execution problem. We consider a single asset and denote by $P_t$ its price at time $t$. We assume that we can write it as the sum of a « fundamental price » component $S_t$ and a « mesoscopic price deviation » $D_t$ :

$$P_t = \underbrace{S_t}_{\text{fundamental price}} + \underbrace{D_t}_{\text{mesoscopic price deviation}}. \tag{1.1}$$

Typically, these quantities are respectively related to the permanent and the transient impact of the market orders. We now specify the model and consider the framework of Obizhaeva and Wang [93] where these impacts are linear. Let $N_t$ be the sum of the signed volumes of past market orders on the book between time 0 and time $t$. By convention, a buy order is counted positively in $N$ while a sell order makes $N$ decrease, and we assume besides that $N$ is a càdlàg (right continuous with left limits) process. We assume that an order modifies the price proportionally to its size, which would correspond to a block-shaped limit order book. A proportion $\nu \in [0,1]$ of the price impact is permanent, while the remaining proportion $1-\nu$ is transient with an exponential decay of speed $\rho > 0$. This mean-reversion effect can be seen as the feedback of market makers, who affect the price using limit orders and cancellations. Namely, we consider the following dynamics for $S$ and $D$ :

$$\begin{aligned} \mathrm{d}S_t &= \frac{\nu}{q} \underbrace{\mathrm{d}N_t}_{\text{market orders}} \\ \mathrm{d}D_t &= \underbrace{-\rho\, D_t\, \mathrm{d}t}_{\text{market resilience}} + \frac{1-\nu}{q} \underbrace{\mathrm{d}N_t}_{\text{market orders}}, \end{aligned}$$

with $q > 0$. One should note that in this model, the variations in the fundamental value of the asset are revealed in its price through the process $S$. Indeed, we assume that the impact of each incoming market order, modeled through the process $N$, contains of proportion $\nu$ of « real » or « exogenous » information, and that the remaining proportion $1 - \nu$ is of endogenous origin and will vanish over time.

**Remark 1.2.1.** *This model assumes a linear price impact with an exponential resilience. As mentioned in the introduction, these assumptions are challenged by empirical facts, and it would be for sure interesting and relevant to enrich the model by considering a non linear price impact and a more general decay of the impact. However, the new feature of the model with respect to the literature on optimal execution is to add a flow of market orders issued by other traders. This is why we afford to make these simplifying assumptions that give analytical tractability, which is important to calculate the optimal execution strategy in real time. Thus, the model is meant to constitute a first step in dynamic optimal execution with the price driven by point processes, and we plan to confront it to market data in a future work.*

As usual, we consider $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space where $\mathbb{P}$ weights the probability of the market events. We assume that the process $(N_t)_{t\geq 0}$ has bounded variation and is square integrable, i.e. $\sup_{s\in[0,t]} \mathbb{E}[N_s^2] < \infty$ for any $t \geq 0$, and we define $(\mathcal{F}_t)_{t\geq 0}$ the natural filtration of $N$, $\mathcal{F}_t = \sigma(N_s, s \leq t)$ for $t \geq 0$. We will specify in Section 1.2.3 which dynamics we consider for $N$ in this paper.

### 1.2.2 Optimal execution framework

We now consider a particular trader who wants to buy or sell a given quantity of assets on the time interval $[0, T]$. Through the paper, we will call this trader the "strategic trader" to make the distinction between his market orders and all the other market orders, that are described by $N$. We will denote by $X_t$ the number of assets owned by the strategic trader at time $t$. We assume that the process is $(\mathcal{F}_t)$-adapted, with bounded variation and càglàd (left continuous with right limits) which means that the strategic trader observes all the information available on the market, and that he can react instantly to the market orders issued by other traders. Besides, a strategy that liquidates $x_0$ assets on $[0, T]$ should satisfy $X_0 = x_0$ and $X_{T^+} = 0 : x_0 > 0$ (resp. $x_0 < 0$) corresponds to to a sell (resp. buy) program.

**Definition 1.2.1.** *A liquidating strategy $X$ for the position $x_0 \in \mathbb{R}$ on $[0, T]$ is admissible if it is $(\mathcal{F}_t)$-adapted, càglàd, square integrable, with bounded variation and such that $X_0 = x_0$ and $X_{T^+} = 0$, a.s.*

**Remark 1.2.2.** *An admissible strategy $X$ has a countable set $\mathcal{D}_X$ of times of discontinuity on $[0, T]$, and can have a non-zero continuous part $X_t^c = X_t - \sum_{\tau\in\mathcal{D}_X\cap[0,t)} (X_{\tau^+} - X_\tau),\ t \in [0, T]$.*

One then has to specify how the strategic trader modifies the price, as well as the cost induced by his trading strategy. Again, we will consider the Obizhaeva and Wang model [93] with the same price impact as above. However, we let the possibility that the proportion $\epsilon \in [0, 1]$ of permanent impact of the strategic trader could be different from the one of the other traders, which we note $\nu \in [0, 1]$. Of course, a reasonable choice would be to set $\epsilon = \nu$ to consider all orders equally, but the model

allows for more generality. We then assume the following dynamics

$$\mathrm{d}S_t = \frac{1}{q}\left(\nu \mathrm{d}N_t + \epsilon \mathrm{d}X_t\right), \tag{1.2}$$

$$\mathrm{d}D_t = -\rho\, D_t\, \mathrm{d}t + \frac{1}{q}\left((1-\nu)\mathrm{d}N_t + (1-\epsilon)\mathrm{d}X_t\right). \tag{1.3}$$

With the assumptions on $N$ and $X$, the price processes $P$, $S$ and $D$ have left and right limits. More precisely, in case of discontinuity at time $t$, (1.2) and (1.3) have to be read here as follows

$$S_t - S_{t-} = \frac{\nu}{q}(N_t - N_{t-}),\ S_{t+} - S_t = \frac{\epsilon}{q}(X_{t+} - X_t),$$
$$D_t - D_{t-} = \frac{1-\nu}{q}(N_t - N_{t-}),\ D_{t+} - D_t = \frac{1-\epsilon}{q}(X_{t+} - X_t).$$

For the sake of tractability only, we make the assumption of a block-shaped Limit Order Book. Thus (see [93]), when the strategic trader places at time $t$ an order of size $v \in \mathbb{R}$ ($v > 0$ for a buy order and $v < 0$ for a sell order), it has the following cost

$$\pi_t(v) = \int_0^v \left[P_t + \frac{1}{q}y\right] \mathrm{d}y = \underbrace{P_t\, v}_{\text{cost at the current price}} + \underbrace{\frac{v^2}{2q}}_{\text{impact cost}}.$$

Since $P_{t+} = P_t + \frac{v}{q}$, this cost amounts to trade all the assets at the average price $(P_t + P_{t+})/2$. We stress here that if an order has just occurred, i.e. $N_t - N_{t-} \neq 0$, the value of $P_t$ is different from $P_{t-}$ and takes into account the price impact of this order. Therefore, the cost of an admissible strategy $X$ is given by

$$\begin{aligned} C(X) &= \int_{[0,T)} P_u\, \mathrm{d}X_u + \frac{1}{2q}\sum_{\tau\in\mathcal{D}_X\cap[0,T)} (\Delta X_\tau)^2 - P_T X_T + \frac{1}{2q}\, X_T^2 \\ &= \int_{[0,T)} P_u\, \mathrm{d}X_u^c + \sum_{\tau\in\mathcal{D}_X\cap[0,T)} P_\tau(\Delta X_\tau) + \frac{1}{2q}\sum_{\tau\in\mathcal{D}_X\cap[0,T)} (\Delta X_\tau)^2 - P_T X_T + \frac{1}{2q}\, X_T^2, \end{aligned} \tag{1.4}$$

since at time $T$ all the remaining assets have to be liquidated. Here, the sum brings on the countable times of discontinuity $\mathcal{D}_X$ of $X$, and the jumps $\Delta X_\tau = X_{\tau+} - X_\tau \neq 0$ for $\tau \in \mathcal{D}_X$. We note that all the terms involved in the cost function are integrable, thanks to the assumption on the square integrability of $X$ and $N$.

**Remark 1.2.3.** *With the initial market price $P_0$ taken as a reference, $-P_0 \times x_0$ is the mark-to-market liquidation cost. Thus, $C(X) + P_0 \times x_0$ can be seen as an additional liquidity cost of it is positive. If it is negative, its absolute value can be seen as the gain associated to the strategy $X$.*

**Remark 1.2.4.** *The cost defined by (1.4) in the price model (1.1), (1.2) and (1.3) is a deterministic function of $(X_t)_{t\in[0,T]}$, $(N_t)_{t\in[0,T]}$, $S_0$, $D_0$ and the parameters $q$, $\nu$, and $\epsilon$. In this remark, we denote by $C(X, N, S_0, D_0, q)$ this function when $\nu$ and $\epsilon$ are given. From (1.2), (1.3) and (1.4), we have the straightforward property*

$$C(X, N, S_0, D_0, q) = C(-X, -N, -S_0, -D_0, q). \tag{1.5}$$

*Observing that* $qC(X) = \int_{[0,T)} qP_u \mathrm{d}X_u + \frac{1}{2}\sum_{0\leq\tau<T}(\Delta X_\tau)^2 - (qP_T)X_T + \frac{1}{2}(X_T)^2$, *and remarking that* $qS$ *and* $qD$ *satisfy* (1.2) *and* (1.3) *with* $q = 1$, *we also get*

$$qC(X, N, S_0, D_0, q) = C(X, N, qS_0, qD_0, 1). \tag{1.6}$$

**Remark 1.2.5.** *Since* $X$ *is a càglàd process and* $N$ *is a càdlàg process, we will have to work with làdlàg (with finite right-hand and left-hand limits) processes. When* $Z$ *is a làdlàg process, we set* $\Delta^- Z_t = Z_t - Z_{t-}$ *and* $\Delta^+ Z_t = Z_{t^+} - Z_t$ *the left and right jumps of* $Z$, *and* $Z^c_t = Z_t - \sum_{0\leq\tau<t}\Delta^+ Z_\tau - \sum_{0<\tau\leq t}\Delta^- Z_\tau$ *the continuous part of* $Z$. *We also set* $\Delta Z_t = Z_{t+} - Z_{t-}$ *and use the shorthand notation* $\mathrm{d}Z_t = \mathrm{d}Z^c_t + \Delta Z_t$. *If* $\mathrm{d}Z_t = \mathrm{d}\tilde{Z}_t$ *for some other làdlàg process* $\tilde{Z}$, *this means that* $\mathrm{d}Z^c_t = \mathrm{d}\tilde{Z}^c_t$ *and* $\Delta Z_t = \Delta\tilde{Z}_t$. *In particular, when* $Z$ *is càdlàg and* $\tilde{Z}$ *is càglàd, this means that* $Z_t - Z_{t-} = \tilde{Z}_{t^+} - \tilde{Z}_t$ *at the jump times.*

Then, the optimal execution problem consists in finding an admissible strategy $X$ that minimizes the expected cost $\mathbb{E}[C(X)]$ for a given initial position $x_0 \in \mathbb{R}$. This problem for $x_0 = 0$ is directly related to the existence of Price Manipulation Strategies as defined below.

**Definition 1.2.2.** *A Price Manipulation Strategy (PMS) in the sense of Huberman and Stanzl [78] is an admissible strategy* $X$ *such that* $X_0 = X_{T^+} = 0$ *a.s. for some* $T > 0$ *and* $\mathbb{E}[C(X)] < 0$.

We have the following result that gives a necessary and sufficient condition to exclude PMS.

**Theorem 1.2.1.** *The model does not admit PMS if, and only if the process* $P$ *is a* $(\mathcal{F}_t)$*-martingale when* $X \equiv 0$. *In this case, the optimal strategy* $X^{\mathrm{OW}}$ *is the same as in the Obizhaeva and Wang [93] model. It is given by*

$$\Delta X^{\mathrm{OW}}_0 = -\frac{x_0}{2+\rho T}, \quad \Delta X^{\mathrm{OW}}_T = -\frac{x_0}{2+\rho T}, \quad \mathrm{d}X^{\mathrm{OW}}_t = -\rho\frac{x_0}{2+\rho T}\,\mathrm{d}t \quad \text{for } t \in (0,T), \tag{1.7}$$

*and has the expected cost* $\mathbb{E}[C(X^{\mathrm{OW}})] = -P_0 x_0 + \left[\frac{1-\epsilon}{2+\rho(T-t)} + \frac{\epsilon}{2}\right] x_0^2/q$.

This theorem is proved in Appendix 1.8. Similar results are standard in financial mathematics, but to the best of our knowledge, it has not yet been formulated as such in the literature in a context with price impact and with respect to the notion of Price Manipulation Strategies. In usual optimal execution frameworks, the unaffected price is assumed *a priori* to be a martingale, which is not the case here. Note that if $P$ is a martingale, the optimal strategy is very robust in the sense that it does not depend on $N$, and is therefore the same as the one in the Obizhaeva and Wang model [93] that corresponds to $N \equiv 0$ and $D_0 = 0$. In fact, it does not depend either on $\epsilon$ and $\nu$, and only depends on $\rho$.

Theorem 1.2.1 indicates that suitable models for the order flow $N$ should be such that $P$ is, roughly speaking, close to a martingale when the strategic trader is absent, so that arbitrage opportunities are short-lived and not too visible. This raises at least three questions. Which "simple" processes $N$ can lead to a martingale price $P$? Can we characterize the optimal strategy when $P$ is not a martingale? In particular, in the latter case, how does the optimal strategy take the market orders issued by other participants into account? In this paper, we study these questions when $N$ follows a Hawkes process.

**Remark 1.2.6.** *The model can be generalized by adding a càdlàg $(\mathcal{F}_t)$-martingale $S^0$ to the price process $P$, i.e. if we replace (1.1) by $P_t = S_t + D_t + S^0_t$, with $S^0_0 = 0$. This does not change the optimal execution problem since, using an integration by parts, $S^0$ adds the following term to the cost*

$$\begin{aligned}\int_{[0,T)} S^0_t \, \mathrm{d}X_t \; - \; S^0_T X_T \;\; &= \;\; S^0_T X_T \; - \; S^0_0 X_0 \; - \; \int_{[0,T)} X_t \, \mathrm{d}S^0_t \; - \; S^0_T X_T \\ &= \; - \int_{[0,T)} X_t \, \mathrm{d}S^0_t,\end{aligned}$$

*which has a zero expected value from the martingale property. Let us note that there is no covariation between the processes $X$ and $S_0$ since they do not jump simultaneously and $X$ has bounded variations.*

**Remark 1.2.7.** *Similarly, when $N$ is a càdlàg $(\mathcal{F}_t)$-martingale and $X$ is an admissible liquidating strategy for $X_0 = x_0$, we have*

$$\mathbb{E}[C(X)] = \mathbb{E}\left[\int_{[0,T)} D_u \, \mathrm{d}X_u \; + \; \frac{1-\epsilon}{2q} \sum_{0 \le \tau < T} (\Delta X_\tau)^2 \; - \; D_T X_T \; + \; \frac{1-\epsilon}{2q} \, X_T^2\right] + \frac{\epsilon}{2q} x_0^2,$$

*since $x_0^2 = \int_{[0,T+]} \mathrm{d}[(X_t - X_0)^2] = 2\int_{[0,T)}(X_u - X_0)\mathrm{d}X_u + \sum_{0\le\tau<T}(\Delta X_\tau)^2 - 2(X_T - X_0)X_T + X_T^2$. When $\epsilon \in [0,1)$, we set $X^\epsilon_t = (1-\epsilon)X_t$ and get*

$$\mathbb{E}[C(X)] = \frac{1}{q(1-\epsilon)}\mathbb{E}\left[\int_{[0,T)} qD_u \mathrm{d}(X^\epsilon_u) + \frac{1}{2}\sum_{0\le\tau<T}(\Delta X^\epsilon_\tau)^2 - qD_T X^\epsilon_T + \frac{1}{2}(X^\epsilon_T)^2\right] + \frac{\epsilon}{2q}x_0^2. \quad (1.8)$$

*Therefore, $X$ is optimal if, and only if $X^\epsilon$ is optimal in the model with $\epsilon = \nu = 0$, $q = 1$ and an incoming flow of market orders equal to $(1-\nu)N$.*

### 1.2.3 The MIH model

#### Definitions and notations

We introduce the MIH (Mixed-market-Impact Hawkes) price model, where

$$N_t = N^+_t - N^-_t,$$

the process $(N^+, N^-)$ being a symmetric two-dimensional marked Hawkes process of intensity $(\kappa^+, \kappa^-)$. The process $(N^+, N^-, \kappa^+, \kappa^-)$ is càdlàg and jumps when $N$ jumps. We note $n^+(\mathrm{d}t, \mathrm{d}v)$ and $n^-(\mathrm{d}t, \mathrm{d}v)$ the Poisson measures on $\mathbb{R}^+ \times \mathbb{R}^+$ associated to $N^+$ and $N^-$ respectively, where the variable $v$ stands for the amplitudes of the jumps, i.e. the volumes of incoming market orders. We restrain to the case of i.i.d. unpredictable marks of common law $\mu$ on $\mathbb{R}^+$, i.e. for any $A \in \mathcal{B}(\mathbb{R}^+)$ and $t \ge 0$,

$$\kappa^\pm_t \; \mu(A) = \lim_{h \to 0^+} \frac{1}{h}\mathbb{E}[n^\pm([t, t+h], A)|\mathcal{F}_t],$$

where $\mathcal{F}_t = \sigma\left(N_u^+, N_u^-, u \leq t\right) = \sigma\left(N_u, u \leq t\right)$ as defined earlier. In other words, at time $t$, the conditional instantaneous jump intensity of $N^\pm$ is given by $\kappa_t^\pm$, and the amplitudes of the jumps are i.i.d. variables of law $\mu$ which are independent from the past, i.e. from $\mathcal{F}_{t^-}$. We also define

$$m_k = \int_{\mathbb{R}^+} v^k \mu(\mathrm{d}v), \ k \in \mathbb{N},$$

assuming moreover that $m_2 < \infty$. We choose the Hawkes kernel to be the exponential $t \mapsto \exp(-\beta t)$, $\beta \geq 0$, so that $(N^+, N^-, \kappa^+, \kappa^-)$ is Markovian. Thus, we set

$$\begin{pmatrix}\kappa_t^+\\ \kappa_t^-\end{pmatrix} = \begin{pmatrix}\kappa_\infty\\ \kappa_\infty\end{pmatrix} + \left[\begin{pmatrix}\kappa_0^+\\ \kappa_0^-\end{pmatrix} - \begin{pmatrix}\kappa_\infty\\ \kappa_\infty\end{pmatrix}\right]\exp(-\beta t) + \int_0^t \int_{(\mathbb{R}^+)^2} \exp(-\beta(t-u)) \begin{pmatrix}\varphi_{\mathrm{s}}(v^+/m_1) & \varphi_{\mathrm{c}}(v^-/m_1)\\ \varphi_{\mathrm{c}}(v^+/m_1) & \varphi_{\mathrm{s}}(v^-/m_1)\end{pmatrix} \cdot \begin{pmatrix}n^+(\mathrm{d}u, \mathrm{d}v^+)\\ n^-(\mathrm{d}u, \mathrm{d}v^-)\end{pmatrix},$$

where $\kappa_\infty \geq 0$ is the common baseline intensity of $N^+$ and $N^-$, and $\varphi_{\mathrm{s}}, \varphi_{\mathrm{c}} : \mathbb{R}^+ \to \mathbb{R}^+$ are measurable positive functions that satisfy

$$\iota_{\mathrm{s}} := \int_{\mathbb{R}^+} \varphi_{\mathrm{s}}(v/m_1)\mu(\mathrm{d}v) < \infty \quad , \quad \iota_{\mathrm{c}} := \int_{\mathbb{R}^+} \varphi_{\mathrm{c}}(v/m_1)\mu(\mathrm{d}v) < \infty.$$

We assume besides that

$$\int_{\mathbb{R}^+} \varphi_{\mathrm{s}}^2(v/m_1)\mu(\mathrm{d}v) < \infty, \ \int_{\mathbb{R}^+} \varphi_{\mathrm{c}}^2(v/m_1)\mu(\mathrm{d}v) < \infty$$

to have $\sup_{s\in[0,t]} \mathbb{E}[N_s^2] < \infty$, and we note that this property is automatically satisfied when $\varphi_{\mathrm{s}}$ and $\varphi_{\mathrm{c}}$ have a sublinear growth since we have assumed $m_2 < \infty$. From the modeling point ov view, we may expect that the functions $\varphi_{\mathrm{s}}$ and $\varphi_{\mathrm{c}}$ are nondecreasing : the larger an order is, the more other orders it should trigger. However, we do not need this monotonicity assumption in the mathematical analysis.

Equivalently, in this Markovian setting, the intensities $\kappa_t^+$ and $\kappa_t^-$ follow the dynamics

$$\begin{aligned}\mathrm{d}\kappa_t^+ &= -\beta\left(\kappa_t^+ - \kappa_\infty\right)\mathrm{d}t + \varphi_{\mathrm{s}}(\mathrm{d}N_t^+/m_1) + \varphi_{\mathrm{c}}(\mathrm{d}N_t^-/m_1),\\ \mathrm{d}\kappa_t^- &= -\beta\left(\kappa_t^- - \kappa_\infty\right)\mathrm{d}t + \varphi_{\mathrm{c}}(\mathrm{d}N_t^+/m_1) + \varphi_{\mathrm{s}}(\mathrm{d}N_t^-/m_1),\end{aligned} \tag{1.9}$$

where formally, $\int_0^t \varphi_{\mathrm{s}}(\mathrm{d}N_u^+/m_1) = \int_0^t \int_{\mathbb{R}^+} \varphi_{\mathrm{s}}(v/m_1)\ n^+(\mathrm{d}u, \mathrm{d}v)$ for $t \geq 0$. As pointed out in Hardiman, Bercot and Bouchaud [68] and Bacry and Muzy [10] for instance (in a slightly different context since in our framework, $N$ models market orders only), a power-law Hawkes kernel is more in accordance with market data than an exponential one. It is possible in principle to approximate a completely monotone decaying kernel with a multi-exponential one while preserving a Markovian framework, at the cost of increasing the dimension of the state space, see for example Alfonsi and Schied [4]. This investigation is left for future research.

Note that $N^+$ and $N^-$ boil down to independent composed Poisson processes in the case $\beta = 0$, $\varphi_{\mathrm{s}} = \varphi_{\mathrm{c}} \equiv 0$. The meaning of the parameters is rather clear : $\kappa^+$ and $\kappa^-$ are mean reverting processes, and $\iota_{\mathrm{s}}$ and $\iota_{\mathrm{c}}$ respectively describe how a market buy order increases the instantaneous probability of buy (resp. sell) orders. More precisely, $\iota_{\mathrm{s}}$ encodes both the splitting of meta-orders, and the fact that participants tend to follow market trends (which is called the herding effect). On the other hand, $\iota_{\mathrm{c}}$ describes opportunistic traders that sell (resp. buy) after a sudden rise (resp. fall) of the price. The functions $\varphi_{\mathrm{s}}$ and $\varphi_{\mathrm{c}}$ allow respectively the self and cross-excitations in the order flow to depend on the volumes of the orders. For instance, for constant functions $\varphi_{\mathrm{s}} \equiv \iota_{\mathrm{s}}$ and $\varphi_{\mathrm{c}} \equiv \iota_{\mathrm{c}}$, the

model boils down to the standard Hawkes model where $\kappa^{\pm}$ makes jumps of constant size when $N^{\pm}$ jumps.

Hawkes processes have been recently used in the literature to model the price. In particular, Bacry et al. [11] consider a similar model where $N$ models all price moves, with $\nu = 1$, $\iota_{\mathrm{s}} = 0$ and deterministic jumps (i.e. $\mu$ is a Dirac mass). More recently, Bacry and Muzy [10] have proposed an four-dimensional Hawkes process to model the market buy and sell orders together with the up and down events on the price. In contrast, the model that we study here determines the price impact of an order in function of its size. For the reader who is not accustomed to Hawkes processes, we point the original paper [73], the paper by Embrechts et al. [51] for an overview of multivariate marked Hawkes processes and the book of Daley and Vere-Jones [46] for a more detailed account.

**Remark 1.2.8.** *As one can see in Equation (1.9), the orders of the strategic trader do not impact the jump rates $\kappa^+$ and $\kappa^-$ (there is no $\mathrm{d}X_t$ term), as opposed to the market orders issued by other traders. The first reason for this modeling choice is tractability. However, it is found empirically by Tòth et al. [105] that the main contribution to the self-excitation of the order flow comes from the splitting effect. Each individual trader tends to post several orders of the same nature (buy or sell) in a row, which creates auto-correlation in the signs of trades, and this effect is significantly stronger than the mutual excitation between different traders. Thus, it is an acceptable approximation to neglect the excitation coming from the orders of the strategic trader. Of course, it would be nice to find in the future a tractable model that gives a unified framework for the mutual excitation that considers equally all the market orders.*

### Stationarity and low-frequency asymptotics of the MIH model

Up to now, we have presented the MIH model without assuming stationarity. In most models featuring Hawkes processes, stationarity is an *a priori* assumption, but here, we do not need it to derive the optimal strategy. However, if one wishes to use the MIH model with constant parameters on a large time period, it may be reasonable to consider parameters that satisfy stationarity. This is why we present here a few results that are standard in the literature of Hawkes processes.

We consider the MIH model when the strategic trader is absent, i.e. $X \equiv 0$.

**Proposition 1.2.1.** *The process $(\kappa_t^+, \kappa_t^-)$ converges to a stationary law if, and only if $\iota_{\mathrm{s}} + \iota_{\mathrm{c}} < \beta$.*

*Proof.* We can apply the results of the existing literature on marked Hawkes processes with unpredictable marks (for instance Hawkes and Oakes [75], Brémaud ans Massoulié [33] or Daley and Vere-Jones [46]) to obtain that $(\kappa_t^+, \kappa_t^-)$ converges to a stationary law if the largest eigenvalue of

$$\int_{\mathbb{R}^+\times\mathbb{R}^+} \exp(-\beta t) \begin{pmatrix} \varphi_{\mathrm{s}}(v/m_1) & \varphi_{\mathrm{c}}(v/m_1) \\ \varphi_{\mathrm{c}}(v/m_1) & \varphi_{\mathrm{s}}(v/m_1) \end{pmatrix} \mathrm{d}t\ \mu(\mathrm{d}v) = \frac{1}{\beta} \begin{pmatrix} \iota_{\mathrm{s}} & \iota_{\mathrm{c}} \\ \iota_{\mathrm{c}} & \iota_{\mathrm{s}} \end{pmatrix}$$

is strictly below unity. Conversely, if $\iota_{\mathrm{s}} + \iota_{\mathrm{c}} \geq \beta$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\kappa_t^+ + \kappa_t^-] = 2\beta\kappa_\infty + (\iota_{\mathrm{s}} + \iota_{\mathrm{c}} - \beta)\mathbb{E}[\kappa_t^+ + \kappa_t^-] \geq 2\beta\kappa_\infty$$

and the process cannot be stationary.

We now study the low-frequency asymptotics of the price process $P$ in the MIH model. We consider the sequence $P_t^{(n)} = P_{nt}/\sqrt{n}$ for $n \geq 1$. We have $P_t^{(n)} = S_t^{(n)} + D_t^{(n)}$, where we also set $S_t^{(n)} = S_{nt}/\sqrt{n}$ and $D_t^{(n)} = D_{nt}/\sqrt{n}$. To study the behaviour of $D^{(n)}$, we need the following lemma.

**Lemma 1.2.1.** *When $\iota_s + \iota_c < \beta$, the expectation $\mathbb{E}[D_t^2]$ converges to a finite positive value as $t \to +\infty$.*

The proof of this lemma is rather straightforward. We just have to calculate $\mathbb{E}[\delta_t^2]$, $\mathbb{E}[\delta_t D_t]$ and $\mathbb{E}[D_t^2]$ and check that these expectations converge when $\iota_s + \iota_c < \beta$. This result implies that $(D_{t_1}^{(n)}, \dots, D_{t_k}^{(n)})$ converges to zero for the $L^2$ norm for any $0 \leq t_1 \leq \cdots \leq t_k$. This gives that the process $D^{(n)}$ converges to zero.

We thus focus on the convergence of $S_t^{(n)} = \frac{\nu}{q} \frac{N_{nt}^+ - N_{nt}^-}{\sqrt{n}}$. If the jumps of $N$ are bounded, i.e. $\mu$ has bounded support, and $v \mapsto \varphi_s(v/m_1)$ and $v \mapsto \varphi_c(v/m_1)$ are bounded on the support of $\mu$ (which are reasonable assumptions in practice), a straightforward adaptation of Corollary 1 of Bacry et al. [12] gives the convergence in law of $S^{(n)}$ to a non-standard Brownian motion with zero drift.

## 1.3 Main results

Now that the whole framework is set up, we present the main results of the present paper.

- The optimal execution problem can be solved explicitly in the MIH model and the optimal strategy has still a quite simple form, see Theorem 1.4.1. Of course, this result relies on the assumptions of linear price impact and exponential decay kernel, which are not in accordance with empirical facts, see for example Potters and Bouchaud [95] and Bouchaud et al. [31]. We mention here that it would be possible to keep an affine structure of the optimal strategy by considering complete monotone decay kernels as in Alfonsi and Schied [4]. However, we believe that the optimal strategy is interesting at least from a qualitative point of view, since it gives clear insights on how to react optimally to observed market orders and on the role of the different parameters of the model.
- Price Manipulation Strategies necessarily appear when the flow of market orders is Poissonian, and they are rather robust in the sense that they can be implemented without knowing the model parameters. Namely, the strategy which consists in trading instantly a small proportion of the volume of each incoming market order in the opposite direction is profitable on average, see Proposition 1.5.2. This justifies to consider more elaborate dynamics for the order arrivals.
- Even in a non-Poissonian MIH setup, Price Manipulation Strategies can arise. Depending on the parameters of the model and on the size of each observed market order, one should either trade instantly in the opposite direction to take market resilience into account, or in the same direction to take advantage of the self-excitation property of Hawkes processes. However, our framework allows for a specific equilibrium to take place, that we call the Mixed-market-Impact Hawkes Martingale (MIHM) model, where PMS disappear.
- In the MIHM model, one has in particular $\iota_s > \iota_c$, $\nu < 1$ and $\beta = \rho$, and the self-excitation property of the order flow exactly compensates the price resilience induced by market makers. The resulting price process is a martingale even at high frequencies, and in this case we find that the optimal strategy and cost function are those of Obizhaeva and Wang [93]. The conditions of this model imply that if $\iota_c = 0$, the norm $\iota_s/\beta$ of the Hawkes kernel that symbolizes the

endogeneity ratio of the market, see Filimonov and Sornette [58], should be equal to $1-\nu$, i.e. the proportion of market impact which is transient.

– The fact of reacting to the market orders of other traders with instantaneous market orders can trigger chain reactions and lead to market instability. We show that in the MIH framework, the conditions under which it is profitable for the strategic trader to react instantaneously to other trades are quite equivalent to the existence of PMS. Although the model is clearly a simplified view of the market, it is remarkable to obtain in this case such a clear connection between market stability and free profits. It would be interesting for further reasearch to investigate if this conclusion still holds in a more general model.

## 1.4 The optimal strategy

We need to introduce some notations to present the main results on the optimal execution. Instead of working with $\kappa_t^+$ and $\kappa_t^-$, we will rather use $\delta_t = \kappa_t^+ - \kappa_t^-$ and $\Sigma_t = \kappa_t^+ + \kappa_t^-$ that satisfy from (1.9)

$$\mathrm{d}\delta_t \;=\; -\beta\;\delta_t\;\mathrm{d}t \;+\; \mathrm{d}I_t \quad , \quad \mathrm{d}\Sigma_t \;=\; -\beta\;(\Sigma_t - 2\kappa_\infty)\;\mathrm{d}t \;+\; \mathrm{d}\overline{I}_t, \tag{1.10}$$

where

$$\begin{aligned} I_t &= \int_0^t \left[(\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(\mathrm{d}N_u^+/m_1) - (\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(\mathrm{d}N_u^-/m_1)\right], \\ \overline{I}_t &= \int_0^t \left[(\varphi_{\mathrm{s}} + \varphi_{\mathrm{c}})(\mathrm{d}N_u^+/m_1) + (\varphi_{\mathrm{s}} + \varphi_{\mathrm{c}})(\mathrm{d}N_u^-/m_1)\right]. \end{aligned} \tag{1.11}$$

The processes $I$ and $\overline{I}$ are càdlàg processes which describe intensity jumps, and their jump times are those of $N$. In the standard Hawkes framework where $\varphi_{\mathrm{s}}$ and $\varphi_{\mathrm{c}}$ are constant, one has $\varphi_{\mathrm{s}} \equiv \iota_{\mathrm{s}}$ and $\varphi_{\mathrm{c}} \equiv \iota_{\mathrm{c}}$, and when $N$ jumps, $I$ jumps of $(\iota_{\mathrm{s}} - \iota_{\mathrm{c}})\ \mathrm{sgn}(\Delta N_t)$ and $\overline{I}$ of $\iota_{\mathrm{s}} + \iota_{\mathrm{c}}$.

We note $(\tau_i)_{i\geq 1}$ the ordered random jump times of $N$ and set $\tau_0 = 0$. For $t \in [0,T]$, we also note $\chi_t$ the total number of jumps of $I$ that occurred between time 0 and time $t$. From (2.43), we have

$$\delta_t \;=\; \delta_0 \exp(-\beta t) \;+\; \sum_{l=1}^{\chi_t} \exp(-\beta(t-\tau_l))\Delta I_{\tau_l} \;=\; \delta_0\ \exp(-\beta t) + \exp(-\beta t)\ \Theta_{\chi_t},$$

where we define $\Theta_0 = 0$ and

$$\Theta_i \;=\; \sum_{l=1}^{i} \exp(\beta\tau_l)\Delta I_{\tau_l} \;=\; \sum_{0<\tau\leq\tau_i} \exp(\beta\tau)\ \Delta I_\tau, \quad i \geq 1.$$

For $i \geq 0$ and $t \in [\tau_i, \tau_{i+1})$, we obtain that $\delta_t \exp(\beta t) = \delta_0 + \Theta_i$ only depends on $t$ through the integer $i = \chi_t$. We introduce the useful quantities

$$\alpha = \iota_{\mathrm{s}} - \iota_{\mathrm{c}}, \quad \eta = \beta - \alpha,$$

and the two continuously differentiable functions $\zeta, \omega : \mathbb{R} \to \mathbb{R}^+$ defined by

$$\zeta(0) = 1 \quad \text{and} \quad \forall y \neq 0, \quad \zeta(y) = \frac{1-\exp(-y)}{y}, \tag{1.12}$$

$$\zeta'(0) = -1/2 \quad \text{and} \quad \forall y \neq 0, \quad \zeta'(y) = \frac{(1+y)\exp(-y)-1}{y^2} = \frac{\exp(-y)-\zeta(y)}{y},$$

$$\omega(0) = 1/2 \quad \text{and} \quad \forall y \neq 0, \quad \omega(y) = \frac{\exp(-y)-1+y}{y^2} = \frac{1-\zeta(y)}{y}, \tag{1.13}$$

$$\omega'(0) = -1/6 \quad \text{and} \quad \forall y \neq 0, \quad \omega'(y) = \frac{2(1-\exp(-y))-y(1+\exp(-y))}{y^3} = \frac{2\zeta(y)-1-\exp(-y)}{y^2}.$$

Both functions non-increasing, diverge to $+\infty$ at negative infinity and vanish at positive infinity. Let us now enounce the main theorem for the optimal execution problem.

**Theorem 1.4.1.** *Let $\epsilon \in [0,1)$. The optimal strategy $X^*$ that minimizes the expected cost $\mathbb{E}[C(X)]$ among admissible strategies that liquidate $x_0$ assets is explicit. It is a linear combination of $(x_0, D_0, \delta_0, I, N)$ and can be written as*

$$X^* = X^{\mathrm{OW}} + X^{\mathrm{trend}} + X^{\mathrm{dyn}},$$

*where*
- *$X^{\mathrm{OW}}$ is the optimal strategy in the Obizhaeva and Wang [93] model, given by (1.7) in Theorem 1.2.1,*
- *$X^{\mathrm{trend}}$ is the « trend strategy », given by (1.19).*
- *$X^{\mathrm{dyn}}$ is the « dynamic strategy », given by (1.20).*

*The strategy $X^{\mathrm{OW}}$ is a linear function of $x_0$, $X^{\mathrm{trend}}$ is a linear function of $(D_0, \delta_0)$ while $X^{\mathrm{dyn}}$ is a linear function of the processes $I$ and $N$. The discontinuity times of $X^{\mathrm{dyn}}$ are those of $N$, and if $N$ jumps at time $\tau \in (0,T)$, we have*

$$(1-\epsilon)\Delta X^{\mathrm{dyn}}_\tau = \frac{1+\rho(T-\tau)}{2+\rho(T-\tau)}\left\{\frac{m_1}{\rho}\,\Delta I_\tau - (1-\nu)\,\Delta N_\tau\right\} + \frac{m_1}{2\rho}(\nu\rho-\eta)\frac{\rho(T-\tau)^2 \times \omega(\eta(T-\tau))}{2+\rho(T-\tau)}\,\Delta I_\tau. \tag{1.14}$$

*All explicit formulas are given in Appendix 1.6. The value function of the problem is given by*

$$\begin{aligned} q \times \mathcal{C}(t,x,d,z,\delta,\Sigma) &= -q(z+d)x + \left[\frac{1-\epsilon}{2+\rho(T-t)} + \frac{\epsilon}{2}\right]x^2 + \frac{\rho(T-t)}{2+\rho(T-t)}\left[qd - \mathcal{G}_\eta(T-t)\,\frac{\delta m_1}{\rho}\right]x \\ &\quad - \frac{1}{1-\epsilon} \times \frac{\rho(T-t)/2}{2+\rho(T-t)}\left[qd - \mathcal{G}_\eta(T-t)\,\frac{\delta m_1}{\rho}\right]^2 + \hat{c}_\eta(T-t)\left(\frac{\delta m_1}{\rho}\right)^2 \\ &\quad + e(T-t)\,\Sigma + g(T-t), \end{aligned}$$

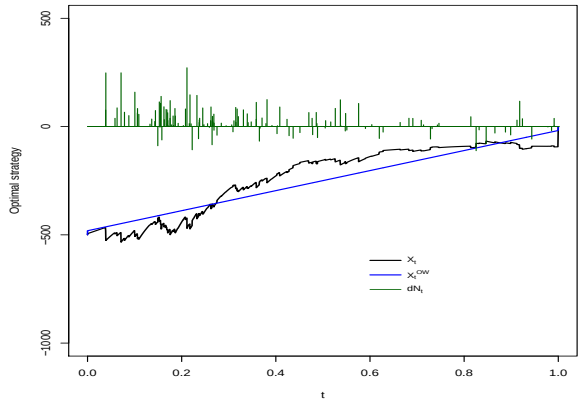
*where for $u \in [0,T]$,*

$$\begin{aligned} \mathcal{G}_\eta(u) &= \zeta(\eta u) + \nu\rho u\,\omega(\eta u), \\ \hat{c}_\eta(u) &= \frac{1}{1-\epsilon} \times (\eta - \nu\rho)^2 \frac{\rho u^3}{8}\omega'(\eta u)\zeta(\eta u). \end{aligned}$$

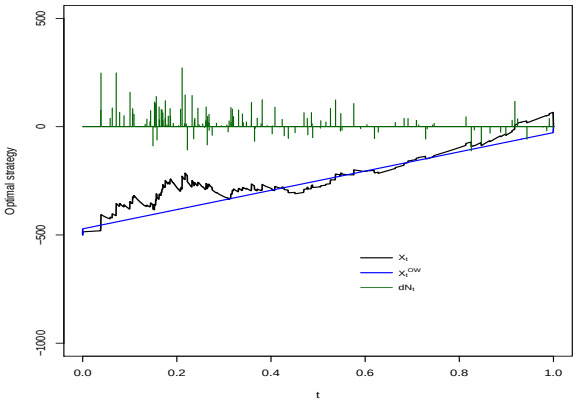
*The functions $e$ and $g$ are the unique solution of the differential equations (1.32) and (1.33) with $e(0) = g(0) = 0$.*

The proof of this theorem is given in Appendix 1.7. Let us mention here that the functions $e$ and $g$ admit explicit forms by the mean of the exponential integral function, that are very cumbersome. They can be obtained by using a formal calculus software such as Mathematica. Since they do not play any role to determine the optimal strategy and require several pages to be displayed, we do not give these explicit formulas. Note that they are simpler in the case $\eta = 0$, for which the explicit formulas are given by Equations (1.38) and (1.39).

The optimal strategy $X^*$ is illustrated on Figure 1.1 for two different sets of parameters. It is worth to notice that the strategy is linear with respect to $x_0$, $D_0$, $\delta_0$, $I$ and $N$. This property is due to the affine structure of the model and the quadratic costs. In particular, the reaction of the optimal strategy to the other trades does not depend on $x_0$. The strategy $X^{\text{trend}}$ is the part of the strategy which is proportional to $D_0$ and $\delta_0$ and thus takes advantage of temporary price trends that are known at time 0. The strategy $X^{\text{dyn}}$ is proportional to the processes $I$ and $N$ and describes the optimal reactions to observed price jumps. Last, let us stress that having an explicit formula for the optimal strategy is an important feature to use it in practice. Since the strategy reacts to each market order (or at least to those which trigger price moves), its computation time should be significantly lower than the typical duration between two of these orders.




(a) $\rho = 25$




(b) $\rho = 16$

FIGURE 1.1 – Optimal strategy in the Hawkes model, in black, for $q = 100$, $T = 1$, $\beta = 20$, $\iota_s = 16$, $\iota_c = 2$, $\kappa_\infty = 12$, $\epsilon = 0.3$, $\nu = 0.3$, $D_0 = 0.1$, $\kappa_0^+ = \kappa_0^- = 60$, $m_1 = 50$, $X_0 = -500$, $\mu = \text{Exp}(1/m_1)$, $\varphi_s(y) = 1.2 \times y^{0.2} + 0.5 \times y^{0.7} + 14.4 \times y$, $\varphi_c(y) = 1.2 \times y^{0.2} + 0.5 \times y^{0.7} + 0.4 \times y$ for all $y > 0$. The strategy of the Obizhaeva and Wang model is given in blue as a benchmark, and the jumps of $(N_t)$ are plotted in green (with the same trajectory for the two graphs). On the left graph, $\iota_s < \beta < \rho$ and the strategy is based on mean-reversion : each time $N$ jumps, $X$ jumps in the opposite direction. On the right graph, $\rho = \iota_s < \beta$ and the strategy is trend-following.

Let us make some comments on the optimal strategy, and more precisely on how the strategic trader reacts to the orders issued by other traders. First, we observe from (1.20) that the block trades that immediately follow jumps of $N$ are then compensated by the continuous trading rate. When $\varphi_s = \varphi_c$, we have $I \equiv 0$ and these block trades, as given by (1.14), are always opposed in sign to the market orders that they follow. For general functions $\varphi_s$ and $\varphi_c$, the signs of these trades depend on the size of the last preceding jumps of $N$. For example, in the case where $\eta = \nu\rho$, the strategic trader makes a trade in the opposite direction if $|\mathrm{d}N_t| > \frac{m_1}{\rho(1-\nu)}(\varphi_s - \varphi_c)(|\mathrm{d}N_t|/m_1)$, but trades in the same way otherwise. The same conclusion holds for any parameter value when $T - t \to 0$ since

$\rho(T-t)^2 \times \omega(\eta(T-t))$ vanishes. We now consider the asymptotics when the trading horizon is large : in this case, it is reasonable to assume that $\eta > 0$ which is required to get stationary intensities $\kappa^+$ and $\kappa^-$, see Section 1.2.3. Then, when $T - t \to +\infty$, the jump part of $X^{\mathrm{dyn}}$ given by (1.14) can be well approximated by

$$\frac{m_1}{2\rho}\left(1 + \frac{\nu\rho}{\eta}\right) \mathrm{d}I_t - (1-\nu)\ \mathrm{d}N_t.$$

Therefore, the strategic trader makes a trade in the opposite direction if $|\mathrm{d}N_t| > \frac{m_1}{2\rho(1-\nu)}(1+\frac{\nu\rho}{\eta})(\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(|\mathrm{d}N_t|/m_1)$ and trades in the same direction otherwise. In the case $\iota_{\mathrm{c}} = 0$ and $\varphi_{\mathrm{s}} \equiv \iota_{\mathrm{s}}$ where there is only volume-independent self-excitation, we can interpret this behavior as follows : if a market buy order is relatively small, it may be a part of a big split order, and thus be followed by other buy orders that will make the price go up, and the strategic trader has interest to follow this trend. However, if a market buy order is relatively big, the price resilience effect is likely to dominate and the strategic trader has interest to trade in the opposite way.

Last, it is interesting to notice that the optimal strategy only depends on $(\varphi_{\mathrm{s}}, \varphi_{\mathrm{c}})$ through $\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}}$. This key self-excitation function tunes the way that the strategic trader should react to other market orders.

**Remark 1.4.1.** *The MIH model with $\eta = 0$ includes the particular case of independent Poisson processes when $\beta = 0$ and $\varphi_{\mathrm{s}} = \varphi_{\mathrm{c}} \equiv 0$. In that case, if $N$ jumps at time $\tau \in (0,T)$, we get from (1.14)*

$$(1-\epsilon)\Delta X_\tau^{\mathrm{dyn}} \ = \ -\frac{1+\rho(T-\tau)}{2+\rho(T-\tau)} \times (1-\nu)\ \Delta N_\tau.$$

*Since the self-excitation effect is removed, the price is a mean-reverting process when the strategic trader is passive. Thus, each time a market order is observed, the optimal strategy consists in posting immediately a market order in the opposite direction, to arbitrage the resilience of the price. Such an obvious Price Manipulation Strategy is unrealistic, therefore modeling the order flow with Poisson processes is not satisfactory. We refer to Section 1.5.2 for more details.*

**Remark 1.4.2.** *Following Remark 1.2.3, a natural question is to look at the quantity $x_0$ that minimizes $\mathbb{E}[C(X)] + P_0 \times x_0$, i.e. the expected liquidation cost with respect to the mark-to-market value. From Theorem 1.4.1 we obtain easily that, at time 0, this quantity is minimal for*

$$x_0^* = \frac{\rho T[qD_0 - \mathcal{G}_\eta(T)\frac{\delta_0 m_1}{\rho}]}{2\left(1 + \frac{\epsilon}{2}\rho T\right)}.$$

*We can give a simple heuristic for the sign of $x_0^*$ : when $D_0 \geq 0$ and $\delta_0 \leq 0$ the price trend is negative and it is more favorable to sell ($x_0 \geq 0$) since $\mathcal{G}_\eta$ is nonnegative.*

## 1.5 Price Manipulation Strategies in the MIH model

In this section, we study Price Manipulation Strategies (PMS), as introduced by Definition 1.2.2, in the context of the MIH model. As a matter of fact, the value function given in Theorem 1.4.1 can be negative even for $x_0 = 0$, which would constitute a PMS. We first determine necessary and sufficient conditions on the parameters of the model to exclude such strategies. Then, we study the particular case of Poisson processes, which may seem natural to model the order flow but allow for robust arbitrages to arise in this framework.

### 1.5.1 The Mixed-market-Impact Hawkes Martingale (MIHM) model

Theorem 1.2.1 gives a necessary and sufficient condition on $N$ to exclude Price Manipulation Strategies. Here, we apply this result to identify which parameters in the Hawkes model exclude PMS. We recall the notation

$$\alpha = \iota_{\mathrm{s}} - \iota_{\mathrm{c}} = \int_{\mathbb{R}^+} (\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(v/m_1)\mu(\mathrm{d}v),$$

and define the (normalized) support of $\mu$

$$\mathcal{S}(\mu) = \{y \geq 0 \ s.t. \ \forall \varepsilon > 0, \ \mu((m_1 \times y - \varepsilon, m_1 \times y + \varepsilon)) > 0\}.$$

**Proposition 1.5.1.** *The MIH model does not admit PMS if, and only if the following conditions hold*

$$\beta = \rho, \ \alpha = (1-\nu)\rho, \ \varphi_{\mathrm{s}}(x) - \varphi_{\mathrm{c}}(x) = \alpha x \text{ for } x \in \mathcal{S}(\mu) \ (i.e. \ m_1 I = \alpha N), \ \text{and } qD_0 = \frac{m_1}{\rho}\delta_0, \quad (1.15)$$

*or $\mu = Dirac(0)$ with $D_0 = 0$. In both cases, the optimal execution strategy is given by (1.7).*

Note that in the case $\mu = \mathrm{Dirac}(m_1)$ where all the jumps have the same size, one has $\mathcal{S}(\mu) = \{1\}$ thus $\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}}$ is necessarily linear on $\mathcal{S}(\mu)$ and $\Delta I_t = \alpha \ \mathrm{sgn}(\Delta N_t)$. If moreover $m_1 = 0$, we have $N \equiv 0$ and the MIH model does not depend any longer on the parameters $\alpha$ and $\beta$, that can then be fixed arbitrarily. *Proof.* From Theorem 1.2.1, PMS are excluded if, and only if the price $P$ is a martingale when $X \equiv 0$. In this case, we have from (1.1), (1.2), (1.3) and (2.43)

$$\mathrm{d}P_t = -\rho D_t dt + \frac{1}{q}\mathrm{d}N_t = \frac{1}{q}(\mathrm{d}N_t - \delta_t m_1 \mathrm{d}t) + \left(\frac{m_1}{q}\delta_t - \rho D_t\right)\mathrm{d}t.$$

Therefore, $P$ is a martingale if, and only if $\frac{m_1}{\rho}\delta_t = qD_t$ $\mathbb{P}$-a.s., $dt$ a.e. This condition is equivalent to $qD_0 = \frac{m_1}{\rho}\delta_0$ and $qdD_t = \frac{m_1}{\rho}d\delta_t$. From (1.3) and (2.43), the latter condition is equivalent to

$$\rho q D_t = \frac{m_1}{\rho}\beta\delta_t \text{ and } (1-\nu)\mathrm{d}N_t = \frac{m_1}{\rho}\mathrm{d}I_t.$$

Using (1.11), the second condition is equivalent to $(1-\nu)\rho \ v = m_1(\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(v/m_1)$ for all $v$ in the support of $\mu$, which implies the linearity of $\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}}$ on $\mathcal{S}(\mu)$ and leads to (1.15). Conversely, (1.15) implies $\frac{m_1}{\rho}\delta_t = qD_t$, and $P$ is then a martingale.

**Remark 1.5.1.** *When $\beta = \rho$, $\alpha = (1-\nu)\rho$, and $\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}}$ is linear on $\mathcal{S}(\mu)$, we get from the previous calculations that $d(\frac{m_1}{q}\delta_t - \rho D_t) = -\rho(\frac{m_1}{q}\delta_t - \rho D_t)dt$, and therefore $\frac{m_1}{q}\delta_t - \rho D_t$ converges exponentially to zero. The condition $qD_0 = \frac{m_1}{\rho}\delta_0$ simply means that the model starts from this steady state.*

One can also check directly that the optimal strategy and its cost given by Theorem 1.4.1 coincide with those of Theorem 1.2.1 when (1.15) holds. For clear reasons, we call Mixed-Impact Hawkes Martingale (MIHM) model the MIH model if these conditions are satisfied. Proposition 1.5.1 is very interesting since it makes connections between the model parameters of the MIH model in a perfect market without PMS. First, the condition $\beta = \rho$ means that the mean-reverting action of liquidity providers compensates the autocorrelation in the signs of the trades of liquidity takers ; we thus reach

a conclusion similar to Bouchaud et al. [31]. The condition $\alpha = (1-\nu)\beta$ gives a link between the Hawkes kernel and the proportion $1-\nu$ of transient price impact. When $\iota_c = 0$, $\alpha/\beta$ represents the average number of child orders coming from one market order, and is thus equal to the proportion of endogenous orders (i.e. triggered by other orders) in the market. What we obtain here is that this ratio should be equal to $1-\nu$, which is a *a priori* different measure of endogeneity, since it gives the proportion of market impact that does not influence the low-frequency price (see Section 1.2.3). The positivity of $\alpha$ reflects the fact that the parameter $\iota_c$ tuning opportunistic trading should be small to avoid market instability. It is interesting to notice that if (1.15) holds, the stationarity condition $\iota_s + \iota_c < \beta$ derived in Section 1.2.3 is equivalent to $2\iota_c < \nu\rho$, which can be seen as a reasonable upper bound for $\iota_c$. Last, we see that $\varphi_s - \varphi_c$ should be linear. Let us recall that $\varphi_s$ and $\varphi_c$ encode the dependence of the self-excitation (resp. the cross-excitation) effect on the volumes of incoming market orders. Condition (1.15) implies that they should have roughly the same functional form, except for a linear part which should be stronger for $\varphi_s$. However, we remind here that these conclusions are obtained in the MIH model and should be confronted to market data. We leave this empirical investigation for further research.

Of course, in practice, it would be miraculous if the calibration of the MIH model on real financial data led to parameters satisfying exactly (1.15). One may rather expect these parameters to be close but not exactly equal to those of the MIHM model, for the following reasons. First, there is no guarantee that fitting a model to a market with no PMS leads to a model with no PMS. Second, the MIH model ignores market frictions such as the bid-ask spread and gives some advantages to the strategic trader such as the possibility to post orders immediately after the other ones (see Stoikov and Waeber [104] for a study on the latency to execute an order). These facts make the existence of PMS more likely in the model than in reality. Third, we know that in practice, temporary arbitrage may exist at high frequencies. Therefore, there is no reason that fitted parameters follow exactly the MIHM condition (1.15). This justifies the potential practical usefulness of the strategy given by Theorem 1.4.1 to reduce execution costs when the estimated parameters deviate from the MIHM model. Let us note that Figure 1.1 illustrates such a case : all the parameters satisfy (1.15) but $\rho$ (which should be equal to $\beta = 20$). The estimation of the MIH model on market data is left for future research.

The framework of the MIH model also gives some interesting insights for the characterization of the existence of short-time arbitrages. Let us introduce the following definition.

**Definition 1.5.1.** *We say that a market admits weak Price Manipulation Strategies (wPMS) if the cost of a liquidation strategy can be reduced by posting a block trade as an immediate response to a market order issued by another trader.*

**Corollary 1.5.1.** *In the MIH model, the market does not admit wPMS if, and only if,*

$$\beta = \rho, \ \alpha = (1-\nu)\rho \text{ and } \varphi_s(x) - \varphi_c(x) = \alpha x \text{ for } x \in \mathcal{S}(\mu) \tag{1.16}$$

*or* $\mu = Dirac(0)$.

*Proof.* The proof is quite straightforward from Theorem 1.4.1. The case $\mu = \text{Dirac}(0)$ is trivial and we consider $m_1 > 0$. The jump term of the strategy (1.14) should be equal to zero for any $\tau \in [0, T]$. By taking $\tau = T$, we get that $\varphi_s(x) - \varphi_c(x) = (1-\nu)\rho x$ for $x \in \mathcal{S}(\mu)$. Integrating this indentity with respect to $\mu$ leads to $\alpha = \iota_s - \iota_c = (1-\nu)\rho$. Then, from (1.14), we should have

$(\nu\rho-\eta)\times\rho u^2\times\omega(\eta u)=0$ for $u\in[0,T]$ which implies $\nu\rho=\eta$. Since $\eta=\beta-\alpha=\beta-(1-\nu)\rho$, we get $\beta=\rho$. The converse implication is obvious. By Remark 1.5.1, the condition $qD_0=\frac{m_1}{\rho}\delta_0$ means that the model has reached its equilibrium, which is basically the case after some time. Therefore, the conditions that exclude wPMS and PMS in the MIH model are quite the same. This is an interesting link between two different point of views. The condition "no PMS" means that there is no free source of income. The condition "no wPMS" rather brings on market stability, since it excludes trading volume coming from the response to other trades. Corollary 1.5.1 is a mathematical formulation of this link in our specific model.

### 1.5.2 The Poisson model

Poisson processes are often used to model the arrival of the customers in queuing theory. It is therefore natural to use them to model the flow of market orders, as it has been made for example by Bayraktar and Ludkovski [21] or Cont and de Larrard [40] in different frameworks.

Here, in the Poisson model, $N^+$ and $N^-$ are two i.i.d. independent compound Poisson processes of respective constant jump rates $\kappa_0^+$ and $\kappa_0^-$, with the same jump law $\mu$. It is a particular case of the MIH model when $\beta=0$, $\varphi_{\mathrm{s}}\equiv 0$ and $\varphi_{\mathrm{c}}\equiv 0$, which implies $\eta=0$. Thus, the optimal strategy and value function in this case can be deduced from Theorem 1.4.1 (see also Remark 1.4.1).

First, let us note that the Poisson model cannot satisfy the condition (1.15), except in the case $\rho=0$, where there is only permanent price impact, which is not relevant in this context. Thus, we know *a priori* that PMS are possible. However, we specify in what follows that a Poisson order flow creates very simple and robust arbitrages. First, we put aside the case $\kappa_0^+\neq\kappa_0^-$ where the trend on the price leads to obvious PMS, and consider now the more interesting case $\kappa_0^+=\kappa_0^-$, and we simply denote by $\kappa_0$ the common intensity.

A natural choice to get a PMS is of course to consider the optimal strategy given by Theorem 1.4.1 when liquidating $x_0=0$ assets. A remarkable feature of this optimal strategy in the Poisson case is that it only depends on the process $N$, and does not depend directly on the law of the jumps and their intensity. Then, when applying the optimal strategy, mainly two quantities have to be known : $qD_0$ and $\rho$. We denote by $\mathcal{C}_0(D_0)$ the cost of the optimal strategy and obtain from Theorem 1.4.1 in this case :

$$(1-\epsilon)q\times\mathcal{C}_0(D_0)=-\frac{\rho T/2}{2+\rho T}\,q^2D_0^2\;-\;(1-\nu)^2\,2\kappa_0 m_2\left[\frac{T}{2}-\frac{1}{\rho}\ln\left(1+\frac{\rho T}{2}\right)\right]. \tag{1.17}$$

In fact, PMS are very robust in this framework. The following proposition shows that even if $qD_0$ and $\rho$ are unknown, one can construct a such a strategy. This indicates that in our framework with a linear price impact and an exponential resilience, compound Poisson processes are not suitable to model the order flow.

**Proposition 1.5.2.** *Let $\kappa_0^+=\kappa_0^-=\kappa_0>0$ and $\lambda\in(0,1)$. The following round-trip strategy $X_0^\lambda=X_{T+}^\lambda=0$ defined by*

$$X_{\tau+}^\lambda-X_\tau^\lambda=-\frac{1-\nu}{1-\epsilon}\times\lambda(N_\tau-N_{\tau-})$$

*at each jump of $N$ is a PMS. Its average cost is given by*

$$\mathbb{E}[C(X^\lambda)]=2\lambda(1-\lambda)\frac{\kappa_0 m_2(1-\nu)^2}{q(1-\epsilon)}\left[\frac{1-\exp(-\rho T)}{\rho}-T\right]<0,$$

*and the best choice is to take* $\lambda = 1/2$.

*Proof.* From Remark 1.2.7, it is sufficient to focus on the case $\nu = \epsilon = 0$ and $q = 1$. In this case, we have

$$C(X) = \int_{[0,T)} D_u \, \mathrm{d}X_u^\lambda \; + \; \frac{1}{2} \sum_{0 \le \tau < T} (\Delta X_\tau^\lambda)^2 \; - \; D_T X_T^\lambda \; + \; \frac{1}{2} \, (X_T^\lambda)^2,$$

with $D_t = D_0 + \int_0^t \exp(-\rho(t-s)) dN_s + \int_0^t \exp(-\rho(t-s)) dX_s^\lambda$.
From $\int_{[0,T)} D_u \mathrm{d}X_u^\lambda = -\lambda \sum_{0 \le \tau < T} \left[ D_{\tau -} \Delta N_\tau + (\Delta N_\tau)^2 \right]$,
we get $\mathbb{E}[\int_{[0,T)} D_u \, \mathrm{d}X_u^\lambda] = -\lambda \mathbb{E}[\sum_{0 \le \tau < T} (\Delta N_\tau)^2] = -2\lambda\kappa_0 m_2 T$. Since $X_T^\lambda = -\lambda N_T$
and $D_T = D_0 + (1-\lambda) \int_0^T \exp(-\rho(T-s)) dN_s$ a.s., we have $\mathbb{E}[(X_T^\lambda)^2] = 2\lambda^2 \kappa_0 m_2 T$ and

$$\mathbb{E}[-D_T X_T^\lambda] = \lambda(1-\lambda)\mathbb{E}\left[ N_T \int_0^T \exp(-\rho(T-s)) dN_s \right] = 2\lambda(1-\lambda)\kappa_0 m_2 \frac{1-\exp(-\rho T)}{\rho}.$$

This eventually yields

$$\begin{aligned} \mathbb{E}[C(X^\lambda)] &= -2\lambda\kappa_0 m_2 T + \lambda^2 \kappa_0 m_2 T + 2\lambda(1-\lambda)\kappa_0 m_2 \frac{1-\exp(-\rho T)}{\rho} + \lambda^2 \kappa_0 m_2 T \\ &= 2\lambda(1-\lambda)\kappa_0 m_2 \left( \frac{1-\exp(-\rho T)}{\rho} - T \right). \end{aligned}$$

## 1.6 Appendix : Explicit formulas for the optimal strategy

We use the function

$$L(r, \lambda, t) \; := \; r \int_0^t \frac{\exp(\lambda s)}{2 + rs} \, \mathrm{d}s \; = \; \exp(-2\lambda/r) \; \left[ \mathcal{E}\left( \frac{\lambda}{r}(2 + rt) \right) - \mathcal{E}\left( \frac{2\lambda}{r} \right) \right], \tag{1.18}$$

where $\mathcal{E}(y) = -\int_{-y}^{+\infty} \frac{e^{-u}}{u} \, \mathrm{d}u$ is the exponential integral of $y$, in terms of Cauchy principal value if $y > 0$. Since we only consider differences $\mathcal{E}(y) - \mathcal{E}(y')$ with either $y, y' > 0$ or $y, y' < 0$, we will only consider proper integrals. The function $\mathcal{E}$ is standard and is implemented in many packages such as the Boost C++ library. Thus, $L$ can be evaluated as a closed formula.

We refer to (1.12) and (1.13) for the definitions of $\zeta$ and $\omega$.

**Auxiliary functions :** For $0 \le s \le t \le T$,

$$\begin{aligned} \phi_\eta(t) = \; & \frac{1}{2(2 + \rho(T-t))} \times \Big[ 1 + \exp(-\eta(T-t)) + \nu\rho(T-t)\zeta(\eta(T-t)) \\ & + \frac{\beta}{\rho} \left[ 2 + \rho(T-t) \times \{ 1 + \zeta(\eta(T-t)) + \nu\rho(T-t) \; \omega(\eta(T-t)) \} \right] \Big], \end{aligned}$$

$$\begin{aligned}\Phi_0(s,t) &= \left[\frac{\beta}{\rho}+\frac{\nu}{2}\left(\frac{1}{2}-\frac{\beta}{\rho}\right)\right]\times\frac{\exp(-\beta s)-\exp(-\beta t)}{\beta}\\ &\quad+(1-\nu)\left(1-\frac{\beta}{\rho}\right)\times\frac{\exp(-\beta T)}{\rho}\times[L(\rho,\beta,T-s)-L(\rho,\beta,T-t)]\\ &\quad+\frac{\nu}{4}\left[(T-s)\exp(-\beta s)-(T-t)\exp(-\beta t)\right],\end{aligned}$$

and for $\eta\neq 0$,

$$\begin{aligned}\Phi_\eta(s,t) &= \frac{1}{2}\left(\frac{1}{\rho}+\frac{\nu}{\eta}\right)\times[\exp(-\beta s)-\exp(-\beta t)]\\ &\quad+\frac{\exp(-\beta T)}{2\rho}\times\left[1+\frac{\nu(\rho-2\beta)}{\eta}+\frac{\beta}{\eta}\left(1-\frac{\nu\rho}{\eta}\right)\right]\times[L(\rho,\beta,T-s)-L(\rho,\beta,T-t)]\\ &\quad+\frac{\exp(-\beta T)}{2\rho}\times\left[1-\frac{\nu\rho}{\eta}-\frac{\beta}{\eta}\left(1-\frac{\nu\rho}{\eta}\right)\right]\times[L(\rho,\alpha,T-s)-L(\rho,\alpha,T-t)].\end{aligned}$$

We now give the explicit formulas for the whole optimal strategy. They are valid for all $\eta\in\mathbb{R}$.

**Trend strategy :**

$$\begin{aligned}(1-\epsilon)\Delta X_0^{\text{trend}} &= \frac{\frac{\delta_0 m_1}{2\rho}\times[2+\rho T\times\{1+\zeta(\eta T)+\nu\rho T\;\omega(\eta T)\}]-[1+\rho T]qD_0}{2+\rho T},\\ (1-\epsilon)\Delta X_T^{\text{trend}} &= \frac{\delta_0 m_1}{2\rho}\times\left[\frac{2+\rho T\times\{1+\zeta(\eta T)+\nu\rho T\;\omega(\eta T)\}}{2+\rho T}-2\rho\;\Phi_\eta(0,T)-2\exp(-\beta T)\right]\\ &\quad+\frac{qD_0}{2+\rho T},\end{aligned}\tag{1.19}$$

and, on $(0,T)$,

$$\begin{aligned}(1-\epsilon)\mathrm{d}X_t^{\text{trend}} &= \frac{\delta_0 m_1}{2\rho}\times\left[\frac{2+\rho T\times\{1+\zeta(\eta T)+\nu\rho T\;\omega(\eta T)\}}{2+\rho T}-2\rho\;\Phi_\eta(0,t)-2\phi_\eta(t)\exp(-\beta t)\right]\;\rho\mathrm{d}t\\ &\quad+\frac{qD_0}{2+\rho T}\;\rho\mathrm{d}t.\end{aligned}$$

**Dynamic strategy :**

$$\begin{aligned}(1-\epsilon)\Delta X_0^{\text{dyn}} &= 0,\\ (1-\epsilon)\Delta X_T^{\text{dyn}} &= -\,m_1\left[\Theta_{\chi_T}\;\Phi_\eta\left(\tau_{\chi_T},T\right)\;+\;\sum_{i=1}^{\chi_T-1}\Theta_i\;\Phi_\eta(\tau_i,\tau_{i+1})\right]\;+\;\sum_{0<\tau\leq T}\frac{(1-\nu)\;\Delta N_\tau}{2+\rho(T-\tau)}\\ &\quad+\frac{m_1}{2\rho}\times\sum_{0<\tau\leq T}\frac{2+\rho(T-\tau)\times\{1+\zeta(\eta(T-\tau))+\nu\rho(T-\tau)\;\omega(\eta(T-\tau))\}}{2+\rho(T-\tau)}\;\Delta I_\tau\\ &\quad-\frac{m_1}{\rho}\;\Theta_{\chi_T}\exp(-\beta T),\end{aligned}\tag{1.20}$$

and, on $(0,T)$,

$$
\begin{aligned}
(1-\epsilon)\mathrm{d}X_t^{\mathrm{dyn}} \;=\; & -\, m_1\, \phi_\eta(t)\, \Theta_{\chi_t} \exp(-\beta t)\, \mathrm{d}t \;+\; \left[\sum_{0<\tau\le t} \frac{(1-\nu)\Delta N_\tau}{2+\rho(T-\tau)}\right] \rho \mathrm{d}t \\
& + \left[\sum_{0<\tau\le t} \frac{2+\rho(T-\tau)\times\{1+\zeta(\eta(T-\tau))+\nu\rho(T-\tau)\,\omega(\eta(T-\tau))\}}{2+\rho(T-\tau)}\,\Delta I_\tau\right] \frac{m_1}{2}\,\mathrm{d}t \\
& - \left[\Theta_{\chi_t}\,\Phi_\eta\left(\tau_{\chi_t},t\right) \;+\; \sum_{i=1}^{\chi_t-1}\Theta_i\,\Phi_\eta(\tau_i,\tau_{i+1})\right]\rho m_1\,\mathrm{d}t \\
& + \frac{1+\rho(T-t)}{2+\rho(T-t)}\left\{\frac{m_1}{\rho}\,\mathrm{d}I_t-(1-\nu)\,\mathrm{d}N_t\right\} \;+\; \frac{m_1}{2\rho}(\nu\rho-\eta)\times\frac{\rho(T-t)^2\times\,\omega(\eta(T-t))}{2+\rho(T-t)}\,\mathrm{d}I_t.
\end{aligned}
$$

## 1.7 Appendix : Proof for the optimal control problem (results of Theorem 1.4.1 and Appendix 1.6)

### 1.7.1 Notations and methodology

The jump intensity of the process $(N_t)$ is characterized by the càdlàg Markovian process $(\delta_t,\Sigma_t)$ defined by (2.43), taking values in $\mathbb{R}\times\mathbb{R}^+$. The state variable of the problem is then $(X_t,D_t,S_t,\delta_t,\Sigma_t)$, and the control is $X_t-x_0$, i.e. the variation of the position of the strategic trader, $(X_t)_{t\in[0,T]}$ being an admissible strategy as described in Definition 1.2.1. The control program is thus to minimize $\mathbb{E}\left[C(0,X)\right]$ over all admissible strategies, where the cost $C(t,X)$ of the strategy $X$ between $t$ and $T$ is given by

$$
C(t,X) \;=\; \int_{[t,T)} P_u\,\mathrm{d}X_u \;+\; \frac{1}{2q}\sum_{t\le\tau<T}(\Delta X_\tau)^2 \;-\; P_T X_T \;+\; \frac{1}{2q}\,X_T^2.
$$

The final value at time $t=T$ is the cost of a market order of signed volume $\Delta X_T=-X_T$ (so that $X_{T^+}=X_T+\Delta X_T=0$). At time $t$, the price $P_t$ depends on $D_t$ and $S_t$ which in turn depend on $(X_u)_{u\in[0,t]}$. Let us define $\mathcal{A}_t$ the set of admissible strategies on $[t,T]$, with $t\in[0,T]$. The value function of the problem is

$$
\mathcal{C}(t,x,d,z,\delta,\Sigma) \;=\; \inf_{X\in\mathcal{A}_t}\mathbb{E}\left[C(t,X)\right]
$$

with $X_t=x,\ D_t=d,\ S_t=z,\ \delta_t=\delta$ and $\Sigma_t=\Sigma$. In order to determine analytically the value function and the optimal control of the problem, we use the probabilistic formulation of the verification theorem. We determine *a priori* a continuously differentiable function $\mathcal{C}(t,x,d,z,\delta,\Sigma)$ and an admissible strategy $X^*$ and then we verify that

$$
\Pi_t(X) \;:=\; \int_0^t P_u\,\mathrm{d}X_u \;+\; \frac{1}{2q}\sum_{0\le\tau<t}(\Delta X_\tau)^2 \;+\; \mathcal{C}(t,X_t,D_t,S_t,\delta_t,\Sigma_t) \tag{1.21}
$$

is a submartingale for any admissible strategy $X$, and that $\Pi_t(X^*)$ is a martingale. We proceed in three steps :

1. We define a suitable function $\mathcal{C}$, and derive a set of ODEs on its coefficients which is a necessary condition for $\mathcal{C}$ to be the value function of the problem.
2. We solve the set of ODEs.
3. Using the results of the previous steps, we derive the strategy $X^*$ such that $\Pi_t(X^*)$ is a martingale.

The verification argument then yields that $\mathcal{C}(t,x,d,z,\delta,\Sigma)$ is the value function and that $X^*$ is optimal. Without loss of generality, we can assume that $q=1$ by using Remark 1.2.4.

### 1.7.2 Necessary conditions on the value function

We search a cost function $\mathcal{C}$ as a generic quadratic form of the variables $x,d,z,\delta,\Sigma$ with time-dependent coefficient (the variable $z$ symbolizes the current value of the fundamental price $S_t$). As we see further, we need $\mathcal{C}$ to verify $\partial_x\mathcal{C}+(1-\epsilon)\partial_d\mathcal{C}+\epsilon\,\partial_z\mathcal{C}+d+z=0$ : it is thus necessary that $\mathcal{C}$ is a quadratic form of $(d-(1-\epsilon)x)$, $(z-\epsilon x)$, $\delta$ and $\Sigma$, plus a term $-(d+z)^2/2$. We define

$$\begin{aligned}\mathcal{C}(t,x,d,z,\delta,\Sigma) &= a(T-t)(d-(1-\epsilon)x)^2 + \frac{1}{2}(z-\epsilon x)^2 + (d-(1-\epsilon)x)(z-\epsilon x) - \frac{(d+z)^2}{2}\\ &\quad + b(T-t)\,\delta\,(d-(1-\epsilon)x) + c(T-t)\,\delta^2 + e(T-t)\,\Sigma + g(T-t), \qquad (1.22)\end{aligned}$$

with $a,b,c,e,g : \mathbb{R}^+ \to \mathbb{R}$ continuously differentiable functions. We choose the limit condition $\mathcal{C}(T,x,d,z,\delta,\Sigma) = -(d+z)x+x^2/2 = \frac{1}{2}(d+z-x)^2-(d+z)^2/2$, which is the cost of a trade of signed volume $-x$. We thus have

$$a(0)=\frac{1}{2},\ b(0)=c(0)=e(0)=g(0)=0.$$

Let us note that other terms should be added in equation (2.44) for $\mathcal{C}$ to be a generic quadratic form. The five terms

$$\begin{aligned}h_1(T-t)\,(d-(1-\epsilon)x) &+ h_2(T-t)\,\Sigma(d-(1-\epsilon)x) + h_3(T-t)\,\delta\Sigma\\ &+ h_4(T-t)\,\delta + h_5(T-t)(z-\epsilon x)\end{aligned}$$

have to be equal to zero since $\mathcal{C}(t,x,d,z,\delta,\Sigma)=\mathcal{C}(t,-x,-d,-z,-\delta,\Sigma)$ by using Remark 1.2.4 and the fact that the buy and sell orders play a symmetric role. For the term in $\Sigma^2$, we checked in prior calculations that it is necessarily associated to a zero coefficient. For $\Delta x\in\mathbb{R}$, we have

$$\mathcal{C}(t,x+\Delta x,d+(1-\epsilon)\Delta x,z+\epsilon\Delta x,\delta,\Sigma)-\mathcal{C}(t,x,d,z,\delta,\Sigma) = -(d+z)\times\Delta x - \frac{(\Delta x)^2}{2}. \qquad (1.23)$$

In what follows, we drop the dependence of $\mathcal{C}(t,X_t,D_t,S_t,\delta_t,\Sigma_t)$ on $(t,X_t,D_t,S_t,\delta_t,\Sigma_t)$ to obtain less cumbersome expressions. The process $\mathcal{C}(t,X_t,D_t,S_t,\delta_t,\Sigma_t)$ is làdlàg, and with the notations of Remark 1.2.5, we have by using (1.23)

$$\begin{aligned}\mathrm{d}\mathcal{C} &= \partial_t\mathcal{C}\,\mathrm{d}t + \partial_x\mathcal{C}\,\mathrm{d}X_t^c + \partial_d\mathcal{C}\Big(-\rho D_t\mathrm{d}t+(1-\epsilon)\mathrm{d}X_t^c\Big) + \partial_z\mathcal{C}\,\epsilon\mathrm{d}X_t^c\\ &\quad -\beta\,\delta_t\,\partial_\delta\mathcal{C}\,\mathrm{d}t-\beta(\Sigma_t-2\kappa_\infty)\,\partial_\Sigma\mathcal{C}\,\mathrm{d}t\\ &\quad +\Big[\mathcal{C}(t,X_t,D_{t^-}+(1-\nu)\Delta N_t,S_{t^-}+\nu\Delta N_t,\delta_{t^-}+\Delta I_t,\Sigma_{t^-}+\Delta\bar{I}_t)-\mathcal{C}(t,X_t,D_{t^-},S_{t^-},\delta_{t^-},\Sigma_{t^-})\Big]\\ &\quad -(D_t+S_t)\,\Delta X_t - \frac{(\Delta X_t)^2}{2}.\end{aligned}$$

where we refer to (1.11) for the definitions of $I$ and $\overline{I}$. The definition of $\Pi(X)$ given by (1.21) yields $\mathrm{d}\Pi_t(X) = (D_t + S_t)\mathrm{d}X_t^{\mathrm{c}} + (D_t + S_t)\Delta X_t + (\Delta X_t)^2/2 + \mathrm{d}\mathcal{C}$. We define the continuous finite variation process $(A_t^X)_{t\in(0,T)}$ such that $A_{0^+}^X = \mathcal{C}(0, X_{0^+}, D_{0^+}, S_{0^+}, \delta_0, \Sigma_0)$ and for $t \in (0,T)$

$$\begin{aligned}\mathrm{d}A_t^X \;=\; & (D_t + S_t)\ \mathrm{d}X_t^{\mathrm{c}} \;+\; Z(t, X_t, D_t, S_t, \delta_t, \Sigma_t)\mathrm{d}t \\ & + \partial_t\mathcal{C}\ \mathrm{d}t \;+\; \partial_x\mathcal{C}\ \mathrm{d}X_t^{\mathrm{c}} \;+\; \partial_d\mathcal{C}\ \Big(-\rho D_t\mathrm{d}t + (1-\epsilon)\mathrm{d}X_t^{\mathrm{c}}\Big) \;+\; \partial_z\mathcal{C}\ \epsilon\mathrm{d}X_t^{\mathrm{c}} \\ & - \beta\ \delta_t\ \partial_\delta\mathcal{C}\ \mathrm{d}t \;-\; \beta\ (\Sigma_t - 2\kappa_\infty)\ \partial_\Sigma\mathcal{C}\ \mathrm{d}t,\end{aligned}$$

where, for $V \sim \mu$, $Z(t,x,d,z,\delta,\Sigma) :=$

$$\begin{aligned}&\frac{\Sigma+\delta}{2} \times \mathbb{E}\big[\mathcal{C}(t,x,d+(1-\nu)V, z+\nu V, \delta + (\varphi_{\mathrm{s}}-\varphi_{\mathrm{c}})(V/m_1), \Sigma + (\varphi_{\mathrm{s}}+\varphi_{\mathrm{c}})(V/m_1)) - \mathcal{C}(t,x,d,z,\delta,\Sigma)\big] \\ +\ &\frac{\Sigma-\delta}{2} \times \mathbb{E}\big[\mathcal{C}(t,x,d-(1-\nu)V, z-\nu V, \delta - (\varphi_{\mathrm{s}}-\varphi_{\mathrm{c}})(V/m_1), \Sigma + (\varphi_{\mathrm{s}}+\varphi_{\mathrm{c}})(V/m_1)) - \mathcal{C}(t,x,d,z,\delta,\Sigma)\big].\end{aligned}$$

Then, $\Pi(X) - A^X$ is a martingale (let us note that almost surely, $\mathrm{d}t$ -a.e. on $(0,T)$, $Z(t, X_t, D_{t^-}, S_{t^-}, \delta_{t^-}, \Sigma_{t^-}) = Z(t, X_t, D_t, S_t, \delta_t, \Sigma_t)$). This yields that $\Pi(X)$ is a submartingale (resp. a martingale) iff $A^X$ is increasing (resp. constant). From (1.23), we obtain $\partial_x\mathcal{C}(t,x,d,z,\delta,\Sigma) + (1-\epsilon)\partial_d\mathcal{C}(t,x,d,z,\delta,\Sigma) + \epsilon\ \partial_z\mathcal{C}(t,x,d,z,\delta,\Sigma) + d + z = 0$, and then

$$\mathrm{d}A_t^X \;=\; \Big\{\partial_t\mathcal{C} - \rho\ D_t\ \partial_d\mathcal{C} + Z(t, X_t, D_t, S_t, \delta_t, \Sigma_t)\ -\beta\ \delta_t\ \partial_\delta\mathcal{C}\ -\beta\ (\Sigma_t - 2\kappa_\infty)\ \partial_\Sigma\mathcal{C}\Big\}\mathrm{d}t. \tag{1.24}$$

Given the quadratic nature of the problem, we search a process $A^X$ of the form

$$\mathrm{d}A_t^X \;=\; \frac{\rho}{1-\epsilon}\mathrm{d}t \times \Big[j(T-t)(D_t - (1-\epsilon)X_t) \;-\; D_t \;+\; k(T-t)\ \delta_t\Big]^2, \tag{1.25}$$

with $j, k : \mathbb{R}^+ \to \mathbb{R}$ continuously differentiable functions, in order to obtain an non-decreasing process $A^X$ that can be constant for a specific strategy $X^*$. Let us note $Y_t := D_t - (1-\epsilon)X_t$, $\Xi_t := S_t - \epsilon X_t$, $y := d - (1-\epsilon)x$, $\xi := z - \epsilon x$. Since $d + z = y + \xi + x = \xi + \frac{d-\epsilon y}{1-\epsilon}$, we have

$$\begin{aligned}\partial_t\mathcal{C}(t,x,d,z,\delta,\Sigma) \;&=\; -\dot{a}\ y^2 \;-\; \dot{b}\ \delta y \;-\; \dot{c}\ \delta^2 \;-\; \dot{e}\ \Sigma \;-\; \dot{g}, \\ -\rho d\ \partial_d\mathcal{C}(t,x,d,z,\delta,\Sigma) \;&=\; -\left(2\rho a + \frac{\rho\epsilon}{1-\epsilon}\right)\ dy \;+\; \frac{\rho}{1-\epsilon}\ d^2 \;-\; \rho b\ \delta d, \\ -\beta\delta\ \partial_\delta\mathcal{C}(t,x,d,z,\delta,\Sigma) \;&=\; -\beta b\ \delta y \;-\; 2\beta c\ \delta^2, \\ -\beta(\Sigma - 2\kappa_\infty)\ \partial_\Sigma\mathcal{C}(t,x,d,z,\delta,\Sigma) \;&=\; -\beta e\ \Sigma \;+\; 2\beta\kappa_\infty e,\end{aligned}$$

Let $V \sim \mu$. One has

$$\mathbb{E}[(\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(V/m_1)] = \iota_{\mathrm{s}} - \iota_{\mathrm{c}} = \alpha \quad , \quad \mathbb{E}[(\varphi_{\mathrm{s}} + \varphi_{\mathrm{c}})(V/m_1)] = \iota_{\mathrm{s}} + \iota_{\mathrm{c}} = \alpha + 2\iota_{\mathrm{c}}.$$

Thus,

$$\begin{aligned}&\mathbb{E}\big[\mathcal{C}(t,x,d+(1-\nu)V, z+\nu V, \delta + (\varphi_{\mathrm{s}}-\varphi_{\mathrm{c}})(V/m_1), \Sigma + (\varphi_{\mathrm{s}}+\varphi_{\mathrm{c}})(V/m_1)) - \mathcal{C}(t,x,d,z,\delta,\Sigma)\big] \\ =\ & a\ [(1-\nu)^2 m_2 + 2(1-\nu)m_1\ y] \;+\; \frac{\nu^2}{2}m_2 \;+\; \nu m_1\ \xi \\ & + \nu(1-\nu)m_2 \;+\; \nu m_1 y \;+\; (1-\nu)m_1\xi \;-\; \frac{1}{2}\left(m_2 + 2\ m_1\ \xi + \frac{2m_1}{1-\epsilon}\ d - \frac{2\epsilon m_1}{1-\epsilon}\ y\right) \\ & + b\ [(1-\nu)m_1\ \delta \;+\; \alpha\ y \;+\; \tilde{\alpha}(1-\nu)] \;+\; c\ [\alpha_2 \;+\; 2\alpha\ \delta] \;+\; (\alpha + 2\iota_{\mathrm{c}})e,\end{aligned}$$

with

$$\tilde{\alpha} = \mathbb{E}[V \times (\varphi_{\rm s} - \varphi_{\rm c})(V/m_1)] \quad , \quad \alpha_2 = \mathbb{E}[(\varphi_{\rm s} - \varphi_{\rm c})^2 (V/m_1)]. \tag{1.26}$$

These quantities $\tilde{\alpha}$ and $\alpha_2$ are finite by assumption. This gives

$$\begin{aligned} Z(t,x,d,z,\delta,\Sigma) \;=\; & \left( m_1 \times \left[ 2(1-\nu)a + \nu + \frac{\epsilon}{1-\epsilon} \right] \;+\; \alpha b \right) \delta y \;-\; \frac{m_1}{1-\epsilon}\, \delta d \\ & + \left[ (1-\nu) m_1 b \;+\; 2\alpha c \right] \delta^2 \\ & + \left( m_2 \times \left[ (1-\nu)^2 a + \nu(1-\nu/2) - \frac{1}{2} \right] + \tilde{\alpha}(1-\nu) b + \alpha_2 c + (\alpha + 2\iota_{\rm c}) e \right) \Sigma, \end{aligned}$$

where we consider $\mathcal{C}$ as a function of the variables $t, x, d, z, \delta, \Sigma$ as in equation (2.45), and substitute $d-(1-\epsilon)x$ by $y$ and $z-\epsilon x$ by $\xi$ in the results. We then make the change of variables $(x, d, z, \delta, \Sigma) \to (y, d, \xi, \delta, \Sigma)$, and we identify each term of equations (2.45) and (2.46) :

**(Eq. $dy$) :** $\quad -\left( 2\rho a + \frac{\rho\epsilon}{1-\epsilon} \right) = -\frac{2\rho}{1-\epsilon} j.$

**(Eq. $y^2$) :** $\quad -\dot{a} \;=\; \frac{\rho}{1-\epsilon} j^2.$

(Eq. $dy$) yields $j = (1-\epsilon)a + \frac{\epsilon}{2}$. We input this relation in (Eq. $y^2$) and we have $\dot{j} = (1-\epsilon)\dot{a} = -\rho j^2$ thus $j(u) = \frac{1}{2+\rho u}$ since $j(0) = (1-\epsilon)a(0) + \frac{\epsilon}{2} = \frac{1}{2}$. This yields $a(u) = \frac{1}{1-\epsilon}\left( \frac{1}{2+\rho u} - \frac{\epsilon}{2} \right)$ with (Eq. $dy$).

**(Eq. $\delta y$) :** $\quad -\dot{b} \;-\; \beta b \;+\; \alpha b \;+\; m_1 \times \left[ 2(1-\nu)a + \nu + \frac{\epsilon}{1-\epsilon} \right] \;=\; \frac{2\rho}{1-\epsilon} jk.$

**(Eq. $\delta d$) :** $\quad -\rho b \;-\; \frac{m_1}{1-\epsilon} \;=\; -\frac{2\rho}{1-\epsilon} k,$

which yields $k(u) \;=\; \frac{1-\epsilon}{2}\; b(u) \;+\; \frac{m_1}{2\rho}$. Plugging equation (1.28) in (Eq. $\delta y$), we have $\dot{b} = -(\beta - \alpha)b - \frac{2\rho}{1-\epsilon} j \left( \frac{1-\epsilon}{2} b + \frac{m_1}{2\rho} \right) + m_1 \left[ 2(1-\nu)a + \nu + \frac{\epsilon}{1-\epsilon} \right]$, and since $j/(1-\epsilon) \;=\; a + \epsilon/[2(1-\epsilon)]$, we have

$$\dot{b}(u) \;=\; \left[ -(\beta - \alpha) - \frac{\rho}{2+\rho u} \right] b(u) \;+\; \frac{m_1}{1-\epsilon} \times \frac{1+\nu\rho u}{2+\rho u}.$$

**(Eq. $\delta^2$) :** $\quad -\dot{c} \;-\; 2\beta c \;+\; 2\alpha c \;+\; (1-\nu) m_1 b \;=\; \frac{\rho}{1-\epsilon}\; k^2.$

**(Eq. $\Sigma$) :** $\quad -\dot{e} \;-\; \beta e \;+\; (\alpha + 2\iota_{\rm c}) e \;+\; m_2 \times \left[ (1-\nu)^2 a + \nu(1-\nu/2) - \frac{1}{2} \right] \;+\; \tilde{\alpha}(1-\nu) b \;+\; \alpha_2 c \;=\; 0.$

We have $2(1-\epsilon) \times \left[ (1-\nu)^2 a + \nu(1-\nu/2) - \frac{1}{2} \right] = 2(1-\nu)^2/(2+\rho u) - (1-\nu)^2\epsilon + \nu(2-\nu)(1-\epsilon) - (1-\epsilon)$, thus

$$\dot{e}(u) \;=\; -(\beta - \alpha - 2\iota_{\rm c}) e(u) \;+\; \tilde{\alpha}(1-\nu) b(u) \;+\; \alpha_2 c(u) \;+\; \frac{(1-\nu)^2\; m_2}{1-\epsilon} \times \left[ \frac{1}{2+\rho u} - \frac{1}{2} \right]$$

**(Eq. constant) :** $\quad -\dot{g} \;+\; 2\beta\kappa_\infty e \;=\; 0.$

We obtain two conditions on the coefficients of the process $A^X$

$$j(u) \;=\; \frac{1}{2+\rho u}, \tag{1.27}$$

$$k(u) \;=\; \frac{1-\epsilon}{2}\; b(u) \;+\; \frac{m_1}{2\rho}, \tag{1.28}$$

and the following set of necessary conditions on the coefficients of $\mathcal{C}$

$$
\begin{aligned}
a(u) &= \frac{1}{1-\epsilon}\left(\frac{1}{2+\rho u} - \frac{\epsilon}{2}\right), && (1.29)\\
\dot{b}(u) &= \left[-(\beta-\alpha) - \frac{\rho}{2+\rho u}\right] b(u) \;+\; \frac{m_1}{1-\epsilon} \times \frac{1+\nu\rho u}{2+\rho u}, && (1.30)\\
\dot{c}(u) &= -2(\beta-\alpha)\; c(u) \;+\; (1-\nu) m_1\; b(u) \;-\; \frac{\rho}{1-\epsilon}\; k(u)^2, && (1.31)\\
\dot{e}(u) &= -(\beta-\alpha-2\iota_c)e(u) + \tilde{\alpha}(1-\nu)b(u) + \alpha_2 c(u) + \frac{(1-\nu)^2\; m_2}{1-\epsilon} \times \left[\frac{1}{2+\rho u} - \frac{1}{2}\right], && (1.32)\\
\dot{g}(u) &= 2\beta\kappa_\infty\; e(u), && (1.33)\\
b(0) &= c(0) \;=\; e(0) \;=\; g(0) \;=\; 0.
\end{aligned}
$$

The resolution of this set of equations determines entirely the function $\mathcal{C}(t,x,d,z,\delta,\Sigma)$ defined in (2.44). This is the purpose of the next step of this proof. Let us note that at this stage, we already know that the system given by Equations (1.27) to (1.33) admits a unique solution, and that the function $\mathcal{C}$ which solves the system is the value function of the problem by using the verification argument.

### 1.7.3 Resolution of the system of ODEs

First of all, we use Equation (1.29) to simplify the function $\mathcal{C}$. The constant term (w.r.t. the time variable $t$) in equation (2.44) is $\frac{1}{2}(z-\epsilon x)^2 \;+\; (d-(1-\epsilon)x)(z-\epsilon x) \;-\; \frac{(d+z)^2}{2} \;=\; -zx - \frac{d^2}{2} - \epsilon dx + \left[\frac{\epsilon}{2} + \frac{\epsilon}{2}(1-\epsilon)\right] x^2$, thus the sum of $a(T-t)(d-(1-\epsilon)x)^2$ and this constant term can be rewritten as

$$
-(z+d)x \;+\; \left[\frac{1-\epsilon}{2+\rho(T-t)} + \frac{\epsilon}{2}\right] x^2 \;-\; \frac{1}{1-\epsilon} \times \frac{\rho(T-t)/2}{2+\rho(T-t)}\; d^2 \;+\; \frac{\rho(T-t)}{2+\rho(T-t)}\; dx. \qquad (1.34)
$$

We note $\eta = \beta - \alpha$. To solve equation (2.48), we search a solution of the form $b(u) = \tilde{b}(u) \times \exp(-\eta u)/(2+\rho u)$. This yields $\dot{\tilde{b}}(u) = \frac{m_1}{1-\epsilon} \times (1+\nu\rho u) \times \exp(\eta u)$. Using the respective definitions (1.12) and (1.13) of the functions $\zeta$ and $\omega$, it is easy to see that for all $\eta \in \mathbb{R}$,

$$
\exp(-\eta u) \int_0^u (1+\nu\rho s)\; \exp(\eta s)\; \mathrm{d}s \;=\; u\zeta(\eta u) + \nu\rho u^2 \omega(\eta u).
$$

Since $\tilde{b}(0) = 2b(0) = 0$, we obtain

$$
b(u) \;=\; \frac{m_1 u}{1-\epsilon} \times \frac{\zeta(\eta u) + \nu\rho u\; \omega(\eta u)}{2+\rho u} \;=\; \frac{1}{1-\epsilon} \times \frac{\rho u}{2+\rho u} \times \frac{m_1}{\rho}\; \mathcal{G}_\eta(u), \qquad (1.35)
$$

where

$$
\mathcal{G}_\eta(u) := \zeta(\eta u) + \nu\rho u\; \omega(\eta u).
$$

Equation (1.28) then gives

$$
k(u) \;=\; \frac{m_1}{2\rho} \times \frac{2+\rho u \times \{1 + \zeta(\eta u) + \nu\rho u\; \omega(\eta u)\}}{2+\rho u}. \qquad (1.36)
$$

The remaining functions $c$, $e$ and $g$ do not play any role to determine the optimal strategy, and their expressions are harder to obtain. Let us first consider the case $\eta \neq 0$. After some tedious calculations, we can show that the function $c$ that solves (1.31) with $c(0) = 0$ is given by :

$$c(u) = -\frac{1}{1-\epsilon} \times \frac{\rho u/2}{2+\rho u} \times \frac{m_1^2}{\rho^2}\, \mathcal{G}_\eta(u)^2 - \frac{m_1^2}{8(1-\epsilon)\rho} \times \left(1 - \frac{\nu\rho}{\eta}\right)^2 \times u\zeta(\eta u) \times [1 + \exp(-\eta u) - 2\zeta(\eta u)]\,. \tag{1.37}$$

For the functions $e$ and $g$, we recall here that they admit explicit but very cumbersome formulas that can be obtained by using a formal calculus software. In the case $\eta = 0$, the resolution of the ODEs is easier, and we get

$$\begin{aligned}
c(u) =& -\frac{(1-\nu)^2}{1-\epsilon} \times \frac{m_1^2}{\rho^2} \times \left[\frac{1}{2} - \frac{1}{2+\rho u}\right] - \frac{\nu m_1^2}{\rho^2(1-\epsilon)} \times \left[\left(\frac{1}{2} - \frac{\nu}{4}\right)\rho u + \frac{\nu}{8}\rho^2 u^2 + \frac{\nu}{48}\rho^3 u^3\right], \\
e(u) =& -\frac{(1-\nu)^2}{1-\epsilon} \times \left(m_2 - \frac{m_1(2\tilde{\alpha}\rho - \alpha_2 m_1)}{\rho^2}\right) \times \left[\frac{\mathcal{I}_0(u)}{2} - \frac{\exp(2\iota_c u)}{\rho}\, L(\rho, -2\iota_c, u)\right] \\
&+ \frac{\nu(1-\nu)m_1}{2\rho^2(1-\epsilon)} \times \left(\tilde{\alpha} - \frac{\alpha_2 m_1}{\rho}\right) \times \rho^2 \mathcal{I}_1(u) - \frac{\alpha_2 \nu^2 m_1^2}{4\rho^3(1-\epsilon)} \times \left[\rho^2\mathcal{I}_1(u) + \frac{1}{2}\rho^3\mathcal{I}_2(u) + \frac{1}{12}\rho^4\mathcal{I}_3(u)\right],
\end{aligned} \tag{1.38}$$

$$\begin{aligned}
g(u) =& -2\beta\kappa_\infty \times \frac{(1-\nu)^2}{1-\epsilon} \times \left(m_2 - \frac{m_1(2\tilde{\alpha}\rho - \alpha_2 m_1)}{\rho^2}\right) \left\{\frac{\mathcal{I}_1(u)}{2} - \frac{1}{2\iota_c\rho} \times \left[\exp(2\iota_c u) L(\rho, -2\iota_c, u) - \ln\left(1 + \frac{\rho u}{2}\right)\right]\right\} \\
&+ \frac{\beta\kappa_\infty \nu(1-\nu)m_1}{2\rho^3(1-\epsilon)} \times \left(\tilde{\alpha} - \frac{\alpha_2 m_1}{\rho}\right) \times \rho^3\mathcal{I}_2(u) - \frac{\beta\kappa_\infty \alpha_2 \nu^2 m_1^2}{4\rho^4(1-\epsilon)} \times \left[\rho^3\mathcal{I}_2(u) + \frac{1}{3}\rho^4\mathcal{I}_3(u) + \frac{1}{24}\rho^5\mathcal{I}_4(u)\right],
\end{aligned} \tag{1.39}$$

where, for $p \in \mathbb{N}$ and $u \geq 0$, $\mathcal{I}_p(u) := \exp(2\iota_c u) \int_0^u s^p \exp(-2\iota_c s)\mathrm{d}s$, and $\tilde{\alpha}, \alpha_2$ are defined in (1.26).

### 1.7.4 Determination of the optimal strategy

The final step of the proof is to determine the strategy $X^*$ such that $\Pi(X^*)$ is a martingale, or equivalently such that $A^{X^*}$ is constant. Equations (2.46) and (1.27) yield

$$\begin{aligned}
\mathrm{d}A_t^X &= \frac{\rho}{1-\epsilon}\mathrm{d}t \times \left[\frac{D_t - (1-\epsilon)X_t}{2+\rho(T-t)} - D_t + k(T-t)\,\delta_t\right]^2 \\
&= \frac{\rho/(1-\epsilon)}{[2+\rho(T-t)]^2}\,\mathrm{d}t \times \Big[(1-\epsilon)X_t + [1+\rho(T-t)]\,D_t - [2+\rho(T-t)]\,k(T-t)\,\delta_t\Big]^2.
\end{aligned}$$

Thus, $A^{X^*}$ is constant on $(0,T)$ if, and only if

$$\text{a.s. , } \mathrm{d}t\text{ -a.e. on } (0,T)\ , \quad (1-\epsilon)X_t^* = -[1+\rho(T-t)]\,D_t^* + [2+\rho(T-t)]\,k(T-t)\,\delta_t, \tag{1.40}$$

where $D = D^*$ when the strategy $X^*$ is used by the strategic trader. Then, we characterize the strategy $X^*$ on $[0,T]$ with the three following steps :
- The initial jump $\Delta X_0^*$ of the strategy is such that $(X^*, D^*)$ satisfies equation (2.51) at time $t = 0^+$.
- The strategy $X^*$ on $(0,T)$ is obtained by differentiating equation (2.51).
- The final jump $\Delta X_T^* = -X_T^*$ closes the position of the strategic trader at time $T$.

We need the following lemma in the sequel.

**Lemma 1.7.1.** *Let $\phi : [0,T] \to \mathbb{R}$ be a measurable function, and for $0 \le s \le t \le T$, $\Phi(s,t) := \int_s^t \phi(u)\exp(-\beta u)\ \mathrm{d}u$. We then have for all $t \in [0,T]$*

$$\int_0^t \phi(u)\ \delta_u\ \mathrm{d}u \ = \ \delta_0\ \Phi(0,t)\ +\ \Theta_{\chi_t}\ \Phi\left(\tau_{\chi_t},t\right)\ +\ \sum_{i=1}^{\chi_t-1}\Theta_i\ \Phi(\tau_i,\tau_{i+1})$$

*Proof.* The proof is straightforward since for $u \in [\chi_t, t]$, $\delta_u = \delta_0\ \exp(-\beta u) + \exp(-\beta u)\ \Theta_{\chi_t}$ and for $i \in \{0,\cdots,\chi_t - 1\}$ and $u \in [\tau_i, \tau_{i+1})$, $\delta_u = \delta_0\ \exp(-\beta u) + \exp(-\beta u)\ \Theta_i$.

To determine the optimal strategy, only the function $k$ given by (2.50) comes into play, thus the cases $\eta = 0$ and $\eta \neq 0$ can be treated simultaneously. We also note that

$$\frac{\mathrm{d}}{\mathrm{d}u}[u^2\ \omega(\eta u)] = u\zeta(\eta u) \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}u}[u\ \zeta(\eta u)] = \exp(-\eta u)$$

hold for for all $u \ge 0$ and $\eta \in \mathbb{R}$. We use Equations (2.50) and (2.51) to obtain the following characterization of the strategy $X^*$ : a.s., $\mathrm{d}t$-a.e. on $(0,T)$,

$$(1-\epsilon)X_t^* = -\left[1+\rho(T-t)\right]D_t^* + \frac{m_1}{2\rho}\times[2+\rho(T-t)\times\{1+\zeta(\eta(T-t)) + \nu\rho(T-t)\ \omega(\eta(T-t))\}]\ \delta_t. \tag{1.41}$$

The initial jump of $X^*$ at $t = 0$ is such that (1.41) is verified for $t = 0^+$ :

$$(1-\epsilon)(x_0+\Delta X_0^*) = -[1+\rho T]\left(D_0 + (1-\epsilon)\Delta X_0^*\right) + \frac{m_1}{2\rho}\times[2+\rho T\times\{1+\zeta(\eta T) + \nu\rho T\ \omega(\eta T)\}]\ \delta_0, \tag{1.42}$$

which gives the initial trade at time 0 as given in Appendix 1.6.

We differentiate Equation (1.41) to get

$$\begin{aligned}(1-\epsilon)\mathrm{d}X_t^* \ &= \ \rho D_t^*\mathrm{d}t\ -\ [1+\rho(T-t)]\ \mathrm{d}D_t^*\ -\ \frac{m_1}{2}\times[1+\exp(-\eta(T-t)) + \nu\rho(T-t)\zeta(\eta(T-t))]\ \delta_t\ \mathrm{d}t \\ &\quad + \frac{m_1}{2\rho}\times[2+\rho(T-t)\times\{1+\zeta(\eta(T-t)) + \nu\rho(T-t)\ \omega(\eta(T-t))\}]\ \mathrm{d}\delta_t.\end{aligned}$$

This yields, using $\mathrm{d}\delta_t = -\beta\ \delta_t\ \mathrm{d}t + \mathrm{d}I_t$,

$$\begin{aligned}(1-\epsilon)\mathrm{d}X_t^* \ &= \ \rho D_t^*\ \mathrm{d}t\ -\ m_1\ \phi_\eta(t)\ \delta_t\ \mathrm{d}t\ +\ \frac{1+\rho(T-t)}{2+\rho(T-t)}\left\{\frac{m_1}{\rho}\ \mathrm{d}I_t - (1-\nu)\ \mathrm{d}N_t\right\} \\ &\quad + \frac{m_1}{2\rho}\times\frac{\rho(T-t)\times\{\zeta(\eta(T-t)) - 1 + \nu\rho(T-t)\ \omega(\eta(T-t))\}}{2+\rho(T-t)}\ \mathrm{d}I_t,\end{aligned} \tag{1.43}$$

where for $t \in [0,T]$

$$\begin{aligned}2[2+\rho(T-t)]\times\phi_\eta(t) := 1 &+ \exp(-\eta(T-t)) + \nu\rho(T-t)\zeta(\eta(T-t)) \\ &+ \frac{\beta}{\rho}\left[2+\rho(T-t)\times\{1+\zeta(\eta(T-t)) + \nu\rho(T-t)\ \omega(\eta(T-t))\}\right]\end{aligned}$$

and $\delta_t = \delta_0\ \exp(-\beta t) + \sum_{0<\tau\le t}\exp(-\beta(t-\tau))\ \Delta I_\tau$. For $t \in (0,T)$,

$$\begin{aligned}\mathrm{d}D_t^* \ &= \ -\rho D_t^*\mathrm{d}t\ +\ (1-\epsilon)\mathrm{d}X_t^*\ +\ (1-\nu)\mathrm{d}N_t \\ &= \ -\ m_1\ \phi_\eta(t)\ \delta_t\ \mathrm{d}t \\ &\quad + \frac{(1-\nu)\ \mathrm{d}N_t}{2+\rho(T-t)}\ +\ \frac{m_1}{2\rho}\times\frac{2+\rho(T-t)\times\{1+\zeta(\eta(T-t)) + \nu\rho(T-t)\ \omega(\eta(T-t))\}}{2+\rho(T-t)}\ \mathrm{d}I_t,\end{aligned}$$

and we have

$$D^*_{0+} = D_0 + (1-\epsilon)\Delta X^*_0 = \frac{D_0-(1-\epsilon)x_0}{2+\rho T} + \frac{m_1}{2\rho}\times\frac{2+\rho T\times\{1+\zeta(\eta T)+\nu\rho T\ \omega(\eta T)\}}{2+\rho T}\ \delta_0$$

$$\begin{aligned}\int_{(0,t]} \mathrm{d}D^*_u &= -\ m_1\int_{(0,t]}\phi_\eta(u)\ \delta_u\ \mathrm{d}u + \sum_{0<\tau\le t}\frac{(1-\nu)\ \Delta N_\tau}{2+\rho(T-\tau)}\\ &\quad + \frac{m_1}{2\rho}\times\sum_{0<\tau\le t}\frac{2+\rho(T-\tau)\times\{1+\zeta(\eta(T-\tau))+\nu\rho(T-\tau)\ \omega(\eta(T-\tau))\}}{2+\rho(T-\tau)}\ \Delta I_\tau.\end{aligned}$$

We define $\Phi_\eta(s,t) := \int_s^t \phi_\eta(u)\exp(-\beta u)\ \mathrm{d}u$ for $0\le s\le t\le T$. Lemma 1.7.1 yields for $t\in[0,T]$

$$\int_0^t \phi_\eta(u)\ \delta_u\ \mathrm{d}u = \delta_0\ \Phi_\eta(0,t) + \Theta_{\chi_t}\ \Phi_\eta\left(\tau_{\chi_t},t\right) + \sum_{i=1}^{\chi_t-1}\Theta_i\ \Phi_\eta(\tau_i,\tau_{i+1}).$$

We obtain the expression of $D^*_t$ for $t\in(0,T)$

$$\begin{aligned}D^*_t &= \frac{D_0-(1-\epsilon)x_0}{2+\rho T} + \frac{\delta_0 m_1}{2\rho}\times\left[\frac{2+\rho T\times\{1+\zeta(\eta T)+\nu\rho T\ \omega(\eta T)\}}{2+\rho T} - 2\rho\ \Phi_\eta(0,t)\right]\\ &\quad - m_1\left[\Theta_{\chi_t}\ \Phi_\eta\left(\tau_{\chi_t},t\right) + \sum_{i=1}^{\chi_t-1}\Theta_i\ \Phi_\eta(\tau_i,\tau_{i+1})\right] + \sum_{0<\tau\le t}\frac{(1-\nu)\ \Delta N_\tau}{2+\rho(T-\tau)}\\ &\quad + \frac{m_1}{2\rho}\times\sum_{0<\tau\le t}\frac{2+\rho(T-\tau)\times\{1+\zeta(\eta(T-\tau))+\nu\rho(T-\tau)\ \omega(\eta(T-\tau))\}}{2+\rho(T-\tau)}\ \Delta I_\tau.\end{aligned}$$

From (1.43), the strategy $X^*$ on $(0,T)$ is as given in Appendix 1.6. By using again (1.41), we also get the final trade at time $T$.

We determine the function $\Phi_\eta$ in the case $\eta\neq 0$ (similar and simpler calculations yield the result for $\eta=0$). We write

$$\begin{aligned}\exp(-\eta(T-t))\times\exp(-\beta t) &= \exp(-\beta T)\times\exp(\alpha(T-t)),\\ (T-t)\zeta(\eta(T-t))\times\exp(-\beta t) &= \frac{\exp(-\beta T)}{\eta}\times[\exp(\beta(T-t))-\exp(\alpha(T-t))].\end{aligned}$$

Thus, $\phi_\eta(t)\times\exp(\beta(T-t))$ is equal to

$$\begin{aligned}&\frac{\beta}{2}\left(\frac{1}{\rho}+\frac{\nu}{\eta}\right)\times\exp(\beta(T-t)) + \left[\frac{1}{2}+\frac{\nu(\rho-2\beta)}{2\eta}+\frac{\beta}{2\eta}\left(1-\frac{\nu\rho}{\eta}\right)\right]\frac{\exp(\beta(T-t))}{2+\rho(T-t)}\\ &\qquad + \left[\frac{1}{2}-\frac{\nu\rho}{2\eta}-\frac{\beta}{2\eta}\left(1-\frac{\nu\rho}{\eta}\right)\right]\frac{\exp(\alpha(T-t))}{2+\rho(T-t)},\end{aligned}$$

which yields for $0\le s\le t\le T$,

$$\begin{aligned}\Phi_\eta(s,t) &= \frac{1}{2}\left(\frac{1}{\rho}+\frac{\nu}{\eta}\right)\times[\exp(-\beta s)-\exp(-\beta t)]\\ &\quad + \frac{\exp(-\beta T)}{2\rho}\times\left[1+\frac{\nu(\rho-2\beta)}{\eta}+\frac{\beta}{\eta}\left(1-\frac{\nu\rho}{\eta}\right)\right]\times[L(\rho,\beta,T-s)-L(\rho,\beta,T-t)]\\ &\quad + \frac{\exp(-\beta T)}{2\rho}\times\left[1-\frac{\nu\rho}{\eta}-\frac{\beta}{\eta}\left(1-\frac{\nu\rho}{\eta}\right)\right]\times[L(\rho,\alpha,T-s)-L(\rho,\alpha,T-t)].\end{aligned}$$

with $\eta=\beta-\alpha\neq 0$.

## 1.8 Appendix : Proof of Theorem 1.2.1

Let $X$ be an admissible strategy. We introduce the following processes : $S_t^N = S_0 + \frac{\nu}{q}(N_t - N_0)$, $S_t^X = \frac{\epsilon}{q}(X_t - X_0)$,

$$\mathrm{d}D_t^N = -\rho D_t^N \mathrm{d}t + \frac{1-\nu}{q}\mathrm{d}N_t \text{ and } \mathrm{d}D_t^X = -\rho D_t^X \mathrm{d}t + \frac{1-\epsilon}{q}\mathrm{d}X_t,$$

with $D_0^N = D_0$ and $D_0^X = 0$. Thus, we have $S = S^N + S^X$, $D = D^N + D^X$ and thus $P = P^N + P^X$, where $P^N = S^N + D^N$ and $P^X = S^X + D^X$. From (1.4), we have

$$C(X) = \int_{[0,T)} P_u^N \ \mathrm{d}X_u - P_T^N X_T + C^{\mathrm{OW}}(X),$$

where

$$C^{\mathrm{OW}}(X) = \int_{[0,T)} P_u^X \ \mathrm{d}X_u \ + \ \frac{1}{2q}\sum_{0\le\tau<T} (\Delta X_\tau)^2 \ - \ P_T^X X_T \ + \ \frac{1}{2q}\ X_T^2$$

is a deterministic function of $X$ that corresponds to the cost when $N \equiv 0$, which is the Obizhaeva and Wang model. We now make an integration by parts as in Remark 1.2.6 and get that

$$\int_{[0,T)} P_u^N \ \mathrm{d}X_u - P_T^N X_T = -\int_{[0,T)} X_u \ \mathrm{d}P_u^N.$$

When $P^N$ is a martingale, this term has a null expectation. Therefore, the optimal execution strategy is the same as in the Obizhaeva and Wang model, see Gatheral, Schied and Slynko [64], Example 2.12, and there is no PMS. Otherwise, we can find $0 \le s < t \le T$ such that $\mathbb{E}[P_t^N|\mathcal{F}_s]$ and $P_s^N$ are not almost surely equal. In this case, we consider the strategy $X_u = \mathbb{E}[P_t^N - P_s^N|\mathcal{F}_s]\mathbf{1}_{u\in(s,t]}$ that is a round-trip, i.e. $X_0 = X_{T+} = 0$. We then get

$$\mathbb{E}\left[-\int_{[0,T)} X_u \ \mathrm{d}P_u^N\right] = -\mathbb{E}[(P_t^N - P_s^N)\mathbb{E}[P_t^N - P_s^N|\mathcal{F}_s]] = -\mathbb{E}[\mathbb{E}[P_t^N - P_s^N|\mathcal{F}_s]^2] < 0.$$

Since $C^{\mathrm{OW}}(cX) = c^2 C^{\mathrm{OW}}(X)$, we can find $c$ small enough such that
$E[C(cX)] = -c\mathbb{E}[\mathbb{E}[P_t^N - P_s^N|\mathcal{F}_s]^2] + c^2 C^{\mathrm{OW}}(X) < 0$, and therefore $cX$ is a PMS.

# Chapitre 2

# Extension et calibration d'un modèle d'exécution optimale dynamique

Ce chapitre est un article écrit avec Aurélien Alfonsi.

**Abstract.** We provide some theoretical extensions and a calibration protocol for our former dynamic optimal execution model. The Hawkes parameters and the propagator are estimated independently on financial data from stocks of the CAC40. Interestingly, the propagator exhibits a smoothly decaying form with one or two dominant time scales, but only so after a few seconds that the market needs to adjust after a large trade. Motivated by our estimation results, we derive the optimal execution strategy for a multi-exponential Hawkes kernel and backtest it on the data for round trips. We find that the strategy is profitable on average when trading at the midprice, which is in accordance with violated martingale conditions. However, in most cases, these profits vanish when we take bid-ask costs into account.

## 2.1 Introduction

In the last fifteen years, the literature in quantitative finance has been enriched by many studies on optimal execution problems. The principle is as follows : one considers a particular trader who wants to liquidate a quantity $x_0$ of assets on the time interval $[0,T]$. Thus, if $X_t$ is the position at time $t$, one has $X_0 = x_0$ and $X_{T+} = 0$ : $x_0 > 0$ (resp. $x_0 < 0$) corresponds to to a sell (resp. buy) program. The trader uses an execution strategy of minimal expected cost, which should take into account the fact that large trades have an impact on the market price. The works of Bertsimas and Lo [24] and Almgren and Chriss [9] are pioneers in this area. They have been followed by several authors who suggested extensions of their framework, such as Obizhaeva and Wang [93] who considered a model that includes transient price impact. This feature allows to reproduce the mean-reversion that is observed in intra-day prices. On average, when a large trade impacts the market price, a fraction of this impact vanishes over time.

In Alfonsi and Blanc [3], we introduce a model where other liquidity takers trade the same asset as the large trader, and share the same price impact profile as her. In this model, the volumes of

incoming trades is described by a càdlàg (right continuous left limits) pure jump process $N_t$, and the market price $P_t$ at time $t$ is given by

$$P_t = \sum_{\tau<t} \Delta N_\tau \times \left[\frac{\nu}{q} + \frac{1-\nu}{q} e^{-\rho(t-\tau)}\right], \tag{2.1}$$

where the times $\tau$ are the jump times of the process $N$ and $\Delta N_\tau = N_\tau - N_{\tau-}$ is the signed volume of the order at time $\tau$. Thus, $q > 0$ is a measure of market liquidity, $\nu \in [0,1]$ the proportion of permanent impact, and $\rho > 0$ the resilience speed of the transient part of the price. In [3], the order flow is modeled by a two-dimensional Hawkes process, which allows self and mutual excitation between buy and sell orders. An interesting feature of this model is that it accounts for herding behavior and meta-orders splitting, see Bacry and Muzy [13]. Namely, let $N^+$ and $N^-$ be two nondecreasing càdlàg pure jump processes that describe respectively the volumes of incoming buy and sell orders. We have $N = N^+ - N^-$, and we proposed in [3] the following model for the respective jump intensities of $N^\pm$

$$\kappa_t^+ = \kappa_\infty + \sum_{\tau<t} \left[\mathbb{1}_{\{\Delta N_\tau>0\}} \varphi_{\mathrm{s}}\left(\frac{\Delta N_\tau}{m_1}\right) + \mathbb{1}_{\{\Delta N_\tau<0\}} \varphi_{\mathrm{c}}\left(-\frac{\Delta N_\tau}{m_1}\right)\right] e^{-\beta(t-\tau)}, \tag{2.2}$$

$$\kappa_t^- = \kappa_\infty + \sum_{\tau<t} \left[\mathbb{1}_{\{\Delta N_\tau<0\}} \varphi_{\mathrm{s}}\left(-\frac{\Delta N_\tau}{m_1}\right) + \mathbb{1}_{\{\Delta N_\tau>0\}} \varphi_{\mathrm{c}}\left(\frac{\Delta N_\tau}{m_1}\right)\right] e^{-\beta(t-\tau)}, \tag{2.3}$$

where $\kappa_\infty \geq 0$ is the common baseline intensity of $N^+$ and $N^-$, $\beta$ is the resilience speed of the intensity and $\varphi_{\mathrm{s}}, \varphi_{\mathrm{c}} : \mathbb{R}^+ \to \mathbb{R}^+$ are measurable positive functions that encode intensity feedback. We assume that the sizes of orders are independent variables distributed according to a square integrable probability law $\mu$ on $\mathbb{R}_+$, and $m_1 = \int_0^\infty x\mu(dx)$ is the average amplitude of the jumps of $N$. This price model is called MIH, as Mixed-Impact Hawkes. In this model, we provide a closed-form solution for the optimal liquidation strategy, and determine a set of conditions on $\nu, \rho, \beta, \varphi_{\mathrm{s}}, \varphi_{\mathrm{c}}$ that exclude Price Manipulation Strategies (as defined in [78]) from the model. These are referred to as the MIHM (Mixed Impact Hawkes Martingale) conditions.

One of the benefits of the framework introduced in [3] is that it is possible to calibrate the model on financial data, without effectively trading (which can be costly). One only has to observe the order flow and price process of the market, and to estimate the price impact of trades issued by other participants, which is expected to be similar to the impact that the liquidating trader would have. The aim of the present paper is to conduct such a calibration on real stock data. This enables us to evaluate the realism of the theoretical price model of [3], as well as the performance of the optimal strategy in a practical context. Since our main goal is to confront the model to market data, we test the validity of our calibration protocol on simulations and we leave its mathematical justification for further research.

Many studies have explored the estimation of Hawkes parameters in various contexts (see for instance Bacry et al. [16], Bouchaud and Hardiman [68], Reynaud-Bouret [98], Lemonnier and Vayatis [86]). The present paper focuses on marked Hawkes processes used to model price jumps triggered by transactions in financial markets, where the marks of the jumps are either the traded volumes or the price jumps. As opposed to most Hawkes models in finance, price moves which do not correspond to trades are treated separately through the propagator function. Propagator price models have been studied extensively in theoretical frameworks such as Gatheral [63], Alfonsi et al. [6] and Gatheral

et al. [64], Bouchaud et al. [31] and Farmer et al. [56]. However, to the best of our knowledge, very few empirical studies have described the form of the propagator curve, or only asymptotically. Here, we suggest an estimation protocol for the propagator and discuss the quality of fit of exponential and multi-exponential decays. We also describe the behavior of the curve on the first seconds, where it is found to have an increasing part.

The paper is structured as follows. First, we present the model in Section 2.2. It extends the one considered in [3] to general decay kernels, while preserving most of its properties. Then, in Section 2.3 we describe our dataset and our calibration method. In particular, we explain how we slightly modify the original model to be in accordance with practical considerations. Section 2.4 validates our calibration procedure with simulations and discusses the calibration results on real stock data. Eventually, we test in Section 2.5 the relevance of the optimal execution strategy described in Section 2.2 and discuss whether it may constitute Price Manipulation Strategies, i.e. round trips that are profitable in average.

## 2.2 Model settings

In view of its estimation to market data, we make the model of [3] more general by adding further parameters. First, even if it is appealing to see the price as the pure result of past trades, equation (2.1) is probably too restrictive and one should add some noise. Besides, we know that adding a martingale to the price process does not change the main results on the model, see Remark 2.6 in [3]. Second, we chose the resilience on the price and on the intensity to be exponential, and one may like to consider a priori more general decay functions. Thus, we consider the following propagator model for the price :

$$P_t = \frac{1}{q}\sum_{\tau<t} \Delta N_\tau G(t-\tau) + \sigma W_t. \tag{2.4}$$

The process $W$ is a Brownian motion independent of $N$ that takes into account the non-deterministic noise in limit orders and cancellations. The parameter $\sigma > 0$ tunes the volatility of this noise. The function $G : \mathbb{R}_+ \mapsto \mathbb{R}$ is the propagator function of the market, that encodes the average evolution of the price between two market orders, which takes form through limit orders and cancellations. As before, $q > 0$ describes the market liquidity and allows to normalize $G$ such that $G(0) = 1$. The propagator model is the same as the one considered by Alfonsi et al [6] and Gatheral et al. [64] and generalizes (2.1). Similar models have been considered for instance by Bouchaud et al. [31] and Gatheral [63]. In the same way, we consider a general decay function $K : \mathbb{R}_+ \mapsto \mathbb{R}_+$ for the intensities of $N^+$ and $N^-$. Namely, we assume that the jump intensities of $N^+$ and $N^-$ are respectively given by

$$\begin{aligned}
\kappa_t^+ &= \kappa_\infty + \sum_{\tau<t} \left[ \mathbb{1}_{\{\Delta N_\tau > 0\}} \varphi_{\mathrm{s}} \left( \frac{\Delta N_\tau}{m_1} \right) + \mathbb{1}_{\{\Delta N_\tau < 0\}} \varphi_{\mathrm{c}} \left( -\frac{\Delta N_\tau}{m_1} \right) \right] K(t-\tau), \\
\kappa_t^- &= \kappa_\infty + \sum_{\tau<t} \left[ \mathbb{1}_{\{\Delta N_\tau < 0\}} \varphi_{\mathrm{s}} \left( -\frac{\Delta N_\tau}{m_1} \right) + \mathbb{1}_{\{\Delta N_\tau > 0\}} \varphi_{\mathrm{c}} \left( \frac{\Delta N_\tau}{m_1} \right) \right] K(t-\tau).
\end{aligned} \tag{2.5}$$

with $K(0) = 1$. We also introduce the average self-excitation $\iota_{\mathrm{s}}$ and the average cross-excitation $\iota_{\mathrm{c}}$

$$\iota_{\mathrm{s}} = \int_0^\infty \varphi_{\mathrm{s}}(v/m_1)\mu(\mathrm{d}v) \text{ and } \iota_{\mathrm{c}} = \int_0^\infty \varphi_{\mathrm{c}}(v/m_1)\mu(\mathrm{d}v).$$

Therefore, the model presented in [3] corresponds to the exponential decay functions $G(t) = e^{-\rho t}$ and $K(t) = e^{-\beta t}$. By estimating more general functions $G(t)$ and $K(t)$ in the sequel, we are able to assess the relevance of the exponential decay assumption.

### 2.2.1 Markovian specification of the model

Considering general decay kernels is very natural from a modeling point of view. Unfortunately, it generally leads to drop the Markov property of the price process, which is important in the context of optimal execution. Still, for completely monotone decay kernels, it is possible to get back Markovian dynamics for the price. This has already been studied in Alfonsi and Schied [4] for the price propagator model. Considering completely monotone kernels amounts to assume the existence of probability measures $\tilde{\lambda}(\mathrm{d}\rho)$ and $\tilde{w}(\mathrm{d}\rho)$ on $\mathbb{R}_+^*$ such that

$$G(t) = \nu + (1-\nu)\int_{\mathbb{R}_+} e^{-\rho t}\tilde{\lambda}(\mathrm{d}\rho),\ K(t) = \int_{\mathbb{R}_+} e^{-\rho t}\tilde{w}(\mathrm{d}\rho). \tag{2.6}$$

Here, for the sake of simplicity, we consider probability measures with finite support. We can then assume without loss of generality [1] that

$$G(u) = \nu + \sum_{i=1}^{p} \lambda_i \exp(-\rho_i u), \quad K(u) = \sum_{i=1}^{p} w_i \exp(-\rho_i u), \tag{2.7}$$

with $0 < \rho_1 < \cdots < \rho_p$, $\nu, \lambda_i, w_i \geq 0$ such that $\nu + \sum_{i=1}^p \lambda_i = 1$ and $\sum_{i=1}^p w_i = 1$. For $i \in \{1, \ldots, p\}$, we introduce the following processes

$$\mathrm{d}D_t^i = -\rho_i\ D_t^i\ \mathrm{d}t\ +\ \frac{\lambda_i}{q}\ \mathrm{d}N_t, \tag{2.8}$$

$$\mathrm{d}\kappa_t^{+(i)} =\ -\rho_i\ (\kappa_t^{+(i)} - \kappa_\infty/p)\ \mathrm{d}t\ +\ w_i[\varphi_{\mathrm{s}}(\mathrm{d}N_t^+/m_1) +\ \varphi_{\mathrm{c}}(\mathrm{d}N_t^-/m_1)], \tag{2.9}$$

$$\mathrm{d}\kappa_t^{-(i)} =\ -\rho_i\ (\kappa_t^{-(i)} - \kappa_\infty/p)\ \mathrm{d}t\ +\ w_i[\varphi_{\mathrm{c}}(\mathrm{d}N_t^+/m_1) +\ \varphi_{\mathrm{s}}(\mathrm{d}N_t^-/m_1)]. \tag{2.10}$$

We also define the process $\mathrm{d}S_t = \frac{\nu}{q}\ \mathrm{d}N_t$ that describes the permanent impact component of the price. Then, it is easy to check from (2.7), (2.4) and (2.5) that

$$P_t = S_t + \sum_{i=1}^{p} D_t^i + \sigma W_t,\ \kappa_t^{\pm} = \sum_{i=1}^{p} \kappa_t^{\pm(i)}, \tag{2.11}$$

and the process $(P, S, D^i, \kappa^{\pm(i)})$ satisfies the Markov property.

**Remark 2.2.1.** *In the general setting (2.4) and (2.5), we implicitly assume that the stationarity conditions $(\iota_{\mathrm{s}}+\iota_{\mathrm{c}})\int_0^\infty K(s)ds < 1$, $G$ integrable are satisfied, so that the sums are well-defined. This is no longer required in the Markovian case since the law of $(P_t, S_t, D_t^i, \kappa_t^{\pm(i)}; t \geq 0)$ is determined by the initial condition $(P_0, S_0, D_0^i, \kappa_0^{\pm(i)})$. In the particular case $D_0^i = 0$ for all $i$, and only in this case, we have $P_t = P_0 + \frac{1}{q}\sum_{0<\tau<t} \Delta N_\tau G(t-\tau) + \sigma W_t$. Thus, if $|D_0^i|[G(t) - G(\infty)] \ll P_t$ for all $i \in \{1, \cdots, p\}$ and all $t \geq t_0$, then the approximation $P_t \approx P_0 + \frac{1}{q}\sum_{0<\tau<t} \Delta N_\tau G(t-\tau) + \sigma W_t$ is reasonable for $t \geq t_0$.*

1. Note that $G$ and $K$ may still include different decay speeds : one only has to include all the speeds in the $\rho_i$'s and to set some weights $\lambda_i, w_i$ to zero if necessary.

Besides the Markov property, the particular form (2.7) enables us to calculate explicitly the auto-covariance function of the number of jumps as explained by Hawkes in [72], Section 3. This auto-covariance structure is of empirical interest, and serves as a starting point for our calibration procedure, see Section 2.3.4. The total intensity $\Sigma_t = \kappa_t^+ + \kappa_t^-$ has the dynamics

$$\Sigma_t = 2\kappa_\infty + \iota \int_{-\infty}^{t} K(t-s)\ \mathrm{d}J_s,$$

where $\iota = \iota_\mathrm{s} + \iota_\mathrm{c}$ is the average jump size of $\Sigma_t$, and
$\mathrm{d}J_t = [(\varphi_\mathrm{s} + \varphi_\mathrm{c})(\mathrm{d}N_t^+/m_1) + \ (\varphi_\mathrm{s} + \varphi_\mathrm{c})(\mathrm{d}N_t^-/m_1)]/\iota$ has jumps normalized to unity. We assume that the stationarity condition $\iota \int_0^\infty K(s)ds < 1$ holds, see Theorem 1 in [32], and that the intensity process $(\kappa_t^+, \kappa_t^-)$ in its stationary state. We consider the symmetric auto-covariance function $\mathscr{C}$ of the infinitesimal increments of $J$. It is defined for $\tau > 0$ by

$$\mathscr{C}(\tau) \ = \ \lim_{h\to 0^+} \frac{1}{h^2}\mathbb{E}[(J_{t+h} - J_t)(J_{t-\tau+h} - J_{t-\tau})] - 4\overline{\kappa}^2 \ = \ \lim_{h\to 0^+} \frac{1}{h}\mathbb{E}[\Sigma_t\ (J_{t-\tau+h} - J_{t-\tau})] - 4\overline{\kappa}^2, \quad (2.12)$$

where $\overline{\kappa} = \kappa_\infty/(1-\iota/\beta)$ is the common stationary mean of $\kappa^+$ and $\kappa^-$. As derived in [72], one gets the self-consistent equation on $\mathscr{C}$ : for $\tau > 0$,

$$\mathscr{C}(\tau) = 2\overline{\kappa}\iota K(\tau) + \iota \int_{-\infty}^{\tau} K(\tau - u)\mathscr{C}(u)\mathrm{d}u. \quad (2.13)$$

**Proposition 2.2.1.** *Let us assume that $K$ satisfies (2.7) with $w_1, \ldots, w_p > 0$ and the stationarity condition $\iota \sum_{i=1}^p \frac{w_i}{\rho_i} < 1$. Then, the autocovariance function is given by*

$$\mathscr{C}(\tau) = \sum_{j=1}^{p} a_j \exp(-b_j|\tau|).\ \tau \in \mathbb{R}^*. \quad (2.14)$$

*The coefficients $a_1, \cdots, a_p$ and $b_1, \cdots, b_p$ are positive and determined as follows : $b_1 < \cdots < b_p$ are the distinct roots of the polynomial functions $P(X) = \prod_{i=1}^p(\rho_i - X) - \iota \sum_{i=1}^p w_i \prod_{k\neq i}(\rho_k - X)$ and $(a_1 b_1, \cdots, a_p b_p)^\top = \overline{\kappa}\ B^{-1}\ (1, \cdots, 1)^\top$, where $B$ is the Cauchy matrix $B_{i,j} = \frac{1}{\rho_i^2 - b_j^2}$.*

*Proof.* Equation (2.13) then yields for $\tau > 0$

$$\sum_{j=1}^{p} a_j \exp(-b_j\tau) = 2\iota \sum_{i=1}^{p} w_i \left[\overline{\kappa} - \sum_{j=1}^{p} \frac{a_j b_j}{(\rho_i + b_j)(\rho_i - b_j)}\right] \exp(-\rho_i\tau) \ + \ \iota \sum_{j=1}^{p} a_j \left[\sum_{i=1}^{p} \frac{w_i}{\rho_i - b_j}\right] \exp(-b_j\tau).$$

Therefore, (2.13) holds if we have

$$\forall j,\ \iota \left[\sum_{i=1}^{p} \frac{w_i}{\rho_i - b_j}\right] = 1, \quad \forall i, \sum_{j=1}^{p} \frac{a_j b_j}{\rho_i^2 - b_j^2} = \overline{\kappa}.$$

The first equation gives precisely $P(b_j) = 0$. Since $P(0) > 0$ from the stationarity condition and $P(\rho_l) = -\iota w_l \prod_{k\neq l}(\rho_k - \rho_l)$ has the same sign as $(-1)^l$, we have by the intermediate value theorem that $0 < b_1 < \rho_1 < b_2 < \rho_2 < \cdots < \rho_{p-1} < b_p < \rho_p$. These coefficients are distincts and therefore the Cauchy matrix $B$ is invertible. Let $v = B^{-1}\ (1, \cdots, 1)^\top$ : $v_i$ is the $i^th$ row sum of $B^{-1}$. By Theorem 2

in [100], $v_i = -A(b_i^2)/B'(b_i^2)$, where $A(x) = \prod_i (x - \rho_i^2)$, $B(x) = \prod_i (x - b_i^2)$. This gives in particular $v_i > 0$ and thus $a_i > 0$. Last, it is easy to check (2.14) is the unique function satisfying (2.13). In the mono-exponential case $p = 1$, Proposition 2.2.1 gives $\iota = \rho - b$, $ab = (\rho + b)(\rho - b)\overline{\kappa}$, which yields

$$\mathscr{C}(\tau) = \frac{\iota(2\rho - \iota)}{2\rho} \times \frac{2\kappa_\infty}{(1 - \iota/\rho)^2} \times \exp(-(\rho - \iota)|\tau|),$$

as found in [72].

### 2.2.2 Trading strategies and a generalized no-arbitrage condition

We now specify the trading rules in our model. We denote by $(\mathcal{F}_t)$ the natural filtration generated by the process $(P, S, D^i, \kappa^{\pm(i)})$. As in [3], we consider a particular trader called "strategic trader" and denote by $X_t$ the number of assets she holds at time $t$. We assume that the strategy $X$ is $(\mathcal{F}_t)$-adapted, càglàd, square integrable and with bounded variations. The càglàd (left continuous - right limits) assumption means that the strategic trader is able to react instantly to the flow of trades. For simplicity and tractability, we assume that the trades of the strategic trader affect the price in the same fashion as other trades, but leave unchanged the flow of orders $N$. To be more precise, we now assume that

$$dS_t = \frac{\nu}{q}(\mathrm{d}N_t \ + \ \mathrm{d}X_t), \quad \mathrm{d}D_t^i = -\rho_i \ D_t^i \ \mathrm{d}t \ + \ \frac{\lambda_i}{q}(\mathrm{d}N_t \ + \ \mathrm{d}X_t),$$

but the intensities $\kappa_t^{+(i)}$ and $\kappa_t^{-(i)}$ remain as defined by (2.9) and (2.10). The price as well as the intensities $\kappa_t^+$ and $\kappa_t^-$ of buy and sell orders are still defined by (2.11). Last, the cost of the trade $\Delta X_t = X_{t+} - X_t$ at time $t$ is assumed to be given by

$$\frac{P_t + P_{t+}}{2}\Delta X_t = P_t \Delta X_t + \frac{1}{2q}(\Delta X_t)^2.$$

This yields the following cost for a liquidation strategy $X$ on $[0, T]$ (i.e. such that $X_{T+} = 0$)

$$C(X) = \int_{[0,T)} P_u \ \mathrm{d}X_u \ + \ \frac{1}{2q} \sum_{\tau \in \mathcal{D}_X \cap [0,T)} (\Delta X_\tau)^2 \ - \ P_T X_T \ + \ \frac{1}{2q} \ X_T^2, \tag{2.15}$$

where $\mathcal{D}_X$ is the (countable) set of discontinuities of $X$.

When considering high-frequency trading, a standard approach is to define arbitrages as strategies that can make money on average, with no specific exogenous signal. Roughly speaking, one may expect that by repeating such strategies one obtains a classical almost sure arbitrage. Thus, Huberman and Stanzl have proposed in [78] the following definition of a *Price Manipulation Strategy* : this is a strategy $X$ such that $X_0 = X_{T+} = 0$ and $\mathbb{E}[C(X)] < 0$.

**Theorem 2.2.1.** *The model excludes Price Manipulation Strategies if, and only if $P_t$ is a $(\mathcal{F}_t)$-martingale when $X_t = 0$ for any $t$. In this case, the optimal strategy is the one given by Theorem 2 (see also Section 1.3) in [4].*

*Besides, under the specification (2.9), (2.10) and (2.11) of the order flow $N = N^+ - N^-$, the model does not admit PMS if, and only if,*

$$\forall i \in \{1, \dots, p\}, \quad (\iota_s - \iota_c) w_i = \lambda_i \rho_i, \quad \frac{m_1}{q}(\kappa_0^{+(i)} - \kappa_0^{-(i)}) - \rho_i D_0^i = 0, \tag{2.16}$$

*and* $\varphi_{\mathrm{s}}(y/m_1) - \varphi_{\mathrm{c}}(y/m_1) = (\iota_{\mathrm{s}} - \iota_{\mathrm{s}})y/m_1$ *for all* $y \geq 0$ *such that* $\forall \epsilon > 0,\ \mu((y-\epsilon, y+\epsilon)) > 0$.

This theorem extends Theorem 2.1 and Proposition 5.1 of [3] to completely monotone kernels $G$ and $K$. Its proof relies on the same arguments that we recall briefly in Appendix 2.9.1. An interesting consequence of (2.16) is the connection made between the price propagator and the decay kernel of the intensity. For general completely monotone functions (2.6), this yields in particular the following condition :

$$\forall \rho > 0, \quad (\iota_{\mathrm{s}} - \iota_{\mathrm{c}})\ \tilde{w}(\mathrm{d}\rho) \ = \ (1-\nu)\ \rho\ \tilde{\lambda}(\mathrm{d}\rho). \tag{2.17}$$

Thus, to exclude PMS, $\tilde{w}(\mathrm{d}\rho)$ has to be proportional to $\rho\ \tilde{\lambda}(\mathrm{d}\rho)$ and therefore the decay speed of $K$ should be higher than that of $G$, whatever their functional form (as soon as they are completely monotone). Besides, we can make the two following comments.

First, by dividing both sides of equation (2.17) by $\rho$, integrating on $(0, +\infty)$ and using Fubini's theorem, one gets the necessary (but not sufficient) martingale price condition

$$\begin{aligned} 1-\nu \ &= \ (\iota_{\mathrm{s}} - \iota_{\mathrm{c}}) \int_0^\infty \frac{\tilde{w}(\mathrm{d}\rho)}{\rho} = (\iota_{\mathrm{s}} - \iota_{\mathrm{c}}) \int_0^\infty \left( \int_0^\infty \exp(-\rho t)\ \mathrm{d}t \right) \tilde{w}(\mathrm{d}\rho) \\ &= (\iota_{\mathrm{s}} - \iota_{\mathrm{c}}) \int_0^\infty \left( \int_0^\infty \exp(-\rho t)\ \tilde{w}(\mathrm{d}\rho) \right) \mathrm{d}t \\ &= (\iota_{\mathrm{s}} - \iota_{\mathrm{c}}) \int_0^\infty K(t)\ \mathrm{d}t \ =: \ \mathrm{DBR}. \end{aligned} \tag{2.18}$$

This equation means that the proportion of transient impact should be equal to the *directional branching ratio*, which we define as

$$\mathrm{DBR} \ = \ (\iota_{\mathrm{s}} - \iota_{\mathrm{c}}) \int_0^\infty K(t)\ \mathrm{d}t \ = \ \frac{\iota_{\mathrm{s}} - \iota_{\mathrm{c}}}{\iota_{\mathrm{s}} + \iota_{\mathrm{c}}} \times \mathrm{BR}, \tag{2.19}$$

where BR is the usual branching ratio for Hawkes-based models that count positively price changes of both signs (see for instance Hardiman and Bouchaud [69]). This result is intuitive since the DBR represents the average number of « children trades of the same sign » for each trade, which, to obtain a diffusive price process, should be equal to the proportion of price impact that vanishes over time. Although it is only a necessary condition, equation (2.18) gives a quite general numerical criterion to assess empirically whether an observed price process is compatible with the martingale property, or rather persistent (DBR $> 1-\nu$) or mean-reverting (DBR $< 1-\nu$).

Second, the power-law kernels

$$G(u) = \nu + (1-\nu)(1 + c_G \times t)^{-a}, \quad K(u) = (1 + c_K \times t)^{-(1+\epsilon)}$$

are particular cases of (2.6), with

$$\tilde{\lambda}(\mathrm{d}\rho) = \frac{\rho^{a-1} \exp(-\rho/c_G)}{\Gamma(a)\ c_G^a}\ \mathrm{d}\rho, \quad \tilde{w}(\mathrm{d}\rho) = \frac{\rho^{\epsilon} \exp(-\rho/c_K)}{\Gamma(1+\epsilon)\ c_K^{1+\epsilon}}\ \mathrm{d}\rho.$$

Equation (2.17) then yields

$$a = \epsilon, \quad c_G = c_K = c, \quad \frac{\iota_{\mathrm{s}} - \iota_{\mathrm{c}}}{\epsilon c} = 1 - \nu.$$

Let us recall that if $K$ is a power-law, one must have $\epsilon > 0$ to obtain integrability, which is a necessary condition for the Hawkes process to be stationary. Also, in that case, the process can only have long-memory (i.e. non-integrable auto-covariance) if the Hawkes norm is equal to one [2] and if $\epsilon \in (0, 1/2)$, see Brémaud and Massoulié, Theorem 1 in [34]. In that case, the auto-covariance decays asymptotically as $t^{-(1-2\epsilon)}$. We thus reach exactly the same conclusion as Bouchaud et al. [31], who give the diffusive price condition $\beta = (1-\gamma)/2$, where $\gamma$ is the decay exponent of the auto-correlation of trade signs, and $\beta = a$ is the decay exponent of the propagator. Note that we used a totally different approach (absence of Price Manipulation Strategies), and that equation (2.17) is a possible generalization of their result to a wider class of kernels, within the Hawkes framework.

The calibration results presented in Section 2.4 allow us to confront real stock data to the martingale price condition obtained above. In particular, it is easy to check whether the proportion of transient impact $1-\nu = \sum \lambda_i$ is smaller, equal or greater than the directional branching ratio DBR. Although we do not expect the condition to be exactly satisfied in practice, we find it interesting to evaluate how much (and which way) real data deviate from the theoretical equilibrium.

### 2.2.3 The optimal execution strategy

In [3], we obtained an explicit characterization of the optimal execution strategy that minimzes $\mathbb{E}[C(X)]$ among strategies such that $X_0 \in \mathbb{R}$ and $X_{T+} = 0$ when $G(t) = e^{-\rho t}$ and $K(t) = e^{-\beta t}$. It is of interest to generalize this result to multi-exponential kernels (2.7). This is in principle possible. In fact, the model is still Markovian and Affine with respect to the state variable $(X_t, P_t, S_t, D_t^i, \kappa_t^{\pm(i)})$, and the cost is still quadratic. As in [3], one should first guess the quadratic form of the cost function, then derive necessary conditions on its coefficients, and last run a verification argument. However, we know from Alfonsi and Schied [4] that the optimal strategy without the flow of trades (i.e. $N \equiv 0$) is already quite involved and is characterized through a matrix Riccati equation. In our context, the system of ordinary differential equations that characterize the cost function would be much more intricated, and one would presumably have to solve it with numerical methods, which are less efficient than closed formulas for high-frequency trading. However, in the particular case where the propagator is kept exponential

$$G(u) = \nu + (1-\nu)\exp(-\rho u), \quad K(u) = \sum_{i=1}^{p} w_i \exp(-\beta_i u), \tag{2.20}$$

with $0 < \beta_1 < \cdots < \beta_p$ and $w_1, \ldots, w_p > 0$, it is still possible to derive explicitly the optimal execution strategy. In fact, we can handle the same arguments as in [3] and obtain the following result, proved in Appendix 2.9.2.

**Theorem 2.2.2.** *Let $\alpha_i = w_i(\iota_s - \iota_c)$ and $H$, the square matrix of order $p$ defined by*

$$1 \le i, j \le p, \ H_{i,j} = \mathbb{1}_{\{i=j\}}\beta_i - \alpha_j. \tag{2.21}$$

*We also define the two continuous matrix functions $\zeta, \omega$ by* [3]

$$\zeta(M) = \sum_{k\ge 0}(-1)^k \frac{M^k}{(k+1)!} \text{ and } \omega(M) = \sum_{k\ge 0}(-1)^k \frac{M^k}{(k+2)!}. \tag{2.22}$$

2. We refer to Hardiman et al. [68] for a test of this property on market data, and to Jaisson and Rosenbaum [79] for a study of Hawkes processes with a norm close to one.

3. When $M$ is invertible, $\zeta(M) = M^{-1}[I_p - \exp(-M)]$ and $\omega(M) = M^{-2}[\exp(-M) - I_p + M]$.

*Then, the strategy $X^*$ that minimizes the expected cost $\mathbb{E}[C(X)]$ satisfies a.s. and dt-a.e on $(0,T)$,*

$$(1-\nu)X_t^* = -\left[1+\rho(T-t)\right]D_t + \frac{m_1}{2\rho}\left[2+\rho(T-t)\right] \times \delta_t^\top \left\{ I_p + \frac{\rho(T-t)}{2+\rho(T-t)} \times \left[\zeta((T-t)H) + \nu\rho(T-t)\ \omega((T-t)H)\right] \right\} \cdot (1,\cdots,1)^\top, \tag{2.23}$$

*where $\delta_t^i = \kappa_t^{+(i)} - \kappa_t^{-(i)}$ for $i \in \{1,\cdots,p\}$ are intensity imbalances. Moreover, the optimal strategy is fully characterized by equation (2.23).*

Though restricted to (2.20), we believe that this extension of the result of [3] may be relevant for applications. In fact, on our dataset, there is not much gain to use the multi-exponential price propagator rather than the mono-exponential one, see Figure 2.1. Instead, for the decay kernel of the intensity, considering an exponential mixture allows to produce a richer variety of autocovariance functions, see Figure 2.3.

## 2.3 Calibration method

### 2.3.1 Description of the dataset

We consider tick-by-tick data provided by the French investment bank Natixis, to which we are grateful. The data contains all the changes in prices and volumes of the best bid and best ask, for two actively traded French stocks : BNP Paribas and Total.

The data is selected between 11a.m. and 1p.m., for every trading day between January and September 2012 and 2013. We exclude the three last months of the year, where activity decreases on average, along with the months where the tick size deviates from 0.005 euros. The two-hour window around noon is chosen to obtain a rather stable and uniform behavior of market activity, see e.g. Lehalle and Laruelle [85], p. 112. This way, for each stock separately and with minimal data treatment, we can reasonably assume that each two-hour window of trading is a realization of the same random price process.

In the initial dataset, for each stock separately, each line corresponds either to an update in price and/or quantity at one of the best queues (triggered by a market event such as a market order, a limit order or a cancellation), or a new trade executed for a given volume at a given price. The time stamps for these updates are precise to the millisecond. We reduce this data by aggregating the events happening on the same millisecond : we only keep track of the best prices at the beginning and at the end of each time stamp, which yields the aggregated price impact of the events that happened « simultaneously », i.e. on the same millisecond. Similarly, we sum all the volumes that were executed on the same time stamp. We obtain a simplified sequence of market events, among which a minority is associated to a traded volume and/or to a price change.

A correspondence should be clarified between the theoretical items of the models of [3] and Section 2.2, and actual financial data. Different possibilities may be relevant, but our choices are the following :

– We define the « market price » $P_t$ as the midpoint price, i.e. the average of the best bid price and the best ask price at any time $t$.
– We only consider time stamps where the midpoint price jumps. In other words, we ignore the trades and cancellations that do not empty either the best bid or the best ask, as well as the passive limit orders that do not define a new best price. For the stocks that we consider, this gives an average latency of one to four seconds between two consecutive time stamps. This is in agreement with the time scale that is thought of in the theoretical model of [3], which is not of ultra-high frequency.
– We express the time in hours, and note $T = 2$ the length of the window that we consider for each trading day. Throughout the paper, we note $\tau \in (0, T)$ the time stamps which correspond to midpoint jumps triggered by trades, i.e. by limit orders that cross the spread or by market orders. These correspond to the jumps of the process $N$ of the theoretical model : they are marked by both a price jump $\Delta M_\tau$ (of one or several half-ticks), and an executed volume $\Delta V_\tau > 0$ expressed in number of shares. The time stamps of other price jumps are noted $\theta \in (0, T)$. They are triggered by cancellations and passive limit orders, with no executed volume, and they are assumed to enforce on average the deterministic resilience effect as in [31]. Between two trades, the deviation of the price from this deterministic average is considered as a noise process, modeled using an arithmetic Brownian motion.

Some key statistics for these items are given in Table 2.1 for BNP Paribas and Total.

| Stock | BNP Paribas | | Total | |
|---|---|---|---|---|
| Year | 2012 | 2013 | 2012 | 2013 |
| Average midprice | 32.4 | 44.9 | 38.2 | 39.0 |
| Tick size | 0.005 | 0.005 | 0.005 | 0.005 |
| Number of mid. changes per hour | 1909 | 1699 | 1209 | 929 |
| Proportion due to transactions | 10.0% | 7.9% | 7.6% | 6.9% |
| $m_1$ | 776 | 636 | 978 | 963 |
| $m_2/m_1^2$ | 3.38 | 4.69 | 4.30 | 6.72 |
| Average size of the first queue | 1398 | 1136 | 1710 | 1779 |

TABLE 2.1 – Table of statistics for the stocks BNP Paribas and Total on the periods January-September 2012-2013, between 11 a.m. and 1 p.m. January 2012 is excluded for BNP Paribas because the tick size dropped below 0.005. We give the proportion of midpoint changes which are triggered by trades, the remaining proportion being triggered by cancellations or passive limit orders. $m_1$ is the average volume of transactions that trigger price moves, and $m_2$ is the average squared volume for these transactions. The greater the ratio $m_2/m_1^2$, the more variance in the distribution of traded volumes.

### 2.3.2 Overview of the calibration process

One specificity of the price model given by equation (2.4) is that it is composed of two separate components :
– The point process $N$ for the trades that trigger the price moves, for the which time stamps $\tau$ and the marks (the price jumps $\Delta M_\tau$ and the executed volumes $\Delta V_\tau$) are modeled and estimated jointly,

– The propagator model, which conditionally to the midpoint jumps due to trades, is a continuous-time linear regression model with a Gaussian noise process $\sigma W_t$.

Therefore, the trades are modeled using marked Hawkes processes, and conditionally to them, the price is Gaussian. This segmentation has at least three advantages. First, the calibration process is simpler since the two parts can be estimated independently, which significantly reduces the dimension of the problem. Second, the estimation results on each side are robust to the choices made in the other. For instance, if one wants to modify the Hawkes modeling for the trades, then our results for the propagator are still valid, and vice versa. Eventually, the results of Section 2.2.2 include some theoretical links between the Hawkes parameters and the propagator, and it seems more rigorous to confront these links to our calibration results when the two parts are estimated independently.

Our calibration protocol as a whole being somewhat sophisticated, we test its validity and robustness by running it on simulated data. In Sections 2.4 and 2.5, we give the results of our analysis for these simulations as well as for real financial data.

### 2.3.3 Estimation of the propagator

#### Framework

In this section we explain how the propagator model introduced in Section 2.2 can be adapted for practical applications, in particular for its calibration. This requires to consider the two following points :

– In practice, the price impact of transactions is not proportional to their volumes. It is typically of a few ticks, while the volumes span a wider range of values. Therefore, one must choose between « price resilience » and « volume resilience » as in Alfonsi et al. [5]. The first choice corresponds to modeling the mean-reversion property of market prices, the second describes how liquidity « regenerates » after a trade, and the two are only equivalent for linear price impact.
– The evolution of the price between two transactions is very noisy, and the propagator model only explains a part of its variance. Therefore, we need to control the variance of the estimation to obtain satisfying calibration results.

For the first point, we choose to model price resilience, which is easier to measure in practice and has been considered more often in the literature. This boils down to replacing $\Delta N_\tau/q$ by the midprice jumps $\Delta M_\tau$ in equation (2.4). For the second point, an intuitive possibility consists in restraining the propagator regression to a finite time window $\Delta_{\mathrm{RW}} > 0$, and to assume that the model predicts the price increment $P_t - P_{t-\Delta_{\mathrm{RW}}}$ for $t \geq \Delta_{\mathrm{RW}}$ instead of $P_t - P_0$. If the noise is an additive Brownian term $\sigma W_t$, this fixes the variance of the predicted variables to $\sigma^2 \Delta_{\mathrm{RW}}$ instead of $\sigma^2 t$, $t \in [\Delta_{\mathrm{RW}}, T]$. We obtain the modified price model

$$P_t \;=\; P_{t-\Delta_{\mathrm{RW}}} \;+ \sum_{t-\Delta_{\mathrm{RW}}<\tau\leq t} \Delta M_\tau \; G(t-\tau) \;+\; \sigma(W_t - W_{t-\Delta_{\mathrm{RW}}}). \tag{2.24}$$

Of course, $\Delta_{\mathrm{RW}}$ must be such that $G(\Delta_{\mathrm{RW}}) - G(\infty)$ is small compared to $G(\infty)$ for the model to be a meaningful approximation of the original model (2.4), see Remark 2.2.1. This condition also allows to avoid bias in the estimation of the propagator $G$. We fix $\Delta_{\mathrm{RW}} = 0.5$ hours (30 minutes) throughout the sequel of this paper, basing ourselves on preliminary observations that we do not detail here. Note that within the range $\Delta_{\mathrm{RW}} \in [0.1, 1]$, the choice of this parameter has little impact on the results. One can verify *a posteriori* that our estimations of $G$ are compatible with $G(0.5) - G(\infty) \ll G(\infty)$.

The predicted price increment between $t - \Delta_{\mathrm{RW}}$ and $t$ is given by

$$\hat{P}_t - P_{t-\Delta_{\mathrm{RW}}} = \sum_{t-\Delta_{\mathrm{RW}} \leq \tau \leq t} \Delta M_\tau \; G(t-\tau) \tag{2.25}$$

where $P_{t-\Delta_{\mathrm{RW}}}$ is the real midpoint price at time $t - \Delta_{\mathrm{RW}}$, taken directly from the data. Equation (2.24) becomes

$$P_t = \hat{P}_t + \sigma(W_t - W_{t-\Delta_{\mathrm{RW}}}). \tag{2.26}$$

Conditionally to $P_{t-\Delta_{\mathrm{RW}}}$ and to the process $M$, one has $P_t \sim \mathcal{N}(\hat{P}_t, \sigma^2 \Delta_{\mathrm{RW}})$. In this setting, the Maximum Likelihood Estimator of $G$ is equivalent to the Least Squares Estimator. We thus minimize numerically on the parameters of $G$ the quadratic error

$$\mathcal{E}(G) = \sum_{\Delta_{\mathrm{RW}} < \theta < T} [\hat{P}_\theta(G) - P_\theta]^2, \tag{2.27}$$

where the $\theta$'s are the occurrences of price jumps due to cancellations or passive limit orders. To get a better understanding of the shape of the propagator, we first estimate $G$ in an « unconstrained » manner, i.e. as the linear interpolation of a discrete set of points. Thus, we model $G$ as

$$G(t) = g_l \mathbb{1}_{[t_l, \Delta_{\mathrm{RW}}[}(t) + \sum_{i=0}^{l-1} \frac{(t_{i+1} - t) g_i + (t - t_i) g_{i+1}}{t_{i+1} - t_i} \; \mathbb{1}_{[t_i, t_{i+1}[}(t),$$

where $t_1, \cdots, t_l$ are fixed a priori and $(g_1, \cdots, g_l)$ is the parameter to estimate. We see that the resulting curve, which is given is Section 2.4 for stock data, has an increasing short-range part, and switches to a decreasing mode after a few seconds. One has $G(0) = 1$, but $G$ reaches a point above unity before it enters its decreasing regime. Let us recall that in an idealized model without bid-ask spread, Alfonsi et al. [6] and Gatheral et al. [64] show that $G$ has to be decreasing and convex around zero to exclude PMS and some market instability. This is not the case on our dataset. We interpret this as the fact that after a trade, the new bid-ask is generally formed around the impacted price. Thus, during a few seconds, limit orders and cancellations tend to impact the midprice in the same direction as the trade. This motivates us to distinguish the propagator $G(t)$ and the functional form of its long-range decay that we call the resilience, noted $R(t)$. This way, we can allow $R(0) \geq 1$ and impose that $R$ is decreasing. One can then link $G$ and $R$ with a simple linear interpolation between $t = 0$ and $t = L_{\mathrm{adj}}$, with $L_{\mathrm{adj}} > 0$ the « adjustment lag » of the market

$$G(t) = \left[1 + (R(L_{\mathrm{adj}}) - 1) \; \frac{t}{L_{\mathrm{adj}}}\right] \mathbb{1}_{\{t \leq L_{\mathrm{adj}}\}} + R(t) \mathbb{1}_{\{t > L_{\mathrm{adj}}\}}.$$

This choice has the merit that once $L_{\mathrm{adj}}$ is fixed, only the resilience curve needs to be estimated since $G$ is characterized by $R$. Therefore, to estimate $R$ with an imposed decreasing functional form, we place ourselves in the following version of the price model

$$P_t = P_{t-\Delta_{\mathrm{RW}}} + \sum_{t-\Delta_{\mathrm{RW}} \leq \tau < t - L_{\mathrm{adj}}} \Delta M_\tau \; R(t-\tau) + \sum_{t - L_{\mathrm{adj}} \leq \tau \leq t} \Delta M_\tau \left[1 + (R(L_{\mathrm{adj}}) - 1) \; \frac{t-\tau}{L_{\mathrm{adj}}}\right] + \sigma(W_t - W_{t-\Delta_{\mathrm{RW}}}).$$

We consider two types of parameterization for the resilience $R(t)$ :

– The mono-exponential curve

$$R(t) = \gamma\ [1 - \lambda(1 - \exp(-\rho t))], \tag{2.28}$$

with three parameters $\gamma, \rho > 0, \lambda \in [0, 1]$. $\gamma$ is an amplification factor, $\rho$ is the resilience speed of the market, $\lambda$ is the transient part of the price impact of trades, and $\nu = 1 - \lambda$ is the permanent part. The mono-exponential curve is the type of resilience considered in the theoretical model of [3].

– The multi-exponential curve

$$R(t) = \gamma\ \left[1 - \sum_{i=1}^{p} \lambda_i (1 - \exp(-\rho_i t))\right], \tag{2.29}$$

is a generalization of the previous one, with $2p+1$ parameters $\gamma, \rho_1, \cdots, \rho_p > 0, \lambda_1, \cdots, \lambda_p \in [0, 1]$, $\sum_i \lambda_i \leq 1$. For $1 \leq i \leq p$, $\lambda_i$ is the proportion of transient impact that decays at speed $\rho_i$, and $\nu = 1 - \sum_i \lambda_i$ is the proportion of permanent impact.

For both parameterizations, we estimate a posteriori the volatility $\sigma$ of the Brownian noise with

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \left[P_T^i - P_0^i - \sum_{0<\tau<T} \Delta M_\tau^i\ G(T-\tau)\right]^2}{n \times T}}, \tag{2.30}$$

where $n$ is the number of days of the sample, and for $i \in \{1, \cdots, n\}$, $P^i$ and $M^i$ are respectively the real price and the midprice jump process for day $i$, and the $\tau$'s are the jump times of $M^i$. Also, since the prediction model defined by (2.25) and (2.26) can be seen as a continuous-time linear regression, where the explained variables are the price increments $P_t - P_{t-\Delta_{\text{RW}}}$ and the regressors are the past price jumps $\Delta M_\tau$ triggered by trades, we can evaluate its quality of fit using a usual analysis of variance. We define the $r^2$ value as

$$r^2 \ = \ 1 \ - \ \frac{\sum_{i=1}^{n} \sum_{\Delta_{\text{RW}}<\theta<T} [\hat{P}_\theta^i - P_\theta^i]^2}{\sum_{i=1}^{n} \sum_{\Delta_{\text{RW}}<\theta<T} \left[P_\theta^i - P_{\theta-\Delta_{\text{RW}}}^i - \overline{\Delta P}\right]^2}, \tag{2.31}$$

where for $i \in \{1, \cdots, n\}$, $\hat{P}^i$ is the predicted price process for day $i$, and

$$\overline{\Delta P} = \frac{1}{\sum_{i=1}^{n} \#\theta_i} \sum_{i=1}^{n} \sum_{\Delta_{\text{RW}}<\theta<T} \left(P_\theta^i - P_{\theta-\Delta_{\text{RW}}}^i\right)$$

is the average price move between $\theta - \Delta_{\text{RW}}$ and $\theta$, where the $\theta$'s are the times of price changes with no executed volumes and and $\#\theta_i$ is the number of such price changes on day $i$. Note that since there is no constant in the regression model (2.26), the $r^2$ could theoretically be negative, but this is not the case in practice. The $r^2$ constitutes a useful comparison criterion between different estimated propagators, and we use it in Section 2.4.

Now that the global practical framework is set, the estimation protocol for $G$ needs to be detailed. This is the object of the following section.

### Estimation protocol

We use a multi-step estimation protocol, that mainly resorts to the minimization of the quadratic error $\mathcal{E}$ defined in (2.27). When $G(t)$ is linear with respect to its parameters, $\mathcal{E}$ is quadratic and one step of Newton-Raphson's algorithm is enough to find the minimum (see Appendix 2.7). When the dependency in the parameters is non-linear, we first use grid minimizations to find a suitable starting point for the algorithm.

As a first step, we estimate the « unconstrained » propagator curve. Then, we estimate the resilience curve using the two parameterizations presented in Section 2.3.3.

#### Estimation of the unconstrained propagator curve

We first estimate $G$ by the linear interpolation $\hat{G}$

$$\hat{G}(t) = g_l \mathbb{1}_{[t_l,T[}(t) + \sum_{i=0}^{l-1} \frac{(t_{i+1}-t)g_i + (t-t_i)g_{i+1}}{t_{i+1}-t_i} \mathbb{1}_{[t_i,t_{i+1}[}(t).$$

For $t_1, \cdots, t_l$ fixed a priori, $\hat{G}$ is linear with respect to $(g_1, \cdots, g_l)$. Thus, one step of Newton-Raphson's method (see Appendix 2.7.1) determines the parameters that minimize the quadratic error $\mathcal{E}(\hat{G})$. To approximate the long-range propagator, we choose a uniform grid $t_i = i/l$ with $l = 20$ on the interval $[0, 0.2]$. On the other hand, for a zoom on the beginning of the curve, we concentrate the $t_i$'s near zero.

#### Estimation of the multi-exponential resilience curve

The simultaneous estimation of multiple $\rho_i$'s being too unstable, we choose to fix four components associated to four simple characteristic time scales (the $\rho_i$'s are expressed in inverse hours) : $\rho_1 = 6$ (ten minutes), $\rho_2 = 60$ (one minute), $\rho_3 = 120$ (thirty seconds) and $\rho_4 = 360$ (ten seconds). We then assume that the vector $(\rho_1, \cdots, \rho_4)$ is rich enough to represent all the relevant time scales in our framework, and we focus on the weights $\lambda_1, \cdots, \lambda_4$ associated to each scale to characterize the decay of the curve. The multi-exponential resilience given by equation (2.29) becomes

$$R(t) = \overline{\nu} + \sum_{i=1}^{4} \overline{\lambda}_i \exp(-\rho_i t),$$

where we re-parameterize $\overline{\nu} = \gamma(1 - \sum_{i=1}^{4} \lambda_i) > 0$ and $\overline{\lambda}_i = \gamma\lambda_i > 0$. Reciprocally, one has $\gamma = \overline{\nu} + \sum_{i=1}^{4} \overline{\lambda}_i$ and $\lambda_i = \overline{\lambda}_i/\gamma$. Since the $\rho_i$'s are fixed, the resilience curve $R(t)$ is linear w.r.t. the parameter $(\overline{\nu}, \overline{\lambda}_1, \cdots, \overline{\lambda}_4)$ that remains to be estimated, thus Newton-Raphson's algorithm (see Appendix 2.7.2) converges with a single iteration. We then select the significant $\rho_i$'s as follows :

1. A first estimation yields a « full » parameter $(\overline{\nu}, \overline{\lambda}_1, \cdots, \overline{\lambda}_4)$. Some of the resulting $\overline{\lambda}_i$'s may be non-positive, which is incompatible with the model.
2. While there exists $i$ such that $\overline{\lambda}_i \leq 0$, we remove the $\rho_i$ corresponding to the minimal $\overline{\lambda}_i$, and we launch the algorithm again with one less parameter.
3. Eventually, we have selected one to four « significant » $\rho_i$'s, of associated weights $\overline{\lambda}_i$'s that are positive, and the estimation process is complete.

Since each of these steps only take one iteration of Newton-Raphson's algorithm, the whole estimation protocol for the multi-exponential curve is quite fast. Therefore, in order to estimate the market adjustment lag $L_{\text{adj}}$, we can conduct the estimation several times for $L_{\text{adj}}$ on some discrete grid, and compare the regression $r^2$'s as defined by (2.31). The result associated to the maximal $r^2$ gives the parameters $\gamma_{\text{multi}}$, $\lambda_{\text{multi}}$ and $\rho_{\text{multi}}$ for the multi-exponential resilience, along with the adjustment lag $L_{\text{adj}}$.

**Estimation of the mono-exponential resilience curve**

The multi-exponential estimation presented above serves as a starting point for the following. The market adjustment lag $L_{\text{adj}}$ is already estimated, along with the associated set of parameters $\gamma_{\text{multi}}, \lambda_{\text{multi}}, \rho_{\text{multi}}$ for the multi-exponential resilience curve. We set

$$\gamma = \gamma_{\text{multi}}, \quad \lambda = \sum_i \lambda^i_{\text{multi}}, \quad \rho = \sum_i \frac{\lambda^i_{\text{multi}}}{\lambda} \rho^i_{\text{multi}} \tag{2.32}$$

as a starting parameter for the mono-exponential estimation. As in the multi-exponential case, we re-parameterize (2.28) as

$$R(t) = \overline{\nu} + \overline{\lambda} \exp(-\rho t),$$

with $\overline{\nu} = \gamma(1-\lambda) > 0$ and $\overline{\lambda} = \gamma\lambda > 0$. We then proceed as follows

1. We use Newton-Raphson's algorithm to minimize the quadratic error on the whole parameter $(\overline{\nu}, \overline{\lambda}, \rho)$ (see Appendix 2.7.2 for $p = 1$ exponential component). If the starting point is convex and the algorithm converges to a satisfying level, we proceed directly to Step 6. Else, we go to Step 2.
2. Keeping $\rho$ fixed to its starting value (2.32), the dependency of $R(t)$ on $\overline{\nu}$ and $\overline{\lambda}$ is linear. Thus, with one step of Newton-Raphson's algorithm, we get the optimal values of $\overline{\nu}$ and $\overline{\lambda}$ for the current value of $\rho$.
3. For $\gamma = \overline{\nu} + \overline{\lambda}$ fixed by Step 2, $\lambda$ initialized to $\overline{\lambda}/\gamma$ and $\rho$ as in (2.32), we minimize the quadratic error $(\lambda, \rho) \mapsto \mathcal{E}(\lambda, \rho)$ on a local two-dimensional grid in the vicinity of the starting point.
4. The pair that reaches the minimum of the error grid at Step 3 is again used as a starting point to Newton-Raphson's algorithm, to determine the optimal $(\lambda, \rho)$ for the current fixed value of $\gamma$, using the « unit » mono-exponential parameterization of Appendix 2.7.2. We actualize $(\lambda, \rho)$ to this optimum, along with $\overline{\nu} = \gamma(1-\lambda)$ and $\overline{\lambda} = \gamma\lambda$.
5. The parameter $(\overline{\nu}, \overline{\lambda}, \rho)$ is now in a region where the quadratic error is more likely to be convex. Therefore, we use this new starting point for an error minimization using Newton-Raphson's algorithm on the whole parameter.
6. We obtain the parameter $\gamma_{\text{mono}}, \lambda_{\text{mono}}, \rho_{\text{mono}}$ for the mono-exponential resilience curve.

The above estimation protocol for the mono-exponential resilience curve may seem complicated : in particular, it is more subtle than the multi-exponential estimation. The reason for this is that we want here to determine the most significant characteristic time scale of the resilience through the parameter $\rho$. The dependency of the quadratic error $\mathcal{E}$ on this parameter being non-linear, nothing guarantees a priori that Newton-Raphson's algorithm (or more simply a gradient algorithm) has a convex starting point, which is a necessary condition to ensure its convergence. Hence we have to proceed more carefully and introduce several intermediary steps.

### 2.3.4 Estimation of the Hawkes parameters

#### Framework

Independently of the propagator, we also estimate the parameters of the Hawkes-based model presented in Section 2.2 for the price jumps due to transactions. We choose the self-excitation functions $\varphi_{\mathrm{s}}$ and $\varphi_{\mathrm{c}}$ to be affine, i.e.

$$\varphi_{\mathrm{s}}(x) = \phi^0_{\mathrm{s}} + \phi^1_{\mathrm{s}} x \quad , \quad \varphi_{\mathrm{c}}(x) = \phi^0_{\mathrm{c}} + \phi^1_{\mathrm{c}} x. \tag{2.33}$$

In the standard Hawkes framework, self-excitation in the order flow is not marked, i.e. only the constant terms $\phi^0_{\mathrm{s}}, \phi^0_{\mathrm{c}}$ appear in $\varphi_{\mathrm{s}}$ and $\varphi_{\mathrm{c}}$. In spite of its simplicity, the affine structure allows us to underline the deviation from the standard Hawkes benchmark, and to detect an increasing part in the self-excitation function.

As pointed out in Section 2.3.3, there are two possible interpretations for the marks associated to the jumps of $N$. Since each of these jumps corresponds to a price jump due to a transaction, they are all associated to two positive variables : the price impact on the one hand, and the traded volume on the other hand. Therefore, we estimate three sets of parameters for different versions of the Hawkes model (unit marks, volume marks, and price marks), each with a different practical interpretation of the intensity jump terms. Precisely, we replace $\varphi_{\mathrm{s/c}}(\mathrm{d}N^{\pm}_t)$ in (2.9) and (2.10) at the jump times $t$ by either of the three possibilities

$$\phi^0_{\mathrm{s/c,unit}}, \ \phi^0_{\mathrm{s/c,vol.}} + \phi^1_{\mathrm{s/c,vol.}}|\Delta V_t|/m_1, \ \phi^0_{\mathrm{s/c,price}} + \phi^1_{\mathrm{s/c,price}}|\Delta M_t|/\overline{m}, \tag{2.34}$$

where $m_1$ is the average executed volume and $\overline{m}$ is the average price impact.

#### Estimation protocol

Our estimation protocol for the Hawkes part of the model is then as follows : we first estimate the mono-exponential Hawkes model $K(u) = \exp(-\beta u)$, which allows us to estimate the Hawkes norm and its repartition in terms of self and cross-excitation, and to select the optimal mark type for the jumps. Then we estimate the multi-exponential Hawkes model $K(u) = \sum_{i=1}^{p} w_i \exp(-\beta_i u)$ with the $\beta_i$'s fixed a priori.

#### Mono-exponential kernel

Let us consider the mono-exponential Hawkes model of equation (2.2), for which the Hawkes decay kernel is $K(u) = \exp(-\beta u)$, $\beta > 0$. We first focus on the parameters of the total intensity $\Sigma_t = \kappa^+_t + \kappa^-_t$ by aggregating all the price jumps due to trades, regardless of their signs. In the mono-exponential case, one has

$$\mathrm{d}\Sigma_t = \ -\beta \ (\Sigma_t - 2\kappa_\infty) \ \mathrm{d}t \ + \ \iota \ \mathrm{d}J_t,$$

where $\iota$ is the average excitation, so that the jumps of $J$ have an average of one. We use a Generalized Method of Moments (GMM) to estimate $\beta, \kappa_\infty$ and $\iota$. We divide the time window $[0, T]$ of length $T = 2$ hours in 720 bins of length $h = 1/360$ (i.e. ten seconds). Then, we compute the number $\Delta\tilde{J}^i_l$ of price jumps due to trades in the time bin $[(l-1)h, lh]$, $l \in \{1, \cdots, \lfloor T/h \rfloor\}$ on day $i \in \{1, \cdots, n\}$, for each time bin and each day. If $l$ is the row index and $i$ is the column index, we obtain a $\lfloor T/h \rfloor \times n$

matrix of which the entries are the positive numbers $\Delta\tilde{J}_l^i$. We normalize this dataset by dividing each column by its mean value and multiplying the whole matrix by the original global mean value, so that the global mean is unchanged and each column has the same mean. We first compute the empirical mean $\overline{\Delta\tilde{J}}$ and variance $\mathcal{V}$ of the discrete process $\Delta\tilde{J}$

$$\overline{\Delta\tilde{J}} = \frac{1}{n \times \lfloor T/h \rfloor} \sum_{i=1}^{n} \sum_{l=1}^{\lfloor T/h \rfloor} \Delta\tilde{J}_l^i, \quad \mathcal{V} = \frac{1}{n \times \lfloor T/h \rfloor - 1} \sum_{i=1}^{n} \sum_{l=1}^{\lfloor T/h \rfloor} \left[\Delta\tilde{J}_l^i - \overline{\Delta\tilde{J}}\right]^2.$$

The average jump intensity $2\overline{\kappa}$ of the total jump process is obtained with the formula $2\overline{\kappa} = \overline{\Delta\tilde{J}}/h$. Besides, the empirical auto-correlation function of $\Delta\tilde{J}$ is given by

$$\forall k \in \{1, \cdots, k_{\max}\}, \quad \widehat{\mathscr{C}}(k) = \frac{1}{\mathcal{V}} \times \left\{ \frac{1}{n \times (\lfloor T/h \rfloor - k)} \left[ \sum_{i=1}^{n} \sum_{l=k+1}^{\lfloor T/h \rfloor} \Delta\tilde{J}_l^i \, \Delta\tilde{J}_{l-k}^i \right] - \overline{\Delta\tilde{J}}^2 \right\}, \quad (2.35)$$

where $k_{\max} = 36$ is the maximum lag (so that the maximum range $k_{\max} h = 0.1$ equals six minutes). Using the results of Da Fonseca and Zaatour [45] for mono-exponential Hawkes processes, we have that $\widehat{\mathscr{C}}(k)$ decays as $\exp(-(\beta - \iota)k)$. Therefore, the exponential fit of the empirical curve $\widehat{\mathscr{C}}(k)$ yields an estimate of $d := \beta - \iota$. Then, we also get from [45]

$$\mathcal{V} = 2\overline{\kappa} \left\{ \frac{\beta^2 h}{d^2} + \left(1 - \frac{\beta^2}{d^2}\right) \frac{1 - \exp(-dh)}{d} \right\}.$$

This relation can be inverted to obtain an estimate for $\beta$ : if we note $z_h = (1 - \exp(-dh))/d$, we get

$$\beta \;=\; d \sqrt{\frac{\mathcal{V}/(2\overline{\kappa}) - z_h}{h - z_h}}.$$

Then, $\iota = \beta - d$ and $\kappa_\infty = (1 - \iota/\beta)\ \overline{\kappa}$ can be deduced from the above equation. We also obtain the mono-exponential branching ratio

$$\mathrm{BR}_{\mathrm{mono}} = \iota/\beta.$$

Keeping $\beta, \iota$ and $\kappa_\infty$ fixed to these GMM estimates, we now turn to the bi-dimensional intensity model (2.2). We use Maximum Likelihood Estimation (see Appendix 2.8) on one-dimensional grids to determine the self and cross-excitation parameters :

1. We determine the proportion $u \in [0, 1]$ such that $\iota_{\mathrm{s}} = u\,\iota,\ \iota_{\mathrm{c}} = (1-u)\,\iota$ maximize the likelihood of the two-dimensional intensity $(\kappa^+, \kappa^-)$, where $\iota_{\mathrm{s}}$ and $\iota_{\mathrm{c}}$ are respectively the average self-excitation and cross-excitation parameters.
2. For volume marks and price marks separately, we determine the proportion $u_{\mathrm{s}} \in [0, 1]$ such that $\phi_{\mathrm{s}}^0 = u_{\mathrm{s}}\ \iota_{\mathrm{s}},\ \phi_{\mathrm{s}}^1 = (1 - u_{\mathrm{s}})\ \iota_{\mathrm{s}}$ maximize the likelihood of $(\kappa^+, \kappa^-)$, where $\phi_{\mathrm{s}}^0, \phi_{\mathrm{s}}^1$ are defined in equation (2.33). Similarly, we determine the optimal proportion $u_{\mathrm{c}}$ for $\phi_{\mathrm{c}}^0 = u_{\mathrm{c}}\ \iota_{\mathrm{c}},\ \phi_{\mathrm{c}}^1 = (1 - u_{\mathrm{c}})\ \iota_{\mathrm{c}}$. For $\iota_{\mathrm{s}}$ and $\iota_{\mathrm{c}}$ fixed, we obtain the optimal constant and linear parts for self and cross-excitation, for the two possible types of marks.
3. The likelihoods obtained for the three models are then compared to determine which of the unit / volumes / price marks yield the best model.

Eventually, we obtain estimates for all the parameters $\beta_{\text{mono}}, \kappa_{\infty\,\text{mono}}, \phi^0_{\text{s}\,\text{mono}}, \phi^1_{\text{s}\,\text{mono}}, \phi^0_{\text{c}\,\text{mono}}, \phi^1_{\text{c}\,\text{mono}}$ of the mono-exponential Hawkes model, along with the optimal type of marks.

**Multi-exponential kernel**

We turn to the multi-exponential Hawkes model $K(u) = \sum_{i=1}^{p} w_i \exp(-\beta_i u)$. As in the case of the estimation of the multi-exponential resilience in Section 2.3.3, we fix four $\beta_i$'s associated to four simple characteristic time scales. In fact, we choose the same time scales as for the resilience : $\beta_1 = 6, \beta_2 = 60, \beta_3 = 120$ and $\beta_4 = 360$. We then calibrate the $w_i$'s associated to each $\beta_i$, and these weights tune the shape of the Hawkes kernel.

The results of the mono-exponential estimation are used to select the type of marks (unit, volume or price) and to get a starting point for $\kappa_\infty, \phi^0_{\text{s}}, \phi^1_{\text{s}}, \phi^0_{\text{c}}, \phi^1_{\text{c}}$ and the branching ratio BR. The starting point for the $w_i$'s is chosen to be uniformly distributed

$$w_i = \frac{\text{BR}_{\text{mono}}}{\sum_{i=1}^{4} \frac{1/4}{\beta_i}} \times \frac{1}{4},$$

with a scaling that matches the initial branching ratio. Then, we maximize the likelihood of the model on the parameter $(\kappa_\infty, w_1, w_2, w_3, w_4)$ using Newton-Raphson's algorithm, as explained in Appendix 2.8. We use the same selection method as for the multi-exponential resilience estimation of Section 2.3.3 : if at least one of the $w_i$'s is non-positive, we delete the $\beta_i$ associated to the minimal $w_i$ and launch the algorithm again, with one less parameter. Finally, we multiply $(\phi^0_{\text{s}}, \phi^1_{\text{s}}, \phi^0_{\text{c}}, \phi^1_{\text{c}})$ by the sum of the remaining $w_i$'s, and scale the latter to one. Without changing the overall model, this imposes $K(0) = 1$ for the Hawkes decay kernel $K$. We obtain the parameters $\beta_{\text{multi}}, w_{\text{multi}}, \kappa_{\infty\,\text{multi}}, \phi^0_{\text{s}\,\text{multi}}, \phi^1_{\text{s}\,\text{multi}}, \phi^0_{\text{c}\,\text{multi}}, \phi^1_{\text{c}\,\text{multi}}$ for the multi-exponential Hawkes model.

## 2.4 Calibration results

### 2.4.1 Description of the results

This section is dedicated to the presentation of our calibration results. The calibration method of Section 2.3 is first applied to simulated data to test its validity, and then to actual financial data from French stocks. We also provide some qualitative comments. For each simulated dataset and each stock, the results are summarized in tables, plus a few graphs for BNP Paribas. The content of the tables is explained below.

**Adjustment lag table :** This table gives the regressions $r^2$'s of the multi-exponential resilience curve, for several values of the market adjustment lag $L_{\text{adj}}$. It is used to select the optimal value of $L_{\text{adj}}$ on a discrete grid.

**Resilience table :** The resilience table gives the estimation results for the propagator. We give the selected adjustment lag $L_{\text{adj}}$ and the estimated parameters for the two types of resilience curve

$$R_{\text{mono}}(t) = \gamma_{\text{mono}}\ [1 - \lambda_{\text{mono}}(1 - \exp(-\rho_{\text{mono}} t))],$$
$$R_{\text{multi}}(t) = \gamma_{\text{multi}}\ \left[1 - \sum \lambda^j_{\text{multi}}(1 - \exp(-\rho^j_{\text{multi}} t))\right],$$

along with the estimated volatility $\sigma$ of the noise and the regression $r^2$, defined respectively by equations (2.30) and (2.31).

**Marks table :** In this table, we give the maximized log-likelihoods per point $\mathcal{L}_{\text{unit}}$, $\mathcal{L}_{\text{vol.}}$ and $\mathcal{L}_{\text{price}}$ for each type of mark (unit, volumes and price jumps), in the mono-exponential Hawkes model. It serves as a selection criterion for the optimal type of mark.

**Intensity table :** This table gives the estimated parameters for the Hawkes model described in Section 2.2.1, for both the mono-exponentiel decay kernel $K(u) = \exp(-\beta u)$ and the multi-exponential one $K(u) = \sum_{i=1}^{p} w_i \exp(-\beta_i u)$. We also give the maximized log-likelihoods per point $\mathcal{L}_{\text{mono}}$ and $\mathcal{L}_{\text{multi}}$, which can be compared to one another or between datasets to quantify the quality of fit of the Hawkes model. Eventually, we give the branching ratio BR and the directional branching ratio DBR defined by equation (2.19), that are obtained with the multi-exponential parameterization.

### 2.4.2 Simulated data

We first give in Tables 2.2 and 2.3 the results of our calibration protocol on two datasets simulated with the price model (2.4). In each table, the first column gives the « real » simulation parameters and the second gives the estimated ones. Both datasets are composed of 150 independent realizations of the price process on two-hour windows, and we choose simulation parameters close to what is found further for stock data in order to obtain relevant benchmarks. Note that Simulation 1 features a non-zero Brownian volatility, whereas Simulation 2 is generated by the « pure » propagator model without noise.

| Year | Simu. | Calib. |
|---|---|---|
| $L_{\text{adj}}$ (sec) | 4 | 4 |
| $\gamma_{\text{multi}}$ | 2.70 | 2.35 |
| $\rho_{\text{multi}}$ | 60/360 | 6/60/360 |
| $\lambda_{\text{multi}}$ | 0.50/0.10 | 0.13/0.35/0.11 |
| $\nu_{\text{multi}}$ | 0.40 | 0.41 |
| $\sigma_{\text{multi}}$ | 0.1000 | 0.1917 |
| $r^2_{\text{multi}}$ | – | 9.554% |
| $\gamma_{\text{mono}}$ | – | 2.38 |
| $\rho_{\text{mono}}$ | – | 68.2 |
| $\lambda_{\text{mono}}$ | – | 0.55 |
| $\sigma_{\text{mono}}$ | – | 0.1923 |
| $r^2_{\text{mono}}$ | – | 9.519% |

| Year | Simu. | Calib. |
|---|---|---|
| Marks type | Volume | Volume |
| $\beta_{\text{multi}}$ | 60/360 | 60/360 |
| $w_{\text{multi}}$ | 0.100/0.900 | 0.102/0.898 |
| $\kappa_{\infty\text{multi}}$ | 15.0 | 15.2 |
| $\phi_{\text{s}_{\text{multi}}}$ | 110.5/19.5 | 109.8/20.9 |
| $\phi_{\text{c}_{\text{multi}}}$ | 66.5/3.5 | 59.7/9.7 |
| $\mathcal{L}_{\text{multi}}$ | – | 3.1659 |
| $\beta_{\text{mono}}$ | – | 153.0 |
| $\kappa_{\infty\text{mono}}$ | – | 16.6 |
| $\phi_{\text{s}_{\text{mono}}}$ | – | 68.7/13.1 |
| $\phi_{\text{c}_{\text{mono}}}$ | – | 37.4/6.1 |
| $\mathcal{L}_{\text{mono}}$ | – | 3.1560 |
| BR | 0.833 | 0.839 |
| DBR | 0.250 | 0.257 |

TABLE 2.2 – Calibration of the resilience (left) and intensity (right) for Simulation 1. For the $\phi$'s, the first entry is the constant term and the second one is the linear term.

Overall, the accuracy of the estimation is satisfying. The estimated Hawkes parameters are very close to the real ones, although the dimensionality is high. Importantly, the branching ratios and directional branching ratios are all determined accurately, within a precision of $\pm 0.03$ on our experiments. Concerning the propagator, the results are more noisy for Simulation 1, which is not surprising since

| Year | Simu. | Calib. |
|---|---|---|
| $L_{\text{adj}}$ (sec) | 2 | 2 |
| $\gamma_{\text{multi}}$ | – | 3.05 |
| $\rho_{\text{multi}}$ | – | 6/120 |
| $\lambda_{\text{multi}}$ | – | 0.0005/0.6850 |
| $\nu_{\text{multi}}$ | – | 0.31 |
| $\sigma_{\text{multi}}$ | – | 0.0055 |
| $r^2_{\text{multi}}$ | – | 96.92% |
| $\gamma_{\text{mono}}$ | 3.20 | 3.06 |
| $\rho_{\text{mono}}$ | 130 | 121.3 |
| $\lambda_{\text{mono}}$ | 0.70 | 0.69 |
| $\sigma_{\text{mono}}$ | 0.0000 | 0.0055 |
| $r^2_{\text{mono}}$ | – | 96.92% |

| Year | Simu. | Calib. |
|---|---|---|
| Marks type | Volume | Volume |
| $\beta_{\text{multi}}$ | 120/360 | 6/120/360 |
| $w_{\text{multi}}$ | 0.050/0.950 | 0.0007/0.0505/0.9488 |
| $\kappa_{\infty\text{multi}}$ | 40.0 | 39.1 |
| $\phi_{\text{s multi}}$ | 84.0/36.0 | 72.8/40.9 |
| $\phi_{\text{c multi}}$ | 45.0/5.0 | 47.4/7.7 |
| $\mathcal{L}_{\text{multi}}$ | – | 2.7218 |
| $\beta_{\text{mono}}$ | – | 82.2 |
| $\kappa_{\infty\text{mono}}$ | – | 19.3 |
| $\phi_{\text{s mono}}$ | – | 27.3/15.4 |
| $\phi_{\text{c mono}}$ | – | 17.8/2.9 |
| $\mathcal{L}_{\text{mono}}$ | – | 2.6740 |
| BR | 0.519 | 0.535 |
| DBR | 0.214 | 0.186 |

TABLE 2.3 – Calibration of the resilience (left) and intensity (right) for Simulation 2. For the $\phi$'s, the first entry is the constant term and the second one is the linear term.

it includes some Brownian noise. Still, the proportion of transient impact is nearly exact and the dominant time scale is well determined. Simulation 2 is generated with a mono-exponential propagator, and the resilience speed $\rho_{\text{mono}}$ is slightly underestimated ; however this parameter is less stable than the $\lambda$'s and the accuracy that we obtain seems reasonable. In this second case, the values that we find for the volatility and the regression $r^2$ are satisfyingly close to 0 and 100% respectively.

### 2.4.3 BNP Paribas

Tables 2.4, 2.5 and 2.6 and Figures 2.1, 2.2 and 2.3 present our estimation results for the French stock BNP Paribas on the periods February-September 2012 and January-September 2013.

| $L_{\text{adj}}$ (sec) | 0 | 2 | 4 | 6 |
|---|---|---|---|---|
| $r^2_{\text{multi}}(2012)$ | 24.572% | 24.675% | 24.677% | 24.672% |
| $r^2_{\text{multi}}(2013)$ | 10.607% | 10.674% | 10.668% | 10.649% |

TABLE 2.4 – Regression $r^2$ for the multi-exponential resilience curve, evaluated for several market adjustment lags $L_{\text{adj}} = 0, 2, 4, 6$ seconds, for the stock BNP Paribas.

| Marks type | Unit | Volume | Price jump |
|---|---|---|---|
| $\mathcal{L}_{\text{mono}}(2012)$ | 2.6804 | 2.6826 | 2.6791 |
| $\mathcal{L}_{\text{mono}}(2013)$ | 2.5772 | 2.5794 | 2.5750 |

TABLE 2.5 – Log-likelihood per point for the mono-exponential Hawkes model, evaluated for several types of marks : unit, volumes and price jumps (see eq. (2.34)), for the stock BNP Paribas.

Let us first look at the estimation results for the propagator. Table 2.4 and Figure 2.2 show that the adjustment lag $L_{\text{adj}}$ defined in Section 2.3.3 is positive and thus that the propagator is increasing

| Year | 2012 | 2013 |
|---|---|---|
| $L_{\text{adj}}$ (sec) | 4 | 2 |
| $\gamma_{\text{multi}}$ | 2.69 | 2.99 |
| $\rho_{\text{multi}}$ | 60 | 60/360 |
| $\lambda_{\text{multi}}$ | 0.61 | 0.30/0.53 |
| $\nu_{\text{multi}}$ | 0.39 | 0.17 |
| $\sigma_{\text{multi}}$ | 0.2253 | 0.2153 |
| $r^2_{\text{multi}}$ | 24.677% | 10.674% |
| $\gamma_{\text{mono}}$ | 2.70 | 2.56 |
| $\rho_{\text{mono}}$ | 60.8 | 116.5 |
| $\lambda_{\text{mono}}$ | 0.62 | 0.80 |
| $\sigma_{\text{mono}}$ | 0.2253 | 0.2153 |
| $r^2_{\text{mono}}$ | 24.678% | 10.688% |

| Year | 2012 | 2013 |
|---|---|---|
| Marks type | Volume | Volume |
| $\beta_{\text{multi}}$ | 6/360 | 6/360 |
| $w_{\text{multi}}$ | 0.010/0.990 | 0.011/0.989 |
| $\kappa_{\infty\text{multi}}$ | 15.1 | 12.1 |
| $\phi_{\text{smulti}}$ | 112.8/18.4 | 115.4/15.7 |
| $\phi_{\text{cmulti}}$ | 50.4/2.1 | 46.4/0.9 |
| $\mathcal{L}_{\text{multi}}$ | 2.7720 | 2.6708 |
| $\beta_{\text{mono}}$ | 73.0 | 114.1 |
| $\kappa_{\infty\text{mono}}$ | 13.9 | 14.0 |
| $\phi_{\text{smono}}$ | 38.3/6.2 | 58.5/8.0 |
| $\phi_{\text{cmono}}$ | 17.1/0.7 | 23.5/0.5 |
| $\mathcal{L}_{\text{mono}}$ | 2.6826 | 2.5794 |
| BR | 0.820 | 0.810 |
| DBR | 0.351 | 0.380 |

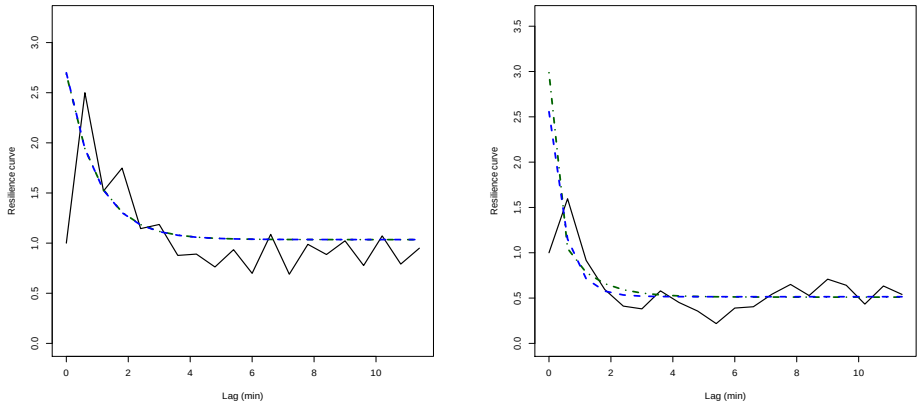
TABLE 2.6 – Calibration of the resilience (left) and intensity (right) for the stock BNP Paribas for the periods February-September 2012 and January-September 2013, between 11 a.m. and 1 p.m. For the $\phi$'s, the first entry is the constant term and the second one is the linear term.

near zero. The estimation yields $L_{\text{adj}} = 4$ sec. for 2012 and $L_{\text{adj}} = 2$ sec. for 2013, and the increasing part lasts indeed longer on Figure 2.2(a) than on Figure 2.2(b). The parameter $\gamma$ given in Table 2.6 tunes the maximum value reached by the propagator at the end of the increasing phase. We find a result between two and three. This means that on average, after a large trade, not only does the bid-ask close around the impacted price (which would yield [4] $\gamma \approx 2$), but cancellations at the new best queue also push the price in the same direction as the trade.

After its short increasing part, the propagator switches to its resilience mode described by Table 2.6 and Figure 2.1. The unconstrained resilience curve is quite smooth, and one can observe on Figure 2.1 that it decays to a non-zero proportion of permanent impact ($\approx 40\%$ for 2012 and $\approx 20\%$ for 2013). Also, the results given in Table 2.6 indicate that the mono-exponential fit for the resilience is good on this dataset. For 2012, only the speed $\rho = 60$ (i.e. a characteristic time scale of one minute) is selected in the multi-exponential estimation. On the other hand, for 2013, there are two selected speeds (corresponding to one minute and ten seconds), but the mono-exponential fit with $\rho_{\text{mono}} = 116.5$ (approximately thirty seconds) yields a higher regression $r^2$. These dominant characteristic time scales motivate the use of the particular case considered for the optimal strategy in Section 2.2.3.

We now focus on the estimation results for the Hawkes parameters. Table 2.5 justifies the selection of volume marks : indeed, they yield a higher likelihood per point than unit marks and price marks. Unit marks are the benchmark model for Hawkes processes, but they fail to reproduce the fact that large orders trigger more activity on the market. Indeed, we see on Table 2.6 that the self-excitation parameter $\phi_{\text{s}}$ and the cross-excitation parameter $\phi_{\text{c}}$ have non-negligible linear parts ($10 - 15\%$ for self-excitation and $2 - 5\%$ for cross-excitation). As for price marks, we think that they give less

4. To be more precise, let us consider for example a buy order that increases the ask of one tick. Then, the midprice jumps of one half tick. If the bid price follows shortly the ask and increases of one tick, this moves again the mid of one half tick upward, which gives $\gamma = 2$.




(a) 2012 (b) 2013

FIGURE 2.1 – The estimated propagator for BNP Paribas. The plain line is the unconstrained propagator, the (blue) dashed line is the mono-exponential resilience curve, and the (green) dot-dashed line is the multi-exponential resilience curve.

information than volume marks since price jumps cluster on a few values (one or two ticks in most cases), while the distribution of volumes is much wider.

Hawkes parameters seem to be quite stable, especially in the multi-exponential case where the estimation results are very similar for 2012 and 2013. Two decay speeds are selected for the intensity, and these are the two extreme ones : the long range $\beta = 6$ (10 minutes) and the short range $\beta = 360$ (10 seconds). The importance of each time scale $\beta_i$ can be measured by the proportion of the norm that it accounts for, given by

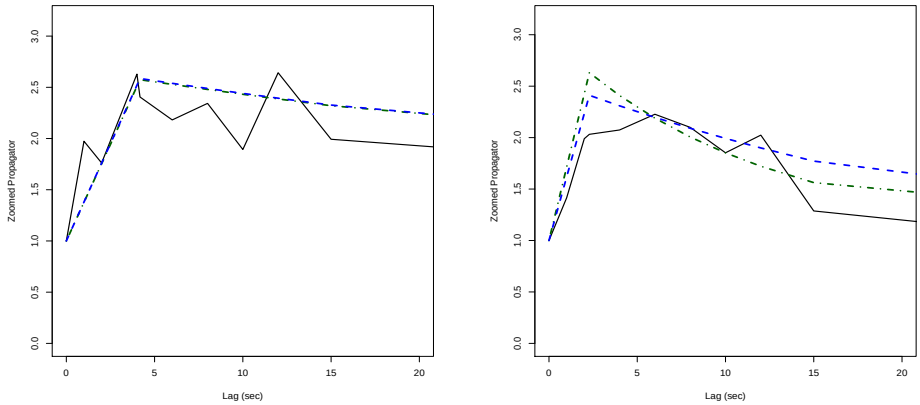
$$\frac{w_i/\beta_i}{\sum_j w_j/\beta_j}.$$

Here, the long-range component $\beta = 6$ accounts for $\approx 40\%$ of the norm, and the short-range one $\beta = 360$ for the remaining 60%. Therefore, both decay speeds are important, which is also reflected by the significant increase from the log-likelihood per point $\mathcal{L}_{\text{mono}}$ of the mono-exponential model to $\mathcal{L}_{\text{multi}}$ for the multi-exponential one. One can deduce that contrary to the propagator, the Hawkes kernel includes at least two exponential components.

Figure 2.3 gives a visual comparison between the data, the mono-exponential Hawkes model and the multi-exponential one through the auto-correlation of the number of events. The formula for the empirical auto-correlation function $\widehat{\mathscr{C}}(k)$ is given by equation (2.35). Using equations (2.12) and (2.14), we have that if $h > 0$ is small and $\tau > 0$, $\widehat{\mathscr{C}}(\tau/h)$ approximates the auto-correlation function $\mathscr{C}(\tau)/\mathscr{C}(0)$ of the total intensity process $\Sigma_t$. For a multi-exponential Hawkes kernel, one has

$$\widehat{\mathscr{C}}(\tau/h) \approx \frac{\mathscr{C}(\tau)}{\mathscr{C}(0)} = \sum_{j=1}^{p} \frac{a_j}{\sum_k a_k} \exp(-b_j|\tau|),$$

where the coefficients $a_1, \cdots, a_p, b_1, \cdots, b_p > 0$ are determined as in Proposition 2.2.1. One can see on Figure 2.3 that the mono-exponential model fits the end of the curve rather well but that its initial




(a) 2012 (b) 2013

FIGURE 2.2 – Zoom on the first twenty seconds of the propagator curve for BNP Paribas. The plain line is the unconstrained curve, the (blue) dashed line is the mono-exponential curve, and the (green) dot-dashed line is the multi-exponential curve. The propagator is increasing during a few seconds, before the resilience effect kicks in.

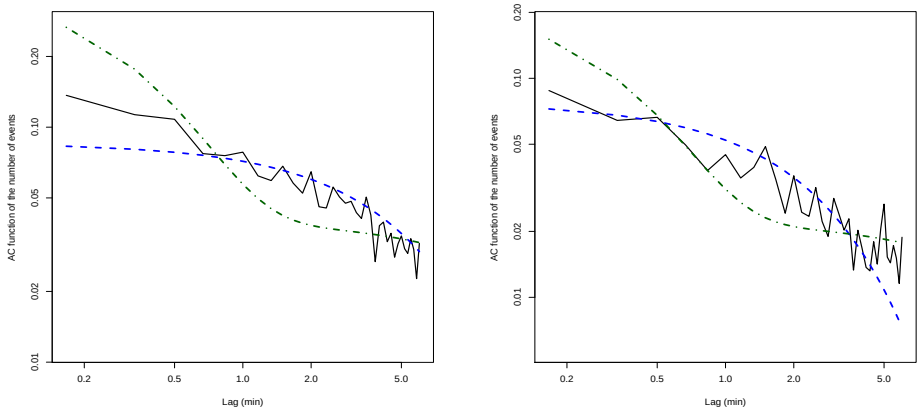
decay is too slow. On the other hand, the multi-exponential model does show a transition between two decay speeds, and captures the short-range behavior of the curve better. Still, the accuracy of the fit is not very satisfactory and it seems that the functional form of the auto-correlation is more subtle than a multi-exponential one.

Finally, we confront our calibration results to the conditions derived in Section 2.2.2 for the absence of Price Manipulation Strategies in the model. It is complicated in practice to quantify the deviation of our set of parameters to the equilibrium using equation (2.16). On the other hand, equation (2.18) gives a simpler criterion : the directional branching ratio DBR and the proportion $1-\nu$ of transient impact should be equal for PMS to be ruled out. Here, the standard branching ratio BR $\approx 80\%$ is high, but the directional branching ratio DBR $\approx 40\%$ is quite low, which is due to a non-negligible part of cross-excitation in the order flow. It implies that the equilibrium condition is violated since $1-\nu \approx 60\%$ for 2012 and $1-\nu \approx 80\%$ for 2013. Since $1-\nu >$ DBR holds in both cases, we find that the price process is mean-reverting on average, rather than diffusive. This should lead to the existence of PMS in practice, which is the object of Section 2.5.

### 2.4.4 Total

Tables 2.7, 2.8 and 2.9 present our estimation results for the French stock Total on the periods January-September 2012 and January-September 2013.

The qualitative interpretation of the results is similar to that of Section 2.4.3. Yet, one should note the following points that are observable on Table 2.9. First, we notice that there is no significant difference between the mono and multi-exponential propagator. Here, contrary to the BNP Paribas case, the fit is slightly better with two time scales. Second, the branching ratio BR $\approx 60\%$ and the directional branching ratio DBR$\approx 30\%$ are smaller for Total, whereas the proportion $1-\nu$ of




(a) 2012 (b) 2013

FIGURE 2.3 – Auto-correlation function of the number of midpoint moves triggered by trades (plain line), in log-log scale, for BNP Paribas. The (blue) dashed line is the auto-correlation generated by the mono-exponential Hawkes model, the (green) dot-dashed line is generated by the multi-exponential Hawkes model.

| $L_{\text{adj}}$ (sec) | 0 | 2 | 4 | 6 |
|---|---|---|---|---|
| $r^2_{\text{multi}}(2012)$ | 23.093% | 23.166% | 23.137% | 23.108% |
| $r^2_{\text{multi}}(2013)$ | 11.604% | 11.613% | 11.608% | 11.606% |

TABLE 2.7 – Regression $r^2$ for the multi-exponential resilience curve, evaluated for several market adjustment lags $L_{\text{adj}} = 0, 2, 4, 6$ seconds, for the stock Total.

transient impact (84% for 2012 and 92% for 2013) is higher, which means that the price has an even stronger mean-reversion tendency.

| Marks type | Unit | Volume | Price jump |
|---|---|---|---|
| $\mathcal{L}_{\text{mono}}(2012)$ | 2.2981 | 2.3034 | 2.2965 |
| $\mathcal{L}_{\text{mono}}(2013)$ | 2.2065 | 2.2127 | 2.2063 |

TABLE 2.8 – Log-likelihood per point for the mono-exponential Hawkes model, evaluated for several types of marks : unit, volumes and price jumps (see eq. (2.34)), for the stock Total.

| Year | 2012 | 2013 |
|---|---|---|
| $L_{\text{adj}}$ (sec) | 2 | 2 |
| $\gamma_{\text{multi}}$ | 3.72 | 2.21 |
| $\rho_{\text{multi}}$ | 60/360 | 6/120/360 |
| $\lambda_{\text{multi}}$ | 0.29/0.55 | 0.004/0.651/0.268 |
| $\nu_{\text{multi}}$ | 0.16 | 0.08 |
| $\sigma_{\text{multi}}$ | 0.1400 | 0.1124 |
| $r^2_{\text{multi}}$ | 23.166% | 11.613% |
| $\gamma_{\text{mono}}$ | 3.84 | 2.65 |
| $\rho_{\text{mono}}$ | 187.2 | 191.3 |
| $\lambda_{\text{mono}}$ | 0.84 | 0.93 |
| $\sigma_{\text{mono}}$ | 0.1399 | 0.1123 |
| $r^2_{\text{mono}}$ | 23.132% | 11.586% |

| Year | 2012 | 2013 |
|---|---|---|
| Marks type | Volume | Volume |
| $\beta_{\text{multi}}$ | 120/360 | 6/60/360 |
| $w_{\text{multi}}$ | 0.052/0.948 | 0.010/0.035/0.955 |
| $\kappa_{\infty\text{multi}}$ | 21.0 | 9.7 |
| $\phi_{\text{s}_{\text{multi}}}$ | 98.7/21.7 | 84.5/18.5 |
| $\phi_{\text{c}_{\text{multi}}}$ | 44.3/3.9 | 36.5/0.7 |
| $\mathcal{L}_{\text{multi}}$ | 2.3801 | 2.2842 |
| $\beta_{\text{mono}}$ | 93.0 | 109.1 |
| $\kappa_{\infty\text{mono}}$ | 9.2 | 9.0 |
| $\phi_{\text{s}_{\text{mono}}}$ | 43.5/9.6 | 47.4/10.4 |
| $\phi_{\text{c}_{\text{mono}}}$ | 19.5/1.7 | 20.4/0.4 |
| $\mathcal{L}_{\text{mono}}$ | 2.3034 | 2.2127 |
| BR | 0.517 | 0.688 |
| DBR | 0.222 | 0.323 |

TABLE 2.9 – Calibration of the resilience (left) and intensity (right) for the stock Total for the periods January-September 2012 and January-September 2013, between 11 a.m. and 1 p.m. For the $\phi$'s, the first entry is the constant term and the second one is the linear term.

## 2.5 Test of some Price Manipulation Strategies

In this section, we apply the optimal strategy derived in [3] and Theorem 2.2.2 to our dataset, with the parameters obtained by our calibration protocol. Essentially, we run the strategy each day with a zero initial and final position. If the model is relevant, this should give some profit on average. This backtest serves as a practical evaluation of our calibration results, and of the model itself.

### 2.5.1 Scaling and discretization of the optimal strategy

The simplest and most natural way is to use the optimal strategy (2.23) is to consider a discrete subset $\Theta$ of $[0,T]$ (possibly made of stopping times) and to trade for each time $t \in \Theta$ the quantity

$$\xi^s_{t,T} = -\frac{[1+\rho(T-t)]qsD_t + X_t}{2+\rho(T-t)} + \frac{m_1}{2\rho} \times \left[(1,\cdots,1)\ .\ \left\{I_p + \frac{\rho(T-t)}{2+\rho(T-t)} \times [\zeta((T-t)H) + \nu\rho(T-t)\ \omega((T-t)H)]\right\}\ .\ s\delta_t\right], \tag{2.36}$$

so that (2.23) holds in $t+$ if $s=1$. Here, $\delta_t = \left(\kappa_t^{+(i)} - \kappa_t^{-(i)}\right)_i$ is the vector of intensity imbalances and we calculate $D$ by using the following formula

$$D_t = \sum_{\tau \le t} \Delta M_\tau\ [G(t-\tau) - G(\infty)].$$

In order to tune the leveraging of the strategy and its discreteness on the market, we introduce a scaling factor $s \in [0,1]$ that multiplies $\delta_t$ and $D_t$. By doing so, we multiply by $s$ the deviation of the whole strategy from the standard Obizhaeva and Wang [93] liquidation scheme. The latter is static since it assumes that the observed price process is always a martingale. The limit $s=0$ thus corresponds to the static strategy, whereas $s=1$ is the optimal strategy given by Theorem 2.2.2, which may be very aggressive in standard market conditions. In fact, using the optimal strategy with $s=1$ may lead to buy and sell repeatedly quantities that exceed the size of the first queues, which is not realistic.

### 2.5.2 Methodology

To backtest the strategy in practice, we choose to update our position when we observe midprice moves. Let us define

$$\Theta = \{\theta \in (\Delta_{\mathrm{RW}}, T),\ \theta - \tau(\theta) > L_{\mathrm{adj}}\},$$

where the $\theta$'s correspond to the times of price jumps due to cancellations and passive limit orders, $\tau(\theta)$ is the time of the last price jump due to a trade before $\theta$, $\Delta_{\mathrm{RW}}$ is the regression window defined in Section 2.3.3 and $L_{\mathrm{adj}}$ is the market adjustment lag. The position of the strategy at time $t \in [0,T]$ is given by

$$X^s_t = \sum_{\theta \in \Theta} \xi^s_{\theta,T}.$$

At time $T$, we close the position with the transaction

$$\Delta X_T^s = -X_T^s.$$

The time horizon is still $T = 2$ hours, where $t = 0$ corresponds to 11 a.m. and $t = T$ to 1 p.m. We choose to apply the strategy on $[\Delta_{\text{RW}}, T]$ instead of $[0, T]$, so that the values of $\delta_t$ and $D_t$ for $t \geq \Delta_{\text{RW}}$ can be accurately computed. Moreover, for each time $\tau \in (\Delta_{\text{RW}}, T)$ where the price jumps because of a transaction, we do not trade on the time interval $[\tau, \tau + L_{\text{adj}}]$. As a matter of fact, the market adjustment lag $L_{\text{adj}}$ corresponds approximately to the time needed for the bid-ask to close after a trade that empties the best bid or the best ask. It would be meaningless to trade at the midprice (or even at the midprice $\pm 1$ half-tick) before the bid-ask is closed, and we would artificially boost the performance of the strategy if we allowed it. However, this constraint is not needed for simulated data, for which we set $\Theta = \{\theta \in (\Delta_{\text{RW}}, T)\}$.

We assume that the scaling $s$ is small enough for the effective impact of the strategy on the market price to be negligible. Although approximative, this assumption allows us to backtest the strategy assuming that we can trade at the observed price.

In the sequel of this section, we apply the optimal strategy for the mono and multi-exponential Hawkes decay kernels and for several stocks. We summarize the results in one table and a few graphs for each stock. We note $Y_i$ is the profit made by the strategy on day $i \in \{1, \cdots, n\}$, $\overline{Y_n} = \frac{1}{n}\sum_{i=1}^n Y_i$ the empirical mean and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n [Y_i - \overline{Y_n}]^2$ the empirical variance of daily profits. The values given in the table are

– The annualized Sharpe ratio of the strategy

$$\text{Sharpe} = \sqrt{n} \times \frac{\overline{Y_n}}{\sqrt{S_n^2}}.$$

– The empirical positivity probability, skew and kurtosis of daily gains

$$\text{Proba.} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i > 0\}}, \quad \text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^n [Y_i - \overline{Y_n}]^3}{S_n^3}, \quad \text{Kurto.} = \frac{\frac{1}{n}\sum_{i=1}^n [Y_i - \overline{Y_n}]^4}{S_n^4}.$$

The choice of the scaling $s$ has no impact on these results, since all the values above are invariant to the multiplication of the strategy by a positive constant. Thus, only the units of the graphs are changed by the scaling, and we fix $s = 0.001$. With this choice, the volumes of individual transactions never exceed 5% of the average volume of the best bid/ask queue, which makes our toy backtest with no impact reasonable.

For each stock and each period, we also evaluate of the « Poisson strategy » that one obtains if trades are modeled with two independent compound Poisson processes, which is equivalent to imposing $\kappa_t^+ \equiv \overline{\kappa}$, $\kappa_t^- \equiv \overline{\kappa}$ and thus $\kappa_t^+ - \kappa_t^- \equiv 0$. More precisely, we trade for $t \in \Theta$ (the same time grid as for the Hawkes model) the quantity

$$\xi_{t,T}^s = -\frac{[1 + \rho(T-t)]qsD_t + X_t}{2 + \rho(T-t)}.$$

This strategy is entirely based on mean-reversion, and the trend-following part disappears. For the Hawkes and the Poisson strategies, we give in the tables the impact of a bid-ask cost of one half-tick on the results. This corresponds to a more realistic implementation of the strategy (which

should trade at the best and not at the midpoint) and we see that this is sufficient to prevent Price Manipulation Strategies in most cases (the Sharpe ratio becomes close to zero or even negative). As a benchmark, we also present in Table 2.10 and 2.11 the results of these strategies on simulated data. These give an idea of the profits that the strategies could reach in theory.

Our findings are the following. On simulated data, the profits made by the strategies are evident and still significant with a half-tick penalty. On real data, the Sharpe ratios remain positive for all the tests, which indicates that the model is not out of scope and captures some characteristics of the real market flow. However, these ratios are lower than for simulated data and may become negative when we take the bid-ask spread into account. Said differently, market participants who use mean-reverting and trend-following strategies already exploit most of the arbitrage opportunities described by our model, and the backtest of our optimal strategy in realistic market conditions does not yield significant gains. Somehow, this justifies the theoretical assumption to consider a market without PMS when dealing with both market impact and the bid-ask spread. Now, let us compare the different strategies used in Tables 2.12 and 2.13. The results are rather similar for the three strategies and none of them seem to outperform the others. Intuitively, this means that the main component of the strategy is the mean-reverting one (which is common to the Poisson and Hawkes strategies), while the trend-following one has a minor contribution. This is confirmed by the statistical facts in Table 2.6 and 2.9 where the directional branching ratio DBR is much lower than the proportion of transient impact $\lambda_{\text{mono}} = 1 - \nu$.

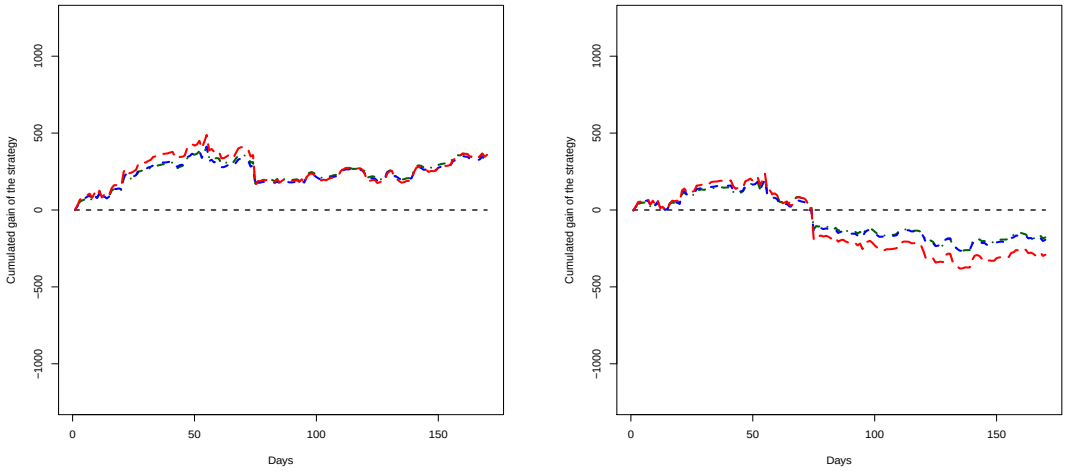
### 2.5.3 Simulated data

Tables 2.10 and 2.11 present the results of the optimal strategy applied to simulated data. The simulation parameters are the same as in Section 2.4.2 (see Tables 2.2 and 2.3), and both datasets are composed of 150 independent two-hour windows. In Tables 2.10 and 2.11, the two first columns contain the results of the strategy computed with the real simulation parameters for the Hawkes model, and the third and fourth columns contain the results for estimated Hawkes parameters. In both cases, the resilience is the estimated mono-exponential curve, since the optimal strategy is known explicitly only in that case.

| Year | Simu. | +bid-ask | Calib. | +bid-ask |
|---|---|---|---|---|
| Sharpe (Multi) | 6.759 | 3.225 | 6.764 | 3.176 |
| Proba. (Multi) | 74.0% | 63.3% | 74.0% | 63.3% |
| Skew (Multi) | 0.55 | 0.23 | 0.57 | 0.24 |
| Kurtosis (Multi) | 4.19 | 4.03 | 4.22 | 4.05 |
| Sharpe (Mono) | – | – | 6.308 | 3.371 |
| Proba. (Mono) | – | – | 74.0% | 62.7% |
| Skew (Mono) | – | – | 0.47 | 0.20 |
| Kurto. (Mono) | – | – | 4.11 | 3.97 |
| Sharpe (Poisson) | – | – | 6.630 | 3.735 |
| Proba. (Poisson) | – | – | 73.3% | 64.0% |
| Skew (Poisson) | – | – | 0.43 | 0.18 |
| Kurto. (Poisson) | – | – | 3.88 | 3.80 |

TABLE 2.10 – Results statistics of the optimal strategy applied on the data of Simulation 1 (simulation parameters of Table 2.2).

### 2.5.4 BNP Paribas



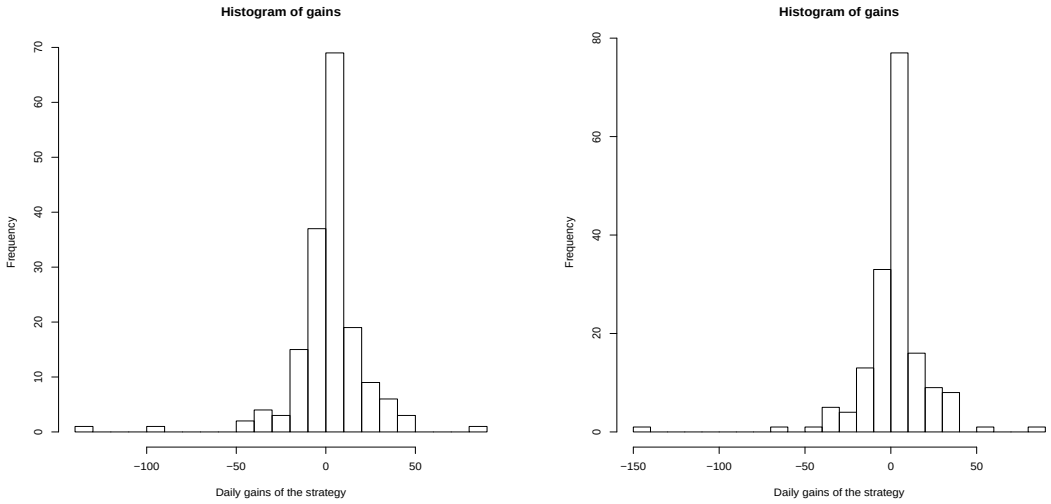

(a) Trading at the midprice

(b) One half-tick penalty

FIGURE 2.4 – Cumulated gains of the strategy applied on BNP Paribas on the period February-September 2012, every day between 11.30a.m. and 1p.m. The (red) long-dashed line is the performance of the Poisson model, the (blue) dashed line is the mono-exponential Hawkes model, and the (green) dot-dashed line is the multi-exponential Hawkes model. Left : we allow the strategy to trade at the midprice. Right : we apply a posteriori a linear cost penalty of one half-tick to account for the bid-ask spread.

| Year | Simu. | +bid-ask | Calib. | +bid-ask |
|---|---|---|---|---|
| Sharpe (Multi) | 33.268 | 27.095 | 32.302 | 25.769 |
| Proba. (Multi) | 100.0% | 100.0% | 100.0% | 99.3% |
| Skew (Multi) | 0.50 | 0.51 | 0.52 | 0.54 |
| Kurtosis (Multi) | 3.22 | 3.35 | 3.25 | 3.40 |
| Sharpe (Mono) | – | – | 34.940 | 28.605 |
| Proba. (Mono) | – | – | 100.0% | 100.0% |
| Skew (Mono) | – | – | 0.45 | 0.46 |
| Kurto. (Mono) | – | – | 3.19 | 3.31 |
| Sharpe (Poisson) | – | – | 34.986 | 28.681 |
| Proba. (Poisson) | – | – | 100.0% | 100.0% |
| Skew (Poisson) | – | – | 0.44 | 0.45 |
| Kurto. (Poisson) | – | – | 3.12 | 3.25 |

TABLE 2.11 – Results statistics of the optimal strategy applied on the data of Simulation 2 (simulation parameters of Table 2.3).

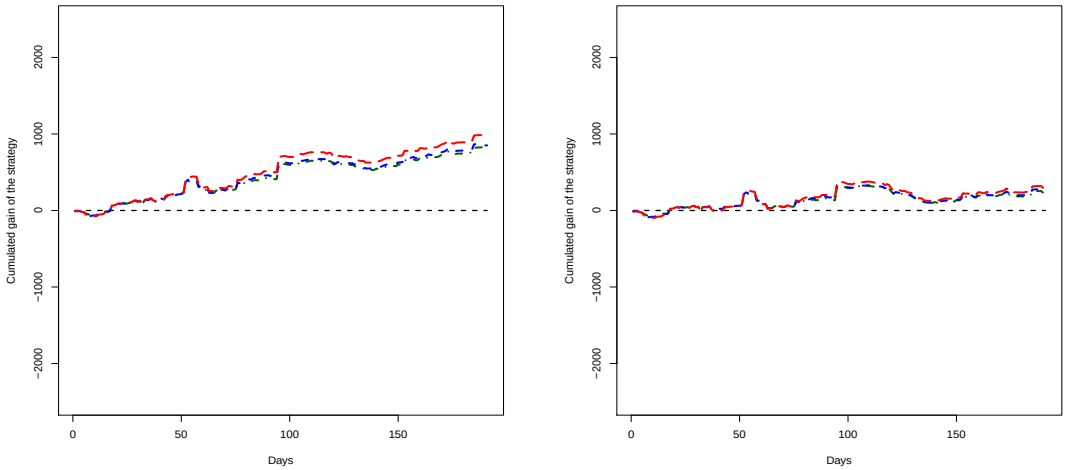



(a) Mono-exponential Hawkes (b) Multi-exponential Hawkes

FIGURE 2.5 – Histogram of the daily gains of the strategy applied on BNP Paribas on the period February-September 2012, between 11.30a.m. and 1p.m. Left : Mono-exponential Hawkes model. Right : Multi-exponential Hawkes model.

| Year | IS 2012 | +bid-ask | IS 2013 | +bid-ask | OS 2013 | +bid-ask |
|---|---|---|---|---|---|---|
| Sharpe (Multi) | 1.382 | −0.675 | 2.454 | 0.725 | 2.248 | 0.418 |
| Proba. (Multi) | 65.9% | 56.5% | 61.3% | 47.1% | 58.1% | 48.2% |
| Skew (Multi) | −2.02 | −2.40 | 3.65 | 3.34 | 4.48 | 4.14 |
| Kurtosis (Multi) | 19.02 | 19.94 | 29.40 | 27.71 | 36.96 | 34.65 |
| Sharpe (Mono) | 1.263 | −0.713 | 2.536 | 0.771 | 2.430 | 0.563 |
| Proba. (Mono) | 62.9% | 57.1% | 62.3% | 48.2% | 58.1% | 49.7% |
| Skew (Mono) | −1.89 | −2.30 | 2.94 | 2.61 | 3.56 | 3.21 |
| Kurto. (Mono) | 16.64 | 17.68 | 23.27 | 21.90 | 26.74 | 24.87 |
| Sharpe (Poisson) | 1.056 | −0.849 | 2.5888 | 0.8077 | 2.513 | 0.630 |
| Proba. (Poisson) | 65.3% | 55.9% | 61.3% | 49.7% | 60.2% | 49.2% |
| Skew (Poisson) | −2.72 | −3.07 | 3.09 | 2.76 | 3.94 | 3.58 |
| Kurto. (Poisson) | 23.46 | 24.68 | 24.41 | 22.82 | 31.13 | 28.86 |

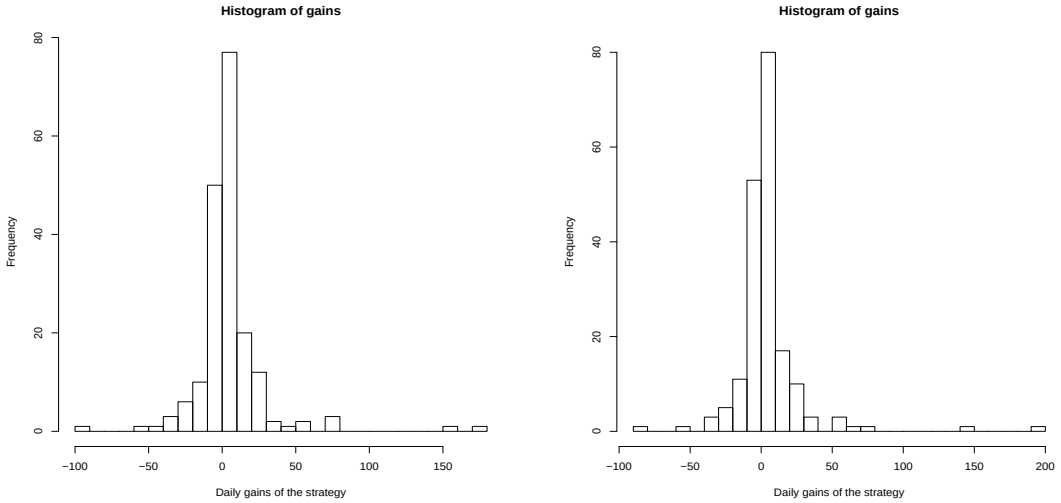
TABLE 2.12 – Results statistics of the optimal strategy applied on BNP Paribas on the periods February-September 2012 and January-September 2013, every day between 11.30a.m. and 1p.m. The first two columns are In-Sample results, i.e. the data used to calibrate the model is the same as the evaluation data. The third column gives Out-of-Sample results, i.e. we calibrate the model on the 2012 data to apply the strategy on the 2013 data.




(a) Trading at the midprice

(b) One half-tick penalty

FIGURE 2.6 – Cumulated gains of the strategy applied on BNP Paribas on the period January-September 2013, every day between 11.30a.m. and 1p.m. The (red) long-dashed line is the performance of the Poisson model, the (blue) dashed line is the mono-exponential Hawkes model, and the (green) dot-dashed line is the multi-exponential Hawkes model. Left : we allow the strategy to trade at the midprice. Right : we apply a posteriori a linear cost penalty of one half-tick to account for the bid-ask spread.




(a) Mono-exponential Hawkes (b) Multi-exponential Hawkes

FIGURE 2.7 – Histogram of the daily gains of the strategy applied on BNP Paribas on the period January-September 2013, between 11.30a.m. and 1p.m. Left : Mono-exponential Hawkes model. Right : Multi-exponential Hawkes model.

### 2.5.5 Total

| Year | IS 2012 | +bid-ask | IS 2013 | +bid-ask | OS 2013 | +bid-ask |
|---|---|---|---|---|---|---|
| Sharpe (Multi) | 0.067 | −0.763 | 2.697 | 1.016 | 2.794 | 1.224 |
| Proba. (Multi) | 57.8% | 44.3% | 66.0% | 51.8% | 65.4% | 51.8% |
| Skew (Multi) | −9.34 | −9.62 | 6.38 | 6.37 | 5.94 | 5.97 |
| Kurtosis (Multi) | 114.76 | 117.75 | 62.86 | 65.85 | 53.84 | 57.93 |
| Sharpe (Mono) | 0.126 | −0.770 | 2.795 | 1.191 | 2.760 | 1.099 |
| Proba. (Mono) | 59.4% | 44.8% | 66.0% | 52.4% | 65.4% | 52.4% |
| Skew (Mono) | −9.52 | −9.82 | 6.01 | 6.02 | 6.18 | 6.18 |
| Kurto. (Mono) | 118.29 | 121.77 | 55.54 | 59.30 | 59.20 | 62.65 |
| Sharpe (Poisson) | 0.001 | −0.810 | 2.807 | 1.259 | 2.790 | 1.224 |
| Proba. (Poisson) | 57.8% | 43.8% | 65.4% | 50.8% | 65.4% | 50.8% |
| Skew (Poisson) | −9.33 | −9.59 | 5.96 | 6.00 | 6.04 | 6.08 |
| Kurto. (Poisson) | 114.39 | 116.97 | 53.37 | 57.35 | 54.90 | 58.87 |

TABLE 2.13 – Results statistics of the optimal strategy applied on Total on the period January-September 2012-2013, every day between 11.30a.m. and 1p.m. The first two columns are In-Sample results, i.e. the data used to calibrate the model is the same as the evaluation data. The third column gives Out-of-Sample results, i.e. we calibrate the model on the 2012 data to apply the strategy on the 2013 data.

## 2.6 Conclusion

In this paper we extend the theoretical model of [3] by allowing more general forms for the propagator and the Hawkes kernel. Moreover, we derive the conditions that exclude Price Manipulation Strategies in the sense of Huberman and Stanzl [78] in the case where both the propagator and the Hawkes part have a multi-exponential decay. This allows us to deduce some interesting links between the propagator and the Hawkes kernel for general completely monotone kernels. Besides, when the price propagator is mono-exponential and the Hawkes kernel is multi-exponential, we can still obtain the optimal strategy as a closed formula. This has some practical interest since the propagator seems to be well approximated by an exponential, while the Hawkes decay kernel clearly includes several characteristic time scales.

We also introduce a calibration protocol for the model, that we apply to tick-by-tick data from French stocks. The results show that the model explains a significant part of the variance of prices. The long-range propagator is a smoothly decaying curve, but the short-range part is increasing during a few seconds (which we think corresponds to the time that the bid-ask needs to close after a large trade). Concerning the estimation of the Hawkes process modeling the flow of trades, we obtain excitation parameters that significantly differ from zero, which shows in particular that the flow is not Poissonian. Also, we find that the main driver of the excitation between trades is volumes rather than price moves. The martingale conditions that prevent PMS are violated in practice, in particular the directional branching ratio is smaller than the proportion of transient price impact. Therefore, in our dataset, the price has a notable mean-reverting tendency.

A series of backtests shows that the optimal strategy used for round trips is profitable on average, therefore the model does offer a relevant prediction for midprice moves. However, a level of transaction costs compatible with the width of the bid-ask spread makes the profits close to zero. This confirms the natural idea that the absence of Price Manipulation Strategies at this frequency stems from both market impact and bid-ask costs.

We eventually draw some applications and perspectives on our study. A first straightforward application is to use the calibrated model for optimal execution, by using the block trades (2.36) on a given (possibly random) time grid $\Theta$. Contrary to most existing models, this strategy takes the flow of trades into account. Another possible use of this model is to detect the instants when it is interesting to trade. In fact, equation (2.46) gives the (theoretical) instantaneous cost of non-trading. One may decide to trade for example only if this cost is above some threshold, or optimize the trade-off between this cost and transaction costs. Such strategies could be interesting in practice, but need to be thoroughly investigated on market data. Let us now consider some possible extensions of our work. First, it would be interesting to handle a calibration of the model on an entire day instead of a two-hour window. This is certainly difficult due to intra-day variations of trading activity between the open and the close. Second, it would be nice to incorporate in our model transaction costs such as the bid-ask spread. A less ambitious goal would be at least to modify our optimal execution strategy to reduce transaction costs in a clever way, maybe by using equation (2.46) as mentioned above.

## 2.7 Appendix : Estimation of the propagator using Newton-Raphson's algorithm

As explained in Section 2.3.3, we resort to Newton-Raphson's algorithm to minimize the quadratic error

$$\mathcal{E}(\hat{G}) = \sum_{\Delta_{\mathrm{RW}}<\theta<T} [\hat{P}_\theta - P_\theta]^2$$

which quantifies the distance between the observed midpoint price $P_t$ and the predicted price

$$\hat{P}_t \;=\; P_{t-\Delta_{\mathrm{RW}}} \;+ \sum_{t-\Delta_{\mathrm{RW}}\leq\tau\leq t} \Delta M_\tau \; \hat{G}(t-\tau).$$

Let us assume that $\pi \in \mathbb{R}^l$, $l \geq 1$, is a parameterization of $\hat{G}$, i.e. $\hat{G} = \hat{G}(\pi)$ is determined by $\pi$, and so is the error $\mathcal{E}(\hat{G}) = \mathcal{E}(\pi)$. For a starting point $\pi_0$, the principle of the algorithm is to approximate $G$ by the sequence $\hat{G}(\pi_n)$ such that

$$\forall n \in \mathbb{N}, \quad \pi_{n+1} \;=\; \pi_n \;-\; \left[\nabla^2\mathcal{E}(\pi_n)\right]^{-1}.\nabla\mathcal{E}(\pi_n)$$

where $\nabla\mathcal{E}(\pi) \in \mathbb{R}^l$ is the gradient of the error $\mathcal{E}$ and $\nabla^2\mathcal{E}(\pi) \in \mathbb{R}^{l\times l}$ is its Hessian matrix, w.r.t. the parameter $\pi$. The convergence of the method is only guaranteed if the starting point $\pi_0$ is « good enough », and if $\nabla^2\mathcal{E}(\pi_n)$ is positive definite for all $n \in \mathbb{N}$.

To apply this method, one needs to compute the gradient $\nabla\mathcal{E}(\pi)$ and the Hessian matrix $\nabla^2\mathcal{E}(\pi)$ of the error $\mathcal{E}$ for each parameterization $\pi$ of $\hat{G}$. One has

$$\begin{aligned}
\nabla\mathcal{E}(\pi) &= 2 \sum_{\Delta_{\mathrm{RW}}<\theta<T} [\hat{P}_\theta(\pi) - P_\theta] \times \nabla\hat{P}_\theta(\pi), \\
\nabla^2\mathcal{E}(\pi) &= 2 \sum_{\Delta_{\mathrm{RW}}<\theta<T} \left\{ [\hat{P}_\theta(\pi) - P_\theta] \times \nabla^2\hat{P}_\theta(\pi) \;+\; \nabla\hat{P}_\theta(\pi).\left(\nabla\hat{P}_\theta(\pi)\right)^\top \right\}.
\end{aligned}$$

The problem boils down to computing $\nabla\hat{P}_\theta(\pi)$ and $\nabla^2\hat{P}_\theta(\pi)$, which can themselves be expressed as

$$\begin{aligned}
\nabla\hat{P}_\theta(\pi) &= \sum_{t-\Delta_{\mathrm{RW}}\leq\tau\leq t} \Delta M_\tau \; \nabla\hat{G}(t-\tau), \\
\nabla^2\hat{P}_\theta(\pi) &= \sum_{t-\Delta_{\mathrm{RW}}\leq\tau\leq t} \Delta M_\tau \; \nabla^2\hat{G}(t-\tau),
\end{aligned}$$

where we drop the dependency of $\hat{G}$ in $\pi$ for clarity. Therefore, only the gradient $\nabla\hat{G}$ and the Hessian $\nabla^2\hat{G}$ of the estimated propagator $\hat{G}$ need to be specifically derived for each parameterization, which is the object of the sequel.

An important particular case is when $\hat{G}$ is linear w.r.t. $\pi$. In that case, $\nabla^2\hat{G} \equiv 0$, thus $\nabla^2\hat{P}_\theta(\pi) \equiv 0$ and

$$\nabla^2\mathcal{E}(\pi) = 2 \sum_{\Delta_{\mathrm{RW}}<\theta<T} \nabla\hat{P}_\theta(\pi).\left(\nabla\hat{P}_\theta(\pi)\right)^\top$$

is positive definite for any $\pi$. Also, in that case, $\nabla\hat{G}$ does not depend on the current values of the parameter $\pi$, and

$$\pi_1 \;=\; \pi_0 \;-\; \left[\nabla^2\mathcal{E}(\pi_0)\right]^{-1}.\nabla\mathcal{E}(\pi_0)$$

is the minimizer of the error $\mathcal{E}(\pi)$ for any $\pi_0$. Therefore, when the propagator is parameterized linearly, the starting point of the algorithm has no importance and one step is enough to find the optimum.

### 2.7.1 Unconstrained propagator

We consider the unconstrained propagator

$$\hat{G}(t) = g_l \mathbb{1}_{[t_l,\Delta_{\mathrm{RW}}[}(t) + \sum_{i=0}^{l-1} \frac{(t_{i+1}-t)g_i + (t-t_i)g_{i+1}}{t_{i+1}-t_i}\; \mathbb{1}_{[t_i,t_{i+1}[}(t),$$

with $l \geq 2$, $0 = t_0 < t_1 < \cdots < t_l$ fixed discretization times, $g_0 = 1$ and $\pi = (g_1, \cdots, g_l) \in [0,+\infty)^l$ the $l$-dimensional parameter to estimate. The dependence of $\hat{G}$ w.r.t. $\pi$ is linear, and we only need to compute the gradient :

$$\begin{aligned}
\frac{\partial\hat{G}(t)}{\partial g_i} &= \frac{t_{i+1}-t}{t_{i+1}-t_i}\mathbb{1}_{[t_i,t_{i+1}[}(t) + \frac{t-t_{i-1}}{t_i-t_{i-1}}\mathbb{1}_{[t_{i-1},t_i[}(t) \quad \text{for } 1 \leq i \leq l-1,\\
\frac{\partial\hat{G}(t)}{\partial g_l} &= \mathbb{1}_{[t_l,\Delta_{\mathrm{RW}}[}(t) + \frac{t-t_{l-1}}{t_l-t_{l-1}}\mathbb{1}_{[t_{l-1},t_l[}(t).
\end{aligned}$$

### 2.7.2 Multi-exponential curve

In this section we consider the multi-exponential resilience curve

$$\hat{R}(t) = \nu + \sum_{i=1}^{p} \lambda_i \exp(-\rho_i t),$$

and the propagator

$$\hat{G}(t) = \left[1 + (\hat{R}(L_{\mathrm{adj}}) - 1)\;\frac{t}{L_{\mathrm{adj}}}\right]\mathbb{1}_{\{t\leq L_{\mathrm{adj}}\}} + \hat{R}(t)\mathbb{1}_{\{t>L_{\mathrm{adj}}\}},$$

determined by $\hat{R}$ for $L_{\mathrm{adj}} \geq 0$ fixed *a priori*. The dependence of $\hat{G}$ is linear w.r.t. the parameters if and only if the $\rho_i$'s are fixed.

#### Unit Multi-exponential curve

The « unit » multi-exponential resilience curve is the case where $\nu = 1 - \sum_{i=1}^{p}\lambda_i$ is imposed. This yields

$$\hat{R}(t) = 1 - \sum_{i=1}^{p} \lambda_i(1 - \exp(-\rho_i t)),$$

and the parameter $\pi = (\lambda_1, \cdots, \lambda_p, \rho_1, \cdots, \rho_p)$ is $2p$-dimensional. One has for $i, j \in \{1, \cdots, p\}$,

$$\frac{\partial \hat{R}(t)}{\partial \lambda_i} = -\left\{1 - \exp(-\rho_i t)\right\}, \qquad \frac{\partial \hat{R}(t)}{\partial \rho_i} = -t\ \lambda_i \exp(-\rho_i t),$$
$$\frac{\partial^2 \hat{R}(t)}{\partial \rho_i^2} = t^2\ \lambda_i \exp(-\rho_i t), \qquad \frac{\partial^2 \hat{R}(t)}{\partial \rho_i \partial \lambda_i} = -t\ \exp(-\rho_i t),$$
$$\frac{\partial^2 \hat{R}(t)}{\partial \lambda_i \partial \lambda_j} = 0, \qquad \frac{\partial^2 \hat{R}(t)}{\partial \rho_i \partial \rho_j} = 0, \quad \frac{\partial^2 \hat{R}(t)}{\partial \lambda_i \partial \rho_j} = 0 \quad \text{if } i \neq j.$$

**General Multi-exponential curve**

If we relax the condition $\nu = 1 - \sum_{i=1}^p \lambda_i$ so that $\hat{R}(0)$ can be greater than unity, we obtain

$$\hat{R}(t) = \overline{\nu} + \sum_{i=1}^{p} \overline{\lambda}_i \exp(-\rho_i t),$$

with $\overline{\nu} \geq 0$, $\overline{\lambda}_i \geq 0$. The parameter $\pi = (\overline{\nu}, \overline{\lambda}_1, \cdots, \overline{\lambda}_p, \rho_1, \cdots, \rho_p)$ is then $(2p+1)$-dimensional. The gradient and Hessian are given by

$$\frac{\partial \hat{R}(t)}{\partial \overline{\nu}} = 1, \qquad \frac{\partial^2 \hat{R}(t)}{\partial \overline{\nu}^2} = 0, \ \frac{\partial^2 \hat{R}(t)}{\partial \overline{\nu} \partial \overline{\lambda}_i} = 0, \ \frac{\partial^2 \hat{R}(t)}{\partial \overline{\nu} \partial \rho_i} = 0,$$
$$\frac{\partial \hat{R}(t)}{\partial \overline{\lambda}_i} = \exp(-\rho_i t), \qquad \frac{\partial \hat{R}(t)}{\partial \rho_i} = -t\ \overline{\lambda}_i \exp(-\rho_i t),$$
$$\frac{\partial^2 \hat{R}(t)}{\partial \rho_i^2} = t^2\ \overline{\lambda}_i \exp(-\rho_i t), \qquad \frac{\partial^2 \hat{R}(t)}{\partial \rho_i \partial \overline{\lambda}_i} = -t\ \exp(-\rho_i t),$$
$$\frac{\partial^2 \hat{R}(t)}{\partial \overline{\lambda}_i \partial \overline{\lambda}_j} = 0, \qquad \frac{\partial^2 \hat{R}(t)}{\partial \rho_i \partial \rho_j} = 0, \quad \frac{\partial^2 \hat{R}(t)}{\partial \overline{\lambda}_i \partial \rho_j} = 0 \quad \text{if } i \neq j.$$

## 2.8 Appendix : Maximum Likelihood Estimation for the Hawkes intensity

The estimation of the Hawkes parameters, as presented in Section 2.3.4, resorts to Maximum Likelihood Estimation. The use of the MLE for Hawkes processes is well known, see for instance Ozaki [94], and has been recently considered by Da Fonseca and Zaatour [45] in a similar financial framework. In this section, we give the formula of the log-likelihood for Hawkes processes, and we derive its gradient and Hessian matrix which are necessary to use Newton-Raphson's algorithm.

We define the jump processes $J_t^+ = \sum_{0<\tau<t} \mathbb{1}_{\{\Delta N_\tau > 0\}}$ and $J_t^- = \sum_{0<\tau<t} \mathbb{1}_{\{\Delta N_\tau < 0\}}$, i.e. $J^+$ (resp. $J^-$) makes a unit jump when $N^+$ (resp. $N^-$) jumps. Say that we observe the realization of the process on the time interval $[0, T]$, and that we want to maximize its log-likelihood on $[t_0, T]$, with $t_0 \in [0, T)$. Conditionally to $(\kappa_t^\pm)_{t\in[0,T]}$, the log-likelihood of a trajectory $(J_t^\pm)_{t\in[t_0,T]}$ on the time interval $[t_0, T]$ is (see [46], Section III Proposition 7.2)

$$\ln \mathcal{L}(J^\pm | \kappa^\pm) = \int_{t_0}^{T} \ln(\kappa_{t^-}^\pm)\ \mathrm{d}J_t^\pm \ - \ \int_{t_0}^{T} \kappa_t^\pm\ \mathrm{d}t \ + \ T. \tag{2.37}$$

Moreover, conditionally to $(\kappa_t^+, \kappa_t^-)_{t\in[0,T]}$, the global log-likelihood of the model is

$$\ln \mathcal{L}(J|\kappa) = \ln \mathcal{L}(J^+|\kappa^+) + \ln \mathcal{L}(J^-|\kappa^-). \tag{2.38}$$

We now compute $\ln \mathcal{L}(J^+|\kappa^+)$. Since we do not know the history of the process before time $t = 0$, it is impossible to compute $\kappa_t^+$ exactly using equation (2.5) since it requires to know all the jumps. However, a reasonable approximation is to choose $t_0 \in (0, T)$ such that

$$\forall u \geq t_0,\ K(u) \ll 1,$$

which yields

$$\kappa_t^+ \approx \kappa_\infty + \sum_{0<\tau<t} K(t-\tau)\left[\mathbb{1}_{\{\Delta N_t>0\}}\varphi_{\mathrm{s}}(\Delta N_t/m_1) + \mathbb{1}_{\{\Delta N_t<0\}}\varphi_{\mathrm{c}}(-\Delta N_t/m_1)\right] \tag{2.39}$$

for $t \in [t_0, T]$. Let us assume in the sequel of this section that $t_0$ is such that (2.39) can be considered as an equality.

We define $\tau_0 = 0$ and $\tau_i,\ i \geq 1$ the ordered combined jump times of $N^+$ and $N^-$ on $[0, T]$, and $\chi(t) = \max\{i \geq 0, \tau_i \leq t\}$ for $t \in [0, T]$. We also define for $i \geq 1$

$$\theta_i^+ = \varphi_{\mathrm{s}}(\Delta N_{\tau_i}^+/m_1)k_i^+ + \varphi_{\mathrm{c}}(\Delta N_{\tau_i}^-/m_1)k_i^-,$$

where $k_i^+ = 1$ if $\tau_i$ is a jump time of $N^+$, $k_i^+ = 0$ otherwise, and $k_i^-$ is defined similarly with $N^-$. One has for $t \in [t_0, T]$

$$\kappa_t^+ = \kappa_\infty + \sum_{j=1}^{\chi(t)} \theta_j^+ K(t-\tau_j).$$

Distinguishing the jumps before and after $t_0$, we get

$$\int_{t_0}^T \kappa_t^+ \,\mathrm{d}t \;=\; \kappa_\infty(T-t_0) \;+\; \sum_{j=1}^{\chi(t_0)} \theta_j^+ \left[\underline{K}(T-\tau_j) - \underline{K}(t_0-\tau_j)\right] \;+\; \sum_{j=\chi(t_0)+1}^{\chi(T)} \theta_j^+ \left[\underline{K}(T-\tau_j) - \underline{K}(0)\right], \tag{2.40}$$

where $\underline{K}$ is the antiderivative of $K$. Let us turn to the other term of the log-likelihood. We set $A_1^+ = 0$ and for $i \geq 2$

$$A_i^+ = \sum_{j=1}^{i-1} \theta_j^+ K(\tau_i - \tau_j),$$

and we have

$$\int_{t_0}^T \ln(\kappa_{t^-}^+) \,\mathrm{d}J_t^+ \;=\; \sum_{i=\chi(t_0)+1}^{\chi(T)} k_i^+ \ln\left(\kappa_\infty + A_i^+\right). \tag{2.41}$$

We have the explicit expression of the log-likelihood $\ln \mathcal{L}(J^+|\kappa^+)$ from (2.38), (2.37), (2.40) and (2.41). Thus, it can be evaluated on a discrete set of points, for instance to estimate one or several parameters with a grid search. Now, to maximize the likelihood using Newton-Raphson's algorithm, one must also determine the gradient and Hessian matrix of $\ln \mathcal{L}(J^+|\kappa^+)$.

For given parameterizations of $\varphi_{\mathrm{s}}, \varphi_{\mathrm{c}}$ and $K$, we note $\pi$ an arbitrary parameter, and we have

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}(J^+|\kappa^+)}{\partial \kappa_\infty} &= \sum_{i=\chi(t_0)+1}^{\chi(T)} \frac{k_i^+}{\kappa_\infty + A_i^+} \; - \; (T - t_0), \\
\frac{\partial \ln \mathcal{L}(J^+|\kappa^+)}{\partial \pi} &= \sum_{i=\chi(t_0)+1}^{\chi(T)} \frac{k_i^+ \partial_\pi A_i^+}{\kappa_\infty + A_i^+} \\
&\quad - \sum_{j=1}^{\chi(t_0)} \partial_\pi \{\theta_j^+ \left[\underline{K}(T - \tau_j) - \underline{K}(t_0 - \tau_j)\right]\} \; - \sum_{j=\chi(t_0)+1}^{\chi(T)} \partial_\pi \{\theta_j^+ \left[\underline{K}(T - \tau_j) - \underline{K}(0)\right]\},
\end{aligned}$$

which yields the gradient of the log-likelihood. For the Hessian matrix, let us note $\pi, \pi'$ two parameters (distinct or not) of $\varphi_{\mathrm{s}}, \varphi_{\mathrm{c}}$ or $K$. We have

$$\begin{aligned}
\frac{\partial^2 \ln \mathcal{L}(J^+|\kappa^+)}{\partial \kappa_\infty^2} &= - \sum_{i=\chi(t_0)+1}^{\chi(T)} \frac{k_i^+}{[\kappa_\infty + A_i^+]^2}, \qquad \frac{\partial^2 \ln \mathcal{L}(J^+|\kappa^+)}{\partial \kappa_\infty \partial \pi} = - \sum_{i=\chi(t_0)+1}^{\chi(T)} \frac{k_i^+ \partial_\pi A_i^+}{[\kappa_\infty + A_i^+]^2}, \\
\frac{\partial^2 \ln \mathcal{L}(J^+|\kappa^+)}{\partial \pi \partial \pi'} &= \sum_{i=\chi(t_0)+1}^{\chi(T)} k_i^+ \left( \frac{\partial^2_{\pi\pi'} A_i^+}{\kappa_\infty + A_i^+} - \frac{\partial_\pi A_i^+ \, \partial_{\pi'} A_i^+}{[\kappa_\infty + A_i^+]^2} \right) \\
&\quad - \sum_{j=1}^{\chi(t_0)} \partial^2_{\pi\pi'} \{\theta_j^+ \left[\underline{K}(T - \tau_j) - \underline{K}(t_0 - \tau_j)\right]\} \; - \sum_{j=\chi(t_0)+1}^{\chi(T)} \partial^2_{\pi\pi'} \{\theta_j^+ \left[\underline{K}(T - \tau_j) - \underline{K}(0)\right]\}.
\end{aligned}$$

As soon as $\underline{K}$ is known and $\varphi_{\mathrm{s}}, \varphi_{\mathrm{c}}, K, \underline{K}$ are twice differentiable w.r.t. the parameterization, it is straightforward to deduce the analytical expressions of the gradient and Hessian matrix of the log-likelihood from the preceding equations.

## 2.9 Appendix : Optimal execution with a multi-exponential Hawkes kernel

### 2.9.1 Proof of Theorem 2.2.1

First, let us remark that $\mathbb{E}\left[\int_0^T W_t dX_t - W_T X_T\right] = 0$, and we can assume without loss of generality that $\sigma = 0$. We decompose the price process as follows. We introduce $dS_t^N = \frac{\nu}{q} \mathrm{d}N_t$, $\mathrm{d}D_t^{N,i} = -\rho_i D_t^{N,i} \mathrm{d}t + \frac{\lambda_i}{q} \mathrm{d}N_t$, $dS_t^X = \frac{\nu}{q} \mathrm{d}X_t$ and $\mathrm{d}D_t^{X,i} = -\rho_i D_t^{X,i} \mathrm{d}t + \frac{\lambda_i}{q} \mathrm{d}X_t$, with $S_0^N = S_0$, $D_0^{N,i} = D_0^i$, $S_0^X = D_0^{X,i} = 0$. We have

$$P_t = P_t^X + P_t^N, \text{ with } P_t^N = S_t^N + \sum_{i=1}^p D_t^{N,i}, \; P_t^X = S_t^X + \sum_{i=1}^p D_t^{X,i}.$$

Then, we can write the cost (2.15) as

$$C(X) = \int_{[0,T)} P_u^N \mathrm{d}X_u - P_T^N X_T + \bar{C}(X),$$

where $\bar{C}(X) = \int_{[0,T)} P_u^X \mathrm{d}X_u + \frac{1}{2q} \sum_{\tau \in \mathcal{D}_X \cap [0,T)} (\Delta X_\tau)^2 - P_T^X X_T + \frac{1}{2q} X_T^2$. We note that $\bar{C}(X)$ is a deterministic function of $X$ and is precisely the cost function considered in [4]. Besides, it satisfies $\bar{C}(cX) = c^2 \bar{C}(X)$ for $c \in \mathbb{R}$. By the same argument as in the proof of Theorem 2.1 in [3], we get that there is no PMS if, and only if $P_t$ is a $(\mathcal{F}_t)$-martingale when $X_t = 0$ for any $t$.

We now consider that $X \equiv 0$ and write the martingale condition for $P$ under the Hawkes model (2.9), (2.10) and (2.11). We have

$$\mathrm{d}P_t = \mathrm{d}S_t + \mathrm{d}D_t + \sigma \, \mathrm{d}W_t \;=\; \frac{1}{q} \, \mathrm{d}N_t - \sum_{i=1}^{p} \rho_i D_t^i \, \mathrm{d}t + \sigma \, \mathrm{d}W_t \;=\; \frac{1}{q} \, \mathrm{d}\tilde{N}_t + \sigma \, \mathrm{d}W_t + \; \mathrm{d}t \sum_{i=1}^{p} A_t^i,$$

where

$$A_t^i = \frac{m_1}{q} \delta_t^i - \rho_i D_t^i, \;\; \delta_t^i = \kappa_t^{+(i)} - \kappa_t^{-(i)},$$

and $\tilde{N}_t = N_t - m_1 \int_0^t \delta_u \mathrm{d}u$ is a martingale. Then, $(P_t)$ is a martingale if and only if almost surely and $\mathrm{d}t$-almost everywhere, $\sum_{i=1}^p A_t^i = 0$. We have

$$\mathrm{d}A_t^i = -\rho_i A_t^i \, \mathrm{d}t + \frac{m_1}{q} w_i \, \mathrm{d}I_t - \frac{\lambda_i \rho_i}{q} \, \mathrm{d}N_t,$$

with

$$I_t = \int_0^t \left[ (\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(\mathrm{d}N_u^+ / m_1) - (\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(\mathrm{d}N_u^- / m_1) \right]. \tag{2.42}$$

In particular, $\mathrm{d}A_t^i = -\rho_i A_t^i \mathrm{d}t$ between two consecutive jumps $\tau$ and $\tau'$ of $N$. Therefore, we have $\sum_{i=1}^p A_t^i = \sum_{i=1}^p A_\tau^i e^{-\rho_i (t-\tau)} = 0$ for $t \in [\tau, \tau')$ and therefore $A_\tau^i = 0$ for all $i$ (the equality for $t = \tau + k(\tau' - \tau)/p, k \in \{0, \dots, p-1\}$ gives a Vandermonde system). Thus, we necessarily have $A_t^i = 0$ for $t \geq 0$ for any $i$. Then, $\mathrm{d}A_t^i = 0$ gives

$$\frac{m_1}{q} w_i \, \left[ (\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(\mathrm{d}N_t^+ / m_1) - (\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}})(\mathrm{d}N_t^- / m_1) \right] \;=\; \frac{\lambda_i \rho_i}{q} \, \left[ \mathrm{d}N_t^+ - \mathrm{d}N_t^- \right]$$

for all $t \geq 0$ and all $i \in \{1, \cdots, p\}$. Thus, $\varphi_{\mathrm{s}} - \varphi_{\mathrm{c}}$ must be linear on the support of the law $\mu$ of the jumps of $N^\pm$, and besides, we must have $\forall i, \quad (\iota_{\mathrm{s}} - \iota_{\mathrm{c}}) w_i = \lambda_i \rho_i$. This precisely gives (2.16). Conversely, it is clear that (2.16) ensures that $P$ is a martingale by the same calculations.

### 2.9.2 Proof of Theorem 2.2.2

As in Section (2.9.1), we assume without loss of generality that $\sigma = 0$. We first introduce some notations to present the main results on the optimal execution. We define $\delta_t^i = \kappa_t^{+(i)} - \kappa_t^{-(i)}$ and $\Sigma_t^i = \kappa_t^{+(i)} + \kappa_t^{-(i)}$. From (2.9), (2.10) and (2.20), we have

$$\mathrm{d}\delta_t^i \;=\; -\beta_i \, \delta_t^i \, \mathrm{d}t \;+\; w_i \, \mathrm{d}I_t \quad , \quad \mathrm{d}\Sigma_t^i \;=\; -\beta_i \, (\Sigma_t^i - 2\kappa_\infty / p) \, \mathrm{d}t \;+\; w_i \, \mathrm{d}\bar{I}_t, \tag{2.43}$$

for all $i \in \{1, \cdots, p\}$, where $\bar{I}_t = \int_0^t [(\varphi_{\mathrm{s}} + \varphi_{\mathrm{c}})(\mathrm{d}N_u^+ / m_1) + (\varphi_{\mathrm{s}} + \varphi_{\mathrm{c}})(\mathrm{d}N_u^- / m_1)]$ and $I_t$ is defined by (2.42).

We now proceed exactly as in [3], Appendix B, and only give here the main lines and use similar notations. We assume without loss of generality $q = 1$. For $t \in [0,T]$, $x, d, z \in \mathbb{R}$ and $\delta, \Sigma \in \mathbb{R}^p$, we denote by $\mathcal{C}(t,x,d,z,\delta,\Sigma)$ the minimal cost to liquidate $X_t = x$ over the time interval $[t,T]$ when $D_t = d$, $S_t = z$, $\delta_t = \delta$ and $\Sigma_t = \Sigma$. We look for a function that has the following form

$$\begin{aligned}\mathcal{C}(t,x,d,z,\delta,\Sigma) &= a(T-t)(d-(1-\nu)x)^2 \;+\; \frac{1}{2}(z-\nu x)^2 \;+\; (d-(1-\nu)x)(z-\nu x) \;-\; \frac{(d+z)^2}{2} \\ &\quad + (d-(1-\nu)x)\sum_{i=1}^{p} b_i(T-t)\,\delta_i \;+\; \sum_{i=1}^{p}\sum_{i=1}^{p} c_{i,j}(T-t)\,\delta_i\delta_j \\ &\quad + \sum_{i=1}^{p} e_i(T-t)\,\Sigma_i \;+\; g(T-t), \end{aligned} \tag{2.44}$$

with $a, b_i, c_{i,j}, e_i, g : \mathbb{R}_+ \to \mathbb{R}$ continuously differentiable functions. We have the limit condition $\mathcal{C}(T,x,d,z,\delta,\Sigma) = -(d+z)x + x^2/2 = \frac{1}{2}(d+z-x)^2 - (d+z)^2/2$, which is the cost of a trade of signed volume $-x$. We thus have

$$a(0) = \frac{1}{2},\; b_i(0) = c_{i,j}(0) = e(0) = g(0) = 0.$$

For an arbitrary strategy $X$, we define $\Pi_t(X) = \int_0^t P_u dX_u + \frac{1}{2}\sum_{0\le\tau<t}(\Delta X_\tau)^2 + \mathcal{C}(t, X_t, D_t, S_t\delta_t, \Sigma_t)$. This is the cost of the strategy which is equal to $X$ up to time $t$ and is then optimal. Therefore, $(\Pi_t(X), t \in [0,T])$ has to be a submartingale and is a martingale if, and only if, $X$ is optimal. We define

$$\mathrm{d}A_t^X = \left[ Z(t, X_t, D_t, S_t, \delta_t, \Sigma_t) + \partial_t\mathcal{C} - \rho D_t\partial_d\mathcal{C} - \sum_{i=1}^{p}\beta_i\delta_t^i\partial_{\delta_i}\mathcal{C} - \sum_{i=1}^{p}\beta_i(\Sigma_t^i - 2\kappa_\infty/p)\partial_{\Sigma_i}\mathcal{C}\right]\mathrm{d}t, \tag{2.45}$$

where the derivatives of $\mathcal{C}$ are taken in $(t, X_t, D_t, S_t\delta_t, \Sigma_t)$ and $Z(t,x,d,z,\delta,\Sigma) :=$

$$\begin{aligned}&\left(\frac{1}{2}\sum_{i=1}^{p}[\Sigma_i + \delta_i]\right)\mathbb{E}\big[\mathcal{C}(t,x,d+(1-\nu)V, z+\nu V, \delta+\varphi_{s-c}(V/m_1)w, \Sigma+\varphi_{s+c}(V/m_1)w) - \mathcal{C}(t,x,d,z,\delta,\Sigma)\big] \\ +\;&\left(\frac{1}{2}\sum_{i=1}^{p}[\Sigma_i - \delta_i]\right)\mathbb{E}\big[\mathcal{C}(t,x,d-(1-\nu)V, z-\nu V, \delta-\varphi_{s-c}(V/m_1)w, \Sigma+\varphi_{s+c}(V/m_1)w) - \mathcal{C}(t,x,d,z,\delta,\Sigma)\big],\end{aligned}$$

with $V \sim \mu$, $\varphi_{s-c} = \varphi_{\mathrm{s}} - \varphi_{\mathrm{c}}$ and $\varphi_{s+c} = \varphi_{\mathrm{s}} + \varphi_{\mathrm{c}}$. The process $A_t^X$ is continuous and such that $\Pi_t(X) - A_t^X$ is a martingale. Given the quadratic nature of the problem, we search a process $A^X$ of the form

$$\mathrm{d}A_t^X \;=\; \frac{\rho}{1-\nu}\mathrm{d}t \times \left[j(T-t)(D_t - (1-\nu)X_t) \;-\; D_t \;+\; \sum_{i=1}^{p} k_i(T-t)\,\delta_t^i\right]^2. \tag{2.46}$$

We now introduce the variables $y = d-(1-\nu)x$ and $\xi = z - \nu x$ and work with $(y, d, \xi, \delta, \Sigma)$ instead

of $(x, d, z, \delta, \Sigma)$. From (2.44) and the definition of $Z$, we have

$$\partial_t \mathcal{C}(t,x,d,z,\delta,\Sigma) = -\dot{a}\ y^2 \ - \ y\sum \dot{b}_i\ \delta_i \ - \ \sum\sum \dot{c}_{i,j}\ \delta_i\delta_j \ - \ \sum \dot{e}_i\ \Sigma_i \ - \ \dot{g},$$

$$-\rho d\ \partial_d \mathcal{C}(t,x,d,z,\delta,\Sigma) = -\left(2\rho a + \frac{\rho\nu}{1-\nu}\right)\ dy \ + \ \frac{\rho}{1-\nu}\ d^2 \ - \ \rho d\sum b_i\delta_i,$$

$$-\beta_i\delta_i\ \partial_{\delta_i}\mathcal{C}(t,x,d,z,\delta,\Sigma) = -\beta_i b_i\ \delta_i y \ - \ \beta_i\delta_i\left[2c_{i,i}\delta_i + \sum_{j\neq i} c_{i,j}\delta_j\right],$$

$$-\beta_i(\Sigma_i - 2\kappa_\infty/p)\ \partial_{\Sigma_i}\mathcal{C}(t,x,d,z,\delta,\Sigma) = -\beta_i e_i\ \Sigma_i \ + \ 2\beta_i\kappa_\infty e_i/p,$$

$$\begin{aligned} Z(t,x,d,z,\delta,\Sigma) = &\left(m_1 \times \left[2(1-\nu)a + \nu + \frac{\nu}{1-\nu}\right] \ + \ \sum_{k=1}^p \alpha_k b_k\right)\ y\sum_{i=1}^p \delta_i \ - \ \frac{m_1}{1-\nu}\ d\sum_{i=1}^p \delta_i \\ &+ \sum_{i=1}^p\sum_{j=1}^p \left[(1-\nu)m_1 b_i \ + \ 2\sum_{k=1}^p c_{i,k}\alpha_k\right]\ \delta_i\delta_j \\ &+ \sum_{i=1}^p \left(m_2 \times \left[(1-\nu)^2 a + \nu(1-\nu/2) - \frac{1}{2}\right] + (1-\nu)\sum_{k=1}^p \tilde{\alpha}_k b_k + \hat{\alpha}\sum_{k=1}^p\sum_{l=1}^p c_{k,l}w_k w_l + \sum_{k=1}^p(\alpha_k + 2w_k\iota_c)e_k\right)\Sigma_i, \end{aligned}$$

with $\tilde{\alpha} = \mathbb{E}[V\times(\varphi_s - \varphi_c)(V/m_1)]$, $\hat{\alpha} = \mathbb{E}[(\varphi_s-\varphi_c)^2(V/m_1)]$. We now identify each term of equations (2.45) and (2.46).

**(Eq. $dy$) :** $-\left(2\rho a + \frac{\rho\nu}{1-\nu}\right) = -\frac{2\rho}{1-\nu}j$, **(Eq. $y^2$) :** $-\dot{a} \ = \ \frac{\rho}{1-\nu}j^2$.

These two equations are the same as in [3] and give

$$j(u) = \frac{1}{2+\rho u} \text{ and } a(u) = \frac{1}{1-\nu}\left(\frac{1}{2+\rho u} - \frac{\nu}{2}\right). \tag{2.47}$$

**(Eq. $\delta_i y$) :** $-\dot{b}_i \ - \ \beta_i b_i \ + \ \sum_{k=1}^p \alpha_k b_k \ + \ m_1 \times \left[2(1-\nu)a + \nu + \frac{\nu}{1-\nu}\right] \ = \ \frac{2\rho}{1-\nu}jk_i.$

**(Eq. $\delta_i d$) :** $-\rho b_i \ - \ \frac{m_1}{1-\nu} \ = \ -\frac{2\rho}{1-\nu}k_i,$

which yields $k_i(u) \ = \ \frac{1-\nu}{2}\ b_i(u) \ + \ \frac{m_1}{2\rho}$. Plugging this in (Eq. $\delta_i y$), we have $\dot{b}_i = -\beta_i b_i + \sum_{k=1}^p \alpha_k b_k - \frac{2\rho}{1-\nu}j\left(\frac{1-\nu}{2}b_i + \frac{m_1}{2\rho}\right) + m_1\left[2(1-\nu)a + \nu + \frac{\nu}{1-\nu}\right]$, and since $j/(1-\nu) \ = \ a + \nu/[2(1-\nu)]$, we have $\dot{b}_i(u) \ = \ -\beta_i b_i(u) + \sum_{k=1}^p \alpha_k b_k(u) - \frac{\rho}{2+\rho u}b_i(u) \ + \ \frac{m_1}{1-\nu} \times \frac{1+\nu\rho u}{2+\rho u}$. We rewrite it as

$$\dot{b}(u) \ = \ \left[-H - \frac{\rho}{2+\rho u}I_p\right] b(u) \ + \ \frac{m_1}{1-\nu} \times \frac{1+\nu\rho u}{2+\rho u} \times (1,\cdots,1)^\top, \tag{2.48}$$

where $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix and $H \in \mathbb{R}^{p\times p}$ is given by (2.21). To solve equation (2.48), we search a solution of the form $b(u) = \frac{1}{2+\rho u} \times [\exp(-uH)\ .\ \tilde{b}(u)]$ for $u \geq 0$. This yields

$$\frac{1}{2+\rho u} \times [\exp(-uH)\ .\ \dot{\tilde{b}}(u)] \ = \ \frac{m_1}{1-\nu} \times \frac{1+\nu\rho u}{2+\rho u} \times (1,\cdots,1)^\top,$$

thus

$$\dot{\tilde{b}}(u) \ = \ \frac{m_1}{1-\nu} \times (1+\nu\rho u) \times [\exp(uH)\ .\ (1,\cdots,1)^\top].$$

From the definition (2.22), we have $\exp(-uH).\left[\int_0^u (1+\nu\rho s)\times\exp(sH)\mathrm{d}s\right] = u\zeta(uH)+\nu\rho u^2\omega(uH)$ for $u\geq 0$. Since $\tilde{b}(0)=2b(0)=0$, we obtain

$$b(u) = \frac{m_1 u}{1-\nu}\times\frac{1}{2+\rho u}\times[\{\zeta(uH)+\nu\rho u\ \omega(uH)\}\ .\ (1,\cdots,1)^\top]. \tag{2.49}$$

Equation (Eq : $\delta_i d$) then gives the vector function $k(u)$

$$k(u) = \frac{m_1}{2\rho}\times\left\{I_p+\frac{\rho u}{2+\rho u}\times[\zeta(uH)+\nu\rho u\ \omega(uH)]\right\}\ .\ (1,\cdots,1)^\top. \tag{2.50}$$

Thus, the functions $j$ and $k$ involved in (2.46) are explicit, which guarantees that the optimal strategy is obtained as a closed formula.

The remaining functions $c_{i,j}$, $e_i$ and $g$ do not play any role to determine the optimal strategy. By identifying the terms in $\delta_i\delta_j$, $\Sigma_i$ and the constant term, we check that they solve a system of linear ODEs. They are thus uniquely determined and well-defined on $\mathbb{R}_+$, and the cost function $\mathcal{C}$ is well-defined. Thes ODEs are also important to run the verification argument, i.e. to check that $\mathcal{C}$ is indeed the optimal cost function and that the strategy $X^*$ described below is the optimal one.

We now determine the strategy $X^*$ such that $\Pi(X^*)$ is a martingale, or equivalently such that $A^{X^*}$ is constant. Equations (2.46) and (2.47) yield

$$\begin{aligned}\mathrm{d}A_t^X &= \frac{\rho}{1-\nu}\mathrm{d}t\times\left[\frac{D_t-(1-\nu)X_t}{2+\rho(T-t)}\ -\ D_t\ +\ \sum_{i=1}^p k_i(T-t)\ \delta_t^i\right]^2\\ &= \frac{\rho/(1-\nu)}{[2+\rho(T-t)]^2}\ \mathrm{d}t\times\left[(1-\nu)X_t\ +\ [1+\rho(T-t)]\ D_t\ -\ [2+\rho(T-t)]\ \sum_{i=1}^p k_i(T-t)\ \delta_t^i\right]^2.\end{aligned}$$

Thus, $A^{X^*}$ is constant on $(0,T)$ if, and only if

$$\text{a.s. , } \mathrm{d}t\text{ -a.e. on }(0,T)\ ,\quad (1-\nu)X_t^* = -\ [1+\rho(T-t)]\ D_t\ +\ [2+\rho(T-t)]\ \sum_{i=1}^p k_i(T-t)\ \delta_t^i. \tag{2.51}$$

This equation characterizes the optimal strategy. In particular, we obtain its initial jump $\Delta X_0^*$ at time $t=0$

$$(1-\nu)\Delta X_0^* = -\frac{[1+\rho T]qD_0+x_0}{2+\rho T}+\frac{m_1}{2\rho}\times\left[(1,\cdots,1)\ .\ \left\{I_p+\frac{\rho T}{2+\rho T}\times[\zeta(TH)+\nu\rho T\ \omega(TH)]\right\}\ .\ \delta_0\right].$$

where $\delta_0=(\delta_0^1,\cdots,\delta_0^p)^\top\in\mathbb{R}^p$.

# Deuxième partie

# Modèles auto-régressifs de volatilité intra-journalière

# Chapitre 3

# Structure fine de la volatilité rétroactive : effets intra-journaliers et nocturnes

Ce chapitre est un article écrit avec Rémy Chicheportiche et Jean-Philippe Bouchaud [27] et publié dans la revue *Physica A : Statistical Mechanics and its Applications*.

**Abstract.** We decompose, within an ARCH framework, the daily volatility of stocks into overnight and intra-day contributions. We find, as perhaps expected, that the overnight and intra-day returns behave completely differently. For example, while past intra-day returns affect equally the future intra-day and overnight volatilities, past overnight returns have a weak effect on future intra-day volatilities (except for the very next one) but impact substantially future overnight volatilities. The exogenous component of overnight volatilities is found to be close to zero, which means that the lion's share of overnight volatility comes from feedback effects. The residual kurtosis of returns is small for intra-day returns but infinite for overnight returns. We provide a plausible interpretation for these findings, and show that our Intra-day/Overnight model significantly outperforms the standard ARCH framework based on daily returns for Out-of-Sample predictions.

## 3.1 Introduction

The ARCH (auto-regressive conditional heteroskedastic) framework was introduced in [53] to account for volatility clustering in financial markets and other economic time series. It posits that the current relative price change $r_t$ can be written as the product of a "volatility" component $\sigma_t$ and a certain random variable $\xi_t$, of zero mean and unit variance, and that the dynamics of the volatility is self-referential in the sense that it depends on the past returns themselves as :

$$\sigma_t^2 = s^2 + \sum_{\tau=1}^{q} K(\tau)\, r_{t-\tau}^2 \equiv s^2 + \sum_{\tau=1}^{q} K(\tau)\, \sigma_{t-\tau}^2 \xi_{t-\tau}^2, \tag{3.1}$$

where $s^2$ is the "baseline" volatility level, that would obtain in the absence of any feedback from the past, and $K(\tau)$ is a kernel that encodes the strength of the influence of past returns. The model is well defined and leads to a stationary time series whenever the feedback is not too strong, i.e. when

$\sum_{\tau=1}^{q} K(\tau) < 1$. A very popular choice, still very much used both in the academic and professional literature, is the so called "GARCH" (Generalized ARCH), that corresponds to an exponential kernel, $K(\tau) = ge^{-\tau/\tau_p}$, with $q \to \infty$. However, the long-range memory nature of the volatility correlations in financial markets suggests that a power-law kernel is more plausible — a model called "FIGARCH", see [19, 29] and below.

Now, the ARCH framework implicitly singles out a *time scale*, namely the time interval over which the returns $r_t$ are defined. For financial applications, this time scale is often chosen to be one day, i.e. the ARCH model is a model for daily returns, defined for example as the relative variation of price between two successive closing prices. However, this choice of a day as the unit of time is often a default imposed by the data itself. A natural question is to know whether other time scales could also play a role in the volatility feedback mechanism. In our companion paper [38], we have studied this question in detail, focusing on time scales *larger than* (or equal to) the day. We have in fact calibrated the most general model, called "QARCH" [102], that expresses the squared volatility as a quadratic form of past returns, i.e. with a two-lags kernel $K(\tau, \tau')r_{t-\tau}r_{t-\tau'}$ instead of the "diagonal" regression (3.1). This encompasses all models where returns defined over arbitrary time intervals could play a role, as well as (realized) correlations between those — see [38] and references therein for more precise statements, and [30, 92, 103, 111] for earlier contributions along these lines.

The main conclusion of our companion paper [38] is that while other time scales play a statistically significant role in the feedback process (interplay between $r_{t-\tau}$ and $r_{t-\tau'}$ resulting in non-zero off-diagonal elements $K(\tau, \tau')$), the *dominant effect* for daily returns is indeed associated with past daily returns. In a first approximation, a FIGARCH model based on daily returns, with an exponentially truncated power-law kernel $K(\tau) = g\tau^{-\alpha}e^{-\tau/\tau_p}$, provides a good model for stock returns, with $\alpha \approx 1.1$ and $\tau_p \approx 50$ days. This immediately begs the question : if returns on time scales larger than a day appear to be of lesser importance,[1] what about returns on time scales smaller than a day ? For one thing, a trading day is naturally decomposed into trading hours, that define an 'Open to Close' (or 'intra-day') return, and hours where the market is closed but news accumulates and impacts the price at the opening auction, contributing to the 'Close to Open' (or 'overnight') return. One may expect that the price dynamics is very different in the two cases, for several reasons. One is that many company announcements are made overnight, that can significantly impact the price. The profile of market participants is also quite different in the two cases : while low-frequency participants might choose to execute large volumes during the auction, higher-frequency participants and market-makers are mostly active intra-day. In any case, it seems reasonable to distinguish two volatility contributions, one coming from intra-day trading, the second one from overnight activity. Similarly, the feedback of past returns should also be disentangled into an intra-day contribution and an overnight contribution. The calibration of an ARCH-like model that distinguishes between intra-day and overnight returns is the aim of the present paper, and is the content of Section 3.2. We have in fact investigated the role of higher frequency returns as well. For the sake of clarity we will not present this study here, but rather summarize briefly our findings on this point in the conclusion.

The salient conclusions of the present paper are that the intra-day and overnight dynamics are indeed completely different — for example, while the intra-day (Open-to-Close) returns impact both the

1. Note a possible source of confusion here since a FIGARCH model obviously involves many time scales. We need to clearly distinguish time *lags*, as they appear in the kernel $K(\tau)$, from time scales, that enter in the definition of the returns themselves.

future intra-day and overnight volatilities in a slowly decaying manner, overnight (Close-to-Open) returns essentially impact the next intra-day but very little the following ones. However, overnight returns have themselves a slowly decaying impact on future overnights. Another notable difference is the statistics of the residual factor $\xi_t$, which is *nearly Gaussian* for intra-day returns, but has an *infinite kurtosis* for overnight returns. We discuss further the scope of our results in the conclusion Section 3.4, and relegate several more technical details to appendices.

## 3.2 The dynamics of Close-to-Open and Open-to-Close stock returns and volatilities

Although the decomposition of the daily (Close-to-Close) returns into their intra-day and overnight components seems obvious and intuitive, very few attempts have actually been made to model them jointly (see [61, 106]). In fact, some studies even discard overnight returns altogether. In the present section, we define and calibrate an ARCH model that explicitly treats these two contributions separately. We however first need to introduce some precise definitions of the objects that we want to model.

### 3.2.1 Definitions, time-line and basic statistics

We consider equidistant time stamps $t$ with $\Delta t = 1$ day. Every day, the prices of traded stocks are quoted from the opening to the closing hour, but we only keep track of the first and last traded prices. For every stock name $a$, $O_t^a$ is the open price and $C_t^a$ the close price at date $t$. (In the following, we drop the index $a$ when it is not explicitly needed). We introduce the following definitions of the geometric returns, volatilities, and residuals :

$$\text{Intra-day return} : r_t^{\mathrm{D}} = \ln(C_t/O_t) \equiv \sigma_t^{\mathrm{D}} \xi_t^{\mathrm{D}} \tag{3.2a}$$

$$\text{Overnight return} : r_t^{\mathrm{N}} = \ln(O_t/C_{t-1}) \equiv \sigma_t^{\mathrm{N}} \xi_t^{\mathrm{N}} \tag{3.2b}$$

$$\text{Daily return} : r_t = \ln(C_t/C_{t-1}) = r_t^{\mathrm{D}} + r_t^{\mathrm{N}} \equiv \sigma_t \xi_t. \tag{3.2c}$$

The following time-line illustrates the definition of the three types of return :

$$\ldots \longrightarrow C_{t-1} \;\Big|\; \underbrace{\xrightarrow{\text{Night } t}}_{r_t^{\mathrm{N}}} O_t \underbrace{\xrightarrow{\text{Day } t}}_{r_t^{\mathrm{D}}} C_t \;\Big|\; \underbrace{\xrightarrow{\text{Night } t+1}}_{r_{t+1}^{\mathrm{N}}} O_{t+1} \longrightarrow \ldots \tag{3.3}$$

$r_t$

To facilitate the reading of our tables and figures, intra-day returns are associated with the green color (or light gray) and overnight returns with blue (or dark gray).

Before introducing any model, we discuss the qualitative statistical differences in the series of Open-Close returns $r_t^{\mathrm{D}}$ and Close-Open returns $r_t^{\mathrm{N}}$. First, one can look at Fig. 3.1 for a visual impression of the difference : while the intra-day volatility is higher than the overnight volatility, the relative importance of « surprises » (i.e. large positive or negative jumps) is larger for overnight returns. This

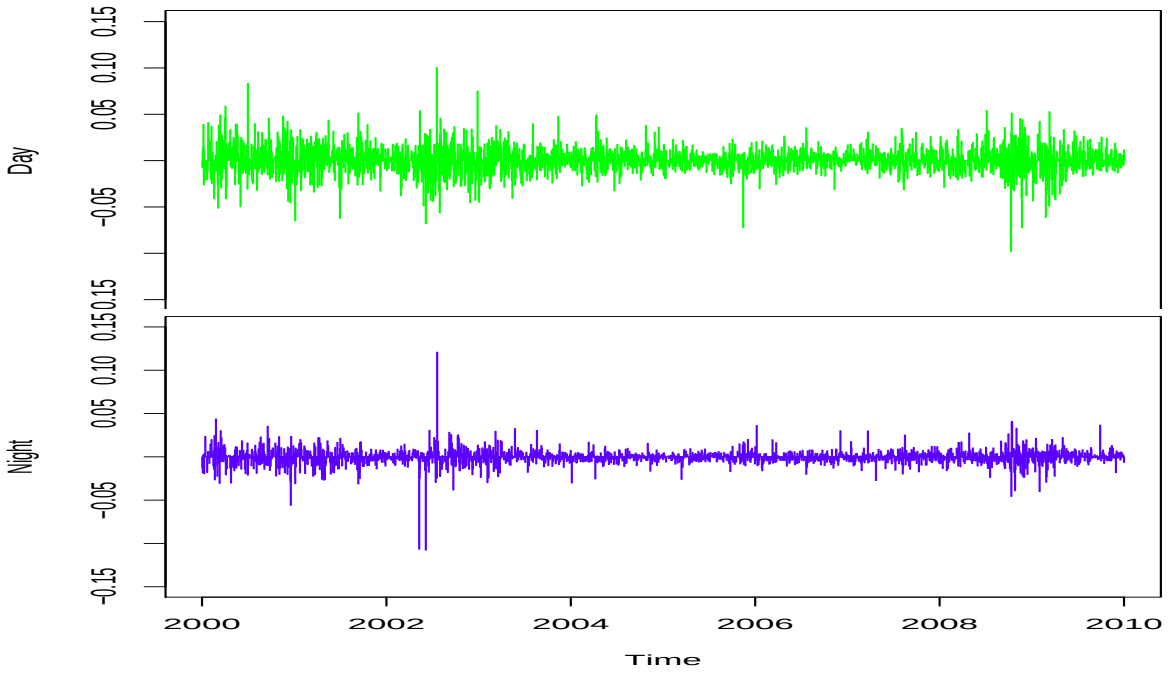



FIGURE 3.1 – Example of a historical time series of stock day returns (top) and overnight returns (bottom).

| J | $[\langle r^{\mathrm{J}}\rangle]$ | $\sigma^{\mathrm{J}} = \sqrt{[\langle r^{\mathrm{J}2}\rangle]}$ | $\left[\langle r^{\mathrm{J}3}\rangle / \langle r^{\mathrm{J}2}\rangle^{\frac{3}{2}}\right]$ | $\left[\langle r^{\mathrm{J}4}\rangle / \langle r^{\mathrm{J}2}\rangle^{2}\right]$ |
|---|---|---|---|---|
| D | $1.2\ 10^{-4}$ | 0.022 | −0.12 | 12.9 |
| N | $-1.0\ 10^{-4}$ | 0.013 | −1.5 | 62.6 |

TABLE 3.1 – Distributional properties of intra-Day and overNight returns (first four empirical moments). $\langle\cdot\rangle$ means average over all dates, and $[\cdot]$ average over all stocks.

is confirmed by the numerical values provided in Tab. 3.1 for the volatility, skewness and kurtosis of the two time series $r_t^{\mathrm{D}}$ and $r_t^{\mathrm{N}}$.

It is also visible on Fig. 3.1 that periods of high volatilities are common to the two series : two minor ones can be observed in the middle of year 2000 and at the beginning of year 2009, and an important one in the middle of year 2002.

An important quantity is the correlation between intra-day and overnight returns, which can be measured either as $\left[\langle r_t^{\mathrm{N}} r_t^{\mathrm{D}}\rangle\right] / \sigma^{\mathrm{N}}\sigma^{\mathrm{D}}$ (overnight leading intra-day) or as $\left[\langle r_t^{\mathrm{D}} r_{t+1}^{\mathrm{N}}\rangle\right] / \sigma^{\mathrm{N}}\sigma^{\mathrm{D}}$ (intra-day leading overnight). The statistical reversion revealed by the measured values of the above correlation coefficients ($-0.021$ and $-0.009$, respectively) is slight enough (compared to the amplitude and reach of the feedback effect) to justify the assumption of i.i.d. residuals. If there were no linear correlations between intra-day and overnight returns, the squared volatilities would be exactly additive, i.e. $\sigma_t^2 \equiv \sigma_t^{\mathrm{D}2} + \sigma_t^{\mathrm{N}2}$. Deviations from this simple addition of variance rule are below 2%.

### 3.2.2 The model

The standard ARCH model recalled in the introduction, Eq. (3.1), can be rewritten identically as :

$$\sigma_t^2 \equiv s^2 + \sum_{\tau=1}^{q} K(\tau) {r_{t-\tau}^{\mathrm{N}}}^2 + \sum_{\tau=1}^{q} K(\tau) {r_{t-\tau}^{\mathrm{D}}}^2 + 2\sum_{\tau=1}^{q} K(\tau) r_{t-\tau}^{\mathrm{N}} r_{t-\tau}^{\mathrm{D}}, \tag{3.4}$$

meaning that there is a unique kernel $K(\tau)$ describing the feedback of past intra-day and overnight returns on the current volatility level.

If however one believes that these returns are of fundamentally different nature, one should expand the model in two directions : first, the two volatilities $\sigma^{\mathrm{D}2}$ and $\sigma^{\mathrm{N}2}$ should have separate dynamics. Second, the kernels describing the feedback of past intra-day and overnight returns should *a priori* be different. This suggests to write the following generalized model for the intra-day volatility :

$$\begin{aligned}\sigma_t^{\mathrm{D}2} = s^{\mathrm{D}2} &+ \sum_{\tau=1}^{\infty} L^{\mathrm{D}\to\mathrm{D}}(\tau) r_{t-\tau}^{\mathrm{D}} + \sum_{\tau=1}^{\infty} K^{\mathrm{DD}\to\mathrm{D}}(\tau) {r_{t-\tau}^{\mathrm{D}}}^2 + 2\sum_{\tau=1}^{\infty} K^{\mathrm{ND}\to\mathrm{D}}(\tau) r_{t-\tau}^{\mathrm{D}} r_{t-\tau}^{\mathrm{N}} \\ &+ \sum_{\tau=0}^{\infty} L^{\mathrm{N}\to\mathrm{D}}(\tau+1) r_{t-\tau}^{\mathrm{N}} + \sum_{\tau=0}^{\infty} K^{\mathrm{NN}\to\mathrm{D}}(\tau+1) {r_{t-\tau}^{\mathrm{N}}}^2 + 2\sum_{\tau=0}^{\infty} K^{\mathrm{DN}\to\mathrm{D}}(\tau+1) r_{t-\tau-1}^{\mathrm{D}} r_{t-\tau}^{\mathrm{N}},\end{aligned} \tag{3.5}$$

where we have added the possibility of a "leverage effect", i.e. terms linear in past returns that can describe an asymmetry in the impact of negative and positive returns on the volatility. The notation used is, we hope, explicit : for example $K^{\mathrm{DD}\to\mathrm{D}}$ describes the influence of squared intra-Day past

returns on the current intra-Day volatility. Note that the mixed effect of intra-Day and overNight returns requires two distinct kernels, $K^{\mathrm{DN}\to\mathrm{D}}$ and $K^{\mathrm{ND}\to\mathrm{D}}$, depending on which comes first in time. Finally, the time-line shown above explains why the $\tau$ index starts at $\tau = 1$ for past intra-day returns, but at $\tau = 0$ for past overnight returns. We posit a similar expression for the overnight volatility :

$$\begin{aligned}\sigma_t^{\mathrm{N}2} = s^{\mathrm{N}2} &+ \sum_{\tau=1}^{\infty} L^{\mathrm{N}\to\mathrm{N}}(\tau) r_{t-\tau}^{\mathrm{N}} + \sum_{\tau=1}^{\infty} K^{\mathrm{NN}\to\mathrm{N}}(\tau) {r_{t-\tau}^{\mathrm{N}}}^2 + 2\sum_{\tau=1}^{\infty} K^{\mathrm{ND}\to\mathrm{N}}(\tau) r_{t-\tau}^{\mathrm{D}} r_{t-\tau}^{\mathrm{N}} \\ &+ \sum_{\tau=1}^{\infty} L^{\mathrm{D}\to\mathrm{N}}(\tau) r_{t-\tau}^{\mathrm{D}} + \sum_{\tau=1}^{\infty} K^{\mathrm{DD}\to\mathrm{N}}(\tau) {r_{t-\tau}^{\mathrm{D}}}^2 + 2\sum_{\tau=1}^{\infty} K^{\mathrm{DN}\to\mathrm{N}}(\tau) r_{t-\tau-1}^{\mathrm{D}} r_{t-\tau}^{\mathrm{N}}.\end{aligned} \tag{3.6}$$

The model is therefore fully characterized by two base-line volatilities $s^{\mathrm{D}}, s^{\mathrm{N}}$, four leverage (linear) kernels $L^{\mathrm{J'}\to\mathrm{J}}$, eight quadratic kernels $K^{\mathrm{J'J''}\to\mathrm{J}}$, and the statistics of the two residual noises $\xi^{\mathrm{D}}, \xi^{\mathrm{N}}$ needed to define the returns, as $r^{\mathrm{J}} = \sigma^{\mathrm{J}}\xi^{\mathrm{J}}$. We derive in Appendix 3.5 conditions on the coefficients of the model under which the two volatility processes remain positive at all times. The model as it stands has a large number of parameters ; in order to ease the calibration process and gain in stability, we in fact choose to parameterize the $\tau$ dependence of the different kernels with some simple functions, namely an exponentially truncated power-law for $K^{\mathrm{J'J''}\to\mathrm{J}}$ and a simple exponential for $L^{\mathrm{J'}\to\mathrm{J}}$ :

$$K(\tau) = g_{\mathrm{p}}\tau^{-\alpha}\,\exp(-\omega_{\mathrm{p}}\,\tau)\,; \qquad L(\tau) = g_{\mathrm{e}}\,\exp(-\omega_{\mathrm{e}}\,\tau)\,. \tag{3.7}$$

The choice of these functions is not arbitrary, but is suggested by a preliminary calibration of the model using a generalized method of moments (GMM), as explained in the companion paper, see Appendix C.2 in Ref. [38].

As far as the residuals $\xi_t^{\mathrm{D}}, \xi_t^{\mathrm{N}}$ are concerned, we assume them to be i.i.d. centered Student variables of unit variance with respectively $\nu^{\mathrm{D}} > 2$ and $\nu^{\mathrm{N}} > 2$ degrees of freedom. Contrarily to many previous studies, we prefer to be agnostic about the kurtosis of the residuals rather than imposing a priori Gaussian residuals. It has been shown that while the ARCH feedback effect accounts for volatility clustering and for some positive kurtosis in the returns, this effect alone is not sufficient to explain the observed heavy tails in the return distribution (see for example [38]). These tails come from true 'surprises' (often called jumps), that cannot be anticipated by the predictable part of the volatility, and that can indeed be described by a Student (power-law) distribution of the residuals.

### 3.2.3 Dataset

The dataset used to calibrate the model is exactly the same as in our companion paper [38]. It is composed of US stock prices (four points every day : Open, Close, High and Low) for $N = 280$ stocks present permanently in the S&P-500 index from Jan. 2000 to Dec. 2009 ($T = 2515$ days). For every stock $a$, the daily returns ($r_t^a$), intra-day returns ($r_t^{\mathrm{D}a}$) and overnight returns ($r_t^{\mathrm{N}a}$) defined in Eq. (3.2) are computed using only Open and Close prices at every date $t$. In order to improve the statistical significance of our results, we consider the pool of stocks as a statistical ensemble over which we can average. This means that we assume a universal dynamics for the stocks, a reasonable assumption as we discuss in Appendix 3.6.

Bare stock returns are "polluted" by several obvious and predictable events associated with the life of the company, such as quarterly announcements. They also reflect low-frequency human activity,

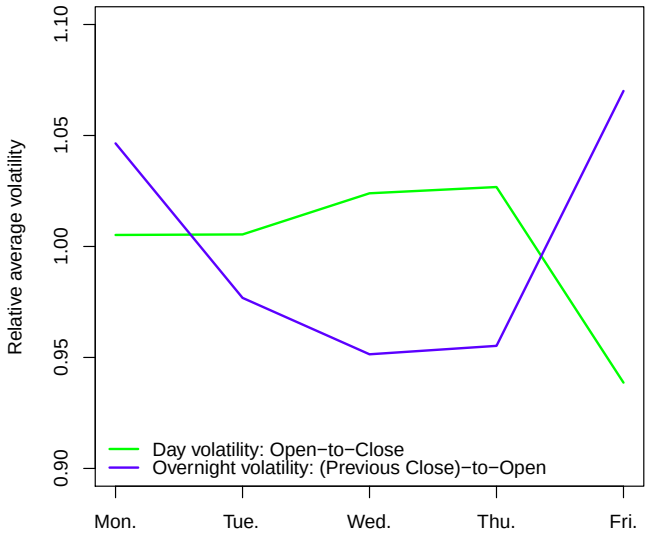



FIGURE 3.2 – Normalized weekly seasonality of the volatility. The overnight volatility is that of the previous night (i.e. the volatility of the weekend for Monday and that of Thursday night for Friday).

such as a weekly cyclical pattern of the volatility, which is interesting in itself (see Fig. 3.2). These are of course real effects, but the ARCH family of models we investigate here rather focuses on the endogenous self-dynamics *on top of* such seasonal patterns. For example, the quarterly announcement dates are responsible for returns of typically much larger magnitude (approximately three times larger on average for daily returns) that have a very limited feedback in future volatility.

We therefore want to remove all obvious seasonal effects from the dataset, and go through the following additional steps of data treatment before estimating the model. For every stock $a$, the average over time is denoted $\langle r_\cdot^a \rangle := \frac{1}{T}\sum_t r_t^a$, and for each date $t$, the cross-sectional average over stocks is $[r_t^\cdot] := \frac{1}{N}\sum_a r_t^a$. All the following normalizations apply both (and separately) for intra-day returns and overnight returns.

- The returns series are first centered around their temporal average : $r_t^a \leftarrow r_t^a - \langle r_\cdot^a \rangle$. In fact, the returns are already nearly empirically centered, since the temporal average is small, see Tab. 3.1 above.
- We then divide the returns by the cross-sectional dispersion :

$$r_t^a \leftarrow r_t^a / \sqrt{\left[r_t^{\cdot \neq a^2}\right]}.$$

This normalization [2] removes the historical low-frequency patterns of the volatility, for example the weekly pattern discussed above (Fig. 3.2). In order to predict the "real" volatility with the model, one must however re-integrate these patterns back into the $\sigma^{\mathrm{J}}$'s.

2. If the element $a$ is not excluded in the average, the tails of the returns are artificially cut-off : when $|r_t^a| \to \infty$, $|r_t^a|/\sqrt{[r_t^{\cdot 2}]}$ is capped at $\sqrt{N} < \infty$.

– Finally, we normalize stock by stock all the returns by their historical standard deviation : for all stock $a$, for all $t$,

$$r_t^a \leftarrow r_t^a / \sqrt{\langle r_\cdot^{a\,2} \rangle},$$

imposing $\langle r_\cdot^{\mathrm{D}\,a\,2} \rangle = \langle r_\cdot^{\mathrm{N}\,a\,2} \rangle \equiv 1$.

This data treatment allows to consider that the residual volatility of the returns series is independent of the effects we do not aim at modeling, and that the series of all stocks can reasonably be assumed to be homogeneous (i.e. identically distributed), both across stocks and across time. This is necessary in order to calibrate a model that is translational-invariant in time (i.e. only the time *lag* $\tau$ enters Eqs. (3.5,3.6)), and also to enlarge the calibration dataset by averaging the results over all the stocks in the pool — see the discussion in Appendix 3.6.

### 3.2.4 Model estimation

Assuming that the distribution of the residuals is a Student law, the log-likelihood per point of the model can be written as ($\mathrm{J} = \mathrm{D}, \mathrm{N}$) :

$$\mathcal{L}(\Theta^{\mathrm{J}}, \nu^{\mathrm{J}} | \{r^{\mathrm{D}}, r^{\mathrm{N}}\}) = \frac{\nu^{\mathrm{J}}}{2} \ln \left| (\nu^{\mathrm{J}} - 2)\, \sigma_t^2(\Theta^{\mathrm{J}}) \right| - \frac{\nu^{\mathrm{J}} + 1}{2} \ln \left| (\nu^{\mathrm{J}} - 2) \sigma_t^2(\Theta^{\mathrm{J}}) \; + \; r_t^{\mathrm{J}\,2} \right|, \tag{3.8}$$

where $\sigma_t^{\mathrm{J}\,2} = \sigma_t^2(\Theta^{\mathrm{J}} | \{r^{\mathrm{D}}, r^{\mathrm{N}}\})$ are defined in Eqs. (3.5,3.6), $\nu^{\mathrm{J}}$ are the degrees of freedom of the Student residuals, and $\Theta^{\mathrm{J}}$ denote generically the sets of volatility feedback parameters.

Conditionally on the dataset, we maximize numerically the likelihood of the model, averaged over all dates and all stocks.

**Calibration methodology :** As mentioned above, we in fact choose to parameterize the feedback kernels as suggested by the results of the method of moments, i.e. exponentially truncated power-laws for $K$'s and simple exponentials for $L$'s. Imposing these simple functional forms allows us to gain stability and readability of the results. However, the functional dependence of the likelihood on the kernel parameters is not guaranteed to be globally concave, as is the case when it is maximized « freely », i.e. with respect to all individual kernel coefficients $K(\tau)$ and $L(\tau)$, with $1 \leq \tau \leq q_{\mathrm{free}}$. This is why we use a three-step approach :

1. A first set of kernel estimates is obtained by the Generalized Method of Moments (GMM), see [38], and serves as a starting point for the optimization algorithm.
2. We then run a Maximum Likelihood Estimation (MLE) of the unconstrained kernels based on Eq. (3.8), over $6 \times q_{\mathrm{free}}$ parameters for both $\Theta^{\mathrm{D}}$ and $\Theta^{\mathrm{N}}$, with a moderate value of maximum lag $q_{\mathrm{free}} = 63 \simeq$ three months. Taking as a starting point the coefficients of step 1 and maximizing with a gradient descent, we obtain a second set of (short-range) kernels.
3. Finally, we perform a MLE estimation of the parametrically constrained kernels with the functional forms (3.7) for $K$'s and $L$'s, which only involves $4 \times 3 + 2 \times 2$ parameters in every set $\Theta^{\mathrm{D}}$ and $\Theta^{\mathrm{N}}$, with now a large value of the maximum lag $q_{\mathrm{constr}} = 512 \simeq$ two years. Taking as a starting point the functional fits of the kernels obtained at step 2, and maximizing with a gradient descent, we obtain our final set of model coefficients, shown in Tabs. 3.2,3.3.

Thanks to step 2, the starting point of step 3 is close enough to the global maximum for the likelihood to be locally concave, and the gradient descent algorithm converges in a few steps. The Hessian matrix of the likelihood is evaluated at the maximum to check that the dependency on all coefficients is indeed concave.

The numerical maximization of the likelihood is thus made on 2 or 3 parameters per kernel, independently of the chosen maximum lag $q_{\text{constr}}$, that can thus be arbitrarily large with little additional computation cost.

Finally, the degrees of freedom $\nu$ of the Student residuals are determined using two separate one-dimensional likelihood maximizations (one for $q = q_{\text{free}}$ and one for $q = q_{\text{constr}}$) and then included as an additional parameter in the MLE of the third step of the calibration. Note that $\nu$ does not vary significantly at the third step, which means that the estimation of the volatility parameters at the second step can indeed be done independently from that of $\nu$.

This somewhat sophisticated calibration method was tested on simulated data, obtaining very good results.

**The special case $s^2 = 0$ :** We ran the above calibration protocol on intra-day and overnight volatilities separately.

For the overnight model, this led to a slightly negative value of the baseline volatility $s^{\text{N}2}$ (statistically compatible with zero). But of course negative values of $s^2$ are excluded for a stable and positive volatility process. For overnight volatility only, we thus add a step to the calibration protocol, which includes the constraint $s^{\text{N}2} = 0$ in the estimation of $K^{\text{DD}\to\text{N}}$ and $K^{\text{NN}\to\text{N}}$ (which are the two main contributors to the value of the baseline volatility). For simplicity, we consider here that $\langle\sigma^{\text{N}2}\rangle = \langle r^{\text{D}2}\rangle = \langle r^{\text{N}2}\rangle = 1$. We take the results of the preceding calibration as a starting point and freeze all the kernels but $K^{\text{DD}\to\text{N}}$ and $K^{\text{NN}\to\text{N}}$, expressed (in this section only) as follows :

$$K^{\text{DD}\to\text{N}}(\tau) = g\,\tau^{-\alpha_1}\,\exp(-\omega_1\tau)\,, \qquad K^{\text{NN}\to\text{N}}(\tau) = \gamma\; g\,\tau^{-\alpha_2}\,\exp(-\omega_2\tau)\,, \tag{3.9}$$

where $g = g(\gamma, \alpha_1, \omega_1, \alpha_2, \omega_2)$ is fixed by the constraint $s^{\text{N}2} = 0$ :

$$g(\gamma, \alpha_1, \omega_1, \alpha_2, \omega_2) = \frac{1-c}{h(\alpha_1,\omega_1) + \gamma\; h(\alpha_2,\omega_2)}; \qquad h(\alpha,\omega) = \sum_{\tau=1}^{q} \tau^{-\alpha}\exp(-\omega\tau)\,, \tag{3.10}$$

with $\gamma > 0$ the ratio of the two initial amplitudes, and $c$ the (low) contribution of the fixed 'cross' kernels $K^{\text{ND}\to\text{N}}$ and $K^{\text{DN}\to\text{N}}$ to $s^{\text{N}2}$. We then maximize the likelihood of the model with respect to the five parameters $\gamma, \alpha_1, \omega_1, \alpha_2$ and $\omega_2$, for which a gradient vector and a Hessian matrix of dimension 5 can be deduced from equations (3.9) and (3.10). The coefficients and confidence intervals of the kernels $K^{\text{DD}\to\text{N}}$ and $K^{\text{NN}\to\text{N}}$ are replaced in Sect. 3.3.1 by the results of this final step, along with the corresponding value of the overnight baseline volatility, $s^{\text{N}2} = 0$ in Sect. 3.3.3.

For intra-day volatility instead, the results are given just below, in Sect. 3.3.1.

## 3.3 Intra-day vs. overnight : results and discussions

The calibration of our generalized ARCH framework should determine three families of parameters : the feedback kernels $K$ and $L$, the statistics of the residuals $\xi$ and finally the "baseline volatilities" $s^2$. We discuss these three families in turn in the following sections.

### 3.3.1 The feedback kernels

In this section, we give the results of the ML estimation of the regression kernels for a maximum lag $q = 512$ : estimates of the parameters are reported in Tabs. 3.2,3.3, and the resulting kernels are shown in Fig. 3.3.

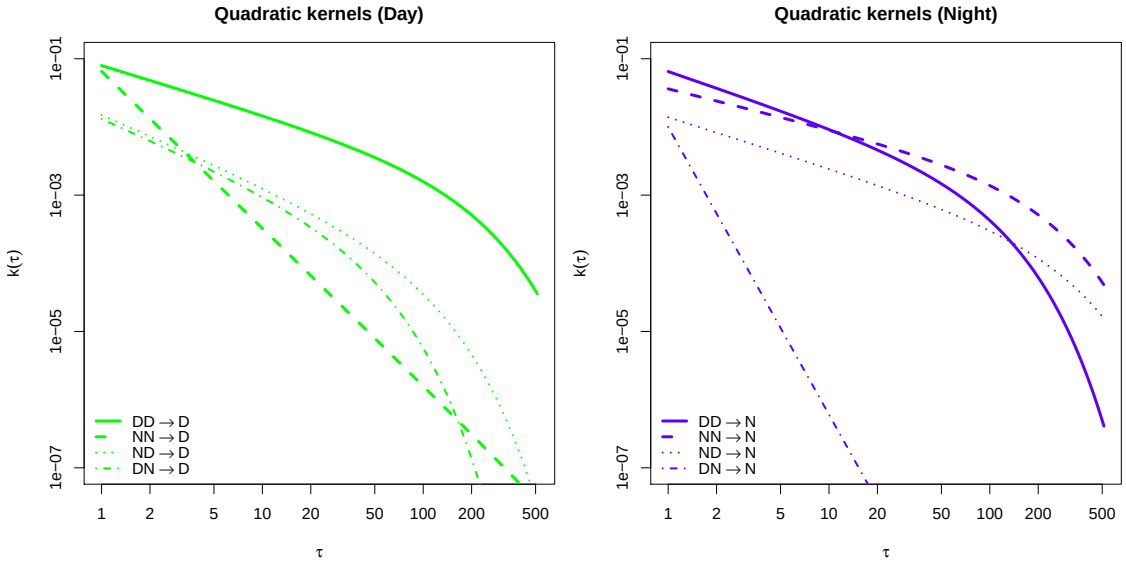
| Kernels | $K(\tau) = g_{\rm p}\,\tau^{-\alpha}\,e^{-\omega_{\rm p}\,\tau}$ | | | $L(\tau) = g_{\rm e}\,e^{-\omega_{\rm e}\,\tau}$ | |
|---|---|---|---|---|---|
| | $g_{\rm p} \times 10^2$ | $\alpha$ | $\omega_{\rm p} \times 10^2$ | $g_{\rm e} \times 10^2$ | $\omega_{\rm e} \times 10^2$ |
| $K^{\rm DD\to D}$ | $7.99 \pm 0.06$ | $0.71 \pm 0.003$ | $0.64 \pm 0.02$ | $--$ | $--$ |
| $K^{\rm NN\to D}$ | $6.53 \pm 0.22$ | $2.30 \pm 0.07$ | $0.04 \pm 0.97$ | $--$ | $--$ |
| $K^{\rm ND\to D}$ | $1.52 \pm 0.17$ | $1.03 \pm 0.11$ | $1.3 \pm 1.2$ | $--$ | $--$ |
| $K^{\rm DN\to D}$ | $1.35 \pm 0.22$ | $1.03 \pm 0.17$ | $3.0 \pm 4.6$ | $--$ | $--$ |
| $L^{\rm D\to D}$ | $--$ | $--$ | $--$ | $-4.97 \pm 0.25$ | $18.3 \pm 1.3$ |
| $L^{\rm N\to D}$ | $--$ | $--$ | $--$ | $-2.83 \pm 0.30$ | $22.3 \pm 2.5$ |

TABLE 3.2 – Day volatility : estimated kernel parameters for $K$'s and $L$'s, with their asymptotic confidence intervals of level 95%, as computed using the Fisher Information matrix with the Gaussian quantile (1.98).

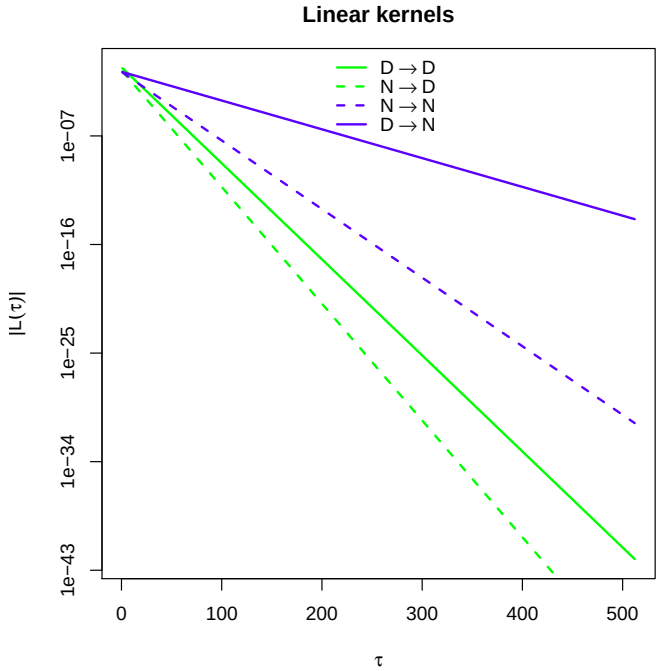
| Kernels | $K(\tau) = g_{\rm p}\,\tau^{-\alpha}\,e^{-\omega_{\rm p}\,\tau}$ | | | $L(\tau) = g_{\rm e}\,e^{-\omega_{\rm e}\,\tau}$ | |
|---|---|---|---|---|---|
| | $g_{\rm p} \times 10^2$ | $\alpha$ | $\omega_{\rm p} \times 10^2$ | $g_{\rm e} \times 10^2$ | $\omega_{\rm e} \times 10^2$ |
| $K^{\rm DD\to N}$ | $6.59 \pm 0.33$ | $0.80 \pm 0.02$ | $1.4 \pm 0.4$ | $--$ | $--$ |
| $K^{\rm NN\to N}$ | $3.64 \pm 0.17$ | $0.58 \pm 0.01$ | $0.58 \pm 0.04$ | $--$ | $--$ |
| $K^{\rm ND\to N}$ | $1.39 \pm 0.11$ | $0.74 \pm 0.03$ | $0.42 \pm 0.12$ | $--$ | $--$ |
| $K^{\rm DN\to N}$ | $-1.00 \pm 0.29$ | $4.22 \pm 2.44$ | $0.02 \pm 23$ | $--$ | $--$ |
| $L^{\rm D\to N}$ | $--$ | $--$ | $--$ | $-2.09 \pm 0.05$ | $5.5 \pm 0.2$ |
| $L^{\rm N\to N}$ | $--$ | $--$ | $--$ | $-2.03 \pm 0.20$ | $13.1 \pm 2.2$ |

TABLE 3.3 – Overnight volatility : estimated kernel parameters for $K$'s and $L$'s, with their asymptotic confidence intervals of level 95%, as computed using the Fisher Information matrix with the Gaussian quantile (1.98).

We define the exponential characteristic times $\tau_{\rm p} := 1/\omega_{\rm p}$ and $\tau_{\rm e} := 1/\omega_{\rm e}$, for which qualitative interpretation is easier than for $\omega_{\rm p}$ and $\omega_{\rm e}$. In the case of the quadratic kernels (of type $K$), $\tau_{\rm p}$ represents the lag where the exponential cut-off appears, after which the kernel decays to zero quickly. One should note that in three cases, we have $\omega_{\rm p} - \Delta\omega_{\rm p} < 0$. These correspond to kernels with $\alpha > 1$, which means that the power-law decays quickly by itself. In these cases the identifiability of $\omega_{\rm p}$ is more difficult and cut-off times are ill-determined, since the value of $\omega_{\rm p}$ only matters in a region where the kernels are already small. The exponential term $\exp(-\omega_{\rm p}\,\tau)$ could be removed from




(a) Quadratic $K(\tau)$ kernels in log-log scale. For 'DN→N', the absolute value of the (negative) kernel is plotted.




(b) Linear $L(\tau)$ kernels in lin-log scale. All leverage kernels are negative, so $-L(\tau)$ are plotted.

FIGURE 3.3 – Estimated kernels, impacting intra-Days in (light) green, overNights in (dark) blue.

the functional form of equation (3.7), for these kernels only (the calibration would then modify very slightly the value of the power-law exponent $\alpha$).

**Intra-day volatility :** From Tab. 3.2 and Fig. 3.3(a), we see that all intra-day quadratic kernels are positive. However, a clear distinction is observed between intra-day feedback and overnight feedback : while the former is strong and decays slowly ($\alpha = 0.71$ and $\tau_p = 157$ days), the latter decays extremely steeply ($\alpha = 2.30$) and is quickly negligible, except for the intra-day immediately following the overnight, where the effect is as strong as that of the previous intra-day. The cross kernels ($\text{ND} \to \text{D}$ or $\text{DN} \to \text{D}$) are both statistically significant, but are clearly smaller, and decay faster, than the $\text{DD} \to \text{D}$ effect.

As far as the leverage effect is concerned, both $L$'s are found to be negative, as expected, and of similar decay time : $\tau_e \approx 5$ days (one week). However, the amplitude for their immediate impact is two times smaller for past overnight returns : $L^{\text{N}\to\text{D}} = -0.0283$ versus $L^{\text{D}\to\text{D}} = -0.0497$.

In summary, the most important part of the feedback effect on the intra-day component of the volatility comes from the past intra-days themselves, except for the very last overnight, which also has a strong impact — as intuitively expected, a large return overnight leads to a large immediate reaction of the market as trading resumes. However, this influence is seen to decay very quickly with time. Since a large fraction of company specific news release happen overnight, it is tempting to think that large overnight returns are mostly due to news. Our present finding would then be in line with the general result reported in [82] : volatility tends to relax much faster after a news-induced jump than after endogenous jumps.

**Overnight volatility :** In the case of overnight volatility, Tab. 3.3 and Fig. 3.3(a) illustrate that the influence of past intra-days and past overnights is similar : $K^{\text{DD}\to\text{N}}(\tau) \approx K^{\text{NN}\to\text{N}}(\tau)$, in particular when both are large. The cross kernels now behave quite differently : whereas the behavior of $K^{\text{ND}\to\text{N}}$ is not very different from that of $K^{\text{DD}\to\text{N}}$ or $K^{\text{NN}\to\text{N}}$ (although its initial amplitude is four times smaller), $K^{\text{DN}\to\text{N}}(\tau)$ is negative and small, but is hardly significant for $\tau \geq 2$. Interestingly, as pointed out above, the equality $K^{\text{ND}\to\text{N}} = K^{\text{DD}\to\text{N}} = K^{\text{NN}\to\text{N}}$ means that it is the full Close-to-Close return that is involved in the feedback mechanism on the next overnight. What we find here is that this equality very roughly holds, suggesting that, as postulated in standard ARCH approaches, the daily close to close return is the fundamental object that feedbacks on future volatilities. However, this is only approximately true for the *overnight volatility*, while the intra-day volatility behaves very differently (as already said, for intra-day returns, the largest part of the feedback mechanism comes from past intra-days only, and the very last overnight).

Finally, the leverage kernels behave very much like for the intra-day volatility. In fact, the $\text{N} \to \text{N}$ leverage kernel is very similar to its $\text{N} \to \text{D}$ counterpart, whereas the decay of the $\text{D} \to \text{N}$ kernel is slower ($\tau_e \approx 18$ days, nearly one month).

**Stability and positivity :** We checked that these empirically-determined kernels are compatible with a stable and positive volatility process. The first obvious condition is that the system is stable with positive baseline volatilities $s^{\text{J}2}$. The self-consistent equations for the average volatilities read :

(neglecting small cross correlations) :

$$\langle {\sigma^{\mathrm{D}}}^2 \rangle = {s^{\mathrm{D}}}^2 + \langle {\sigma^{\mathrm{D}}}^2 \rangle \sum_{\tau=1}^{\infty} K^{\mathrm{DD}\to\mathrm{D}}(\tau) + \langle {\sigma^{\mathrm{N}}}^2 \rangle \sum_{\tau=1}^{\infty} K^{\mathrm{NN}\to\mathrm{D}}(\tau), \tag{3.11}$$

$$\langle {\sigma^{\mathrm{N}}}^2 \rangle = {s^{\mathrm{N}}}^2 + \langle {\sigma^{\mathrm{D}}}^2 \rangle \sum_{\tau=1}^{\infty} K^{\mathrm{DD}\to\mathrm{N}}(\tau) + \langle {\sigma^{\mathrm{N}}}^2 \rangle \sum_{\tau=1}^{\infty} K^{\mathrm{NN}\to\mathrm{N}}(\tau). \tag{3.12}$$

This requires that the two eigenvalues of the $2 \times 2$ matrix of the corresponding linear system are less than unity, i.e.

$$\frac{1}{2}\left| \widehat{K}^{\mathrm{DD}\to\mathrm{D}} + \widehat{K}^{\mathrm{NN}\to\mathrm{N}} \pm \sqrt{(\widehat{K}^{\mathrm{DD}\to\mathrm{D}} - \widehat{K}^{\mathrm{NN}\to\mathrm{N}})^2 + 4\widehat{K}^{\mathrm{NN}\to\mathrm{D}}\widehat{K}^{\mathrm{DD}\to\mathrm{N}}} \right| < 1, \tag{3.13}$$

where the hats denote the integrated kernels, schematically $\widehat{K} = \sum_{\tau=1}^{\infty} K(\tau)$. This is indeed verified empirically, the eigenvalues being $\lambda_1 \simeq 0.94$ , $\lambda_2 \simeq 0.48$.

Moreover, for intra-day and overnight volatilities separately, we checked that our calibrated kernels are compatible with the two positivity conditions derived in Appendices 3.5.2,3.5.3 : the first one referring to the cross kernels $K^{\mathrm{ND}\to}$ and $K^{\mathrm{DN}\to}$, and the second one to the leverage kernels $L^{\mathrm{D}\to}$ and $L^{\mathrm{N}\to}$. For $q = 512$, the criteria fail for two spurious reasons. Firstly, for lags greater than their exponential cut-offs, the quadratic kernels vanish quicker than the 'cross' kernels, which makes the « $\tau$ by $\tau$ » criterion fail. Secondly, the criterion $L^{\dagger}K^{-1}L \leq 4s^2$ cannot be verified for overnight volatility with ${s^{\mathrm{N}}}^2 = 0$ (for lower values of $q$, using the same functional forms and coefficients for the kernels, ${s^{\mathrm{N}}}^2$ rises to a few percents). These two effects can be considered spurious because the long-range contributions have a weak impact on the volatilities and cannot in deed generate negative values, as we checked by simulating the volatility processes with $q = 512$. We thus restricted the range to $q = 126$ (= six months) in order to test our results with the two positive volatility criteria (again, see Appendix 3.5). For the ill-determined exponential decay rates $\omega_{\mathrm{p}}$, the upper bounds of the confidence intervals are used. The two criteria are then indeed verified for both intra-day and overnight volatilities.

### 3.3.2 Distribution of the residuals

As mentioned above, we assume that the residuals $\xi$ (i.e. the returns divided by the volatility predicted by the model) are Student-distributed. This is now common in ARCH/GARCH literature and was again found to be satisfactory in our companion paper [38]. The fact that the $\xi$ are not Gaussian means that there is a residual surprise element in large stock returns, that must be interpreted as true 'jumps' in the price series.

The tail cumulative distribution function (CDF) of the residuals is shown in Fig. 3.4 for both intra-day and overnight returns, together with Student best fits, obtained with long feedback kernels ($q = 512$). This reveals a clear difference in the statistical properties of $\xi^{\mathrm{D}}$ and $\xi^{\mathrm{N}}$. First, the Student fit is better for overnight residuals than for intra-day residuals, in particular far in the tails. More importantly, the number of degrees of freedom $\nu$ is markedly different for the two types of residuals : indeed, our MLE estimation yields $\nu^{\mathrm{D}} = 13.5$ and $\nu^{\mathrm{N}} = 3.61$ as reported in Tab. 3.4, resulting in values of the residual kurtosis $\kappa^{\mathrm{D}} = 3 + 6/(\nu^{\mathrm{D}} - 4) = 3.6$ and $\kappa^{\mathrm{N}} = \infty$. This result must be compared

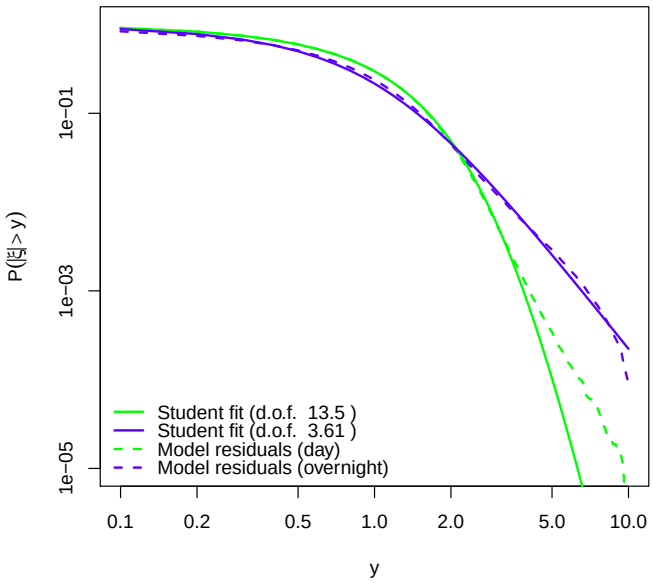



FIGURE 3.4 – CDF $\mathbb{P}(|\xi| > y)$ of the residuals $\xi^{\mathrm{D}}$ and $\xi^{\mathrm{N}}$, in log-log scale.

| $q$ | 21 | 42 | 512 |
|---|---|---|---|
| $\nu^{\mathrm{D}}(\pm 0.3)$ | 10.7 | 12.6 | 13.5 |
| $\nu^{\mathrm{N}}(\pm 0.02)$ | 3.49 | 3.58 | 3.61 |
| $\rho^{\mathrm{D}}$ | 18.1% | 12.8% | 8.5% |
| $\rho^{\mathrm{N}}$ | 14.0% | 7.3% | 0.0% |

TABLE 3.4 – Baseline volatility and tail index of the Student residuals for several maximum lags $q$.

with the empirical kurtosis of the returns that was measured directly in Sect. 3.2. The intuitive conclusion is that the large (infinite ?) kurtosis of the overnight returns cannot be attributed to fluctuations in the volatility, but rather, as mentioned above, to large jumps related to overnight news. This clear qualitative difference between intra-day and overnight returns is a strong argument justifying the need to treat these effects separately, as proposed in this paper.

We have also studied the evolution of $\nu^{\mathrm{D}}$ and $\nu^{\mathrm{N}}$ as a function of the length $q$ of the memory of the kernels, see Tab. 3.4. If longer memory kernels allow to account for more of the dynamics of the volatility, less kurtic residuals should be found for larger $q$'s. This is indeed what we find, in particular for intra-day returns, for which $\nu^{\mathrm{D}}$ increases from 10.7 for $q = 21$ to 13.5 for $q = 512$. The increase is however much more modest for overnight returns. We propose below an interpretation of this fact.

### 3.3.3 Baseline volatility

Finally, we want to study the ratio $\rho^{\mathrm{J}} = s^{\mathrm{J}2}/\langle\sigma^{\mathrm{J}2}\rangle$, which can be seen as a measure of the relative importance of the baseline component of the volatility, with respect to the endogenous feedback component. [3] The complement $1-\rho^{\mathrm{J}}$ gives the relative contribution of the feedback component, given by $\widehat{K}^{\mathrm{DD}\to\mathrm{J}} + \widehat{K}^{\mathrm{NN}\to\mathrm{J}}$ in the present context. [4]

The results for $\rho^{\mathrm{J}}$ are given in Tab. 3.4 for $q = 21, 42$ and $512$ : as mentioned above, a larger $q$ explains more of the volatility, therefore reducing the value of both $\rho^{\mathrm{D}}$ and $\rho^{\mathrm{N}}$. While $\rho^{\mathrm{D}}$ is small ($\sim 0.1$) and comparable to the value found for the daily ARCH model studied in the companion paper [38], the baseline contribution is *nearly zero* for the overnight volatility. We find this result truly remarkable, and counter-intuitive at first sight. Indeed, the baseline component of the volatility is usually associated to exogenous factors, which, as we argued above, should be dominant for $\sigma^{\mathrm{N}}$ since many unexpected pieces of news occur overnight !

Our interpretation of this apparent paradox relies on the highly kurtic nature of the overnight residual, with a small value $\nu^{\mathrm{N}} \approx 3.6$ as reported above. The picture is thus as follows : most overnights are news-less, in which case the overnight volatility is completely fixed by feedback effects, set by the influence of past returns themselves. The overnights in which important news is released, on the other hand, contribute to the tails of the residual $\xi^{\mathrm{N}}$, because the large amplitude of these returns could hardly be guessed from the previous amplitude of the returns. Furthermore, the fact that $K^{\mathrm{NN}\to\mathrm{D}}$ decays very quickly is in agreement with the idea, expressed in [82], that the impact of news (chiefly concentrated overnight) on volatility is short-lived. [5]

In conclusion, we find that most of the predictable part of the overnight volatility is of endogenous origin, but that the contribution of unexpected jumps reveals itself in the highly non-Gaussian statistics of the residuals. The intra-day volatility, on the other hand, has nearly Gaussian residuals but still a very large component of endogenous volatility ($1-\rho^{\mathrm{D}} \approx 90\%$).

### 3.3.4 In-Sample and Out-of-Sample tests

In order to compare our bivariate Intra-day/Overnight volatility prediction model with the standard ARCH model for daily (close-to-close) volatility, we ran In-Sample (IS) and Out-of-Sample (OS) likelihood computations with both models. Of course, in order to compare models, the same quantities must be predicted. A daily ARCH model that predicts the daily volatility $\sigma_t^2$ at date $t$ can predict intra-day and overnight volatilities as follows :

$$\widehat{\sigma^{\mathrm{D}}}_t^2 = \frac{\left[\langle r^{\mathrm{D}2}\rangle\right]}{[\langle r^2\rangle]}\,\sigma_t^2 \ , \quad \widehat{\sigma^{\mathrm{N}}}_t^2 = \frac{\left[\langle r^{\mathrm{N}2}\rangle\right]}{[\langle r^2\rangle]}\,\sigma_t^2 \ , \tag{3.14}$$

where $[\langle\cdot\rangle]$ is the average over all dates and all stocks, and as in Sect. 3.2, $r_t = r_t^{\mathrm{D}} + r_t^{\mathrm{N}}$ is the daily (close-to-close) return of date $t$. Similarly, our bivariate intra-day/overnight model that provides

3. In fact, the stability criterion for our model reads $\rho^{\mathrm{J}} > 0$, which is found to be satisfied by our calibration, albeit marginally for the overnight volatility.

4. There is a contribution of the cross terms $K^{\mathrm{J'J''}\to\mathrm{J}}$ since intra-day/overnight and overnight/intra-day correlations are not exactly zero, but this contribution is less than one order of magnitude lower than the $\widehat{K}^{\mathrm{J'J'}\to\mathrm{J}}$.

5. This effect was confirmed recently in [43] using a direct method : the relaxation of volatility after large overnight jumps of either sign is very fast, much faster than the relaxation following large intra-day jumps.

predictions for intra-day and overnight volatilities separately can give the following estimation of the daily volatility :

$$\widehat{\sigma}_t^2 \;=\; \sigma_t^{\mathrm{D}\,2} \;+\; \sigma_t^{\mathrm{N}\,2} \;+\; 2\,[\langle r^{\mathrm{D}} r^{\mathrm{N}}\rangle]. \tag{3.15}$$

For each of the 6 predictions (of the intra-day, overnight and daily volatilities by the two models separately, bivariate Intra-day/Overnight and standard ARCH), we use the following methodology :

- The pool of stock names is split in two halves, and the model parameters are estimated separately on each half.
- The "per point" log-likelihood (3.8) is computed for both sets of parameters, once with *the same* half dataset as used for the calibration (In-Sample), and once with *the other* half dataset (Out-of-Sample, or "control"). We compute the log-likelihoods for intra-day and overnight volatilities ($\mathrm{J} \in \{\mathrm{D}, \mathrm{N}\}$), and for daily volatility :

$$\begin{aligned} \text{Bivariate models :} \quad & \mathcal{L}^{\mathrm{biv}}\left(\sigma^{\mathrm{J}}, \nu^{\mathrm{J}}, r^{\mathrm{J}}\right) \quad \text{and} \quad \mathcal{L}^{\mathrm{biv}}\left(\widehat{\sigma}, \nu^{\mathrm{daily}}, r^{\mathrm{daily}}\right); \\ \text{Standard ARCH models :} \quad & \mathcal{L}^{\mathrm{std}}\left(\widehat{\sigma^{\mathrm{J}}}, \nu^{\mathrm{J}}, r^{\mathrm{J}}\right) \quad \text{and} \quad \mathcal{L}^{\mathrm{std}}\left(\sigma, \nu^{\mathrm{daily}}, r^{\mathrm{daily}}\right); \end{aligned}$$

  where $\mathcal{L}^{\mathrm{biv}}$ is computed in the bivariate model (i.e. with six regressors : four quadratic and two linear) and $\mathcal{L}^{\mathrm{std}}$ is computed in the standard ARCH model (i.e. with two regressors : one quadratic and one linear), and $\widehat{\sigma^{\mathrm{J}}}$ and $\widehat{\sigma}$ are as given by equations (3.14) and (3.15).
- The IS log-likelihood $\mathcal{L}_{\mathrm{IS}}$ of the model is computed as the average of the two In-Sample results, and similarly for the OS log-likelihood $\mathcal{L}_{\mathrm{OS}}$. We call $l_{\mathrm{IS}} = \exp(\mathcal{L}_{\mathrm{IS}})$ and $l_{\mathrm{OS}} = \exp(\mathcal{L}_{\mathrm{OS}})$ the average likelihood per point (ALpp) of the model IS and OS, expressed as percentages, that are two proxies of the « probability that the sample data were generated by the model ».

We then use the values of $l_{\mathrm{IS}}$ and $l_{\mathrm{OS}}$ to compare models. For a « good » model, these values must be as high as possible, but they must also be close to each other. As a matter of fact, if $l_{\mathrm{IS}}$ is significantly greater than $l_{\mathrm{OS}}$, the model may be over-fitting the data. On the contrary, if $l_{\mathrm{OS}}$ is greater than $l_{\mathrm{IS}}$, which seems counter-intuitive, the model may be badly calibrated. The results of this model comparison are given in Tab. 3.5 : the bivariate Intra-day/Overnight model has a higher likelihood than the standard daily ARCH model, both In-Sample (this was to be expected even from pure over-fitting due to additional parameters) and Out-of-Sample (thus outperforming in predicting the "typical" random realization of the returns).

The likelihoods of the predictions obtained with equations (3.14) and (3.15) are marked with the exponent † in Tab. 3.5. For these likelihoods, "In-Sample" simply means that the same half of the stock pool was used for the calibration of the model and for the estimation of the likelihood, although different types of returns are considered. Similarly, "Out-of-Sample" likelihoods are estimated on the other half of the stock pool. These values serve as comparison benchmarks between the two models.

| Prediction | $\sigma^{\mathrm{D}\,2}$ | | $\sigma^{\mathrm{N}\,2}$ | | $\sigma^2$ (daily) | |
|---|---|---|---|---|---|---|
| ALpp [%] | $l_{\mathrm{IS}}$ | $l_{\mathrm{OS}}$ | $l_{\mathrm{IS}}$ | $l_{\mathrm{OS}}$ | $l_{\mathrm{IS}}$ | $l_{\mathrm{OS}}$ |
| Biv. Intra-day/Overnight | 44.423 | 44.418 | 50.839 | 50.826 | $45.512^{\dagger}$ | $45.509^{\dagger}$ |
| Standard ARCH | $44.227^{\dagger}$ | $44.225^{\dagger}$ | $50.598^{\dagger}$ | $50.595^{\dagger}$ | 44.931 | 44.928 |

TABLE 3.5 – In-Sample and Out-of-Sample average per point likelihoods. Figures with † pertain to reconstructed volatilities Eqs. (3.14) or (3.15).

We see that in all cases, the bivariate Intra-day/Overnight significantly outperforms the standard daily ARCH framework, in particular concerning the prediction of the total (Close-Close) volatility.

## 3.4 Conclusion and extension

The main message of this study is quite simple, and in fact to some extent expected : overnight and intra-day returns are completely different statistical objects. The ARCH formalism, that allows one to decompose the volatility into an exogenous component and a feedback component, emphasizes this difference. The salient features are :

- While past intra-day returns affect equally both the future intra-day and overnight volatilities, past overnight returns have a weak effect on future intra-day volatilities (except for the very next one) but impact substantially future overnight volatilities.
- The exogenous component of overnight volatilities is found to be close to zero, which means that the lion's share of overnight volatility comes from feedback effects.
- The residual kurtosis of returns (once the ARCH effects have been factored out) is small for intra-day returns but infinite for overnight returns.
- The bivariate intra-day/overnight model significantly outperforms the standard ARCH framework based on daily returns for Out-of-Sample predictions.

Intuitively, a plausible picture for overnight returns is as follows : most overnights are news-less, in which case the overnight volatility is completely fixed by feedback effects, set by the influence of past returns themselves. Some (rare) overnights witness unexpected news releases, which lead to huge jumps, the amplitude of which could hardly have be guessed from the previous amplitude of the returns. This explains why these exogenous events contribute to residuals with such fat tails that the kurtosis diverge, and *not* to the baseline volatility that concerns the majority of news-less overnights.

These conclusions hold not only for US stocks : we have performed the same study on European stocks obtaining very close model parameter estimates.[6] Notably, the baseline volatilities are found to be $\rho^{\mathrm{D}} \simeq 0.1$ and $\rho^{\mathrm{N}} \simeq 0$ (for intra-day and overnight volatilities, respectively), in line with the figures found on US stocks and the interpretation drawn. The only different qualitative behavior observed on European stocks is the quality of the Student fit for the residuals of the overnight regression : whereas US stocks exhibit a good fit with $\nu^{\mathrm{N}} = 3.61 < 4$ degrees of freedom (hence infinite kurtosis), European stocks have a fit of poorer quality in the tails and a parameter $\nu^{\mathrm{N}} = 5.34$ larger than 4, hence a positive but finite kurtosis.

Having decomposed the Close-Close return into an overnight and an intra-day component, the next obvious step is to decompose the intra-day return into higher frequency bins — say five minutes. We have investigated this problem as well, the results are reported in [26]. In a nutshell, we find that once the ARCH prediction of the intra-day average volatility is factored out, we still identify a causal feedback from past five minute returns onto the volatility of the current bin. This feedback has again a leverage component and a quadratic (ARCH) component. The intra-day leverage kernel is close to an exponential with a decay time of $\approx 1$ hour. The intra-day ARCH kernel, on the other hand, *is still a power law*, with an exponent that we find to be close to unity, in agreement with several studies in the literature concerning the intra-day temporal correlations of volatility/activity — see e.g. [87, 96, 107], and, in the context of Hawkes processes, [15, 71]. It would be very interesting to repeat the analysis of the companion paper [38] on five minute returns and check whether there is also a dominance of the diagonal terms of the QARCH kernels over the off-diagonal ones, as

6. Equities belonging to the Bloomberg European 500 index over the same time span 2000–2009, see Appendix 3.7 for detailed results.

we found for daily returns. This would suggest a microscopic interpretation of the ARCH feedback mechanism in terms of a Hawkes process for the trading activity.

## 3.5 Appendix : Non-negative volatility conditions

In this appendix, we study the mathematical validity of our volatility regression model. The first obvious condition is that the model is stable, which leads to the condition (3.13) in the text above. This criterion is indeed obeyed by the kernels calibrated on empirical results. Secondly, the volatility must remain positive, which is not a priori guaranteed with multiple kernels associated to signed regressors. We now establish a set of sufficient conditions on the feedback kernels to obtain non-negative volatility processes.

### 3.5.1 One correlation feedback kernel, no leverage coefficients

We consider first the simpler model for daily volatility, without linear regression coefficients :

$$\sigma_t^2 = s^2 + \sum_{\tau=1}^{q} K^{\mathrm{DD}\rightarrow}(\tau) {r_{t-\tau}^{\mathrm{D}}}^2 + \sum_{\tau=1}^{q} K^{\mathrm{NN}\rightarrow}(\tau) {r_{t-\tau}^{\mathrm{N}}}^2 + 2\sum_{\tau=1}^{q} K^{\mathrm{ND}\rightarrow}(\tau) r_{t-\tau}^{\mathrm{D}} r_{t-\tau}^{\mathrm{N}}.$$

This modification of the standard ARCH model can lead to negative volatilities if (at least) one term in the last sum takes large negative values. This issue can be studied more precisely with the matrix form of the model :

$$\sigma_t^2 = s^2 + R_t^\dagger K R_t,$$

with

$$K = \begin{pmatrix} K^{\mathrm{DD}\rightarrow}(1) & & & K^{\mathrm{ND}\rightarrow}(1) & & \\ & \ddots & & & \ddots & \\ & & K^{\mathrm{DD}\rightarrow}(q) & & & K^{\mathrm{ND}\rightarrow}(q) \\ K^{\mathrm{ND}\rightarrow}(1) & & & K^{\mathrm{NN}\rightarrow}(1) & & \\ & \ddots & & & \ddots & \\ & & K^{\mathrm{ND}\rightarrow}(q) & & & K^{\mathrm{NN}\rightarrow}(q) \end{pmatrix}, \quad R_t = \begin{pmatrix} r_{t-1}^{\mathrm{D}} \\ \vdots \\ r_{t-q}^{\mathrm{D}} \\ r_{t-1}^{\mathrm{N}} \\ \vdots \\ r_{t-q}^{\mathrm{N}} \end{pmatrix},$$

and where $K^{\mathrm{DD}\rightarrow}$ and $K^{\mathrm{NN}\rightarrow}$ coefficients are assumed to be all positive (which is the case empirically). This formula highlights the fact that the volatility remains positive as soon as the symmetric matrix $K$ is positive semidefinite. We now determine a sufficient and necessary condition under which $K$ has negative eigenvalues. The characteristic polynomial of $K$ is

$$\chi_K(X) = \prod_{\tau=1}^{q} \Big[ (K^{\mathrm{DD}\rightarrow}(\tau) - X)(K^{\mathrm{NN}\rightarrow}(\tau) - X) - K^{\mathrm{ND}\rightarrow}(\tau)^2 \Big],$$

and the eigenvalues of $K$ are the zeros of $\chi_K(X)$, solutions $\chi_K(\lambda) = 0$, i.e. such that

$$\lambda^2 - (K^{\mathrm{DD}\rightarrow}(\tau) + K^{\mathrm{NN}\rightarrow}(\tau))\lambda + K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau) - K^{\mathrm{ND}\rightarrow}(\tau)^2 = 0.$$

Hence, $K$ has at least one negative eigenvalue iff $\exists \tau \in \{1, \dots, q\}$ s.t.

$$K^{\mathrm{DD}\rightarrow}(\tau) + K^{\mathrm{NN}\rightarrow}(\tau) - \sqrt{(K^{\mathrm{DD}\rightarrow}(\tau) - K^{\mathrm{NN}\rightarrow}(\tau))^2 + 4K^{\mathrm{ND}\rightarrow}(\tau)^2} < 0,$$

and finally,

$$K \text{ is positive semidefinite } \Leftrightarrow \ \forall \tau \in \{1,\dots,q\}, \quad K^{\mathrm{ND}\rightarrow}(\tau)^2 \leq K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau). \tag{3.16}$$

When the quadratic kernel $K$ is positive-semidefinite, $\sigma_t^2$ remains positive for all $t$. For example, in the standard ARCH model, the inequality is saturated for all $\tau$ by construction, and the condition (3.16) is satisfied.

The positive-semidefiniteness of $K$, equivalent to

$$\forall v = \left(v_1^{\mathrm{D}}, \dots, v_q^{\mathrm{D}}, v_1^{\mathrm{N}}, \dots, v_q^{\mathrm{N}}\right)^\dagger \in \mathbb{R}^{2q} \ , \quad v^\dagger K v \geq 0,$$

is ensured by the *sufficient* condition that every term in the development of the quadratic form is positive :

$$\begin{aligned}
&K \text{ is positive-semidefinite} \\
&\quad \Leftarrow \forall v, \ \forall \tau \in \{1,\dots,q\}, \quad K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + K^{\mathrm{NN}\rightarrow}(\tau) {v_\tau^{\mathrm{N}}}^2 + 2K^{\mathrm{ND}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_\tau^{\mathrm{N}} \geq 0 \\
&\quad \Leftarrow \forall v, \ \forall \tau \in \{1,\dots,q\}, \quad |K^{\mathrm{ND}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_\tau^{\mathrm{N}}| \leq \frac{1}{2}(K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + K^{\mathrm{NN}\rightarrow}(\tau) {v_\tau^{\mathrm{N}}}^2) \\
&\quad \Leftarrow \forall \tau \in \{1,\dots,q\}, \quad |K^{\mathrm{ND}\rightarrow}(\tau)| \leq \sqrt{K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau)}
\end{aligned}$$

Although more stringent *a priori*, this « $\tau$ by $\tau$ » condition resumes, in this particular case, to the necessary *and* sufficient criterion (3.16). In the next subsection, we use the same method to obtain a sufficient condition for the semidefiniteness of $K$ in the more complicated case with two correlation feedback kernels.

### 3.5.2 Two correlation feedback kernels, no leverage coefficients

With an additional feedback function $K^{\mathrm{DN}\rightarrow}$ and a coefficient $K^{\mathrm{NN}\rightarrow}(0)$ corresponding to the $\tau = 0$ term in the ${r^{\mathrm{N}}}^2$ sum, the model is

$$\begin{aligned}
\sigma_t^2 = s^2 &+ \sum_{\tau=1}^{q} K^{\mathrm{DD}\rightarrow}(\tau) {r_{t-\tau}^{\mathrm{D}}}^2 + \sum_{\tau=0}^{q} K^{\mathrm{NN}\rightarrow}(\tau) {r_{t-\tau}^{\mathrm{N}}}^2 \\
&+ 2\sum_{\tau=1}^{q} K^{\mathrm{ND}\rightarrow}(\tau) r_{t-\tau}^{\mathrm{D}} r_{t-\tau}^{\mathrm{N}} + 2\sum_{\tau=1}^{q} K^{\mathrm{DN}\rightarrow}(\tau) r_{t-\tau}^{\mathrm{D}} r_{t-\tau+1}^{\mathrm{N}},
\end{aligned}$$

or $\sigma_t^2 = s^2 + R_t^\dagger K R_t$, with $K \in \mathbb{R}^{(2q+1)\times(2q+1)}$, $R_t \in \mathbb{R}^{(2q+1)}$ defined by

$$K = \begin{pmatrix}
K^{\mathrm{DD}\rightarrow}(1) & & & K^{\mathrm{DN}\rightarrow}(1) & K^{\mathrm{ND}\rightarrow}(1) & & \\
& \ddots & & & \ddots & \ddots & \\
& & K^{\mathrm{DD}\rightarrow}(q) & & & K^{\mathrm{DN}\rightarrow}(q) & K^{\mathrm{ND}\rightarrow}(q) \\
K^{\mathrm{DN}\rightarrow}(1) & & & K^{\mathrm{NN}\rightarrow}(0) & & & \\
K^{\mathrm{ND}\rightarrow}(1) & \ddots & & & K^{\mathrm{NN}\rightarrow}(1) & & \\
& \ddots & K^{\mathrm{DN}\rightarrow}(q) & & & \ddots & \\
& & K^{\mathrm{ND}\rightarrow}(q) & & & & K^{\mathrm{NN}\rightarrow}(q)
\end{pmatrix}, \quad R_t = \begin{pmatrix}
r_{t-1}^{\mathrm{D}} \\ \vdots \\ r_{t-q}^{\mathrm{D}} \\ r_t^{\mathrm{N}} \\ r_{t-1}^{\mathrm{N}} \\ \vdots \\ r_{t-q}^{\mathrm{N}}
\end{pmatrix}.$$

The positive-semidefiniteness of $K$ is harder to characterize directly, so we use the « $\tau$ by $\tau$ » method to find a criterion for a sufficient condition. For any $\beta \in ]0,1[$ and any vector $v = \left(v_1^{\mathrm{D}}, \ldots, v_q^{\mathrm{D}}, v_0^{\mathrm{N}}, v_1^{\mathrm{N}}, \ldots, v_q^{\mathrm{N}}\right)^\dagger \in \mathbb{R}^{(2q+1)}$, the quadratic form $v^\dagger K v$ is decomposed as follows :

$$\begin{aligned} v^\dagger K v =& K^{\mathrm{NN}\rightarrow}(0) {v_0^{\mathrm{N}}}^2 + \sum_{\tau=1}^{q} \Big[ K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + K^{\mathrm{NN}\rightarrow}(\tau) {v_\tau^{\mathrm{N}}}^2 + 2K^{\mathrm{ND}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_\tau^{\mathrm{N}} + 2K^{\mathrm{DN}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_{\tau-1}^{\mathrm{N}} \Big] \\ =& \quad \beta K^{\mathrm{NN}\rightarrow}(0) {v_0^{\mathrm{N}}}^2 + (1-\beta) K^{\mathrm{NN}\rightarrow}(q) {v_q^{\mathrm{N}}}^2 \\ &+ \sum_{\tau=1}^{q} \Big[ \beta K^{\mathrm{NN}\rightarrow}(\tau) {v_\tau^{\mathrm{N}}}^2 + (1-\beta) K^{\mathrm{NN}\rightarrow}(\tau-1) {v_{\tau-1}^{\mathrm{N}}}^2 \\ &\qquad + K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + 2K^{\mathrm{ND}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_\tau^{\mathrm{N}} + 2K^{\mathrm{DN}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_{\tau-1}^{\mathrm{N}} \Big], \end{aligned}$$

and clearly, a sufficient condition for the sum to be non-negative is that each term is non-negative :

$$\begin{aligned} \forall t,\ \sigma_t^2 \geq 0 \Leftarrow\ & K \text{ is positive-semidefinite} \\ \Leftarrow\ & \exists \beta \in ]0,1[\ ,\ \forall v = \left(v_1^{\mathrm{D}}, \ldots, v_q^{\mathrm{D}}, v_0^{\mathrm{N}}, v_1^{\mathrm{N}}, \ldots, v_q^{\mathrm{N}}\right)^\dagger \in \mathbb{R}^{(2q+1)}, \forall \tau \in \{1, \ldots, q\}, \\ & 0 \leq \beta K^{\mathrm{NN}\rightarrow}(\tau) {v_\tau^{\mathrm{N}}}^2 + (1-\beta) K^{\mathrm{NN}\rightarrow}(\tau-1) {v_{\tau-1}^{\mathrm{N}}}^2 \\ & \quad + K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + 2K^{\mathrm{ND}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_\tau^{\mathrm{N}} + 2K^{\mathrm{DN}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_{\tau-1}^{\mathrm{N}} \\ \Leftarrow\ & \exists \beta,\ \forall v,\ \forall \tau \in \{1, \ldots, q\}, \quad \exists \alpha_\tau \in [0,1], \\ & \bullet\ \ |K^{\mathrm{ND}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_\tau^{\mathrm{N}}| \leq \frac{1}{2} \Big( \alpha_\tau\ K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + \beta K^{\mathrm{NN}\rightarrow}(\tau) {v_\tau^{\mathrm{N}}}^2 \Big) \\ & \bullet\ \ |K^{\mathrm{DN}\rightarrow}(\tau) v_\tau^{\mathrm{D}} v_{\tau-1}^{\mathrm{N}}| \leq \frac{1}{2} \Big( (1-\alpha_\tau)\ K^{\mathrm{DD}\rightarrow}(\tau) {v_\tau^{\mathrm{D}}}^2 + (1-\beta) K^{\mathrm{NN}\rightarrow}(\tau-1) {v_{\tau-1}^{\mathrm{N}}}^2 \Big) \\ \Leftarrow\ & \exists \beta,\ \forall \tau \in \{1, \ldots, q\}, \quad \exists \alpha_\tau \in [0,1], \\ & \bullet\ \ K^{\mathrm{ND}\rightarrow}(\tau)^2 \leq \beta\ \alpha_\tau\ K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau) \\ & \bullet\ \ K^{\mathrm{DN}\rightarrow}(\tau)^2 \leq (1-\beta)(1-\alpha_\tau)\ K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau-1). \end{aligned}$$

The last condition is equivalent to a simpler one, with $\alpha_\tau$ saturating one of the two inequalities : denoting $\delta^{(\mathrm{nn})}(\tau) = K^{\mathrm{NN}\rightarrow}(\tau)/K^{\mathrm{NN}\rightarrow}(\tau-1)$ and $\delta^{(\mathrm{nd})}(\tau) = K^{\mathrm{DN}\rightarrow}(\tau)/K^{\mathrm{ND}\rightarrow}(\tau)$, $K$ is positive-semidefinite if (but not only if), $\exists \beta \in ]0,1[,\ \forall \tau \in \{1, \ldots, q\}$,

$$\begin{aligned} \mathcal{M}(\beta, \tau) \equiv \max \Bigg\{ & \frac{\beta \delta^{(\mathrm{nn})}(\tau)}{1-\beta} \left( \frac{(1-\beta) K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau-1)}{K^{\mathrm{ND}\rightarrow}(\tau)^2} - \delta^{(\mathrm{nd})}(\tau)^2 \right), \\ & \frac{1-\beta}{\beta \delta^{(\mathrm{nn})}(\tau)} \left( \frac{\beta K^{\mathrm{DD}\rightarrow}(\tau) K^{\mathrm{NN}\rightarrow}(\tau)}{K^{\mathrm{DN}\rightarrow}(\tau)^2} - \frac{1}{\delta^{(\mathrm{nd})}(\tau)^2} \right) \Bigg\} \end{aligned}$$

is larger than one, yielding the following a.s. positive volatility criterion :

$$\forall t,\ \sigma_t^2 \geq 0 \Leftarrow 1 \leq \sup_{\beta \in ]0,1[} \min_{1 \leq \tau \leq q} \mathcal{M}(\beta, \tau).$$

### 3.5.3 With leverage coefficients

We now add leverage terms to the volatility equation :

$$\sigma_t^2 = s^2 + R_t^\dagger K R_t + R_t^\dagger L = \widetilde{R}_t^\dagger M \widetilde{R}_t,$$

with

$$M = \begin{pmatrix} K & \frac{1}{2}L \\ \frac{1}{2}L^\dagger & s^2 \end{pmatrix},$$

and appropriate vectors $R_t, L_t, \widetilde{R}_t$. It is easy to show that, assuming a positive-definite $K$,

$$M \text{ is positive-semidefinite } \Leftrightarrow \; L^\dagger K^{-1} L \leq 4s^2. \tag{3.17}$$

## 3.6 Appendix : Universality assumption

To obtain a better convergence of the parameters of the model, the estimates are averaged over a pool of 280 US stocks. The validity of this method lies on the assumption that the model is approximately universal, i.e. that the values of its coefficients do not vary significantly among stocks.

We check that this assumption is relevant by splitting the stock pool in two halves and running the estimation of the model on the two halves independently. We obtain a set $\Theta_1 \in \mathbb{R}^{17}$ of coefficients calibrated on the first half and a set $\Theta_2 \in \mathbb{R}^{17}$ on the second half (each set contains 17 parameters, three for each of the four $K$ kernels, two for each of the two $L$ kernels, plus $\nu$).

If the (normalized) returns series for each stock were realizations of the same process, the differences between the coefficients of the two half stock pools would be explained by statistical noise only. To quantify how close the observed differences are to statistical noise, we run a series of Wald tests and study the obtained p-values. We test $H_0 = \{f(\Theta) = 0\}$ against $H_1 = \{f(\Theta) \neq 0\}$, where $\Theta = (\Theta_1, \Theta_2) \in \mathbb{R}^{34}$ , $f(\Theta) = \Theta_1 - \Theta_2$ , $f : \mathbb{R}^{34} \mapsto \mathbb{R}^{17}$, by comparing the statistic

$$\Xi_n = n\, f(\Theta)^\dagger\, \Sigma(\Theta)^{-1}\, f(\Theta), \quad \text{with} \quad \Sigma(\Theta) = \frac{\partial f}{\partial \Theta}(\Theta)\, I(\Theta)^{-1} \left(\frac{\partial f}{\partial \Theta}(\Theta)\right)^\dagger, \tag{3.18}$$

to the quantiles of a $\chi^2$ variable, where $n = \frac{1}{2} \times 280 \times 2515$ is the sample size for each half stock pool, $I(\Theta)$ is the Fisher Information matrix of the model and $\frac{\partial f}{\partial \Theta}$ is the Jacobian matrix of $f(\Theta)$.

For intra-day volatility, the p-value is close to zero if all the 17 coefficients are included, but becomes very high (p-val = 12.3%) if we exclude $\alpha^{\mathrm{NN}\to\mathrm{D}}$ from the test. One can conclude that all the parameters but $\alpha^{\mathrm{NN}\to\mathrm{D}}$ can be considered universal, with a high significance level of 10%. It is not surprising that at least one coefficient varies slightly among stocks (it would indeed be a huge discovery to find that 280 US stocks can be considered as identically distributed!).

In the case of overnight volatility, we first test the universality of the parameters in $K^{\mathrm{DD}\to\mathrm{N}}$ and $K^{\mathrm{NN}\to\mathrm{N}}$, for which the constraint $s^{\mathrm{N}2} = 0$ is included in the estimation. We then test the other 11 parameters for universality. Once again, a few of them ($g_{\mathrm{p}}^{\mathrm{ND}\to\mathrm{N}}$, $\alpha^{\mathrm{NN}\to\mathrm{N}}$ and $\omega_{\mathrm{p}}^{\mathrm{NN}\to\mathrm{N}}$) must be excluded from the tests to obtain acceptable p-values. We then obtain p-val = 1% for the first test and p-val = 7.5% for the second.

It is then natural to wonder whether the four coefficients that cannot be considered as statistically universal differ significantly in relative values. That is why we compute a second comparison criterion : for a pair $(c^{(1)}, c^{(2)})$ of coefficients estimated on the first and second half stock pools respectively, we compute the relative difference, defined as :

$$\frac{|c^{(1)} - c^{(2)}|}{\max\{|c^{(1)}|, |c^{(2)}|\}}.$$

The values of this criterion are summarized in Tabs. 3.6,3.7. The first observation is that no relative difference exceeds 50% (except for three of the ill-determined $\omega_{\mathrm{p}}$) which indicates that the signs and orders of magnitude of the coefficients of the model are invariant among stocks. The coefficients for which the relative difference is high but the statistical one is low do not contradict the universality assumption : the ML estimation would need a larger dataset to determine them with precision, and the difference between the two halves can be interpreted as statistical noise.

Three of the four « non-universal » coefficients, $\alpha^{\mathrm{NN}\to\mathrm{D}}$, $g_{\mathrm{p}}^{\mathrm{ND}\to\mathrm{N}}$ and $\omega_{\mathrm{p}}^{\mathrm{NN}\to\mathrm{N}}$ also show a significant relative difference between the two stock pools (above 10%). These are thus the only coefficients of the model for which averaging over all stocks is in principle not well-justified, and for which the confidence intervals given in Sect. 3.3.1 should be larger. However, these variations do not impact the global shapes of the corresponding kernels in a major way, and our qualitative comments on the feedback of past returns on future intra-day and overnight volatilities are still valid.

The results of this section indicate that most coefficients of the model are compatible with the assumption of universality. Although some coefficients do show slight variations, our stock aggregation method (with proper normalization, as presented in Sect. 3.2.3) is reasonable.

| Kernels | $K(\tau) = g_{\mathrm{p}}\,\tau^{-\alpha}\,e^{-\omega_{\mathrm{p}}\,\tau}$ | | | Kernels | $L(\tau) = g_{\mathrm{e}}\,e^{-\omega_{\mathrm{e}}\,\tau}$ | |
|---|---|---|---|---|---|---|
| | $g_{\mathrm{p}}$ | $\alpha$ | $\omega_{\mathrm{p}}$ | | $g_{\mathrm{e}}$ | $\omega_{\mathrm{e}}$ |
| $K^{\mathrm{DD}\to\mathrm{D}}$ | 9.4% | 7.3% | 13.1% | $L^{\mathrm{D}\to\mathrm{D}}$ | 10.4% | 1.8% |
| $K^{\mathrm{NN}\to\mathrm{D}}$ | 2.6% | 17.2% | **34.4 %** | | | |
| $K^{\mathrm{ND}\to\mathrm{D}}$ | 13.3% | 9.6% | **77.6 %** | $L^{\mathrm{N}\to\mathrm{D}}$ | 5.8% | 2.3% |
| $K^{\mathrm{DN}\to\mathrm{D}}$ | 2.0% | 9.9% | **94.8 %** | | | |

TABLE 3.6 – Intra-day volatility : relative differences between the two half stock pools ($q = 512$). For $\nu^{\mathrm{D}}$, the value is 7.1%. Bold figures are above 20%.

| Kernels | $K(\tau) = g_{\mathrm{p}}\,\tau^{-\alpha}\,e^{-\omega_{\mathrm{p}}\,\tau}$ | | | Kernels | $L(\tau) = g_{\mathrm{e}}\,e^{-\omega_{\mathrm{e}}\,\tau}$ | |
|---|---|---|---|---|---|---|
| | $g_{\mathrm{p}}$ | $\alpha$ | $\omega_{\mathrm{p}}$ | | $g_{\mathrm{e}}$ | $\omega_{\mathrm{e}}$ |
| $K^{\mathrm{DD}\to\mathrm{N}}$ | 3.1% | 9.7% | 17.0% | $L^{\mathrm{D}\to\mathrm{N}}$ | 18.3% | **32.0 %** |
| $K^{\mathrm{NN}\to\mathrm{N}}$ | 8.3% | 7.6% | **33.0 %** | | | |
| $K^{\mathrm{ND}\to\mathrm{N}}$ | **30.1 %** | 1.7% | **80.0 %** | $L^{\mathrm{N}\to\mathrm{N}}$ | 19.4% | 6.5% |
| $K^{\mathrm{DN}\to\mathrm{N}}$ | **35.5 %** | **21.0 %** | 0.2% | | | |

TABLE 3.7 – Overnight volatility : relative differences between the two half stock pools ($q = 512$). For $\nu^{\mathrm{N}}$ the value is 0.03%. Bold figures are above 20%.

## 3.7 Appendix : The case of European stocks : results and discussions

In order to verify that our conclusions are global and not specific to US stock markets, we also calibrated our model on European stock returns. The dataset is composed of daily prices for 179 European stocks of the Bloomberg European 500 index, on the same period 2000–2009. The data treatment is exactly the same as before. The following sections analyze and compare the results to those obtained on US stocks.

### 3.7.1 The feedback kernels : parameters estimates

ML estimates of the parameters in the regression kernels (for a maximum lag $q = 512$) are reported in Tabs. 3.8,3.9, and the resulting kernels are shown in Figs. 3.5(a),3.5(b).

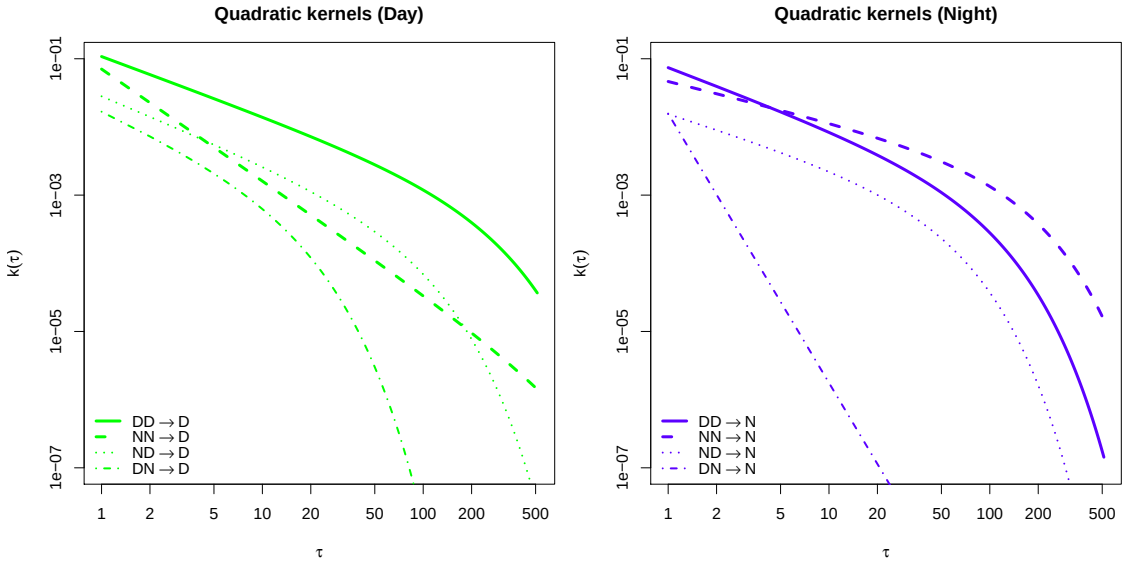
| Kernels | $K(\tau) = g_{\mathrm{p}}\,\tau^{-\alpha}\,e^{-\omega_{\mathrm{p}}\times\tau}$ | | | $L(\tau) = g_{\mathrm{e}}\,e^{-\omega_{\mathrm{e}}\times\tau}$ | |
|---|---|---|---|---|---|
| | $g_{\mathrm{p}} \times 10^2$ | $\alpha$ | $\omega_{\mathrm{p}} \times 10^2$ | $g_{\mathrm{e}} \times 10^2$ | $\omega_{\mathrm{e}} \times 10^2$ |
| $K^{\mathrm{DD}\to\mathrm{D}}$ | $10.83 \pm 0.11$ | $0.87 \pm 0.005$ | $0.50 \pm 0.03$ | $--$ | $--$ |
| $K^{\mathrm{NN}\to\mathrm{D}}$ | $7.06 \pm 0.24$ | $1.64 \pm 0.03$ | $0.12 \pm 0.28$ | $--$ | $--$ |
| $K^{\mathrm{ND}\to\mathrm{D}}$ | $2.86 \pm 0.19$ | $0.98 \pm 0.06$ | $1.51 \pm 0.84$ | $--$ | $--$ |
| $K^{\mathrm{DN}\to\mathrm{D}}$ | $1.83 \pm 0.30$ | $1.08 \pm 0.26$ | $9.02 \pm 8.11$ | $--$ | $--$ |
| $L^{\mathrm{D}\to\mathrm{D}}$ | $--$ | $--$ | $--$ | $-3.20 \pm 0.27$ | $15.29 \pm 1.39$ |
| $L^{\mathrm{N}\to\mathrm{D}}$ | $--$ | $--$ | $--$ | $-3.73 \pm 0.51$ | $35.35 \pm 4.69$ |

TABLE 3.8 – Intra-day volatility : estimated kernel parameters for $K$'s and $L$'s, with their asymptotic confidence intervals of level 95%, as computed using the Fisher Information matrix with the Gaussian quantile (1.98).

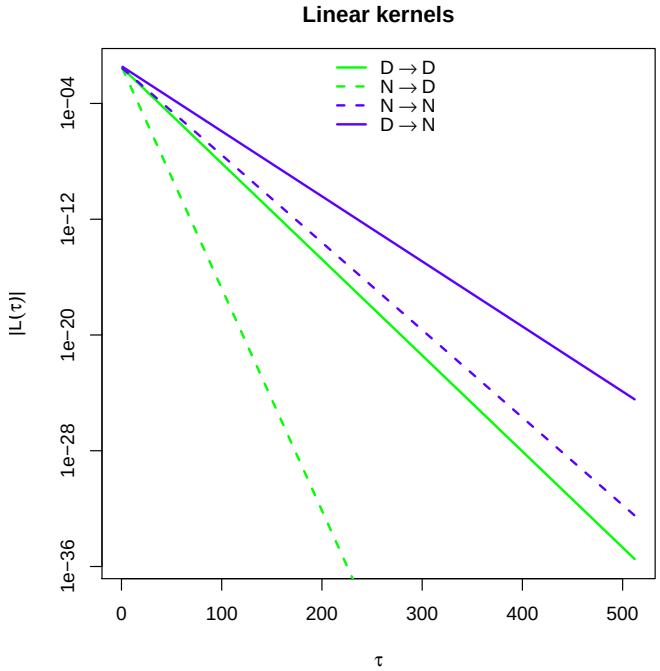
| Kernels | $K(\tau) = g_{\mathrm{p}}\,\tau^{-\alpha}\,e^{-\omega_{\mathrm{p}}\times\tau}$ | | | $L(\tau) = g_{\mathrm{e}}\,e^{-\omega_{\mathrm{e}}\times\tau}$ | |
|---|---|---|---|---|---|
| | $g_{\mathrm{p}} \times 10^2$ | $\alpha$ | $\omega_{\mathrm{p}} \times 10^2$ | $g_{\mathrm{e}} \times 10^2$ | $\omega_{\mathrm{e}} \times 10^2$ |
| $K^{\mathrm{DD}\to\mathrm{N}}$ | $7.53 \pm 0.39$ | $0.89 \pm 0.03$ | $1.49 \pm 0.57$ | $--$ | $--$ |
| $K^{\mathrm{NN}\to\mathrm{N}}$ | $4.69 \pm 0.25$ | $0.58 \pm 0.01$ | $0.86 \pm 0.07$ | $--$ | $--$ |
| $K^{\mathrm{ND}\to\mathrm{N}}$ | $1.59 \pm 0.20$ | $0.75 \pm 0.14$ | $2.62 \pm 2.07$ | $--$ | $--$ |
| $K^{\mathrm{DN}\to\mathrm{N}}$ | $-1.57 \pm 0.33$ | $3.95 \pm 1.32$ | $0.02 \pm 13.67$ | $--$ | $--$ |
| $L^{\mathrm{D}\to\mathrm{N}}$ | $--$ | $--$ | $--$ | $-3.78 \pm 0.23$ | $10.36 \pm 0.71$ |
| $L^{\mathrm{N}\to\mathrm{N}}$ | $--$ | $--$ | $--$ | $-3.09 \pm 0.29$ | $13.93 \pm 1.74$ |

TABLE 3.9 – Overnight volatility : estimated kernel parameters for $K$'s and $L$'s, with their asymptotic confidence intervals of level 95%, as computed using the Fisher Information matrix with the Gaussian quantile (1.98).

**Intra-day volatility :** From Tab. 3.8 and Fig. 3.5(a), we see that all the conclusions drawn previously for the case of US stocks hold for European stocks. The intra-day feedback is stronger and of much longer memory than overnight feedback, which decays very quickly (although more slowly for European stocks, with $\alpha \simeq 1.6$ instead of $\alpha \simeq 2.3$). The cross kernels are still clearly smaller than the two quadratic ones, with $\alpha$ close to unity.




(a) Quadratic $K(\tau)$ kernels in log-log scale. For 'DN→N', the absolute value of the (negative) kernel is plotted.




(b) Linear $L(\tau)$ kernels in lin-log scale. All leverage kernels are negative, so $-L(\tau)$ are plotted.

FIGURE 3.5 – Estimated kernels, impacting Intra-Days in (light) green, OverNights in (dark) blue.

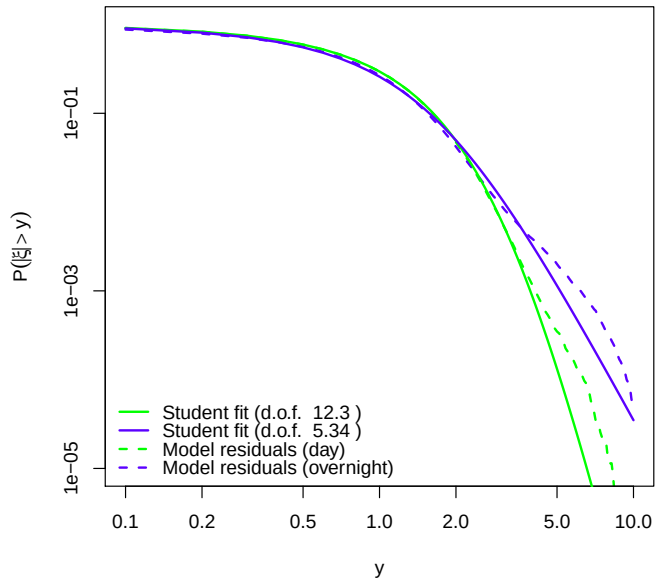



FIGURE 3.6 – CDF $\mathbb{P}(|\xi| > y)$ of the residuals $\xi^{\mathrm{D}}$ and $\xi^{\mathrm{N}}$, in log-log scale.

The leverage effect is similar to the case of US stocks too, although its initial amplitude is approximately equal for past intra-day and overnight returns, whereas past intra-days are stronger than overnights for US stocks.

**Overnight volatility :** In the case of overnight volatility, we see from Tab. 3.9 and Fig. 3.5(a) that not only do all our previous conclusions still hold in the European case (long memory for both intra-day and overnight feedback, second cross kernel nearly equal to zero), but the coefficients of the model are remarkably close to those of the US calibration.

### 3.7.2 Distribution of the residuals

For intra-day returns, the distribution of the residuals is very similar to the case of US stocks. However, for overnight returns, some differences must be pointed out. Firstly, as can be seen on Tab. 3.10, $\nu^{\mathrm{N}}$ is significantly higher for European stock (5.3) than for US stocks (3.6). As a consequence, the kurtosis of overnight residuals is finite : $3 + \frac{6}{\nu^{\mathrm{N}}-4} \simeq 7.5$. The distribution is still highly leptokurtic, but the result is less extreme than for US stocks. Secondly, figure 3.6 shows that the quality of the Student fit is of lesser quality here. For European stocks, both intra-day and overnight residuals seem to be fitted by a lower value of $\nu$ for far tail events, whereas this only held for intra-day residuals in the US case.

| $q$ | 512 |
|---|---|
| $\nu^{\mathrm{D}}(\pm 0.4)$ | 12.3 |
| $\nu^{\mathrm{N}}(\pm 0.06)$ | 5.34 |
| $\rho^{\mathrm{D}}$ | 10.0% |
| $\rho^{\mathrm{N}}$ | 0.0% |

TABLE 3.10 – Baseline volatility and tail index of the Student residuals for $q = 512$.

### 3.7.3 Baseline volatility

Finally, we compare the ratios $\rho^{\mathrm{J}} = {s^{\mathrm{J}}}^2/\langle {\sigma^{\mathrm{J}}}^2 \rangle$ of the two stock pools. Here again, the results are very close to our previous calibration : $\rho^{\mathrm{D}} \simeq 0.1$ for intra-day volatility, $\rho^{\mathrm{N}} \simeq 0$ for overnight volatility. Like in the case of US stocks, the calibration procedure yields a slightly negative ${s^{\mathrm{N}}}^2$, so we add an additional step that includes the constraint ${s^{\mathrm{N}}}^2 = 0$ (for overnight volatility only).

One of our main conclusions for US stocks is that overnight volatility is entirely endogenous, and that the exogeity of overnight returns is contained in the leptokurtic distribution of overnight residuals. This section proves that this is also true for European stocks and suggests that our findings hold quite generally.

# Chapitre 4

# Un modèle de rétroaction quadratique pour la volatilité à haute fréquence

Ce chapitre est issu d'un travail en cours avec Jonathan Donier et Jean-Philippe Bouchaud.

**Abstract.** We introduce the QHawkes (Quadratic Hawkes) model which generalizes the Hawkes price models introduced by Bacry et al., by allowing quadratic feedback effects in the jump intensity. A non-parametric fit on NYSE stock data shows that the quadratic, off-diagonal component has indeed a structure that linear Hawkes models would fail to reproduce. This model exhibits two main properties, that we believe are crucial in the modelling and the understanding of the volatility process : first, the model is time-reversal asymmetric, similar to financial markets whose time evolution has a preferred direction. Second, it generates a fat-tailed volatility process, for which we give the SDE in the simple case of exponentially decaying kernels, and which is linked to Pearson diffusions in the continuous limit.

## 4.1 Introduction

It is very common in the financial literature to assume that the log-price $P_t$ of assets follow a diffusion equation of the form

$$\mathrm{d}P_t = \mu \mathrm{d}t + \sigma \mathrm{d}W_t \tag{4.1}$$

where $W$ is a Wiener process. When it comes to pricing derivatives, only the volatility process $(\sigma_t)$ matters, since the drift term disappears under the risk-neutral measure. More generally, $\sigma$ accounts for market risk : it is therefore crucial to understand its dynamics, for either derivatives pricing [25] or optimal investment [89].

In this context, a flurry of volatility models have emerged, most of them motivated by the need of derivatives traders to price and hedge their portfolios [20, 22, 49, 67, 76]. A common feature that they share, is that they describe the low-frequency dynamics of volatility, in a regime where the price can indeed be considered as a continuous, real-valued stochastic process. However, at high frequencies – i.e. at the scale of market events – there is no such thing as a continuous price process : at these scales, one instead faces a minimal price increment (called the tick) so that the price evolves on a discrete

price grid. A comprehensive understanding of the volatility process, from the scale of the event up to macroscopic scales, would seem very valuable in several respects, in particular that of market design, in order to understand how a change in market microstructural rules (e.g. the tick size) may affect its macroscopic properties (e.g. volatility). That one could find solid, behavioural microscopic foundations to the volatility process seems crucial : when fully understood, simple constraints on the agents might then change the overall, macroscopic market behaviour.

A natural high-frequency counterpart of the class of models (4.1), is the family of point processes (more precisely, the difference between two point processes) with jump intensity $(\lambda_t)$ that plays the role of volatility. In order to reproduce the volatility clustering effect, Bacry et al. have recently proposed the use of Hawkes processes [17], that lead to price models that mimic many empirical properties of high-frequency prices in an intuitive and tractable mathematical framework, and allows for a high-frequency interpretation of the volatility process. More precisely, they consist in self-exciting jump processes $(N_t)_{t\geq 0}$ of intensity $\lambda_t$ that depends on the history of the process via the auto-regressive relation

$$\lambda_t = \lambda_\infty + \int_{-\infty}^{t} \phi(t-s)\, \mathrm{d}N_s, \tag{4.2}$$

where $\lambda_\infty$ is a baseline intensity and $\phi$ is a nonnegative, measurable function. They are called "self-exciting", because every jump $\mathrm{d}N_s$ increases the probability of future events to happen by increasing the future jump intensity $\lambda_t$ for $t > s$ via the kernel $\phi$. Such process however needs to be confronted with empirical findings, in order to assess its practical relevance. After many years of academic research, people have found that the volatility process exhibits three important features that should be included in a complete high-frequency volatility/price model :

- A positive and slowly-decaying auto-correlation in the time series of volatility or number of events [42],
- Leptokurtic returns, which, to be in accordance with empirical findings, should be asymptotically distributed as a power-law $p(r) \sim r^{-\alpha}$ of exponent $\alpha$ that varies typicall between 3 and 5 depending on the asset, the period and the sampling frequency [41, 42, 91],
- Significant time asymmetry (causality) due to the fact that a succession of price moves in the same direction triggers more volatility than a succession of compensated price moves (see [39] for some empirical evidence of this fact).

While very appealing from a mathematical and conceptual point of view, linear Hawkes processes *cannot* as such reproduce the last two bullet points of the previous list (leptokurtic returns and significant time asymmetry). This could explain the fact that the calibration of such processes on real financial data systematically leads to a saturated version of the model (i.e. a norm equal to one, see [71]).

The goal of the present paper is to introduce an extension of the Hawkes framework, the QHawkes model, that palliates some of its weaknesses by replacing the linear feedback term in the intensity process by a more general, quadratic one. The QHawkes process appears as a high frequency, continuous time analog of the QARCH model introduced by Sentana [102]. As we shall see, the QHawkes model seems very promising as it succeeds at reproducing fat-tail returns as well as significant time asymmetry. Moreover, we introduce a particular case in which the continuous-time limit boils down to a simple, tractable Pearson diffusion, which can then be used as a low-frequency proxy for the volatility process, and used e.g. for options pricing. We first introduce the model in Section 4.2,

and highlight some of its general properties. Section 4.3 works out the parallel with QARCH models, which we calibrate on intra-day US stock data using the methodology similar to [39], showing a non-zero off-diagonal structure. Section 4.4 introduces a particular sub-family of QHawkes processes, which we call ZHawkes, and for which the kernel can be factorized. We also show that in the case of exponential kernels the process is Markovian, and we write the corresponding stochastic differential equation as well as its continuous counterpart. Finally, we show that ZHawkes processes achieve significant time asymmetry, with order of magnitude that matches data well, and produce a fat-tailed volatility process. Section 4.5 then concludes.

## 4.2 The QHawkes model

### 4.2.1 General model

Similarly to Hawkes processes (4.2), the QHawkes process $(P_t)_{t\geq 0}$ is a self-exciting point process, whose intensity $\lambda_t$ is dependent on the past realization of the process itself. More precisely, we model the intensity of price changes as

$$\lambda_t = \lambda_\infty \; + \; \frac{1}{\omega}\int_{-\infty}^{t} L(t-s)\; \mathrm{d}P_s \; + \; \frac{1}{\omega^2}\int_{-\infty}^{t}\int_{-\infty}^{t} K(t-s,t-u)\; \mathrm{d}P_s\; \mathrm{d}P_u, \tag{4.3}$$

where $P$ is the high-frequency price, which is a pure jump process with signed increments, $L : \mathbb{R}^+ \to \mathbb{R}$ is a leverage kernel and $K : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is a quadratic feedback kernel. $\lambda_\infty$ and $\omega$ are two positive constants that represent respectively the baseline intensity of the process and the standard deviation of price jumps. Although necessary on daily time scales to account for the leverage effect, we will see later that at the intra-day scale the kernel $L$ is not significant, so for many applications one can focus on the quadratic kernel and write

$$\lambda_t = \lambda_\infty \; + \; \frac{1}{\omega^2}\int_{-\infty}^{t}\int_{-\infty}^{t} K(t-s,t-u)\; \mathrm{d}P_s\; \mathrm{d}P_u. \tag{4.4}$$

It is easy to see that the models (4.3) and (4.4) encompass the linear Hawkes price model of [17] : by taking unit price jumps $\mathrm{d}P_t = \pm\omega$ where $\omega$ can be seen as the tick and discarding the leverage kernel ($L \equiv 0$) as well as the off-diagonal quadratic effects (so that $K(t,s) = \phi(t)\delta_{t-s}$), we recover a Hawkes process of kernel $\phi$.

Let us give some intuition on the last quadratic term in Eq. (4.3). It is well known that the linear Hawkes process (4.2) can be seen as a branching process, where each « immigrant » event from the exogenous intensity $\lambda_\infty$ gives birth to a number of « children » events distributed as a Poisson law of parameter $||\phi||_1$, where $||\phi||_1$ is the $L^1$ norm of the kernel $\phi$. Each of these children in turn gives birth to a second generation of children with the same probability law and so on. The intuition behind the QHawkes in terms of a branching process is very similar, except that now the rate of events also depends on the interaction between the pairs of events. Usually, one will consider a positive feedback such that two mother events with the same sign (i.e. two prices moves in the same direction) increase the probability of a new event to be triggered in the future (i.e. increase future volatility), whereas compensated events have inhibiting effects, in line with (and directly motivated by) the empirical observations of [39] as emphasized in the introduction.

### 4.2.2 Mathematical framework

Let us start by precising the mathematical definition of the objects present in Equation (4.3) :
- $(P_t)_{t\in\mathbb{R}}$ is a pure jump process of stochastic intensity $(\lambda_t)_{t\in\mathbb{R}}$, with unpredictable i.i.d. jump sizes of common law $p$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We assume that $\int_{\mathbb{R}} \xi \; p(\mathrm{d}\xi) = 0$ and $\int_{\mathbb{R}} \xi^2 \; p(\mathrm{d}\xi) = \omega^2 < +\infty$, i.e. that jumps are centred and have a finite variance.
- $\mathcal{F}_t = \sigma(P_s, s \leq t)$ is the natural filtration of $P$.
- $m(\mathrm{d}t, \mathrm{d}\xi)$ is the Punctual Poisson Measure associated to $P$, such that for all $t \in \mathbb{R}$ and $A \in \mathcal{B}(\mathbb{R})$,

$$\lim_{h\to 0} \frac{1}{h} \mathbb{E}\left[m([t, t+h[, A) \middle| \mathcal{F}_t\right] = \lambda_t \; p(A).$$

The quadratic kernel $K : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is assumed to satisfy
- Symmetry : $\forall s, t \geq 0, \; K(t,s) = K(s,t)$,
- Positivity : $\forall f \in L^2(\mathbb{R}^+), \; \int_0^{+\infty} \int_0^{+\infty} K(t,s) f(t) f(s) \; \mathrm{d}t \; \mathrm{d}s \; \geq 0$,
- Non-explosion : $\int_0^{+\infty} |K(t,t)| \; \mathrm{d}t \; < +\infty$.

$K$ defines an integral operator $T_K : L^2(\mathbb{R}^+) \to L^2(\mathbb{R}^+)$ which maps $f \in L^2(\mathbb{R}^+)$ to $T_K f : t \mapsto \int_0^{+\infty} K(t,s) f(s) \; \mathrm{d}s$. If $K$ is continuous, this operator is Hilbert-Schmidt and thus compact and one has $K(t,t) \geq 0$ for all $t \geq 0$ (see [36]). We define the trace of $K$

$$\mathrm{Tr}(K) = \int_0^{+\infty} K(t,t) \; \mathrm{d}t \; < +\infty.$$

The leverage kernel $L : \mathbb{R}^+ \to \mathbb{R}$ is assumed to be a measurable function. By analogy with QARCH models (see [39]) it should be dominated by $K$ in some way to ensure the positivity of the intensity $\lambda_t$. Since the leverage kernel is found empirically negligible in the sequel, we leave this positivity condition for future research.

### 4.2.3 Condition for time stationarity

In the case of linear Hawkes processes, it has been shown that stationarity is obtained as soon as the norm of the kernel verifies $\|\phi\|_1 < 1$. Intuitively, this means that each event triggers on average less than one child event, so that the clusters generated by each ancestor eventually die out. If this condition is violated, the probability that an ancestor generates an infinite number of events is non-zero, which can result in a stationary process only in the case $\|\phi\|_1 = 1$ and $\lambda_\infty = 0$ studied in [35]. Because of the quadratic feedback, the QHawkes process cannot be interpreted as a simple branching process, making things somewhat trickier. The goal of this section is to find a necessary condition for (first order) time stationarity.

We define the jump process $(N_t)$ that has the same jump times as $(P_t)$, with $\Delta N_\tau = (\Delta P_\tau)^2/\omega^2$ ($= 1$ iff $\Delta P_\tau = \pm\omega$) for any jump time $\tau$, and re-write Equation (4.3) as

$$\lambda_t = \lambda_\infty + \mathcal{L}_t + A_t + 2M_t \tag{4.5}$$

with the notations

$$\begin{cases} \mathcal{L}_t = \frac{1}{\omega}\int_{-\infty}^{t} L(t-u)\ \mathrm{d}P_u & \text{(leverage)} \\ A_t = \int_{-\infty}^{t} K(t-u,t-u)\ \mathrm{d}N_u & \text{(diagonal)} \\ M_t = \frac{1}{\omega^2}\int_{-\infty}^{t} \Theta_{t,u}\ \mathrm{d}P_u & \text{(off-diagonal)} \end{cases}$$

where $\Theta_{t,u} = \int_{-\infty}^{u-} K(t-u,t-r)\ \mathrm{d}P_r$ is $(\mathcal{F}_u)_{u\le t}$-adapted for $t$ fixed. Since $P$ is a martingale, one has $\mathbb{E}[M_t] = 0$ and $\mathbb{E}[\mathcal{L}_t] = 0$. Therefore,

$$\begin{aligned} \mathbb{E}[\lambda_t] &= \lambda_\infty + \frac{1}{\omega^2}\mathbb{E}\left[\int_{\mathbb{R}}\int_{-\infty}^{t} K(t-s,t-s)\ \xi^2\ \tilde{m}(\mathrm{d}s,\mathrm{d}\xi)\right] \\ &= \lambda_\infty + \mathbb{E}\left[\int_{-\infty}^{t} K(t-s,t-s)\ \lambda_s\ \mathrm{d}s\right] \end{aligned}$$

by definition of the punctual Poisson measure $m(\mathrm{d}s,\mathrm{d}\xi)$. We obtain

$$\mathbb{E}[\lambda_t] = \lambda_\infty + \int_{-\infty}^{t} K(t-s,t-s)\ \mathbb{E}[\lambda_s]\,\mathrm{d}s.$$

A necessary condition for the process $(\lambda_t)_{t\in\mathbb{R}}$ to be in a stationary state is that its expected value $\overline{\lambda} \equiv \mathbb{E}[\lambda_t]$ is constant, positive and finite. This yields $\overline{\lambda} = \lambda_\infty + \overline{\lambda}\mathrm{Tr}(K)$, thus if $\lambda_\infty > 0$,

$$\overline{\lambda} \;=\; \frac{\lambda_\infty}{1-\mathrm{Tr}(K)}.$$

This leads to the necessary stationarity condition [1]

$$\lambda_\infty > 0 \text{ and } \mathrm{Tr}(K) < 1 \tag{4.6}$$
$$\text{or } \lambda_\infty = 0 \text{ and } \mathrm{Tr}(K) = 1. \tag{4.7}$$

### 4.2.4 Auto-correlation structure in the QHawkes model

It is quite common for such type of models to investigate the link between the input kernels and the auto-correlation functions of the generated process. For linear Hawkes processes, one finds a Wiener-Hopf equation that relates the two-points correlation function to the 1-d kernel [15]. In our case, one also needs to consider the three-points correlation function, since the input kernel is of dimension 2.

#### System of equations

We take the model with no leverage, $L \equiv 0$. Equation (4.5) becomes

$$\lambda_t = \lambda_\infty + A_t + 2M_t.$$

1. In the case of linear Hawkes processes, this condition is also sufficient to obtain stationarity in the case $\mathrm{Tr}(K) < 1$ (whereas the case $\mathrm{Tr}(K) = 1$ is more subtle, see [35]).

We define for $\tau \neq 0$ and $\tau_1 > 0, \tau_2 > 0, \tau_1 \neq \tau_2$, the correlation functions

$$\begin{aligned}\mathcal{C}(\tau) &\equiv \mathbb{E}\left[\frac{\mathrm{d}N_t}{\mathrm{d}t}\frac{\mathrm{d}N_{t-\tau}}{\mathrm{d}t}\right] - \overline{\lambda}^2 \;=\; \mathbb{E}\left[\lambda_t \frac{\mathrm{d}N_{t-\tau}}{\mathrm{d}t}\right] - \overline{\lambda}^2, \\ \mathcal{D}(\tau_1,\tau_2) &\equiv \frac{1}{\omega^2}\,\mathbb{E}\left[\frac{\mathrm{d}N_t}{\mathrm{d}t}\frac{\mathrm{d}P_{t-\tau_1}}{\mathrm{d}t}\frac{\mathrm{d}P_{t-\tau_2}}{\mathrm{d}t}\right] \;=\; \frac{1}{\omega^2}\,\mathbb{E}\left[\lambda_t \frac{\mathrm{d}P_{t-\tau_1}}{\mathrm{d}t}\frac{\mathrm{d}P_{t-\tau_2}}{\mathrm{d}t}\right].\end{aligned} \tag{4.8}$$

$\mathcal{C}$ is then extended continuously at zero, as in [72]. Let us note that $\mathcal{C}$ is even and that $\mathcal{D}$ is symmetric. Then, one finds the following relationship between the autocorrelation functions $(\mathcal{C}, \mathcal{D})$ and the kernel $K$ (cf derivation in Appendix 4.6.1) :

$$\mathcal{C}(\tau) = \kappa\overline{\lambda}K(\tau,\tau) + \int_{-\infty}^{\tau} K(\tau-u,\tau-u)\mathcal{C}(u)\mathrm{d}u + 2\int_{0+}^{\infty}\int_{u+}^{\infty} K(\tau+u,\tau+r)\mathcal{D}(u,r)\mathrm{d}r\mathrm{d}u, \tag{4.9}$$

$$\begin{aligned}\mathcal{D}(\tau_1,\tau_2) =& 2K(\tau_1,\tau_2)[\mathcal{C}(\tau_2-\tau_1) + \overline{\lambda}^2] \\ &+ \int_{(\tau_2-\tau_1)+}^{\tau_2} K(\tau_2-u,\tau_2-u)\mathcal{D}(u-\tau_2+\tau_1,u)\mathrm{d}u \\ &+ 2\int_{-\infty}^{(\tau_2-\tau_1)-} K(\tau_1,\tau_2-u)\mathcal{D}(\tau_2-\tau_1,\tau_2-\tau_1-u)\mathrm{d}u,\end{aligned} \tag{4.10}$$

where $\kappa = \frac{1}{\omega^4}\int_{\mathbb{R}} \xi^4\, p(\mathrm{d}\xi)$ is the kurtosis of price jumps ($\kappa = 1$ for constant price jumps). As $\mathcal{C}(\tau)$ and $\mathcal{D}(\tau_1,\tau_2)$ are directly measurable on the data, one can infer some properties of the kernel $K$ using the above equations.

### Asymptotic behaviour

Whereas the above equations (4.9) and (4.10) are difficult to solve in general, one can investigate the joint tail behaviours as $\tau \to \infty$ when both the kernel and the auto-correlation functions have power law decays. Let us assume that

$$\begin{cases} K(\tau v_1, \tau v_2) & \underset{\tau\to\infty}{\sim} \tilde{K}(v_1,v_2)\;\tau^{-2\delta} & \text{(off-diagonal)} \\ K(\tau,\tau) & \underset{\tau\to\infty}{\sim} c_0\;\tau^{-2\delta} & \text{(diagonal)} \\ \mathcal{D}(\tau v_1, \tau v_2) & \underset{\tau\to\infty}{\sim} \tilde{\mathcal{D}}(v_1,v_2)\;\tau^{-\rho} & \text{(3-points AC)} \\ \mathcal{C}(\tau) & \underset{\tau\to\infty}{\sim} c_1\;\tau^{-\beta} & \text{(2-points AC)} \end{cases} \tag{4.11}$$

where $c_0, c_1$ are constants and $\tilde{K}(v_1,v_2), \tilde{\mathcal{D}}(v_1,v_2)$ are bounded functions of $(v_1,v_2)$. The exponents $\delta, \beta$ and $\rho$ can be related by plugging these ansatzs into Eqs. (4.9) and (4.10), to find two possible phases for the auto-covariance structure when $\mathrm{Tr}(K) < 1$ (cf derivation in Appendix 4.6.2) :

$$\delta > 1 \Rightarrow \beta = \rho = 2\delta, \tag{4.12}$$

$$\frac{1}{2} < \delta < 1 \Rightarrow \beta = 4\delta - 2,\;\; \rho = 2\delta. \tag{4.13}$$

The interpretation of these two phases is straightforward. In the first phase (4.12), the tail of the auto-correlation functions directly comes from the tail of the diagonal part : direct effects then dominate

quadratic feedback effects. In the second phase (4.13) however, a more sophisticated phenomenon enters into play, as the off-diagonal effects feedback in such a way that they generate correlation with fatter tails than that of the diagonal part of the kernel. In this phase, if $\frac{1}{2} < \delta < \frac{3}{4}$ then $\beta$ is below unity ,which corresponds to a long memory process. This result is important as it means that the QHawkes process need not be critical (i.e. $\mathrm{Tr}(K) = 1$) to generate long memory, unlike standard, linear Hawkes processes.

Note that, with little incidence on the results, we could choose $K(\tau, \tau) \sim c_0 \ \tau^{-\alpha}$ for the diagonal part, with $\alpha \neq 2\delta$. Also, to complete the analysis, one should study separately the critical case $\mathrm{Tr}(K) = 1$ that yields different equations, but this is not our focus here since we mainly consider non-critical QHawkes processes in the sequel.

## 4.3 The intra-day QARCH model

### 4.3.1 QHawkes as a limit of QARCH

In this section we investigate the link between the QHawkes model given by (4.3) and the discrete QARCH model introduced by Sentana in [102]. For a fixed time step $\Delta > 0$, we define for all $t \in \mathbb{R}$ :
– the price (or log-price) increment between time $t$ and time $t + \Delta$ : $r_t^\Delta = P_{t+\Delta} - P_t$,
– the volatility at time $t$ : $\sigma_t^\Delta = \sqrt{\mathbb{E}\left[{r_t^\Delta}^2 \middle| \mathcal{F}_t\right]}$.
The QHawkes model appears as the limit (in some sense) when $\Delta \to 0^+$ of the QARCH model

$$ {\sigma_t^\Delta}^2 = {\sigma_\infty^\Delta}^2 \ + \sum_{\tau \geq 1} L^\Delta(\tau) \ r_{t-\tau\Delta}^\Delta \ + \sum_{\tau, \tau' \geq 1} K^\Delta(\tau, \tau') \ r_{t-\tau\Delta}^\Delta r_{t-\tau'\Delta}^\Delta, \tag{4.14} $$

where ${\sigma_\infty^\Delta}^2 = \omega^2 \lambda_\infty \Delta, L^\Delta(\tau) = L(\tau\Delta) \ \Delta$ and $K^\Delta(\tau, \tau') = K(\tau\Delta, \tau'\Delta) \ \Delta$. Indeed, for $t \in \mathbb{R}$,

$$ \begin{aligned} \mathbb{E}\left[{r_t^\Delta}^2 \middle| \mathcal{F}_t\right] &= \omega^2 \ \mathbb{P}\left(P_{t+\Delta} - P_t \neq 0 \middle| \mathcal{F}_t\right) \ + \ o(\Delta) \\ &= \omega^2 \ \lambda_t \ \Delta \ + \ o(\Delta), \end{aligned} $$

which implies the scaling :

$$ \frac{{\sigma_t^\Delta}^2}{\Delta} \underset{\Delta \to 0^+}{\longrightarrow} \omega^2 \ \lambda_t. $$

Thanks to this link between the two models, it is possible to calibrate a QARCH model on intra-day bin returns, and obtain some qualitative insight on the structure of the underlying QHawkes model. Indeed, the direct calibration of the latter would be significantly harder, more noisy and computationally more demanding, and is therefore beyond the scope of this introductory paper.

### 4.3.2 Intra-day calibration of a QARCH model

QARCH models have mainly been calibrated on daily data so far ([102], [39]). To give a starting point to our study of quadratic effects in high-frequency volatility, we calibrate a discrete QARCH on intra-day five-minute returns.

### Dataset and notations

We consider the same dataset as in [8], which is composed of stock prices on intra-day five-minute bins. It includes 133 stocks of the New York Stock Exchange, that have been traded without interruption between 1 January 2000 and 31 December 2009. This yields 2499 trading days, with 78 five-minute bins per day. For each bin, the open, close, high and low prices $(O, C, H, L > 0)$ are available. We consider the log-price process and define on each bin :

- The return $r = \ln(C/O)$.
- The Rogers-Satchell volatility $\sigma^{\mathrm{RS}} = \sqrt{\ln(H/O) \times \ln(H/C) + \ln(L/O) \times \ln(L/C)}$.

### Normalization procedure

To be able to consider that intra-day prices are (approximately) independent realizations of a stationary stochastic process, we need to normalize the data carefully. As a matter of fact, strong intra-day seasonalities may falsify the calibration results. This can be avoided to some extent through a cross-sectional and historical normalization. We take advantage of our large dataset to compute a cross-sectional intra-day volatility pattern for each trading day and we simplify the returns by this pattern, which dampens the effect of collective shocks. On the other hand, we use the intra-day/overnight model volatility model of [27] to factor out daily feedback effects and focus on pure intra-day dynamics. To explain our normalization protocol, we introduce the following notations :

- The 5-min bin index $1 \leq b \leq 78$, the day index $1 \leq t \leq 2499$ and the stock index $1 \leq u \leq 133$.
- The empirical averages : $\langle x_{u,t,.} \rangle$ means conditional average of $x$ over bins, for stock $u$ and day $t$ fixed ; $\langle x_{u,.,b} \rangle$ and $\langle x_{.,t,b} \rangle$ are defined similarly as the conditional averages over days/stocks ; $\langle x \rangle = \langle x_{.,.,.} \rangle$ means average of $x$ over stocks, days and bins.

We compute the cross-sectional volatility pattern of day $t$, that we use to normalize the data of stock $u$, as :

$$b \in \{1, \cdots, 78\} \;\mapsto\; v_{u,t}(b) \equiv \sqrt{\langle r^2_{u' \neq u,t,b} \rangle}.$$

For stock $u$, the value $r^2_{u,t,b}$ is excluded from the average, so that the normalization protocol does not cap the large returns of stock $u$ artificially. We also consider the open-to-close volatility $\sigma^{\mathrm{D}}_{u,t}$ of day $t$ for stock $u$, as computed by the intra-day/overnight model of [27] with the data of stock $u$ over the days $\{1, \cdots, t-1\}$. For $t = 1$, we fix $\sigma^{\mathrm{D}}_{u,1} = 1$.

The normalization protocol is as follows : $\forall u, t, b$,

- $r_{u,t,b} \leftarrow r_{u,t,b}/\sigma^{\mathrm{D}}_{u,t}, \quad \sigma^{\mathrm{RS}}_{u,t,b} \leftarrow \sigma^{\mathrm{RS}}_{u,t,b}/\sigma^{\mathrm{D}}_{u,t}, \qquad$ (normalization by open-to-close volatility)
- $r_{u,t,b} \leftarrow r_{u,t,b}/v_{u,t}(b), \quad \sigma^{\mathrm{RS}}_{u,t,b} \leftarrow \sigma^{\mathrm{RS}}_{u,t,b}/v_{u,t}(b). \qquad$ (cross-sectional intra-day normalization)

Then, we exclude the trading days which include at least one bin where the absolute return is greater than the average plus six standard deviations. This represents approximately 7% of trading days, i.e. one day every three weeks. Combined with the cross-sectional pattern normalization, this data treatment strongly dampens the impacts of exceptional news events, which we do not aim to model here. Eventually, we set the mean of the squares to one and the average return to zero to make the stock universe more homogeneous : $\forall u, t, b$,

- $r_{u,t,b} \leftarrow r_{u,t,b}/\sqrt{\langle r^2_{u,.,.} \rangle}$, so that $\langle r^2 \rangle = 1$,
- $\sigma^{\mathrm{RS}}_{u,t,b} \leftarrow \sigma^{\mathrm{RS}}_{u,t,b}/\sqrt{\langle {\sigma^{\mathrm{RS}}}^2_{u,.,.} \rangle}$, so that $\langle {\sigma^{\mathrm{RS}}}^2 \rangle = 1$,

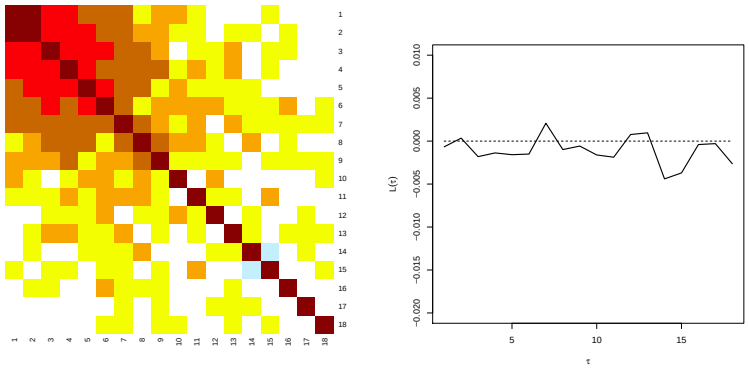


FIGURE 4.1 – QARCH kernels calibrated on five-minute intra-day returns for US stocks. The maximum lag is 18 bins, i.e. one hour and a half of trading time. Left : heatmap of the quadratic kernel. White coefficients are close to zero, blue ones are negative and yellow/orange/red ones are positive, with darker shades as they increase in absolute value. We see that all the significant coefficients are positive, with a non-negligible off-diagonal component. Right : leverage kernel. It is hardly distinct from zero and can be considered as pure noise (as opposed to daily models where it is significantly negative).

– $r_{u,t,b} \leftarrow r_{u,t,b} - \langle r_{u,.,.} \rangle$ so that $\langle r \rangle = 0$.

### Calibration results

The calibration process is similar to [39] and [27]. A first estimate of the kernels is obtained with the Generalized Method of Moments, which uses a set of correlation functions that are empirically observable. Then, using this estimate as a starting point, we use Maximum Likelihood Estimation, assuming that the residuals are t-distributed (which allows to account for possible fat tails that would remain in the residuals). This second step significantly improves the precision of the calibration results, compared to a solo GMM estimation.

We find it worth to notice that as opposed to the daily calibration results of [39], a clear off-diagonal structure appears in the feedback matrix in the intra-day case (see Figure 4.1). Also, the intra-day leverage kernel is found to be close to zero, justifying the fact that we mainly consider $L \equiv 0$ throughout the paper. The spectral decomposition of quadratic kernel (see Figure 4.2) suggests that $K$ is the superposition of a positive rank-one matrix and a diagonal one. Indeed, we obtain to a good approximation (see Figure 4.3)

$$K(\tau, \tau') \;\approx\; \phi(\tau)\delta_{\tau-\tau'} + k(\tau)k(\tau')$$

where

$$\phi(\tau) = g\tau^{-\alpha} \quad , \quad k(\tau) = \gamma \exp(-\delta\tau),$$

with $g = 0.09,\ \alpha = 0.60,\ \gamma = 0.14,\ \delta = 0.15$. Note that $\delta = 0.15$ corresponds to a characteristic time of about thirty minutes for the decay of the off-diagonal component. We then fix the off-diagonal part of the kernel $K$ as its fitted value $k(\tau)k(\tau') = \gamma^2 \exp(-\delta(\tau+\tau'))$, and we calibrate the diagonal

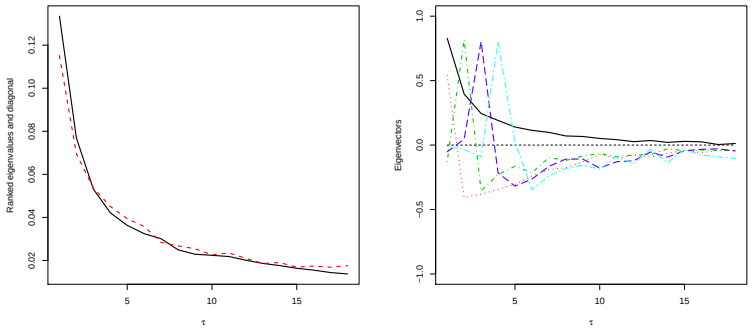



FIGURE 4.2 – Spectral decomposition of the quadratic QARCH kernel. Left : ranked eigenvalues (plain dark line) and diagonal coefficients (dashed). One can see that the diagonal coefficients are very close to the eigenvalues, except for the first eigenvalue which is significantly larger than the maximum of the diagonal. Right : eigenvectors corresponding to the five largest eigenvalues. The first eigenvector (plain dark line) is a positive decaying kernel, the others are close to the canonical vectors $e_i(\tau) = \delta_{i-\tau}$.

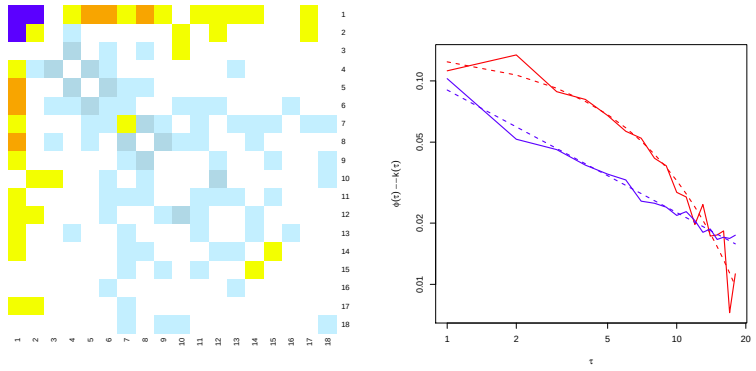



FIGURE 4.3 – Fit of the kernel $K$ by the sum of a power-law diagonal matrix and an exponential rank-one matrix. Left : heatmap of the difference between the fitted matrix and the original one. The coefficients are small (white or lightly-colored) except for the upper-left corner : the original matrix features a stronger short-term feedback. Right : kernels $\phi(\tau)$ and $k(\tau)$ that minimize the matrix distance $\sum[K(\tau,\tau') - \phi(\tau)\delta_{\tau-\tau'} - k(\tau)k(\tau')]^2$. The rank-one kernel $k$ is plotted in red (and is larger for small $\tau$'s), and the diagonal kernel $\phi$ is plotted in blue, both in log-log scale. The dashed lines are the power-law fit for $\phi(\tau)$ with exponent 0.6, and the exponential fit for $k(\tau)$ with characteristic time about 30 min.

of $K$ with a higher maximum lag of 60 bins (five hours of trading). We obtain

$$\phi_{\mathrm{lr}}(\tau) = g_{\mathrm{lr}}\tau^{-\alpha_{\mathrm{lr}}}$$

with the long-range coefficients $g_{\mathrm{lr}} = 0.09,\ \alpha_{\mathrm{lr}} = 0.76$. The residuals $\xi_t$ of the QARCH model, defined by

$$r_t = \sigma_t \xi_t,$$

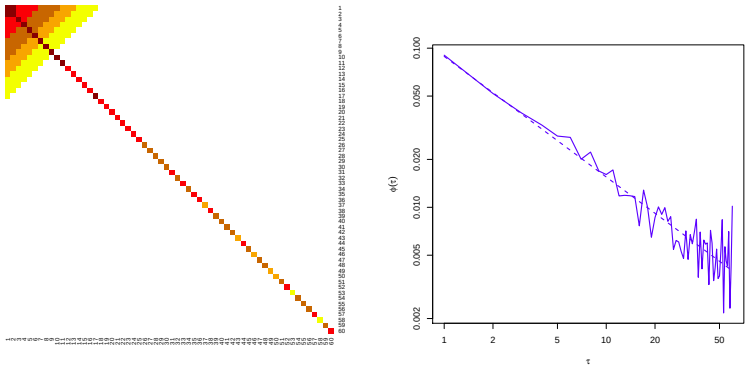



FIGURE 4.4 – Long-range kernel $K$. Left : heatmap of the long-range kernel, with the off-diagonal fixed as its exponential rank-one fit, and with the diagonal calibrated with no constraints. Right : long-range kernel $\phi_{\rm lr}(\tau) = K(\tau,\tau) - k^2(\tau)$. The kernel $\phi(\tau)$ is plotted in log-log scale, with its power-law fit with exponent 0.76 (dashed).

where $r_t$ is the five-minute return and $\sigma_t$ is the QARCH volatility, are modeled with Student's t-distribution. The calibration of the model with $K(\tau,\tau') = \phi(\tau)\delta_{\tau-\tau'} + k(\tau)k(\tau')$ yields $\nu \approx 7.9$ degrees of freedom for the residuals, which gives a kurtosis $\kappa \approx 4.5$. Thus, the QARCH model with this specific form of $K$ explains to a good extent the fat tails of five-minute returns.

In the QARCH model, the endogeneity ratio of the volatility (i.e. the proportion of the volatility that stems from feedback effects) is given by the trace $\mathrm{Tr}(K)$ of the quadratic kernel. With our parameterization and a maximum lag of $q \geq 1$, one has

$$\mathrm{Tr}(K) = \sum_{\tau=1}^{q} \phi(\tau) + \sum_{\tau=1}^{q} k^2(\tau).$$

We use the fits $k(\tau) = \gamma \exp(-\delta\tau)$ and $\phi_{\rm lr}(\tau) = g_{\rm lr}\tau^{-\alpha_{\rm lr}}$ to compute $\mathrm{Tr}(K)$ for $q = 78$, which is the total number of five-minute bins in a trading day. We obtain

$$\sum_{\tau=1}^{q} \phi(\tau) \simeq 0.74, \quad \sum_{\tau=1}^{q} k^2(\tau) \simeq 0.06 \quad \Rightarrow \quad \mathrm{Tr}(K) \simeq 0.80.$$

This endogeneity ratio may seem high, since it implies that 80% of the intra-day volatility is due to endogenous feedback effects. In fact, it is close to the value obtained for QARCH and ARCH models at a daily time scale, see [39] and [27]. These results plead in favor of a model in which the endogeneity ratio is constant across time scales, with values in the range $0.7 - 0.9$ depending on periods and asset classes. In particular, this range is significantly below the critical limit $\mathrm{Tr}(K) = 1$. We investigate in this direction in Section 4.4.2.

## 4.4 The ZHawkes model

### 4.4.1 Definition

Motivated by the results of the previous section, we consider the particular case of the QHawkes model where there is no leverage ($L \equiv 0$) and the quadratic feedback kernel $K$ is of the form

$$K(t,s) = \phi(t)\delta_{t-s} + k(t)k(s),$$

i.e. the sum of a diagonal Hawkes component and of a factorisable, rank one kernel. We assume that $\phi, k : \mathbb{R}^+ \to \mathbb{R}^+$ are two measurable functions that satisfy

$$||\phi|| \equiv \int_0^{+\infty} \phi(u)\ \mathrm{d}u < +\infty \quad , \quad ||k^2|| \equiv \int_0^{+\infty} k(u)^2\ \mathrm{d}u < +\infty.$$

The endogeneity ratio of the process is then

$$\mathrm{Tr}(K) = ||\phi|| + ||k^2|| \equiv n_H + n_Z,$$

where $n_H \equiv ||\phi||$ is the « Hawkes norm » and $n_Z \equiv ||k^2||$ is the « Zumbach norm ». Moreover, Equation (4.3) becomes in that case

$$\lambda_t = \lambda_\infty + H_t + Z_t^2, \tag{4.15}$$

where

– The « Hawkes term » is given by

$$H_t = \int_{-\infty}^{t} \phi(t-s)\ \mathrm{d}N_s,$$

where we recall the notation $\Delta N_\tau = (\Delta P_\tau)^2/\omega^2$ for a jump time $\tau$ of $P$.

– The « Zumbach term » is given by $Z_t^2$ where

$$Z_t = \frac{1}{\omega}\int_{-\infty}^{t} k(t-s)\ \mathrm{d}P_s.$$

Its name is inspired by empirical observations made by Gilles Zumbach on the volatility process ([110], [111]). In particular, the author finds that persistence in the signs of returns triggers more future volatility than compensated returns, as we explain in more detail in Section 4.4.3 [2]

We call this particular case of the QHawkes model, the ZHawkes model. Besides its empirical motivations, its factorization property significantly reduces the risk of over-fitting, since we are left with two one-dimensional kernels instead of the two-dimensional kernel in Eq. 4.3. As we see below, this simplified setup still captures the main phenomenology of price volatility, with in particular time-reversal asymmetry and fat tails, even for short-ranged kernels.

2. Although Zumbach describes this effect at the daily time scale.

### 4.4.2 Distribution of the volatility in the ZHawkes model

#### SDE in the exponential case

If the kernels $\phi$ and $k$ of the ZHawkes model have an exponential form, the process is Markovian and one can write a stochastic differential equation to describe its evolution. For the sake of simplicity, let us assume that the law of the jumps of $P$ is $p = (\delta_{-\omega} + \delta_\omega)/2$, where $\omega > 0$ is the typical size of a midprice jump (e.g. the tick size). Besides, we note $k(t) = \gamma \exp(-\delta t)$ and $\phi(t) = \alpha \exp(-\beta t)$ with $\gamma, \alpha \geq 0$, $\delta, \beta > 0$, which yields $n_H = \alpha/\beta$ for the Hawkes norm, $n_Z = \gamma^2/(2\delta)$ for the Zumbach norm and thus

$$\mathrm{Tr}(K) = \frac{\alpha}{\beta} + \frac{\gamma^2}{2\delta} < 1.$$

Then one has $\lambda_t = \lambda_\infty + H_t + Z_t^2$ where

$$\left\{ \begin{array}{lll} \mathrm{d}H_t & = -\beta\ H_t\ \mathrm{d}t & +\ \alpha\ \mathrm{d}N_t, \\ \mathrm{d}Z_t & = -\delta\ Z_t\ \mathrm{d}t & +\ \gamma\ \mathrm{d}P_t/\omega. \end{array} \right. \tag{4.16}$$

The processes $N$ and $P$ jump simultaneously with intensity $\lambda_t$ and amplitudes $\Delta N_\tau = 1$ and $\Delta P_\tau = \pm\omega$ with equal probability. Although quite simple, this system of jump SDEs lacks tractability compared to a continuous diffusion. Thus, we turn to the low-frequency asymptotics that one obtains as the number of jumps in a given time window becomes large, while their amplitudes are scaled down accordingly. This is the object of the following section.

#### Low-frequency asymptotics

The low-frequency asymptotics of nearly critical Hawkes processes with short-ranged kernels have been investigated by Jaisson and Rosenbaum [80], who show that for suitable scaling and convergence to the critical point $\|\phi\|_1 = 1$, the Hawkes-based price process of Bacry et al. [14] converges towards a Heston process (since the Hawkes intensity converges towards a CIR volatility process). The same authors [81] show that when the kernel exhibits power-law behaviour $\phi(t) \sim t^{-1-\alpha}$ with $1/2 < \alpha < 1$, the limiting process for the intensity is a fractional Brownian motion with Hurst exponent $H = \alpha - \frac{1}{2}$. The roughness of the latter process is well in agreement with the empirical results of [65] who find a Hurst exponent $H \simeq 0.1$ on financial data for the *log*-volatility. However, it is unclear how the Hawkes process intensity could be identified with the log-volatility, rather than the volatility itself, and a fat-tailed behaviour can by no means be reproduced by a simple, linear Hawkes process. Therefore, we consider in the present paper the low-frequency asymptotics of the ZHawkes model, which opens new modeling possibilities through the use of quadratic feedback effects.

For a time scale $T > 0$, we define the processes $\bar{H}_t^T = H_{tT}$, $\bar{Z}_t^T = Z_{tT}$, $\bar{N}_t^T = N_{tT}$ and $\bar{P}_t^T = P_{tT}$, where the parameters $\alpha_T, \beta_T, \gamma_T$ and $\delta_T$ may depend on $T$. Equation (4.16) gives

$$\left\{ \begin{array}{lll} \mathrm{d}\bar{H}_t^T & = -\beta_T\ \bar{H}_t^T\ T\mathrm{d}t & +\ \alpha_T\ \mathrm{d}\bar{N}_t^T, \\ \mathrm{d}\bar{Z}_t^T & = -\delta_T\ \bar{Z}_t^T\ T\mathrm{d}t & +\ \gamma_T\ \mathrm{d}\bar{P}_t^T/\omega, \end{array} \right. \tag{4.17}$$

where the common jump intensity of $\bar{N}^T$ and $\bar{P}^T$ is $T \times [\lambda_\infty + \bar{H}_t^T + (\bar{Z}_t^T)^2]$. Since the signs of the jumps of $\bar{P}^T$ are assumed to be unpredictable and uniformly distributed on $\{-\omega, \omega\}$, the infinitesimal

generator of the process is given by

$$\begin{aligned}\mathcal{A}^T f(h,z) \;=\; & -\beta_T\, h\, T\, \partial_h f(h,z) \;-\; \delta_T\, z\, T\, \partial_z f(h,z) \\ & + T\left[\lambda_\infty + h + z^2\right] \times \left\{\frac{1}{2} f\left(h+\alpha_T,\; z+\gamma_T\right) + \frac{1}{2} f\left(h+\alpha_T,\; z-\gamma_T\right) - f\left(h,z\right)\right\}\end{aligned} \tag{4.18}$$

for all functions $f$ twice continuously differentiable on $(0,+\infty)\times\mathbb{R}$. We now consider the spacial scaling

$$\alpha_T = \overline{\alpha}/T,\;\; \beta_T = \overline{\beta}/T,\;\; \gamma_T = \overline{\gamma}/\sqrt{T},\;\; \delta_T = \overline{\delta}/T, \tag{4.19}$$

with $\overline{\alpha},\overline{\gamma} \geq 0$ and $\overline{\beta},\overline{\delta} > 0$. It is chosen so that the Hawkes norm $n_H = \alpha_T/\beta_T = \overline{\alpha}/\overline{\beta}$ and the Zumbach norm $n_Z = \gamma_T^2/(2\delta_T) = \overline{\gamma}^2/(2\overline{\delta})$ are independent of the time scale $T$. This can be considered as the « scaling of constant endogeneity » (as opposed to the scaling used by Jaisson and Rosenbaum in [80] and [81], where the endogeneity ratio $\|\phi\|_1$ of the process needs to converge to unity as $T$ goes to infinity). Our choice is motivated by the calibration results of Section 4.3.2 for intra-day returns, that yield an endogeneity ratio in the range $0.7-0.9$ which is close to what is obtained at the daily time scale in [39] and [27], and significantly smaller than one. Equations (4.18) and (4.19) combine as

$$\begin{aligned}\mathcal{A}^T f(h,z) \;=\; & -\overline{\beta}\, h\, \partial_h f(h,z) \;-\; \overline{\delta}\, z\, \partial_z f(h,z) \\ & + \left[\lambda_\infty + h + z^2\right] \times T \times \left\{\frac{1}{2} f\left(h+\overline{\alpha}/T,\; z+\overline{\gamma}/\sqrt{T}\right) + \frac{1}{2} f\left(h+\overline{\alpha}/T,\; z-\overline{\gamma}/\sqrt{T}\right) - f\left(h,z\right)\right\}.\end{aligned}$$

We turn to the low-frequency asymptotics. As $T$ goes to infinity, one has

$$\frac{1}{2} f\left(h+\overline{\alpha}/T,\; z+\overline{\gamma}/\sqrt{T}\right) + \frac{1}{2} f\left(h+\overline{\alpha}/T,\; z-\overline{\gamma}/\sqrt{T}\right) - f\left(h,z\right) = \frac{\overline{\alpha}}{T}\partial_h f(h,z) + \frac{\overline{\gamma}^2}{2T}\partial^2_{zz} f(h,z) + \;\mathrm{o}\left(\frac{1}{T}\right),$$

therefore $\mathcal{A}^T f(h,z)$ converges to

$$\mathcal{A}^\infty f(h,z) \;=\; \left[-(\overline{\beta}-\overline{\alpha})h + \overline{\alpha}(\lambda_\infty + z^2)\right]\partial_h f(h,z) \;-\; \overline{\delta} z \partial_z f(h,z) \;+\; \frac{\overline{\gamma}^2}{2}\left[\lambda_\infty + h + z^2\right]\partial^2_{zz} f(h,z).$$

The operator $\mathcal{A}^\infty$ is the infinitesimal generator of the diffusion

$$\left\{\begin{array}{lll} \mathrm{d}\bar{H}^\infty_t & = \left[-(\overline{\beta}-\overline{\alpha})\;\bar{H}^\infty_t + \overline{\alpha}\left(\lambda_\infty + \left(\bar{Z}^\infty_t\right)^2\right)\right]\mathrm{d}t, \\ \mathrm{d}\bar{Z}^\infty_t & = -\overline{\delta}\;\bar{Z}^\infty_t\;\mathrm{d}t \;+\; \overline{\gamma}\;\sqrt{\lambda_\infty + \bar{H}^\infty_t + \left(\bar{Z}^\infty_t\right)^2}\mathrm{d}W_t, \end{array}\right. \tag{4.20}$$

where $W$ is a standard Brownian motion. A standard argument of Kallenberg [83] (Theorem 19.25) then gives the convergence of the process $(\bar{H}^T, \bar{Z}^T)$ to $(\bar{H}^\infty, \bar{Z}^\infty)$ as $T$ goes to infinity. Hence, one does *not* need that the norm of the process tends to 1 (i.e. that the process is nearly critical) for a non-degenerate limit process to be obtained.

Let us note that there is no Brownian part in the SDE for $\bar{H}^\infty$ and that it solves explicitly as a deterministic function of $(\bar{Z}^\infty_s)_{s\leq t}$ :

$$\bar{H}^\infty_t \;=\; \frac{\overline{\alpha}\lambda_\infty}{\overline{\beta}-\overline{\alpha}} \;+\; \overline{\alpha}\int_{-\infty}^t \exp(-(\overline{\beta}-\overline{\alpha})(t-s))\left(\bar{Z}^\infty_s\right)^2 \mathrm{d}s.$$

In the considered limit, $\bar{H}^\infty$ can thus be written as the sum of a constant term and an exponential moving average of the square of $\bar{Z}^\infty$. We get the autonomous, but non-Markovian SDE for $\bar{Z}^\infty$ :

$$\mathrm{d}\bar{Z}^\infty_t = -\overline{\delta}\;\bar{Z}^\infty_t\;\mathrm{d}t + \overline{\gamma}\;\sqrt{\frac{\lambda_\infty}{1-\overline{\alpha}/\overline{\beta}} + \left(\bar{Z}^\infty_t\right)^2 + \overline{\alpha}\left[\int_{-\infty}^t \exp(-(\overline{\beta}-\overline{\alpha})(t-s))\left(\bar{Z}^\infty_s\right)^2 \mathrm{d}s\right]}\;\mathrm{d}W_t. \tag{4.21}$$

We first consider the case where the Hawkes term is zero, i.e. $\overline{\alpha} = 0$. This corresponds to the case where only the Zumbach term is present in the starting model, i.e. $\lambda_t = \lambda_\infty + Z_t^2$ in Equation (4.15). As we see in the sequel, this simpler model is rich enough to reproduce some interesting empirical properties of the volatility process. One gets

$$\mathrm{d}\bar{Z}_t^\infty = -\overline{\delta}\ \bar{Z}_t^\infty\ \mathrm{d}t\ +\ \overline{\gamma}\ \sqrt{\lambda_\infty + \left(\bar{Z}_t^\infty\right)^2}\ \mathrm{d}W_t, \tag{4.22}$$

which is a particular case of Pearson diffusions. These are extensively described by Forman and Sorensen [59]. The process $\bar{Z}^\infty/\sqrt{\lambda_\infty}$ fits in Case 3 of the classification of Section 2.1. in their paper, with $\mu = 0, \theta = \overline{\delta}$ and $a = \overline{\gamma}^2/(2\overline{\delta}) = n_Z$. Therefore, $\bar{Z}^\infty$ is ergodic and its stationary law is a t-distribution with $1+1/n_Z$ degrees of freedom and scale parameter $\sqrt{n_Z\lambda_\infty/(1+n_Z)}$. This implies that stationary law of the square of $\bar{Z}^\infty$ is a F-distribution with 1 and $1+1/n_Z$ degrees of freedom, and scale parameter $n_Z\lambda_\infty/(1+n_Z)$. We note

$$V_t\ =\ \omega^2\left[\lambda_\infty + \left(\bar{Z}_t^\infty\right)^2\right]$$

the low-frequency squared volatility of the price. A straightforward change of variables yields the stationary density $q(v)$ of the process $V$

$$q(v) = \frac{\Gamma\left(1+\frac{1}{2n_Z}\right)}{\Gamma\left(\frac{1}{2}+\frac{1}{2n_Z}\right)\sqrt{\pi v_\infty}} \times \frac{1}{\sqrt{v-v_\infty}} \times \left(\frac{v}{v_\infty}\right)^{-\left(1+\frac{1}{2n_Z}\right)} \mathbb{1}_{\{v>v_\infty\}} \tag{4.23}$$

where $v_\infty = \lambda_\infty\omega^2$ is the baseline level of the squared volatility. For the tail exponent of the distribution of $V_t$, we get

$$q(v) \underset{v\to+\infty}{\sim} C\ v^{-\left(\frac{3}{2}+\frac{1}{2n_Z}\right)} \tag{4.24}$$

with $C$ an explicit constant.

We find this result quite remarkable for two reasons. First, one obtains a power-law behavior that emerges naturally from the fact that in Equation (4.22), the volatility coefficient behaves as $|\bar{Z}_t^\infty|$ for large values of $\bar{Z}_t^\infty$, so that locally the process is a multiplicative Brownian motion with drift. This is at variance with the « diagonal » counterpart [80] where the volatility coefficient scales as a square root, which inevitably leads to a process that has a characteristic scale. Second, the stationary distribution of $V$ only depends on the parameters $\overline{\gamma}$ and $\overline{\delta}$ through the Zumbach norm $n_Z = \overline{\gamma}^2/(2\overline{\delta})$, that can be seen as the *endogeneity* of the process. This last result suggests that, similar to Hawkes processes where the asymptotic properties only depend on the norm $||\phi||_1$ as soon as the kernel is short-ranged, the distribution (4.23) of the squared volatility may hold for any short-ranged kernel.

Another remark is that when $n_Z \geq 1/3$, the variance of the squared volatility $V$ explodes while its mean remains finite. Therefore, when fitting it by a simple Hawkes process for which the norm verifies $||\phi||_1 \simeq 1-\sqrt{\mu_W/\sigma_W^2}$ for a suitable choice of window size $W$ (see [70]), the vanishing mean/variance ratio necessarily imposes that the process is critical, i.e. $||\phi||_1 = 1$. What we argue here is that this vanishing ratio may only be due to quadratic feedback effects, and not to the criticality of the process.

In the diffusive limit where the price process satisfies the equation $\mathrm{d}\bar{P}_t^\infty = \sqrt{V_t}\mathrm{d}W_t$, the asymptotic stationary distribution for the returns is

$$g(r) \underset{|r|\to\infty}{\sim} \frac{C'}{|r|^{2+\frac{1}{n_Z}}}.$$

The fat-tail volatility that is generated by our model naturally produces a fat-tail distribution of instantaneous returns, with exponent $2+1/n_Z \geq 3$. This lower bound, reached for a critical process $n_Z = 1$, seems indeed to be an empirical limit [91]. When the criticality decreases the tail exponent of returns becomes larger, and the exponent 4 for instance (observed on a large universe of traded products) is obtained for $n_Z = 0.5$. The more endogenous, the fatter the tails for the returns : this interpretation seems consistent with empirical observations, that show that the returns exponent tends to become more negative as the market gains in maturity, corresponding to a decrease in endogeneity.

Now, if we go back to Equation (4.21) with $\overline{\alpha} > 0$ to take the Hawkes term into account, the analytic study of the process is more subtle. We solve the extreme cases $\overline{\beta}-\overline{\alpha} \ll 2\overline{\delta}-\overline{\gamma}^2$ and $\overline{\beta}-\overline{\alpha} \gg 2\overline{\delta}-\overline{\gamma}^2$ (where $2\overline{\delta}-\overline{\gamma}^2$ is the characteristic time scale of the square of $\bar{Z}^\infty$) to get some intuition. In the first case, one has

$$\int_{-\infty}^t (\overline{\beta}-\overline{\alpha})\exp(-(\overline{\beta}-\overline{\alpha})(t-s))\left(\bar{Z}_s^\infty\right)^2 \mathrm{d}s \;\approx\; \mathbb{E}\left[(\bar{Z}_\infty^\infty)^2\right] = \frac{\overline{\gamma}^2}{2\overline{\delta}} \times \frac{\lambda_\infty}{1-\overline{\alpha}/\overline{\beta}-\overline{\gamma}^2/(2\overline{\delta})},$$

and adding the Hawkes term boils down to multiplying $v_\infty$ by $(1-n_Z)/(1-n_H-n_Z)$ in Equation (4.23), which inflates the baseline volatility increasingly with $n_H$ but does not affect the tail exponent $3/2+1/(2n_Z)$. On the other hand, the case $\overline{\beta}-\overline{\alpha} \gg 2\overline{\delta}-\overline{\gamma}^2$ yields

$$\int_{-\infty}^t (\overline{\beta}-\overline{\alpha})\exp(-(\overline{\beta}-\overline{\alpha})(t-s))\left(\bar{Z}_s^\infty\right)^2 \mathrm{d}s \;\approx\; \left(\bar{Z}_t^\infty\right)^2.$$

Here, the Hawkes contribution divides $n_Z$ by $1-n_H \in (0,1]$ in Equations (4.23) and (4.24). This impacts the tail exponent of $V_t$ which becomes $3/2+(1-n_H)/(2n_Z)$. One finds the expected result that the tails are fattened by the extra feedback term.

Between these two extreme cases, it is clear that the Hawkes contribution modifies the distribution (4.23) by making the small values less probable, and increases the tail exponent of (4.24) up to the maximal value $3/2+(1-n_H)/(2n_Z)$.

### Empirical results and simulations

In this section, we compare numerically the volatility process generated by the ZHawkes model, with a standard Hawkes-based price model on the one hand, and the financial dataset introduced in Section 4.3.2 on the other hand. We simulate a ZHawkes model with an exponential Zumbach part and a power-law Hawkes part, with parameters inspired by the QARCH calibration of Section 4.3.2 : for $t$ expressed in minutes,

$$\phi(t) = 0.0016 \times (1+0.01\times t)^{-1.2}, \quad k(t) = 0.003 \times \exp(-0.03\times t),$$

so that $n_H = 0.8$, $n_Z = 0.1$ and $\mathrm{Tr}(K) = 0.9$. Note that to obtain integrability, we choose a decay exponent above 1 for $\phi$, although the QARCH calibration suggests a slower decay for small $t$'s. As a benchmark, we also simulate a standard Hawkes-based price process with decay $(1 + 0.01 \times t)^{-1.3}$ and norm 0.99, which is close to the calibration results of [71].

It is important to note that to simulate the ZHawkes and the Hawkes model, we choose constant price jumps $\Delta P_\tau = \pm\omega$. Therefore, our numerical results for the distribution of the volatility can by no means be linked to the kurtosis of price jumps, which is fixed to one.

For both simulated and real data, we consider the Rogers-Satchell volatility times series for five-minute bins. We use the Hill exponent of the empirical distribution of the volatility

$$h = 1 + \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \log(\sigma_i/\sigma_{\min})}$$

where $\sigma_{\min}$ is a cutoff and $\sigma_i \geq \sigma_{\min}$ are the selected volatilities, to compare the far tails of the distribution. One obtains $h = 4.50$ for real data, $h = 5.07$ for ZHawkes and $h = 12.41$ for the standard Hawkes-based model. Even with a norm close to one and a slowly-decaying kernel, the standard Hawkes model cannot reproduce the tails observed on US stock data. Instead, the ZHawkes model, with a norm strictly below unity and a short-lived Zumbach effect, naturally produces fat tails. These observations are illustrated by Figures 4.5 and 4.6.

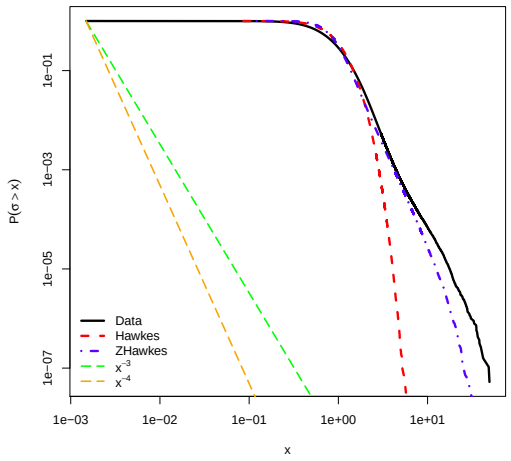



FIGURE 4.5 – Cumulative density function of the Rogers-Satchell volatility for US stock data (plain line), simulated Hawkes data (red dashed line), and simulated ZHawkes data (blue dot-dashed line).

### 4.4.3 Time-reversal asymmetry of the ZHawkes process

Another noticeable feature of financial markets is the time-reversal asymmetry of the volatility process. In [39], the authors study this feature for financial data on the one hand, and for a simulated ARCH volatility process on the other hand. They compare the cross-correlation of present Rogers-Satchell volatilities with past squared returns, to that of present squared returns with past

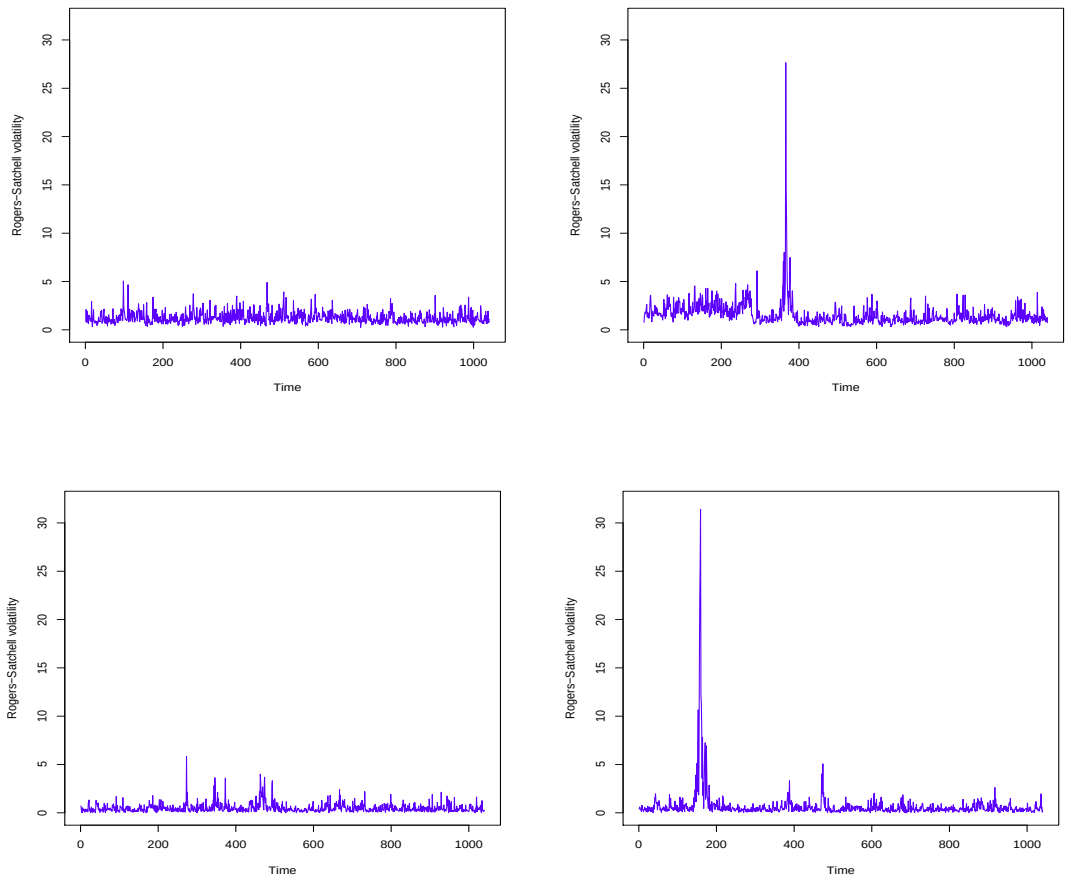



FIGURE 4.6 – Time series of Rogers-Satchell volatility. Above : real data ; below : simulated ZHawkes data ; left : period of calm ; right : cluster of intense activity.

volatilities. They find that the first is significantly larger than the second for both real data and ARCH processes. This can be interpreted as follows. Whereas the Rogers-Satchell volatility only measures the « agitation » of the price, the square of the return between time $t_1$ and time $t_2$ can only be large if price moves are *persistent* on the time interval $[t_1, t_2]$. As a matter of fact, if many price moves occur on $[t_1, t_2]$ but exactly compensate one another, the return is zero, while the volatility is high. Therefore, the difference observed between the two cross-correlations indicates that *price persistence increases future volatilities*, whereas high volatilities only generate high future volatilities, not necessarily with some price persistence. In terms of trading psychology, this could be explained by the fact that persistent price moves (in either direction) generate both opportunities and panic, thus increasing the number of transactions and the volatility more than compensated price moves. This observation is one of the main motivations for the model introduced in the present paper.

The standard models that use Brownian SDEs cannot reproduce this asymmetry, since they are time-reversal invariant by construction. In this section, we measure this feature for the simulated

ZHawkes process and the Hawkes benchmark described in Section 4.4.2, and for the financial dataset introduced in Section 4.3.2.

As in Sections 4.3.2 and 4.4.2, we consider the returns and Rogers-Satchell volatilities for intra-day five-minute bins. Here, the maximum lag $q$ is fixed to 36 (36 bins of 5 minutes = 3 hours of trading) and the lag index $\tau$ varies between 1 and $q$. We introduce

– The cross-correlation function of the Rogers-Satchell volatility and absolute returns

$$C(\tau) = \frac{\langle \sigma^{\mathrm{RS}}_{\cdot,\cdot,\cdot} \times |r_{\cdot,\cdot,\cdot,-\tau}| \rangle - \langle \sigma^{\mathrm{RS}} \rangle \langle |r| \rangle}{\sqrt{\langle {\sigma^{\mathrm{RS}}}^2 \rangle - \langle \sigma^{\mathrm{RS}} \rangle^2} \sqrt{\langle r^2 \rangle - \langle |r| \rangle^2}}.$$

– The time asymmetry ratio

$$\Delta(\tau) = \frac{\sum_{\tau'=1}^{\tau} [C(\tau') - C(-\tau')]}{2 \sum_{\tau'=1}^{q} \max(|C(\tau')|, |C(-\tau')|)} \in [-1, 1].$$

Note that we choose to compute the cross-correlation function using the *absolute* returns instead of the *squared* returns, since it yields results that are less noisy and more robust to tail events (and thus less sensible to the normalization method).

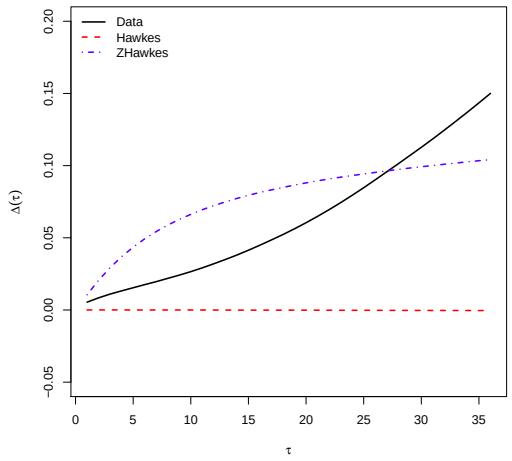



FIGURE 4.7 – Time asymmetry ratio $\Delta(\tau)$ for US stock data (plain line), simulated Hawkes data (red dashed line), and simulated ZHawkes data (blue dot-dashed line).

We compare the time asymmetry ratios $\Delta(\tau)$ for real stock returns, returns simulated with the ZHawkes model and returns simulated with a standard Hawkes-based price model. The results are illustrated by Figure 4.7. For the standard Hawkes, one has $|\Delta(\tau)| < 10^{-3}$ for all $\tau$. It is clear that the Hawkes model with no off-diagonal quadratic feedback cannot reproduce the time asymmetry observed in intra-day volatility, for which $\Delta(\tau)$ is one hundred times larger. On the other hand, the ZHawkes model with parameters in line with the QARCH calibration of Section 4.3 features some time asymmetry, which is of the correct sign and order of magnitude. However, $\tau \mapsto \Delta(\tau)$ is concave

for ZHawkes and convex for stock data. Even with a thorough normalization protocol, intra-day returns are not rigorously stationary, and we believe that the convexity of $\tau \mapsto \Delta(\tau)$ observed on real data is spurious, as it should become concave at some point and saturate below unity. Such convexity would probably be hard to reproduce with a simple model, unless it is non-stationary by construction.

## 4.5 Conclusion

In this paper we propose a quadratic feedback model for high-frequency volatility, the QHawkes model, which is an extension of Hawkes-based price models and reproduces several important empirical facts known so far about volatility. The calibration results for an intra-day QARCH model (which is shown to converge in some sense to QHawkes) on 133 NYSE stocks, shows that an off-diagonal feedback component is indeed present, and that its structure corresponds to a particular case of the model, called ZHawkes, for which the expression of the intensity is simpler and more tractable. This model has some interesting properties that standard Hawkes processes lack, namely : (i) the quadratic feedback naturally produces power-law tails for the volatility and the returns, that we are able to characterize in a specific Markovian case, (ii) it reproduces the time-reversal asymmetry at levels that are compatible with what is measured on actual financial data, in accordance with the idea that financial markets are causal and (iii) it can generate long memory without necessarily be critical. We finally derive the continuous limit SDE in the case of exponential kernels, that is found to be closely linked to Pearson diffusions. These mathematically tractable diffusions are very reminiscent of the volatility processes considered in [23] and [65]. Whereas we limited the present study to introducing the QHawkes model and establishing its empirical relevance, we believe that future research deriving more analytical properties of the general QHawkes model would be highly valuable. Also, a more precise calibration of the model itself, instead of its discrete QARCH counterpart, would be of empirical interest.

## 4.6 Appendix : Relation between the kernel and the auto-correlation functions

To alleviate the notations, we note (in this appendix only) $\varphi(t) = K(t,t)$.

### 4.6.1 Exact integral relation

For $s < t$, one has $\mathcal{C}(t-s) = \lambda_\infty \overline{\lambda} - \overline{\lambda}^2 + \mathbb{E}\left[A_t \frac{\mathrm{d}N_s}{\mathrm{d}s}\right] + 2\mathbb{E}\left[M_t \frac{\mathrm{d}N_s}{\mathrm{d}s}\right]$.

$$\mathbb{E}\left[A_t \frac{\mathrm{d}N_s}{\mathrm{d}s}\right] = \int_{-\infty}^{t} \varphi(t-u)\mathbb{E}\left[\frac{\mathrm{d}N_u}{\mathrm{d}u}\frac{\mathrm{d}N_s}{\mathrm{d}s}\right]\mathrm{d}u.$$

For $u \neq s$, $\mathbb{E}\left[\frac{\mathrm{d}N_u}{\mathrm{d}u}\frac{\mathrm{d}N_s}{\mathrm{d}s}\right]\mathrm{d}u = [\mathcal{C}(u-s)+\overline{\lambda}^2]\mathrm{d}u$, and for $u = s$, $\mathbb{E}\left[\left(\frac{\mathrm{d}N_u}{\mathrm{d}u}\right)^2\right]\mathrm{d}u = \kappa\mathbb{E}\left[\frac{\mathrm{d}N_u}{(\mathrm{d}u)^2}\right]\mathrm{d}u = \kappa\overline{\lambda}$, where $\kappa$ is the kurtosis of the law $\mu$ of the jumps of $P$ ($\kappa = 1$ if $\Delta P_\tau = \pm\omega$). Thus,

$$\mathbb{E}\left[A_t\frac{\mathrm{d}N_s}{\mathrm{d}s}\right] = \mathrm{Tr}(K)\overline{\lambda}^2 + \kappa\overline{\lambda}\varphi(t-s) + \int_{-\infty}^{t}\varphi(t-u)\mathcal{C}(u-s)\mathrm{d}u.$$

On the other hand,

$$\begin{aligned}\mathbb{E}\left[M_t\frac{\mathrm{d}N_s}{\mathrm{d}s}\right] &= \frac{1}{\omega^2}\int_{-\infty}^{t}\mathbb{E}\left[\Theta_{t,u}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}N_s}{\mathrm{d}s}\right]\mathrm{d}u\\ &= \frac{1}{\omega^2}\int_{-\infty}^{t}\int_{-\infty}^{u-}K(t-u,t-r)\mathbb{E}\left[\frac{\mathrm{d}N_s}{\mathrm{d}s}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r\mathrm{d}u\\ &= \int_{-\infty}^{s-}\int_{-\infty}^{u-}K(t-u,t-r)\mathcal{D}(s-u,s-r)\mathrm{d}r\mathrm{d}u,\end{aligned}$$

since $\Delta P_\tau$ and $(\Delta P_\tau)^3$ are centered, which implies that $\mathbb{E}\left[\frac{\mathrm{d}N_s}{\mathrm{d}s}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right] = 0$ for $u \geq s$. Taking $t = \tau > 0$ and $s = 0$, we obtain

$$\mathcal{C}(\tau) = \kappa\overline{\lambda}\varphi(\tau) + \int_{-\infty}^{\tau}\varphi(\tau-u)\mathcal{C}(u)\mathrm{d}u + 2\int_{0+}^{\infty}\int_{u+}^{\infty}K(\tau+u,\tau+r)\mathcal{D}(u,r)\mathrm{d}r\mathrm{d}u.$$

For $t > t_1 > t_2$, one has $\mathcal{D}(t-t_1,t-t_2) = \frac{1}{\omega^2}\mathbb{E}\left[A_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] + \frac{2}{\omega^2}\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right]$. The first term gives

$$\begin{aligned}\frac{1}{\omega^2}\mathbb{E}\left[A_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] &= \frac{1}{\omega^2}\int_{-\infty}^{t}\varphi(t-u)\mathbb{E}\left[\frac{\mathrm{d}N_u}{\mathrm{d}u}\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right]\mathrm{d}u\\ &= \int_{t_1+}^{t}\varphi(t-u)\mathcal{D}(u-t_1,u-t_2)\mathrm{d}u.\end{aligned}$$

The second term is given by

$$\frac{1}{\omega^2}\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] = \frac{1}{\omega^4}\int_{-\infty}^{t}\int_{-\infty}^{u-}K(t-u,t-r)\mathbb{E}\left[\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r\mathrm{d}u$$

Since $r < u$ in the integral and $t_2 < t_1$, the expected value is zero if $u \neq t_1$. For $u = t_1$, we have $\mathbb{E}\left[\left(\frac{\mathrm{d}P_u}{\mathrm{d}u}\right)^2\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}u = \omega^2\mathbb{E}\left[\frac{\mathrm{d}N_u}{(\mathrm{d}u)^2}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}u = \omega^2\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]$. Thus,

$$\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] = \frac{1}{\omega^2}\int_{-\infty}^{t_1-}K(t-t_1,t-r)\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r.$$

For $r \neq t_2$, one has $\frac{1}{\omega^2}\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r = \mathcal{D}(t_1-t_2,t_1-r)\mathrm{d}r$. On the other hand $r = t_2$ yields $\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}N_r}{(\mathrm{d}r)^2}\right]\mathrm{d}r = \mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}N_{t_2}}{\mathrm{d}t_2}\right] = \mathcal{C}(t_1-t_2)+\overline{\lambda}^2$. We obtain

$$\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] = K(t-t_1,t-t_2)[\mathcal{C}(t_1-t_2)+\overline{\lambda}^2] + \int_{-\infty}^{t_1-}K(t-t_1,t-r)\mathcal{D}(t_1-t_2,t_1-r)\mathrm{d}r.$$

We eventually obtain by taking $\tau_2 = t > \tau_1 = t - t_1, t_2 = 0$,

$$\mathcal{D}(\tau_1,\tau_2) = 2K(\tau_1,\tau_2)[\mathcal{C}(\tau_2-\tau_1)+\overline{\lambda}^2] + \int_{(\tau_2-\tau_1)+}^{\tau_2} \varphi(\tau_2-u)\mathcal{D}(u-\tau_2+\tau_1,u)\mathrm{d}u + 2\int_{-\infty}^{(\tau_2-\tau_1)-} K(\tau_1,\tau_2-u)\mathcal{D}(\tau_2-\tau_1,\tau_2-\tau_1-u)\mathrm{d}u.$$

### 4.6.2 Power-law asymptotics

For $\tau$ large, Equation (4.9) yields

$$c_1\ \tau^{-\beta} = \kappa\overline{\lambda}c_0\tau^{-2\delta} + c_0\tau^{-2\delta}\int_{-\infty}^{\tau}\left(1-\frac{u}{\tau}\right)^{-2\delta}\mathcal{C}(u)\mathrm{d}u + 2\tau^{-2\delta}\int_0^\infty\int_u^\infty \tilde{K}\left(1+\frac{u}{\tau},1+\frac{r}{\tau}\right)\mathcal{D}(u,r)\mathrm{d}r\mathrm{d}u.$$

In both integrals, we make the change of variables $u' = u/\tau$, and then $r' = r/\tau$, to obtain

$$\begin{aligned} c_1\ \tau^{-\beta} &= \kappa\overline{\lambda}c_0\tau^{-2\delta} + c_0\tau^{-2\delta}\int_{-\infty}^{1}\left(1-u'\right)^{-2\delta}\mathcal{C}(\tau u')\tau\mathrm{d}u' + 2\tau^{-2\delta}\int_0^\infty\int_{u'}^\infty \tilde{K}\left(1+u',1+r'\right)\mathcal{D}(\tau u',\tau r')\tau^2\mathrm{d}r'\mathrm{d}u' \\ &= \kappa\overline{\lambda}c_0\tau^{-2\delta} + c_0c_1\tau^{1-2\delta-\beta}\int_{-\infty}^{1}\left(1-u'\right)^{-2\delta}u'^{-\beta}\mathrm{d}u' + 2\tau^{2-2\delta-\rho}\int_0^\infty\int_{u'}^\infty \tilde{K}\left(1+u',1+r'\right)\tilde{\mathcal{D}}(u',r')\mathrm{d}r'\mathrm{d}u' \end{aligned}$$

which can be written

$$c_1\ \tau^{-\beta} = \kappa\overline{\lambda}c_0\tau^{-2\delta} + c_2\tau^{1-2\delta-\beta} + c_3\tau^{2-2\delta-\rho},$$

with $c_2, c_3$ two constants[3]. This leaves only three possibilities :

$$\beta = 2\delta, \quad \beta > 1, \quad \rho > 2, \tag{4.25}$$

$$2\delta = 1, \quad \beta < 1, \quad \rho > \beta + 1, \tag{4.26}$$

$$\beta = 2\delta + \rho - 2, \quad \rho < 2, \quad \rho < \beta + 1. \tag{4.27}$$

Note that in the first case, one also has $\mathcal{C}(\tau)/\varphi(\tau) \underset{\tau\to\infty}{\longrightarrow} \kappa\overline{\lambda}$, which relates the kurtosis $\kappa$ of price jumps and the auto-covariance function $\mathcal{C}$.

Let us now consider Equation (4.10) for $\tau_1 = \tau v_1, \tau_2 = \tau v_2$, and $\tau$ large. One has

$$\begin{aligned} \tilde{\mathcal{D}}(v_1,v_2)\tau^{-\rho} &= 2\tilde{K}(v_1,v_2)\tau^{-2\delta}[c_1(v_2-v_1)^{-\beta}\tau^{-\beta}+\overline{\lambda}^2] \\ &+ c_0\tau^{-2\delta-\rho}\int_{\tau(v_2-v_1)}^{\tau v_2}\left(v_2-\frac{u}{\tau}\right)^{-2\delta}\tilde{\mathcal{D}}\left(\frac{u}{\tau}-v_2+v_1,\frac{u}{\tau}\right)\mathrm{d}u \\ &+ 2\tau^{-2\delta-\rho}\int_{-\infty}^{\tau(v_2-v_1)}\tilde{K}\left(v_1,v_2-\frac{u}{\tau}\right)\tilde{\mathcal{D}}\left(v_2-v_1,v_2-v_1-\frac{u}{\tau}\right)\mathrm{d}u. \end{aligned}$$

Again, the change of variables $u' = u/\tau$ in the two integrals yields

$$\begin{aligned} \tilde{\mathcal{D}}(v_1,v_2)\tau^{-\rho} &= 2\tilde{K}(v_1,v_2)\tau^{-2\delta}[c_1(v_2-v_1)^{-\beta}\tau^{-\beta}+\overline{\lambda}^2] \\ &+ c_0\tau^{1-2\delta-\rho}\int_{v_2-v_1}^{v_2}\left(v_2-u'\right)^{-2\delta}\tilde{\mathcal{D}}\left(u'-v_2+v_1,u'\right)\mathrm{d}u' \\ &+ 2\tau^{1-2\delta-\rho}\int_{-\infty}^{v_2-v_1}\tilde{K}\left(v_1,v_2-u'\right)\tilde{\mathcal{D}}(v_2-v_1,v_2-v_1-u')\mathrm{d}u'. \end{aligned}$$

3. Note that the finiteness of the integral terms (once $\tau$ is factored out) does not matter since we are only interested in the asymptotic dependence in $\tau$. For instance, $(1-u')^{-2\delta}$ could be replaced by $(2-u')^{-2\delta}$ to avoid integrability issues for $u' \to 1$, with no incidence on the final results.

We thus have

$$\tilde{\mathcal{D}}(v_1, v_2)\tau^{-\rho} = 2\overline{\lambda}^2 \tilde{K}(v_1, v_2)\tau^{-2\delta} + f_1(v_1, v_2)\tau^{-(2\delta+\beta)} + f_2(v_1, v_2)\tau^{-(2\delta+\rho-1)}$$

with $f_1, f_2$ two bounded functions of $(v_1, v_2)$. Since $\beta$ is necessarily positive, this only leaves two possibilities :

$$\rho = 2\delta, \quad \beta > 1, \quad \rho > 1, \tag{4.28}$$

$$2\delta = 1, \quad \rho < 1. \tag{4.29}$$

Equations (4.26) and (4.29) are not compatible since $\beta > 0$ implies $\rho > 1$. Thus, the only remaining possibility given by Equation (4.28) yields $\rho = 2\delta > 1$. This implies

$$\forall v_1, v_2, \ \mathcal{D}(\tau v_1, \tau v_2)/K(\tau v_1, \tau v_2) \underset{\tau\to\infty}{\longrightarrow} 2\overline{\lambda}^2.$$

Moreover, the combination of Equations (4.25), (4.27) and (4.28) yields the two possible phases for the auto-covariance structure :

$$\delta > 1 \Rightarrow \beta = \rho = 2\delta,$$

$$\frac{1}{2} < \delta < 1 \Rightarrow \beta = 4\delta - 2, \ \rho = 2\delta.$$

# Bibliographie


[1] Frédéric Abergel and Aymen Jedidi. A mathematical approach to order book modeling. *International Journal of Theoretical and Applied Finance (IJTAF)*, 16(05), 2013. URL http://EconPapers.repec.org/RePEc:wsi:ijtafx:v:16:y:2013:i:05:p:1350025-1-1350025-40.

[2] Aurélien Alfonsi and José Infante Acevedo. Optimal execution and price manipulations in time-varying limit order books. *Applied Mathematical Finance*, 21(3) :201–237, 2014.

[3] Aurélien Alfonsi and Pierre Blanc. Dynamic optimal execution in a mixed-market-impact hawkes price model. *arXiv preprint arXiv :1404.0648*, 2014.

[4] Aurélien Alfonsi and Alexander Schied. Capacitary measures for completely monotone kernels via singular control. *SIAM J. Control Optim.*, 51(2) :1758–1780, 2013. ISSN 0363-0129. doi : 10.1137/120862223. URL http://dx.doi.org/10.1137/120862223.

[5] Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quant. Finance*, 10(2) :143–157, 2010. ISSN 1469-7688. doi : 10.1080/14697680802595700. URL http://dx.doi.org/10.1080/14697680802595700.

[6] Aurélien Alfonsi, Alexander Schied, and Alla Slynko. Order Book Resilience, Price Manipulation, and the Positive Portfolio Problem. *SSRN eLibrary*, 2011.

[7] Aurélien Alfonsi, Alexander Schied, and Alla Slynko. Order book resilience, price manipulation, and the positive portfolio problem. *SIAM Journal on Financial Mathematics*, 3(1) :511–533, 2012.

[8] Romain Allez and Jean-Philippe Bouchaud. Individual and collective stock dynamics : intraday seasonalities. *New Journal of Physics*, 13(2) :025010, 2011.

[9] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3 :5–39, 2000.

[10] E. Bacry and J. F Muzy. Hawkes model for price and trades high-frequency dynamics. *ArXiv e-prints*, January 2013.

[11] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quant. Finance*, 13(1) :65–77, 2013. ISSN 1469-7688. doi : 10.1080/14697688.2011.647054. URL http://dx.doi.org/10.1080/14697688.2011.647054.

[12] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy. Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Process. Appl.*, 123(7) :2475–2499, 2013. ISSN 0304-4149. doi : 10.1016/j.spa.2013.04.007. URL http://dx.doi.org/10.1016/j.spa.2013.04.007.

[13] Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quant. Finance*, 14(7) :1147–1166, 2014. ISSN 1469-7688. doi : 10.1080/14697688.2014.897000. URL http://dx.doi.org/10.1080/14697688.2014.897000.

[14] Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7) :1147–1166, 2014.

[15] Emmanuel Bacry, Khalil Dayri, and Jean-François Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85 :1–12, 2012. ISSN 1434-6028. URL http://dx.doi.org/10.1140/epjb/e2012-21005-8.

[16] Emmanuel Bacry, Khalil Dayri, and Jean-François Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B-Condensed Matter and Complex Systems*, 85(5) :1–12, 2012.

[17] Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-François Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13(1) :65–77, 2013.

[18] Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-Francois Muzy. Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7) :2475–2499, 2013.

[19] Richard T. Baillie, Tim Bollerslev, and Hans O. Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1) :3–30, 1996.

[20] Christian Bayer, Peter K Friz, and Jim Gatheral. Pricing under rough volatility. *Available at SSRN*, 2015.

[21] Erhan Bayraktar and Michael Ludkovski. Optimal trade execution in illiquid markets. *Math. Finance*, 21(4) :681–701, 2011. ISSN 0960-1627. doi : 10.1111/j.1467-9965.2010.00446.x. URL http://dx.doi.org/10.1111/j.1467-9965.2010.00446.x.

[22] Lorenzo Bergomi. Smile dynamics i. *Available at SSRN 1493294*, 2004.

[23] Lorenzo Bergomi. Smile dynamics ii. *Available at SSRN 1493302*, 2005.

[24] Dimitris Bertsimas and Andrew Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1 :1–50, 1998.

[25] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

[26] Pierre Blanc. Modélisation de la volatilité des marchés financiers par une structure ARCH multi-fréquence. Master's thesis, Université de Paris VI Pierre et Marie Curie, Sep 2012. Available upon request.

[27] Pierre Blanc, Rémy Chicheportiche, and Jean-Philippe Bouchaud. The fine structure of volatility feedback ii : overnight and intra-day effects. *Physica A : Statistical Mechanics and its Applications*, 402 :58–75, 2014.

[28] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3) :307–327, 1986.

[29] Tim Bollerslev, Robert F Engle, and Daniel B Nelson. Arch models. *Handbook of econometrics*, 4 :2959–3038, 1994.

[30] Lisa Borland and Jean-Philippe Bouchaud. On a multi-timescale statistical feedback model for volatility fluctuations. *The Journal of Investment Strategies*, 1(1) :65–104, December 2011.

[31] Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets : the subtle nature of "random" price changes. *Quantitative Finance*, 4(2) :176–190, 2004. doi : 10.1080/14697680400000022. URL http://www.tandfonline.com/doi/abs/10.1080/14697680400000022.

[32] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *Ann. Probab.*, 24(3) :1563–1588, 1996. ISSN 0091-1798. doi : 10.1214/aop/1065725193. URL http://dx.doi.org/10.1214/aop/1065725193.

[33] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *Ann. Probab.*, 24(3) :1563–1588, 1996. ISSN 0091-1798. doi : 10.1214/aop/1065725193. URL http://dx.doi.org/10.1214/aop/1065725193.

[34] Pierre Brémaud, Laurent Massoulié, et al. Hawkes branching point processes without ancestors. *Journal of applied probability*, 38(1) :122–135, 2001.

[35] Pierre Brémaud, Laurent Massoulié, et al. Hawkes branching point processes without ancestors. *Journal of applied probability*, 38(1) :122–135, 2001.

[36] Jorge Buescu. Positive integral operators in unbounded domains. *Journal of Mathematical Analysis and Applications*, 296(1) :244–255, 2004.

[37] S.D. Chatterji. *Cours d'analyse Tome 3 : Équations différentielles ordinaires et aux dérivées partielles*. Number vol. 1 in Cours d'analyse. Presses polytechniques et universitaires romandes, 1998. ISBN 9782880743505. URL http://books.google.fr/books?id=nTkXLPD-XuwC.

[38] Rémy Chicheportiche and Jean-Philippe Bouchaud. The fine-structure of volatility feedback 1 : Multi-scale self-reflexivity. *Physica A : Statistical Mechanics and its Applications*, 2014. ISSN 0378-4371. doi : http://dx.doi.org/10.1016/j.physa.2014.05.007. URL http://www.sciencedirect.com/science/article/pii/S0378437114003719.

[39] Rémy Chicheportiche and Jean-Philippe Bouchaud. The fine-structure of volatility feedback i : Multi-scale self-reflexivity. *Physica A : Statistical Mechanics and its Applications*, 410 : 174–195, 2014.

[40] R. Cont and A. de Larrard. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1) :1–25, 2013. doi : 10.1137/110856605. URL http://epubs.siam.org/doi/abs/10.1137/110856605.

[41] Rama Cont. Empirical properties of asset returns : stylized facts and statistical issues. *Quantitative Finance*, 1(2) :223–236, 2001. doi : 10.1080/713665670. URL http://www.tandfonline.com/doi/abs/10.1080/713665670.

[42] Rama Cont. Volatility clustering in financial markets : empirical facts and agent-based models. In *Long memory in economics*, pages 289–309. Springer, 2007.

[43] Nicolas Cosson. Analysis of realized and implied volatility after stock price jumps. Master's thesis, Université de Paris VI Pierre et Marie Curie, Jun 2013. Available upon request.

[44] José Da Fonseca and Riadh Zaatour. Hawkes process : Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, pages n/a–n/a, 2013. ISSN 1096-9934. doi : 10.1002/fut.21644. URL http://dx.doi.org/10.1002/fut.21644.

[45] José Da Fonseca and Riadh Zaatour. Hawkes process : Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6) :548–579, 2014.

[46] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes : volume II : general theory and structure*, volume 2. Springer Science & Business Media, 2007.

[47] Jonathan Donier. Market impact with autocorrelated order flow under perfect competition. Papers, arXiv.org, 2012. URL http://EconPapers.repec.org/RePEc:arx:papers:1212.4770.

[48] Jonathan Donier, Julius Friedrich Bonart, Iacopo Mastromatteo, and Jean-Philippe Bouchaud. A fully consistent, minimal model for non-linear market impact. *Minimal Model for Non-Linear Market Impact (November 29, 2014)*, 2014.

[49] Bruno Dupire et al. Pricing with a smile. *Risk*, 7(1) :18–20, 1994.

[50] Zoltán Eisler, Jean-Philippe Bouchaud, and Julien Kockelkoren. The price impact of order book events : market orders, limit orders and cancellations. *Quant. Finance*, 12(9) :1395–1419, 2012. ISSN 1469-7688. doi : 10.1080/14697688.2010.528444. URL http://dx.doi.org/10.1080/14697688.2010.528444.

[51] Paul Embrechts, Thomas Liniger, and Lu Lin. Multivariate Hawkes processes : an application to financial data. *J. Appl. Probab.*, 48A(New frontiers in applied probability : a Festschrift for Soren Asmussen) :367–378, 2011. ISSN 0021-9002. doi : 10.1239/jap/1318940477. URL http://dx.doi.org/10.1239/jap/1318940477.

[52] Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica : Journal of the Econometric Society*, pages 987–1007, 1982.

[53] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica : Journal of the Econometric Society*, pages 987–1007, 1982.

[54] Robert F. Engle and Daniel L. McFadden, editors. *Handbook of Econometrics*, volume 4. Elsevier/North-Holland, Amsterdam, 1994.

[55] Robert F. Engle and Magdalena E. Sokalska. Forecasting intraday volatility in the US equity market. multiplicative component GARCH. *Journal of Financial Econometrics*, 10(1) :54–83, 2012.

[56] J Doyne Farmer, Austin Gerig, Fabrizio Lillo, and Szabolcs Mike. Market efficiency and the long-memory of supply and demand : is price impact variable and permanent or fixed and temporary ? *Quantitative Finance*, 6(02) :107–112, 2006.

[57] J Doyne Farmer, Austin Gerig, Fabrizio Lillo, and Henri Waelbroeck. How efficiency shapes market impact. *Quantitative Finance*, 13(11) :1743–1758, 2013.

[58] Vladimir Filimonov and Didier Sornette. Quantifying reflexivity in financial markets : Toward a prediction of flash crashes. *Phys. Rev. E*, 85 :056108, May 2012. doi : 10.1103/PhysRevE.85.056108. URL http://link.aps.org/doi/10.1103/PhysRevE.85.056108.

[59] Julie Lyng Forman and Michael Sørensen. The pearson diffusions : A class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics*, 35(3) :438–465, 2008.

[60] A. Fruth, T. Schöneborn, and M. Urusov. Optimal trade execution and price manipulation in order books with time-varying liquidity. *Working Paper Series*, 2011.

[61] Giampiero M. Gallo. Modelling the impact of overnight surprises on intra-daily volatility. *Australian Economic Papers*, 40(4) :567–580, 2001.

[62] A. Gareche, G. Disdier, J. Kockelkoren, and Jean-Philippe Bouchaud. A Fokker-Planck description for the queue dynamics of large tick stocks, April 2013. URL http://arxiv.org/abs/1304.6819.

[63] Jim Gatheral. No-dynamic-arbitrage and market impact. *Quant. Finance*, 10(7) :749–759, 2010. ISSN 1469-7688. doi : 10.1080/14697680903373692. URL http://dx.doi.org/10.1080/14697680903373692.

[64] Jim Gatheral, Alexander Schied, and Alla Slynko. Transient linear price impact and fredholm integral equations. *Mathematical Finance*, 22(3) :445–474, 2012.

[65] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Available at SSRN 2509457*, 2014.

[66] Olivier Guéant. Optimal execution and block trade pricing : a general framework. Papers 1210.6372, arXiv.org, October 2012. URL http://ideas.repec.org/p/arx/papers/1210.6372.html.

[67] Patrick S Hagan, Deep Kumar, Andrew S Lesniewski, and Diana E Woodward. Managing smile risk. *The Best of Wilmott*, page 249, 2002.

[68] Stephen Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. Critical reflexivity in financial markets : a hawkes process analysis. *The European Physical Journal B - Condensed Matter and Complex Systems*, 86(10) :1–9, 2013. URL http://EconPapers.repec.org/RePEc:spr:eurphb:v:86:y:2013:i:10:p:1-9:10.1140/epjb/e2013-40107-3.

[69] Stephen J Hardiman and Jean-Philippe Bouchaud. Branching-ratio approximation for the self-exciting hawkes process. *Physical Review E*, 90(6) :062807, 2014.

[70] Stephen J Hardiman and Jean-Philippe Bouchaud. Branching-ratio approximation for the self-exciting hawkes process. *Physical Review E*, 90(6) :062807, 2014.

[71] Stephen J Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. Critical reflexivity in financial markets : a hawkes process analysis. *The European Physical Journal B*, 86(10) :1–9, 2013.

[72] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1) :83–90, 1971.

[73] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1) :pp. 83–90, 1971. ISSN 00063444. URL http://www.jstor.org/stable/2334319.

[74] Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503, 1974.

[75] Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *J. Appl. Probability*, 11 :493–503, 1974. ISSN 0021-9002.

[76] Steven L Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies*, 6(2) :327–343, 1993.

[77] Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data : The queue-reactive model. Papers, arXiv.org, 2013. URL http://EconPapers.repec.org/RePEc:arx:papers:1312.0563.

[78] Gur Huberman and Werner Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 72(4) :1247–1275, 2004. ISSN 0012-9682. doi : 10.1111/j.1468-0262.2004.00531.x. URL http://dx.doi.org/10.1111/j.1468-0262.2004.00531.x.

[79] Thibault Jaisson and Mathieu Rosenbaum. Limit theorems for nearly unstable Hawkes processes, October 2013. URL http://arxiv.org/abs/1310.2033.

[80] Thibault Jaisson and Mathieu Rosenbaum. Limit theorems for nearly unstable hawkes processes. *arXiv preprint arXiv :1310.2033*, 2013.

[81] Thibault Jaisson and Mathieu Rosenbaum. Rough fractional diffusions as scaling limit of nearly unstable heavy tailed hawkes processes. *to appear*, 2015.

[82] Armand Joulin, Augustin Lefevre, Daniel Grunberg, and Jean-Philippe Bouchaud. Stock price jumps : News and volume play a minor role. *Wilmott Magazine*, pages 1–7, September/October 2008.

[83] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2002.

[84] Jan Kallsen and Johannes Muhle-Karbe. High-resilience limits of block-shaped order books. 2014.

[85] Charles-Albert Lehalle and Sophie Laruelle. *Market Microstructure in Practice*. World Scientific publishing, 2013. URL http://www.worldscientific.com/worldscibooks/10.1142/8967.

[86] Remi Lemonnier and Nicolas Vayatis. Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate hawkes processes. In *Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer, 2014.

[87] Yanhui Liu, Parameswaran Gopikrishnan, Pierre Cizeau, Martin Meyer, Chung-Kang Peng, and H. Eugene Stanley. Statistical properties of the volatility of price fluctuations. *Physical Review E*, 60 :1390–1400, Aug 1999. doi : 10.1103/PhysRevE.60.1390. URL http://link.aps.org/doi/10.1103/PhysRevE.60.1390.

[88] Benoit B Mandelbrot. *The variation of certain speculative prices*. Springer, 1997.

[89] Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1) :77–91, 1952.

[90] Iacopo Mastromatteo, Bence Toth, and Jean-Philippe Bouchaud. Agent-based models for latent liquidity and concave price impact. Papers, arXiv.org, 2013. URL http://EconPapers.repec.org/RePEc:arx:papers:1311.6262.

[91] Guo-Hua Mu, Wei-Xing Zhou, et al. Tests of nonuniversality of the stock return distributions in an emerging market. *Physical Review E*, 82(6) :066103, 2010.

[92] Ulrich A. Müller, Michel M. Dacorogna, Rakhal D. Davé, Richard B. Olsen, Olivier V. Pictet, and Jacob E. von Weizsäcker. Volatilities of different time resolutions — analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2) :213–239, 1997.

[93] Anna Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16 :1–32, 2013.

[94] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Ann. Inst. Statist. Math.*, 31(1) :145–155, 1979. ISSN 0020-3157. doi : 10.1007/BF02480272. URL http://dx.doi.org/10.1007/BF02480272.

[95] Marc Potters and Jean-Philippe Bouchaud. More statistical properties of order books and price impact. *Physica A : Statistical Mechanics and its Applications*, 324(1-2) :133–140, 2003. ISSN 0378-4371. doi : 10.1016/S0378-4371(02)01896-4. URL http://www.sciencedirect.com/science/article/pii/S0378437102018964. Proceedings of the International Econophysics Conference.

[96] Marc Potters, Rama Cont, and Jean-Philippe Bouchaud. Financial markets as adaptive systems. *EPL (Europhysics Letters)*, 41(3) :239, 1998.

[97] Silviu Predoiu, Gennady Shaikhet, and Steven Shreve. Optimal execution in a general one-sided limit-order book. *SIAM J. Financial Math.*, 2 :183–212, 2011. ISSN 1945-497X. doi : 10.1137/10078534X. URL http://dx.doi.org/10.1137/10078534X.

[98] Patricia Reynaud-Bouret, Sophie Schbath, et al. Adaptive estimation for hawkes processes ; application to genome analysis. *The Annals of Statistics*, 38(5) :2781–2822, 2010.

[99] Christian Y. Robert and Mathieu Rosenbaum. A new approach for the dynamics of ultra-high-frequency data : The model with uncertainty zones. *Journal of Financial Econometrics*, 9(2) :344–366, 2011. doi : 10.1093/jjfinec/nbq023. URL http://jfec.oxfordjournals.org/content/9/2/344.abstract.

[100] Samuel Schechter. On the inversion of certain matrices. *Math. Tables Aids Comput.*, 13 : 73–77, 1959. ISSN 0891-6837.

[101] Alexander Schied and Tao Zhang. A hot-potato game under transient price impact and some effects of a transaction tax. *Available at SSRN 2256510*, 2013.

[102] Enrique Sentana. Quadratic ARCH models. *The Review of Economic Studies*, 62(4) :639, 1995.

[103] Yoash Shapira, Dror Y. Kenett, Ohad Raviv, and Eshel Ben-Jacob. Hidden temporal order unveiled in stock market volatility variance. *AIP Advances*, 1(2) :022127–022127, 2011.

[104] Sasha Stoikov and Rolf Waeber. Optimal Asset Liquidation Using Limit Order Book Information. *SSRN eLibrary*, 2012.

[105] Bence Toth, Imon Palit, Fabrizio Lillo, and J. Doyne Farmer. Why is order flow so persistent ? Papers, arXiv.org, 2011. URL http://EconPapers.repec.org/RePEc:arx:papers:1108.1632.

[106] Ilias Tsiakas. Overnight information and stochastic volatility : A study of European and US stock exchanges. *Journal of Banking & Finance*, 32(2) :251–268, 2008.

[107] Fengzhong Wang, Kazuko Yamasaki, Shlomo Havlin, and H. Eugene Stanley. Scaling and memory of intraday volatility return intervals in stock markets. *Physical Review E*, 73(2) : 026117, 2006.

[108] Adriaan Cornelis Zaanen. Linear analysis. 1956.

[109] B. Zheng, F. Roueff, and F. Abergel. Modelling bid and ask prices using constrained hawkes processes : Ergodicity and scaling limit. *SIAM Journal on Financial Mathematics*, 5(1) :99–136, 2014. doi : 10.1137/130912980. URL http://epubs.siam.org/doi/abs/10.1137/130912980.

[110] Gilles Zumbach. Time reversal invariance in finance. *Quantitative Finance*, 9(5) :505–515, 2009.

[111] Gilles Zumbach and Paul Lynch. Heterogeneous volatility cascade in financial markets. *Physica A : Statistical Mechanics and its Applications*, 298(3-4) :521–529, 2001.