Clustering in foreign exchange markets : price, trades and traders

Mehdi Lallouache

CentraleSupélec

Clustering in foreign exchange markets: price, trades and traders

Mehdi Lallouache

Thèse présentée à CentraleSupélec
pour l'obtention du grade de Docteur
Spécialité: Mathématiques Appliquées

Laboratoire de Mathématiques Appliquées aux Systèmes
CentraleSupélec

Soutenue le 10 Juillet 2015 devant un jury composé de :

Directeurs:	Frédéric Abergel	-	CentraleSupélec
	Damien Challet	-	CentraleSupélec
Rapporteurs:	Fabrizio Lillo	-	Scuola Normale Superiore di Pisa
	Fulvio Baldovin	-	Università degli Studi di Padova
Examinateurs:	Vladimir Filimonov	-	ETH Zürich
	Roberto Renò	-	Università di Siena

Remerciements

Mes premiers remerciements sont adressés à Frédéric Abergel qui m'a accueilli au laboratoire MAS en 2010 pour un stage, puis un autre ... puis une thèse ! *Merci pour tous tes conseils et toutes les attentions qui ont rendu ces années au sein de FiQuant très agréables.*

Merci ensuite à Damien Challet qui a encadré une grande partie de la thèse. *Merci pour ton enthousiasme, toutes tes idées et tout le savoir que tu m'as transmis. Mention spéciale aux data.table qui ont changé mon quotidien !*

Merci à Anirban Chakraborti de m'avoir initié à la finance quantitative. Merci à mes rapporteurs Fabrizio Lillo et Fulvio Baldovin pour leurs commentaires judicieux. Merci à Sylvie Dervin d'assurer la bonne ambiance au laboratoire et de m'avoir toujours sauvé lors de mes nombreux démêlés administratifs. Merci à Annie Glomeron et Dany Kouoh-Etamè pour leur sympathie et leur aide précieuse tout au long de cette thèse.

J'ai eu la chance de côtoyer tous les doctorants qui sont passés dans l'équipe depuis sa création ! Merci à Riadh Zaatour, Aymen Jedidi, Rémi Chicheportiche, Rémi Tachet, Nicolas Millot, Nicolas Huth, Fabrizio Pomponio, Marouane Anane, Ahmed Bel Hadj Ayed et Sofiene El Aoud pour la bonne ambiance qui a toujours régné au sein du laboratoire et en dehors.

J'adresse évidemment une pensée particulière à mon ami Joao da Gama Batista qui a partagé mon bureau et bien plus durant ces années de doctorat. La liste ne serait pas complète sans les post-doc aux cotés de qui j'ai eu plaisir à travailler : Stanislao Gualdi, Dahlia Ibrahim et Sophie Laruelle. Enfin, je souhaite bonne chance à Kevin Primicerio pour qui l'aventure ne fait que commencer.

Merci Elise Lepeltier d'avoir toujours été présente, avec ton optimisme indéfectible durant toutes ces années étudiantes.

Un grand merci à ma famille pour son soutien, plus particulièrement à ma mère et à mon frère. Je dédie cette thèse à mon père, qui m'a transmis il y a bien longtemps, le goût du laboratoire.

Leslie, merci de m'accompagner depuis dix ans déjà et d'avoir brillamment supporté les derniers mois plutôt mouvementés. Je t'aime.

Résumé

En utilisant des données haute-fréquence inédites, cette thèse étudie trois types de regroupements (« clusters ») présents dans le marché des changes : la concentration d'ordres sur certains prix, la concentration des transactions dans le temps et l'existence de groupes d'investisseurs prenant les mêmes décisions. Nous commençons par étudier les propriétés statistiques du carnet d'ordres EBS pour les paires de devises EUR/USD et USD/JPY et l'impact d'une réduction de la taille du tick sur sa dynamique. Une grande part des ordres limites est encore placée sur les anciens prix autorisés, entraînant l'apparition de prix-barrières, où figurent les meilleures limites la plupart du temps. Cet effet de congestion se retrouve dans la forme moyenne du carnet où des pics sont présents aux distances entières. Nous montrons que cette concentration des prix est causée par les traders manuels qui se refusent d'utiliser la nouvelle résolution de prix. Les traders automatiques prennent facilement la priorité, en postant des ordres limites un tick devant les pics de volume.

Nous soulevons ensuite la question de l'aptitude des processus de Hawkes à rendre compte de la dynamique du marché. Nous analysons la précision de tels processus à mesure que l'intervalle de calibration est augmenté. Différent noyaux construits à partir de sommes d'exponentielles sont systématiquement comparés. Le marché FX qui ne ferme jamais est particulièrement adapté pour notre but, car il permet d'éviter les complications dues à la fermeture nocturne des marchés actions. Nous trouvons que la modélisation est valide selon les trois tests statistiques, si un noyau à deux exponentielles est utilisé pour fitter une heure, et deux ou trois pour une journée complète. Sur de plus longues périodes la modélisation est systématiquement rejetée par les tests à cause de la non-stationnarité du processus endogène. Les échelles de temps d'auto-excitation estimées sont relativement courtes et le facteur d'endogénéité est élevé mais sous-critique autour de 0.8.

La majorité des modèles à agents suppose implicitement que les agents interagissent à travers du prix des actifs et des volumes échangés. Certains utilisent explicitement un réseau d'interaction entre traders, sur lequel des rumeurs se pro-

pagent, d'autres, un réseau qui représente des groupes prenant des décisions communes. Contrairement à d'autres types de données, de tels réseaux, s'ils existent, sont nécessairement implicites, ce qui rend leur détection compliquée. Nous étudions les transactions des clients de deux fournisseur de liquidités sur plusieurs années. En supposant que les liens entre agents sont déterminés par la synchronisation de leur activité ou inactivité, nous montrons que des réseaux d'interactions existent. De plus, nous trouvons que l'activité de certains agents entraîne systématiquement l'activité d'autres agents, définissant ainsi des relations de type « lead-lag » entre les agents. Cela implique que le flux des clients est prévisible, ce que nous vérifions à l'aide d'une méthode sophistiquée d'apprentissage statistique.

Abstract

The aim of this thesis is to study three types of clustering in foreign exchange markets, namely in price, trades arrivals and investors decisions. We investigate the statistical properties of the EBS order book for the EUR/USD and USD/JPY currency pairs and the impact of a ten-fold tick size reduction on its dynamics. A large fraction of limit orders are still placed right at or halfway between the old allowed prices. This generates price barriers where the best quotes lie for much of the time, which causes the emergence of distinct peaks in the average shape of the book at round distances. Furthermore, we argue that this clustering is mainly due to manual traders who remained set to the old price resolution. Automatic traders easily take price priority by submitting limit orders one tick ahead of clusters, as shown by the prominence of buy (sell) limit orders posted with rightmost digit one (nine).

The clustering of trades arrivals is well-known in financial markets and Hawkes processes are particularly suited to describe this phenomenon. We raise the question of what part of market dynamics Hawkes processes are able to account for exactly. We document the accuracy of such processes as one varies the time interval of calibration and compare the performance of various types of kernels made up of sums of exponentials. Because of their around-the-clock opening times, FX markets are ideally suited to our aim as they allow us to avoid the complications of the long daily overnight closures of equity markets. One can achieve statistical significance according to three simultaneous tests provided that one uses kernels with two exponentials for fitting an hour at a time, and two or three exponentials for full days, while longer periods could not be fitted within statistical satisfaction because of the non-stationarity of the endogenous process. Fitted timescales are relatively short and endogeneity factor is high but sub-critical at about 0.8.

Most agent-based models of financial markets implicitly assume that the agents interact through asset prices and exchanged volumes. Some of them add an explicit trader-trader interaction network on which rumors propagate or that encode groups that take common decisions. Contrarily to other types of data, such networks, if they exist, are necessarily implicit, which makes their determination a

more challenging task. We analyze transaction data of all the clients of two liquidity providers, encompassing several years of trading. By assuming that the links between agents are determined by systematic simultaneous activity or inactivity, we show that interaction networks do exist. In addition, we find that the (in)activity of some agents systematically triggers the (in)activity of other traders, defining lead-lag relationships between the agents. This implies that the global investment flux is predictable, which we check by using sophisticated machine learning methods.

Contents

Résumé détaillé en Français

Le but de cette partie est d'offrir au lecteur non-anglophone un résumé substantiel des travaux de thèse, le reste du manuscrit étant entièrement rédigé en anglais.

Contexte

Le marché des changes, désigné par FX (foreign exchange) dans le reste du résumé est fascinant par bien des aspects, le plus marquant étant sa taille incroyable. Selon la « Bank for International Settlements » BIS [16], le volume moyen journalier de la globalité FX a atteint 5300 milliards en 2013, ce qui représente une augmentation de 35% par rapport à 2010. Dans cette thèse on s'intéressera uniquement au marché « spot », qui à lui seul atteint 2000 milliards de dollars échangés quotidiennement. Pour appréhender un tel chiffre, on peut remarquer que ce volume est équivalent à 36 fois le cumul des imports et exports des 35 plus grandes économies mondiales et à 10 fois le volume échangé sur le marché actions [67]. C'est de loin le plus gros marché financier du monde.

Le complexité de ce marché est la deuxième caractéristique qui frappe immédiatement l'esprit. Il n'y a aucune bourse avec des horaires définies pour changer des devises mais plutôt une grande variété de plate-formes de trading, chacune avec ses propres règles de fonctionnement. Une certaine opacité règne sur ce marché étant donnée l'absence d'obligation de rapporter les opérations de changes à une autorité de régulation (marché dit « over-the-counter »). Les investisseurs du monde entier, du trader haute-fréquence utilisant des techniques de pointe au boursicoteur récréationnel, peuvent effectuer des transactions 24 heures sur 24 et 7 jours sur 7.

L'importance de ce marché invite à l'étude des mécanismes déterminant les taux de change. Une compréhension globale de la dynamique du prix semble impossible au regard de l'hétérogénéité des participants, de l'opacité de la plupart des transactions et de la grande fragmentation de la liquidité. Le but de cette thèse est d'apporter quelques contributions spécifiques du point de vue de la finance

quantitative à ce vaste projet de recherche.

Deux mots composés permettent de délimiter plus précisément le champ de la thèse : *microstructure* et *haute-fréquence*. En finance, la microstructure place l'acte de « trading » au cœur de l'étude et se demande comment les différentes structures imposées aux traders se transforment en dynamique de prix. L'effort de recherche est principalement concentré sur les marchés basés sur un carnet d'ordres. C'est le mode de fonctionnement le plus répandu aujourd'hui. D'immenses progrès ont pu être effectués quand des données haute-fréquence permettant des études statistiques sont devenues disponibles. L'obtention de multiples succès par cette approche est attestée par un nombre croissant d'articles de revues, voir par exemple [19, 26, 50]. Malgré son indiscutable portée, le FX semble avoir été négligé dans cette littérature, ce qui s'explique aisément par le manque de données facilement accessibles sur ses carnet d'ordres. Le travail de thèse met en évidence trois types de regroupements (« clusters ») présents dans le marché des changes : la concentration d'ordres sur certains prix, la concentration des transactions dans le temps et l'existence de groupes d'investisseurs prenant les mêmes décisions. Il vise aussi à combler une partie du manque d'études empiriques sur la microstructure du FX et à répondre à des questions nouvelles liées aux données *identifiées*. Ce type de données permet de connaître les transactions d'un investisseur donné. En effet, tous les deals répertoriés comportent un identifiant (typiquement un nombre) qui les rattachent à leur initiateur, on peut ainsi explorer de nouvelles problématiques. Nous nous intéresserons plus particulièrement à révéler des réseaux de communications implicites entre les traders. Nous présentons maintenant les principaux résultats obtenus chapitre par chapitre.

Chapitre 1 : Organisation du marché FX

Le premier chapitre décrit brièvement l'histoire et l'organisation du marché des changes. Ici nous rappelons juste qu'au prix d'une grossière simplification (les nuances sont apportées au sein du chapitre), on peut diviser le marché en deux segments. D'une part le marché dit *interdealer* où les banques effectuent des transactions de grandes taille (minimum 1 million de la devise de base) pour évacuer leurs inventaires accumulés afin de répondre aux demandes de leurs clients. Les données étudiées dans les chapitres 2 et 3 proviennent de ce segment par l'intermédiaire de la plate-forme EBS. D'autre part le marché client, ou les institutions financières, les entreprises commerciales, les fonds d'investissements et les investisseurs individuels consomment la liquidité offerte par les « dealers » grâce à diverses plate-formes électroniques. C'est l'objet du chapitre 4.

Chapitre 2 : Influence de la taille du tick dans le marché interbancaire

Dans ce chapitre, nous étudions les propriétés statistiques du carnet d'ordres EBS et l'impact d'une réduction de la taille du tick sur sa dynamique. Le trading électronique entre dealers représente environ un tiers du marché spot. Le carnet d'ordres EBS peut donc être considéré comme le plus représentatif du marché global et donc le plus intéressant d'un point de vue académique. Nous nous intéressons à deux paires de devises : EUR/USD et USD/JPY pour lesquelles EBS est la plate-forme dominante. Nos données couvrent une période cruciale au cours de laquelle EBS a décidé de diviser la taille du tick par 10 (procédé appelé décimalisation). Avant d'analyser en détail l'influence de ce changement de résolution, nous vérifions si l'universalité des « faits stylisés » peut être étendu au marché EBS. En effet, à cause du manque de données concernant les carnets d'ordres FX, très peu d'études empiriques s'intéressent à la microstructure du marché des changes au-delà des meilleures limites [77, 78, 68]. Nous retrouvons bien les queues épaisses dans la distribution des retours. La distribution des log-retours calculés à une échelle de 5 min montre une claire déviation à la loi Normale. Une approximation par une loi de Student donne de bons résultats avec des exposants pour les queues de l'ordre de 4 pour l'EUR/USD et de 3.5 pour l'USD/JPY. Les retours ne présentent pas de corrélation et on observe bien le phénomène de « volatility clustering » (mémoire longue pour les retours au carré). Une saisonnalité est clairement présente dans les données, par exemple le nombre moyen de transactions par heure fluctue fortement au cours de la journée. L'activité est haute entre 8h et 18h (heure de Londres), en dehors de ce créneau horaire. On note aussi une chute de l'activité vers 12h à cause de la pause déjeuner et une activité supérieure quand les sessions de New-York et Londres se recoupent. En résumé, les fait stylisés sont aussi valides pour le FX.

Impact sur le carnet

Les données s'étendent du 01-08-2011 au 31-07-2011 et du 01-01-2012 au 31-03-2013 et contiennent un *Quote Record* et un *Deal Record* toutes les 0.1 s. Le premier est un instantané des 10 meilleures limites (prix et quantité) de chaque côté du carnet. Le second donne la transaction qui a atteint le prix le plus élevé côté ask et celle qui a atteint le prix le plus bas côté bid, ainsi que leurs volumes. Pour les données de 2012, le volume total à l'achat et à la vente dans chaque intervalle de temps est aussi fourni. A l'heure actuelle c'est le meilleur jeu de données concernant EBS en termes de fréquence et de profondeur. La taille du

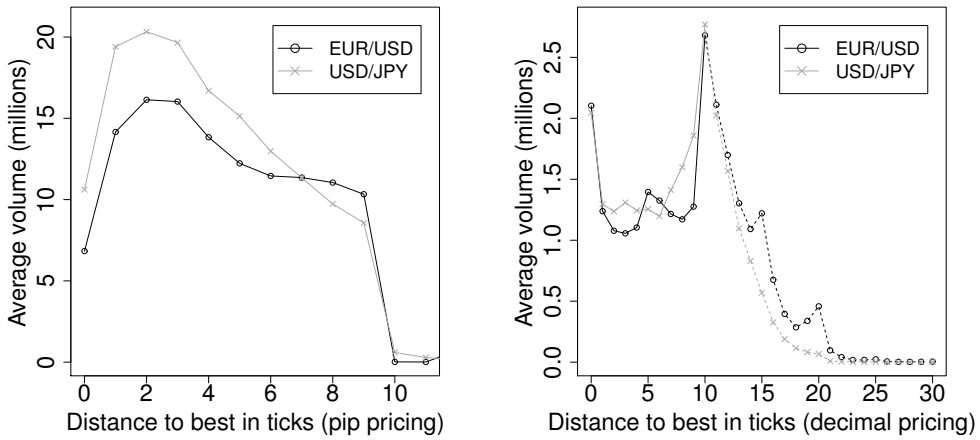

FIGURE 1 – Forme moyenne du carnet côté ask. Forme similaire pour le côté bid. Gauche : Avant décimalisation (Février 2011). Droite : Après décimalisation (Mars 2012). La décimalisation entraîne l'apparition de pics dans la forme moyenne du carnet.

tick EUR/USD est passée de 10^{-4} à 10^{-5} et celle du tick USD/JPY de 10^{-2} à 10^{-3}, impactant considérablement le carnet. Depuis Bouchaud et al. [18], la forme moyenne du carnet, c'est-à-dire le volume moyen présent dans le carnet en fonction de la distance à la meilleure limite, est devenue la façon standard de synthétiser l'information contenue dans ce dernier. Sur la Fig. 1 nous le calculons avant et après la décimalisation. Avant la décimalisation, on retrouve la forme classique de bosse avec un maximum à deux pips tandis qu'après la décimalisation la forme est plus surprenante. Le volume diminue initialement pour augmenter jusqu'à un maximum atteint en 10, correspondant à 1 dans l'ancienne résolution. Pour les deux paires la figure est piquée aux multiples de 10, pour l'EUR/USD on observe aussi des pics de moindre importance à 5 et 15 ticks. Nous montrons dans le prochain paragraphe que ces pics proviennent d'un placement spécifique des ordres limites, ainsi la concentration de volume à des distances particulières dans le carnet est en fait une concentration des ordres en termes de prix absolu.

Concentration des ordres sur les prix entiers

La fréquence élevée de nos données qui les rend presque tick-par-tick dans le cas d'EBS nous permet de reconstruire le flux d'ordres en recoupant les informations des instantanés du carnet et celles sur les transactions. L'analyse du placement

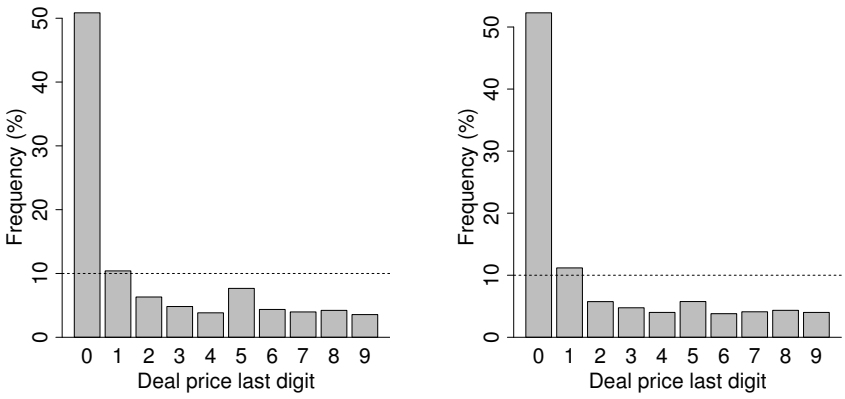

FIGURE 2 – Distribution de la dernière décimale du prix des transactions (côté ask) pour Mars 2012. Les pointillés représentent la fréquence théorique sous hypothèse d'uniformité. EUR/USD à gauche et USD/JPY à droite. Environ 50% des transactions se font sur un prix entier.

des ordres révèle qu'ils sont placés de façon préférentielle sur les prix « entiers » ou « demi-entiers ». C'est-à-dire juste sur ou entre deux prix autorisés par l'ancienne résolution. La distribution de la dernière décimale dans le prix des transactions, qui devrait être uniforme sans effet de concentration, l'illustre clairement sur la figure 2. Le résultat est aussi vrai pour les ordres limites et les annulations. Les chiffres 0 et 5 sont systématiquement favorisés. La concentration est très forte et stable dans le temps, ce qui est surprenant pour un marché si liquide et si mature. Un regroupement similaire des prix a déjà été observé dans différents contextes [48, 103, 84, 97]. L'accumulation d'ordres aux prix entiers les transforme en barrières effectives. En effet, les meilleures limites passent plus de temps sur les prix entiers que sur un prix quelconque. La diffusion est comme localement ralentie, ce qui peut être nuisible au traitement de l'information par les acteurs du marché.

Une explication simple

La concentration des prix s'explique par la forte présence de traders manuels sur ce marché (contrairement au marché actions par exemple). Les traders manuels refusant de s'adapter à la nouvelle grille de prix continuent à utiliser uniquement une échelle grossière. Cela s'explique par le fait qu'au sein des grandes banques, ils sont généralement en charge d'un nombre restreint de transactions, ayant un

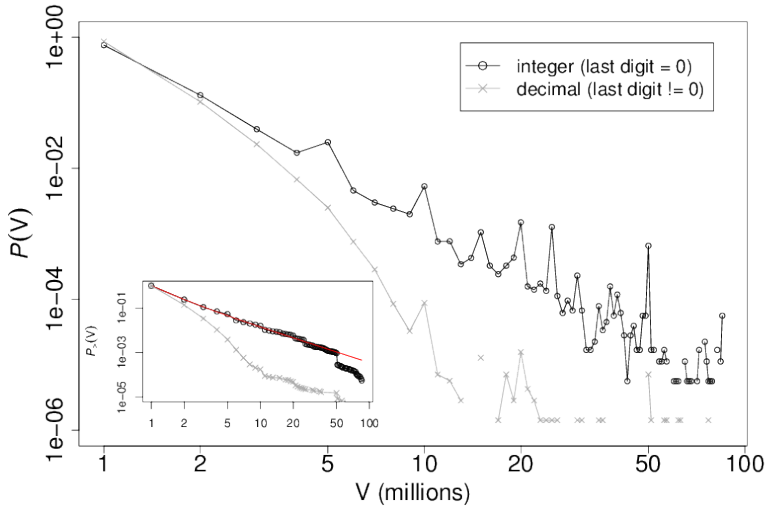

FIGURE 3 – Distribution de la taille des ordres limites pour le cas EUR/USD. La distribution est très différente selon la catégorie du prix. Loi de puissance pour les entiers et loi exponentielle pour les autres. Les exposants de la loi de puissance sont stables dans le temps et valent : 2.6 pour EUR/USD et 2.7 pour USD/JPY. En médaillon : distribution cumulée avec ajustement en loi de puissance. (estimateur du maximum de vraisemblance discret).

gros volume. Ils ne sont donc pas sensibles à des variations de prix aussi petites que 10^{-5}. En revanche, les traders algorithmiques se sont facilement adaptés au nouveau tick et en tirent d'ailleurs avantage en postant des ordres juste devant les prix-barrières pour prendre la priorité tout en étant protégés d'une variation trop importante du prix futur. Cette dualité de comportement vis à vis de résolution du prix est responsable de la concentration des ordres. Cela peut être démontré en séparant les ordres limites en deux catégories : ordres postés sur un entier et ordres postés sur un non-entier et en calculant la distribution de la taille d'un ordre pour ces deux groupes (figure 3). Pour les ordres entiers, la distribution est très large avec une queue en loi de puissance et des pics aux volumes multiples de nombres ronds, deux faits caractéristiques d'un comportement humain. D'autre part la distribution de la taille des ordres décimaux décroît exponentiellement. Les ordres ont principalement la taille minimum d'un million, typique du trading algorithmique, qui utilise un grand nombre de petites transactions.

Chapitre 3 : Processus de Hawkes et dynamique d'arrivée des transactions

Le chapitre 3 s'intéresse à la modélisation de la dynamique d'arrivée des transactions par des processus de Hawkes. Plus généralement, on se posera la question des limites d'une telle modélisation. Beaucoup de fits par des processus de Hawkes dans la littérature ont l'air excellent visuellement (QQplots des résidus) mais ne passent pas les tests statistiques standards. Notre but est de fournir une étude complète de la qualité de l'ajustement aux données quand la fenêtre de calibration est augmentée et ce pour différentes spécifications du noyau. Le marché FX est intéressant pour cette tache grâce à son fonctionnement permanent, évitant ainsi le problème de fermeture nocturne rencontré dans les modélisations du marché actions.

Processus de Hawkes

Un processus de Hawkes est caractérisé par son intensité :

$$\begin{aligned}\lambda_t &= \mu_t + \int_0^t \phi(t-s)dN_s \\ &= \mu_t + \sum_{t_i<t} \phi(t-t_i),\end{aligned}$$

où μ_t est une intensité de base décrivant l'arrivée Poissonnienne d'événements exogènes, et où le second terme est une somme pondérée sur les événement passés. Le noyau $\phi(t-t_i)$ décrit l'impact sur l'intensité actuelle d'un événement précédemment arrivé en t_i. Un processus de Hawkes peut aussi être interprété en termes de processus de branchements, où des événements « mères » exogènes arrivant avec une intensité μ_t donnent naissance à des événements « filles » endogènes. Chacun des événement filles peut donner à son tour naissance à d'autres filles (petites-filles pour l'événement mère original), et ainsi de suite. La quantité $n \equiv \int_0^\infty \phi(s)ds$ est égale au nombre moyen d'enfants pour n'importe quel événement et est appelé coefficient de branchement.

Noyaux paramétriques

La performance des noyaux suivants est comparée dans l'étude.

- Somme d'exponentielles :

$$\phi_M(t) = \sum_{i=1}^{M} \alpha_i e^{-t/\tau_i},$$

où M est le nombre d'exponentielles. Les amplitudes α_i et échelles de temps τ_i sont les paramètres à estimer. Le coefficient de branchement est facilement donné par : $n = \sum_{i=1}^{M} \alpha_i \tau_i = \sum_{i=1}^{M} n_i$.

- Les lois de puissance ont l'avantage de ne demander qu'un nombre restreint de paramètres, facilitant leur calibration. Un noyau en loi de puissance s'approxime par :

$$\phi_M^{\mathrm{PL}}(t) = \frac{n}{Z} \sum_{i=0}^{M-1} a_i^{-(1+\epsilon)} e^{-\frac{t}{a_i}},$$

où

$$a_i = \tau_0 m^i.$$

M contrôle l'étendue de l'approximation et m sa précision. Z est défini tel que $\int_0^\infty \phi_{PL}(t)dt = n$. Les paramètres sont le coefficient de branchement n, l'exposant de la loi ϵ et la plus petite échelle de temps τ_0.

- Loi de puissance approximée avec une coupure aux temps courts [53] :

$$\phi_M^{\mathrm{HBB}}(t) = \frac{n}{Z} \left(\sum_{i=0}^{M-1} a_i^{-(1+\epsilon)} e^{-\frac{t}{a_i}} - S e^{-\frac{t}{a_{-1}}} \right),$$

la définition est la même que ϕ_M^{PL} avec l'ajout d'une décroissance exponentielle pour les temps inférieurs à τ_0. S est défini tel que $\phi_M^{\mathrm{HBB}}(0) = 0$.

- Nous proposons aussi un nouveau type de noyau, construit avec une loi de puissance ϕ_M^{PL} et une exponentielle à paramètres libres. Cela afin d'avoir plus de contrôle dans la structure des échelles de temps. La définition du noyau devient :

$$\phi_M^{\mathrm{PLx}}(t) = \frac{n}{Z} \left(\sum_{i=0}^{M-1} a_i^{-(1+\epsilon)} e^{-\frac{t}{a_i}} + b e^{-\frac{t}{\tau}} \right),$$

où l'exponentielle rajoute deux paramètres b et τ. Les autres variables conservent le même sens qu'au-dessus.

La première forme est très flexible et permet d'approximer n'importe quelle fonction continue, ce au prix du problème connu de l'estimation des paramètres de la somme d'un grand nombre d'exponentielles [110]. Les deuxième et troisième formes tentent de reproduire la longue mémoire observée sur différents marchés mais sont beaucoup moins flexibles. La dernière forme tente d'effectuer la synthèse. Une fois le noyau spécifié, les paramètres sont estimés en maximisant la logvraisemblance. La structure exponentielle de toutes les spécifications permet d'exploiter une relation de récurrence lors du calcul de la vraisemblance, réduisant ainsi la complexité de $\mathcal{O}(N^2)$ à $\mathcal{O}(N)$ (voir Ozaki [98]).

Comparaison des noyaux

Les noyaux sont comparés sur leur capacité à modéliser les temps d'arrivée des transactions sur EBS. Si la description est correcte, les résidus définis par $\theta_i = \int_{t_{i-1}}^{t_i} \hat{\lambda}_t dt$, doivent être i.i.d. exponentiels. Les résidus sont soumis à trois test standards : Kolmogorov-Smirnov, Ljung-Box et Excess-Dispersion. Les log-vraisemblances, les AIC et les poids d'Akaike moyennés sur toutes les fenêtres de calibration sont calculés pour obtenir un classement des noyaux, et ce à trois échelles de temps : horaire, journalière et bi-journalière. Dans ce résumé, nous ne présentons que l'échelle journalière (24h). Les résultats sont synthétisés dans le tableau 1.

Seuls ϕ_2 et ϕ_3 sont capables de passer le test de Ljung-Box. ϕ_3 est le modèle le plus performant selon les poids d'Akaike ; il n'est rejeté par aucun test. On remarque que ϕ_{15}^{PLx} est aussi un bon candidat. On peut se faire une idée générale des résultats sur tous les jours en observant les QQ-plots. En effet, sous l'hypothèse nulle, les résidus possèdent tous la même distribution, indépendamment du jour considéré. On peut donc tous les concaténer et construire les QQ-plots par rapport à la loi exponentielle. La figure 4 montre les performances de 4 noyaux et confirme visuellement les résultats du tableau 1. De plus, elle permet de voir où chaque noyau est bon ou mauvais. Par exemple, ϕ_{30}^{PL} est meilleur dans les extrêmes que dans le cœur de la distribution. Nous voyons aussi les difficultés de ϕ_3 dans ce secteur, résolues en ajoutant une 4ième exponentielle (voir ϕ_4).

Conclusions

Les résultats sont plutôt positifs : les processus de Hawkes peuvent modéliser de façon statistiquement satisfaisante, selon trois tests, une journée complète de données. Ils sont donc capables de décrire précisément un grand nombre d'événements (en moyenne 15000). Les échelles de temps estimés sont assez courtes (de 0.15 s à 300 s), cela montre que les effets endogènes (qui représente 80% des évé-

Kernel	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_{15}^{HBB}	ϕ_{15}^{PL}	ϕ_{30}^{HBB}	ϕ_{30}^{PL}	$\phi_{15}^{\mathrm{PL_x}}$
n	0.48	0.79	0.83	0.85	0.81	0.83	0.98	0.97	0.88
pKS	$7e-13$	0.09	0.13	0.16	4×10^{-6}	2×10^{-7}	6×10^{-4}	$4e-6$	0.04
pED	0	0.10	0.31	0.45	0.61	0.51	0.52	0.49	0.66
pLB	0	0.058	0.056	0.012	$9e-5$	0.001	1×10^{-4}	$2e-4$	0.006
$\log\mathcal{L}_p$	60271.0	61559.3	61596.2	61575.5	61279.6	61340.2	61286.3	61340.6	61468
AIC_p	−120525.4	−123097.7	−123167.4	−123121.8	−122540.4	−122661.7	−122553.9	−122662.5	−122912.0
ϵ	NA	NA	NA	NA	0.090	0.115	0.13	0.14	0.08
w	0	0.13	0.38	0.24	0	0	0	0	0.24
N_{max}	0	9	22	14	0	0	0	0	14

TABLE 1 – Comparaison des noyaux pour des journées complètes. 59 points. pKS, pED et pLB sont les p-values moyennes des test de respectivement Komogorov-Smirnov, Excess-Dispersion et Ljung-Box. $\log\mathcal{L}_p$ est la log-vraisemblance par point pour chaque intervalle, moyennée sur tous les intervalles puis multipliée par le nombre moyen de point par intervalle. Idem pour le critère d'information d'Akaike AIC_p. Les poids d'Akaike normalisés $w[\phi] = \frac{1}{W}\exp\left(-\frac{AIC[\phi]-AIC_{min}}{2}\right)$, sont les probabilités que le noyau ϕ soit le meilleur modèle selon la divergence de Kullback–Leibler [108]. $N_{max}[\phi]$ est le nombre d'intervalles pour lesquels le poids d'Akaike du noyau ϕ est le plus élevé.

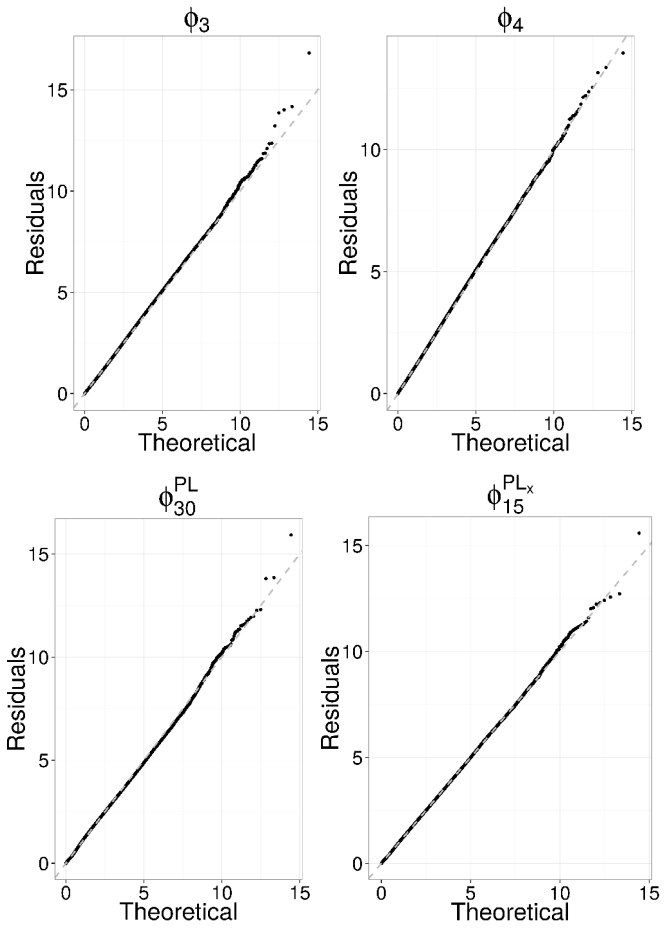

FIGURE 4 – QQplots de la fusion des résidus provenant de tous les jours considérés.

nements) sont limités à des réactions à court terme sur le marché FX. A mesure que la période de calibration est augmentée, le coefficient de branchement estimé devient plus grand. L'effet est attendu car une mesure de longue-mémoire requiert par définition une longue série temporelle. Les tableaux montrent clairement que l'utilisation des noyaux en loi de puissance augmente mécaniquement le coefficient de branchement apparent, certains étant dangereusement proches de la limite de stabilité du processus : $n = 1$ (e.g ϕ_{30}^{HBB} et ϕ_{30}^{PL}). De plus, les meilleurs noyaux ne sont jamais ceux avec le plus grand n.

L'extension des calibrations à des intervalles supérieurs à un jour se heurte aux non-stationnarités des parties endogène et exogène de la dynamique. L'exemple de la chute d'activité autour de l'heure du déjeuner est représentatif de l'impossibilité de considérer le noyau constant sur de grandes périodes de temps.

Chapitre 4 : Réseaux de communications implicites entre traders

Les travaux de thèse se sont concentrés jusque-là sur le marché inter-dealer. Il est temps d'explorer l'autre facette du marché : les flux clients. Les données utilisées dans ce chapitre proviennent de deux sources : une grande banque exerçant un mandat de market-maker : (dénotée LB) et un courtier en ligne suisse : Swiss-Quote (SQ). Chaque ordre provenant d'un de leurs clients est enregistré avec un identifiant client, un prix, une taille et la paire de devises concernée.

Les traders, utilisant des stratégies similaires, observant le même prix, doivent présenter un certain degré de synchronisation au moment de la prise de décision (achat-vente-inaction). Il semble donc possible de les classifier au sein d'un même groupe. Notre but est de catégoriser les clients en groupe de traders agissant de concert, comme si une communication implicite existait entre eux. Pour arriver à nos fins, nous utilisons la méthode développée par Tumminello et al. [106], nommé statistically validated networks (SVN). Ensuite, nous adaptons cette méthode pour découvrir des relations de lead-lag entre les groupes. Finalement, nous exploitons ces relations pour tenter de prédire la signe du flux client à travers une méthode de machine learning : les forêts aléatoires.

Méthode

Les SVN se construisent en mettant un lien entre deux agents si leur degré de synchronisation est supérieur à celui attendu par une prise de décision à des temps aléatoires. Pour chaque heure t, une variable d'état $s_i(t)$ est associée à

chaque agent i. Les séries s_i sont calculées en comparant le déséquilibre achat-vente relatif $(\frac{V_{achat}(t)-V_{vente}(t)}{V_{achat}(t)+V_{vente}(t)})$ à un seuil fixé à 0.01. Si le ratio est supérieur à 0.01, alors $s_i(t) = +1$. Si le ratio est inférieur à -0.01, alors $s_i(t) = -1$. Si le ratio est entre les deux, alors $s_i(t) = 2$. Si l'agent est inactif durant toute l'heure considérée alors $s_i(t) = 0$. Pour chaque paire d'agents et pour chaque combinaison d'états, on peut associer une p-value au nombre de co-occurences. Soit N_P le nombre d'heures où $s_i(t) = P$, N_Q le nombre d'heures où $s_j(t) = Q$ et N_{PQ} le nombre d'heures où les deux situations arrivent simultanément. La p-value, ou la probabilité d'avoir plus de co-occurences par un trading à temps aléatoires tout en respectant le niveau d'activité de chaque agent est donnée par :

$$p(N_{PQ}) = 1 - \sum_{k=0}^{N_{PQ}-1} \frac{\binom{N_P}{k}\binom{T-N_P}{N_Q-k}}{\binom{T}{N_Q}},$$

où T est la longueur des séries temporelles. La p-value est par suite comparée à un niveau donné par le « False Discovery Rate » (FDR), pour tenir compte de la multitude de tests effectués en parallèle. Si la p-value est inférieure au niveau FDR alors un lien de nature PQ est validé entre les deux agents.

Résultats

Actions simultanées

Nous appliquons la méthode précédente sur une période de 4 mois pour obtenir les réseaux représentés sur la figure 5. Les communautés apparaissant en couleurs sont détectées en utilisant l'algorithme InfoMap. Pour les deux jeux de données nous observons un grand groupe entouré de beaucoup de petits groupes satellites.

Actions retardées

Nous adaptons la méthode précédente pour détecter des liens inter-temporels entres les groupes précédemment détectés. Au lieu de considérer les actions des agents individuellement, nous regardons les séries $s_{g_i}(t)$ qui encodent l'état global du groupe g_i. Avant de valider des liens entre groupes, une série est retardée, i.e. les co-occurences examinées sont entre $s_{g_i}(t-1)$ et $s_{g_j}(t)$. Nous révélons ainsi sur la figure 6 des réseaux de lead-lag entre les groupes de traders. L'existence de liens valides sous de sévères critères statistiques démontre l'existence d'une prévisibilité dans la direction de trading des groupes, que nous essayons d'exploiter dans la partie suivante.

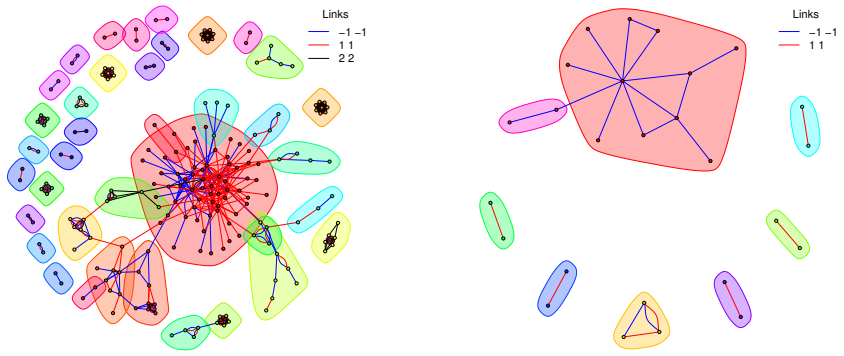

FIGURE 5 – Exemples de réseaux de traders statistiquement valides pour la paire EUR/USD. A gauche : SQ. A droite : LB.

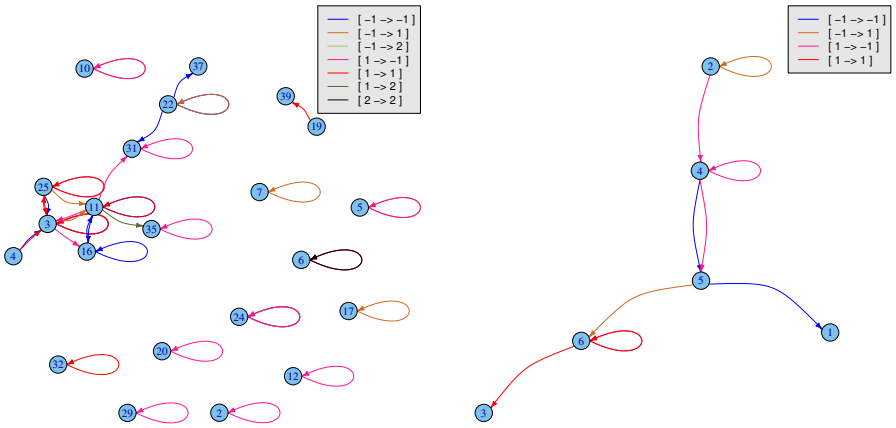

FIGURE 6 – Exemples de réseaux de traders statistiquement valides pour la paire EUR/USD. A gauche : SQ. A droite : LB.

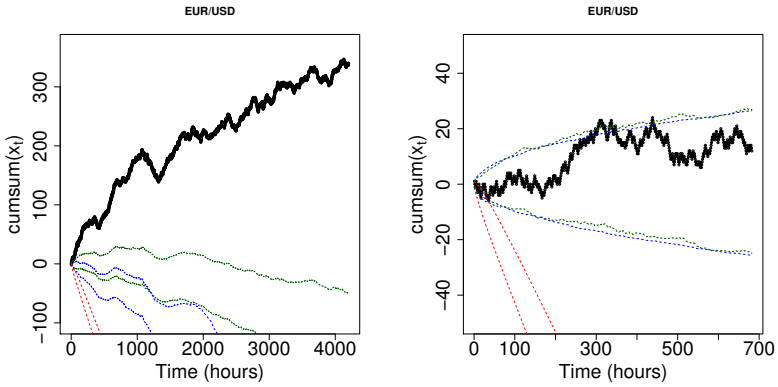

FIGURE 7 – Prédictions du flux client EUR/USD. A gauche : LB. A droite : SQ. La forêt aléatoire (en noir) surpasse les benchmarks (en rouge, bleu et vert) pour LB et obtient une performance remarquable, tandis que les prédictions pour SQ ne dépassent pas les bornes bleues et vertes.

Prédictions du flux

Les forêts aléatoires sont une méthode de machine learning totalement non-paramétrique et non sujette à l'*overfitting*. Elles représentent donc un candidat très intéressant pour la prédiction du signe du flux client. Sur chaque fenêtre glissante de 4 mois, la forêt est entraînée à expliquer le signe du flux à partir des actions des groupes trouvés dans les SVN. La prédiction est ensuite faite *out-of-sample* sur les heures de la journée suivant la période d'apprentissage. La figure 7 représente la somme cumulée d'une variable de « vérité » ($+1$ si la prédiction est correcte, -1 sinon). Notre méthode est clairement supérieure à trois benchmarks décrits dans le corps de la thèse pour LB. En revanche, les résultats sont équivalents à deux des trois benchmarks pour SQ.

Conclusions

De façon surprenante, le comportement pourtant très variable des investisseurs présente un niveau de synchronisation notablement supérieur à celui attendu de pures fluctuations statistiques. Un réseau contemporain et inter-temporel entre les traders a été clairement mis en évidence. L'extension des résultats à d'autres catégories d'actifs est envisageable puisqu'il n'y a aucune raison évidente pour laquelle ils devraient être spécifiques au marché FX.

L'analyse de la stabilité de ces réseaux au cours du temps reste à faire. Ainsi que l'étude systématique des paramètres rentrant en jeu dans la prédiction par

forêts aléatoires. Cela permettra une amélioration et une meilleure compréhension des prévisions, rendant possibles d'éventuelles applications en industrie.

Introduction

The foreign exchange (FX) market possesses many fascinating aspects, one of which is its sheer size. According to the triennial survey of the Bank for International Settlements (BIS) [16], the global FX turnover reached a daily average of $5.3 trillion in 2013, 35% rise more than in 2010. If we restrict ourselves to the spot market, which is the focus of this thesis, we still get an enormous volume of $2 trillion. The rest being constituted mainly by swaps, forwards and options. To fully grasp the meaning of this figure, it is fruitful to notice that the spot market size exceeds 36 times the combined exports and imports for the world's 35 largest economies and 10 times the exchange-traded equity turnover [67]. It is by far the largest financial market in the world, affecting output, employment and inflation. A second remarkable feature of this market is its complexity. There is no official exchange with definite opening hours, but rather a wide variety of trading platforms, each one with its own set of rules and level of transparency. Investors from all other the world can trade 24/7. These investors range from high-frequency traders, using computerized strategies to the recreational individual trader.

Clearly, understanding the mechanisms driving exchange rates is of utmost importance from an academic point of view but also for all the practical benefits that the global economy could draw from such knowledge. However, due to the trading heterogeneity, opaqueness and liquidity fragmentation of this market, a full grasp of the dynamics is an almost impossible task that would mobilize many fields such as mathematics, economics, finance, sociology, psychology, etc.

The purpose of this thesis is to make specific contributions from a quantitative finance perspective to this tremendous research challenge.

Context

Two composed keywords are decisive to pin down the domain of this thesis: microstructure and high-frequency. The object of microstructural analysis is the trading process itself and how the microscopic phenomena emerging from the institutional

structure of trading determine asset prices. The phenomena researchers study in this field are diverse and include the liquidity supply, consumption and resilience, the order book dynamics, the market-impact and the volatility. More and more progress in that direction has been made thanks to the availability of high-frequency data, which made statistically reliable results possible. Many review paper can attest the wealth of results reached by this relatively recent field (see for example [19, 26, 50]).

Although its significance cannot be questioned, the FX market seems to have been sidelined in the above mentioned literature. It is easily explained by its fragmented nature and the lack of easily accessible data on FX order books. This thesis provides precise and mainly empirical contributions to the FX microstructure research effort. In particular, three kinds of clustering are studied. A price clustering due to a tick size reduction, the clustered arrival of trades and finally, thanks to *trader-resolved* data, clusters of investors making the same decision.

A focus on two extremely different market segment

This work can be split into two categories corresponding to the two extreme ends of the FX market. Firstly, the interdealer market, where major dealing banks' traders conduct high volume transaction to offload risky inventory. It is the heart of the FX market and detailed data are available making it the best candidate for order book studies. Secondly, the dealer-customer market, where dealers offers liquidity to their customers. We also include in this part the broker-customer market, where an online broker executes low size transactions for retail investors and small asset managers. This segment has been growing quickly in the past few years and constitutes now an interesting topic. The FX market organization will be detailed in the first chapter.

Outline

The rest of the thesis is organized as follows. Chapter 1 is descriptive and details the market organization. A brief history is given, showing how the electronic revolution has changed the face of the market. Following by the depiction of the actors characteristics and interactions, unveiling the wide diversity of the FX ecology. This chapter is useful from a cultural point of view but also permits to understand the position within the whole of the electronic platforms under scrutiny in the thesis. In each subsequent chapters, we report an original contribution alongside the relevant context and literature.

In Chapter 2, we study the leading platform for EUR/USD and USD/JPY interdealer trading, namely EBS. We perform an empirical investigation of order flow and limit order book dynamics. We provide statistics regarding the average shape of the order book, and the distributions of market order, limit order and cancellations price. We analyze the impact of a ten-fold tick size reduction that occurred on March 2011 on these properties. A large fraction of limit orders are still placed right at or halfway between the old allowed prices. This generates price barriers where the best quotes lie for much of the time, which causes the emergence of distinct peaks in the average shape of the book at round distances. Furthermore, we argue that this clustering is mainly due to manual traders who remained set to the old price resolution.

In Chapter 3, we raise the question of what part of market dynamics Hawkes processes are able to account for exactly. We document the accuracy of such processes as one varies the time interval of calibration and compare the performance of various types of kernels made up of sums of exponentials. Because of their around-the-clock opening times, FX markets are ideally suited to our aim as they allow us to avoid the complications of the long daily overnight closures of equity markets.

In Chapter 4, we analyzes transaction data of all the clients of two brokers, encompassing several years of trading. By assuming that the links between agents are determined by systematic simultaneous trading decisions, we show that implicit interaction networks exist. In addition, we find that the (in)activity of some agents systematically triggers the (in)activity of other traders, defining lead-lag relationships between the agents. This implies that the global investment flux is predictable, which we check by using random forests.

Chapter 1

The FX market anatomy

Most of the information in this chapter borrows from Ref. [67].

1.1 History of the modern FX market

Besides a genuine interest on its own, roaming through the history of the foreign exchange market is enlightening to understand its current multi-layered nature.

1.1.1 The telephone era

The modern era of the FX market began with the introduction of floating rates in the early 1970s. The market was well split into two groups at these times. *Customers* requiring actual currency conversion to conduct international trade in goods and services and *dealers* (usually large banks) fulfilling the customers' need of a counterparty. Dealers captured the bid-ask spread to compensate the trouble of liquidity provision (inventory risk, credit quality risk). This led to the emergence of two "sub-markets" with different characteristics, the customer market where transactions involved a customer and a dealer and the interdealer market where dealers traded among themselves with higher volumes and tighter spreads. This kind of market is called "over-the-counter" (OTC) as opposed to a centralized exchange structure (the NYSE, for example). The vast majority of the transactions were conducted over the phone. Figure 1.1 (top) gives a layout of the market in the 1980s. Typically, a customer (C) would call one or several dealers (D) and ask for his ongoing quotes. He could then decide to buy or to sell the base currency[1] or to "pass" without trading. Dealers keen to set out interdealer

[1]The first currency quoted in a currency pair. For instance, Euro in the EUR/USD case.

transactions could call each other directly (arrow 1 in the figure) or preserve their anonymity by placing an order to the *voice broker* (VB, arrow 2). Voice brokers were a kind of permanently open multi-party telephone line on speaker, where brokers used to yell out quotes from dealers and other dealers would yell back to trade: "mine!" or "yours!".

1.1.2 The electronic revolution

1.1.2.1 Interdealer market

It was very cumbersome for dealers to stay up-to-date with current market prices. They had to make calls very frequently because of the flickering market conditions. In order to improve trading efficiency, Reuters introduced in 1987 an electronic system for bilateral communication called D2000-1, through which dealers could engage instant message conversations to arrange trades (Fig. 1.1 middle, arrow 5). At the same time, another Reuters product increased the information flow: the FXFX page. On this screen, dealers provided indicative (non-binding) quotes. Essentially, this screen showed adverts posted by dealers to show their active interest in the market. However, those bid/ask quotes were not firm prices at which dealers could trade but are rather just indicative quotes. Nevertheless, for approximately 10 years FXFX was the main source of real-time exchange-rate information for traders.

Reuters came up with a decisive innovation in 1992, D2000-2, an electronic platform similar to current limit order books. This automatic order-matching engine became favored by dealers more and more over direct trading and voice brokers because of the operational efficiency and the anonymity. A year later, a group of banks frightened by the possibility of Reuters dwarfing the interdealer trading volume launched a competing platform: the **Electronic Broking Services (EBS)**. EBS is the core object of the two following chapters.

These two electronic brokers (EB), Reuters D2000-2 and EBS, allowed the FX interdealer market to enter a new era and throughout the 1990s their popularity only went up (Fig. 1.1 middle, arrow 4). Electronic platforms nowadays dominate the interdealer trading in liquid currencies. Each of the aforementioned platforms owns a market segment: while EBS is the leader in terms of traded volume for the EUR, JPY and CHF, Reuters prevails in the GBP, AUD, CAD, NOK, SEK and DKK trading. The increased transparency gained from the EB allowed dealers to have a reference price for this decentralized market and to adjust their quotes accordingly. Surprisingly, voice brokers did not completely disappear and remain important for less liquid currencies. In 2010, this seemingly old-fashioned

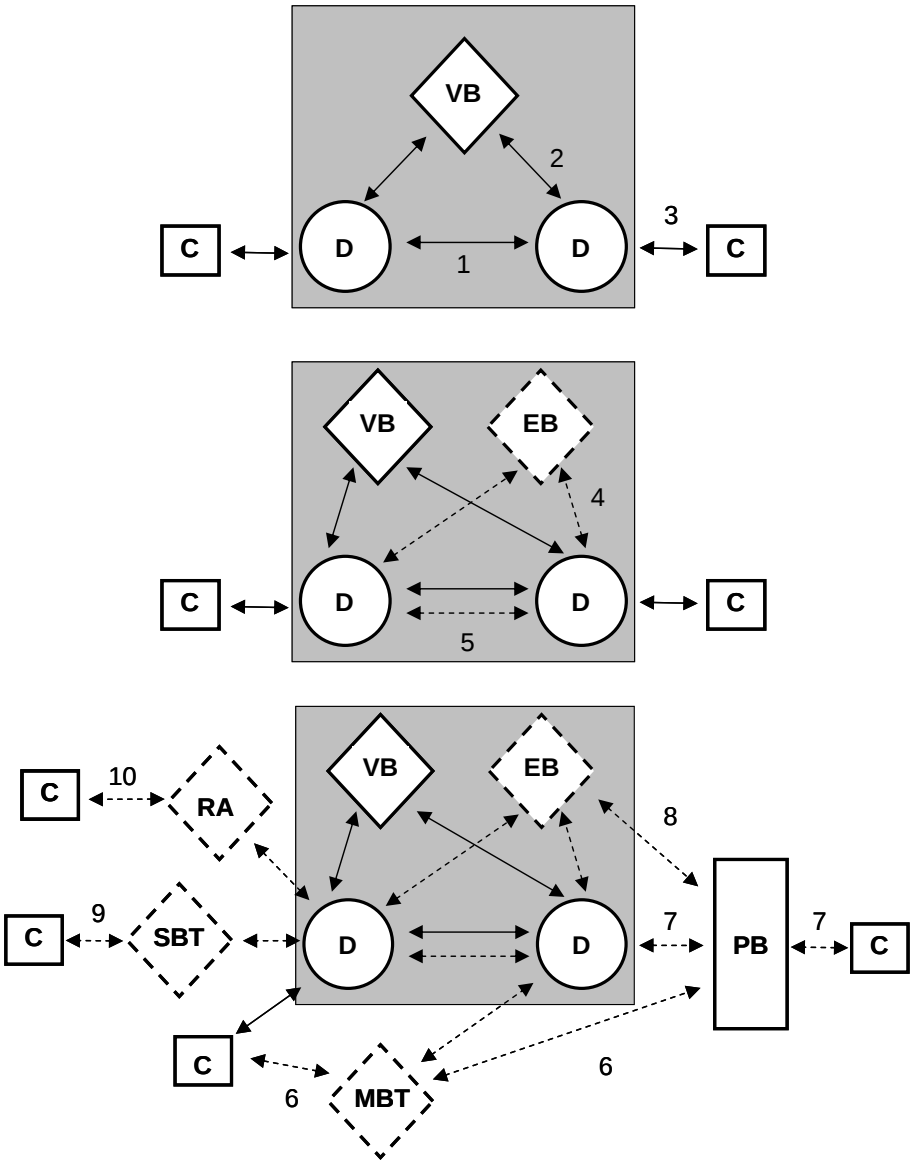

Figure 1.1: Evolution of the FX market anatomy. Top: 1980s. Middle: Early and mid 1990s. Bottom: 2010. The complexity is increasing with time. D = dealer, C = customer, VB = voice broker, EB = electronic broker, PB = prime broker, MBT = multi-bank trading system, SBT = single bank trading system, RA = retail aggregator. Solid arrows = phone transactions. Dashed arrows = electronic transactions. Source: King et al. [67].

technology still represented 10 percent of FX spot trading.

1.1.2.2 Customer market

In the mid 90s, when dealers secured higher profits than before, the customer market did not benefit from the enhanced technology and the tight spreads that went along. This profitability brought heavy competition for customer business boosted by the development of the Internet, a flurry of new electronic platforms targeting end-customers quickly emerged. Figure 1.1 (bottom) attests the diversity of new trading venues for customers. Moreover, it shows that by then the split between customer and interdealer market had become blurred. A crucial transformation occurred when several independent firms developed "multibank trading systems" (MBT, arrow 6), which opened the possibility for customers to trade more directly with several dealers. The first MBT was Currenex in 1999. Customers were able to request quotes from many dealers at the same time without calling them one by one. A consortium of banks launched FXall, a similar platform, in 2001. MBT also came in a limit-order book fashion (e.g.: Hotspot FX in 2000), granting anonymity for customers for the first time. Customers could trade aggressively and passively on these platforms while dealers maintained a continuous quotation (by contract). In the meantime, major banks developed proprietary electronic platforms for their clients, called single bank trading systems (SBT, arrow 9). Clients could then set up trades easily to suit their needs and execute them with that particular bank immediately. Notable examples of SBT were UBS's FX Trader, DB's Autobahn or BNP's Cortex FX. Finally, let us mention prime brokerage (PB, arrows 7 and 8), a service through which banks authorized end-customers (particularly hedge-funds) to trade directly in the interdealer market, thus effectively bringing in liquidity. This spell of innovations brought a fragmented liquidity provision, more transparency, an enhanced operating efficiency, tighter bid-ask spreads and data for the researcher!

1.2 Today

1.2.1 Participants

With the traditional customer-dealer split almost gone, it is imperative to portray the new parties involved. Who are the different actors and their motives in the FX market? First, the "historical" FX customers: corporations and governments, which participate in the market only to support their core business. For example a French company might export goods to Japan and receive yens. It then needs

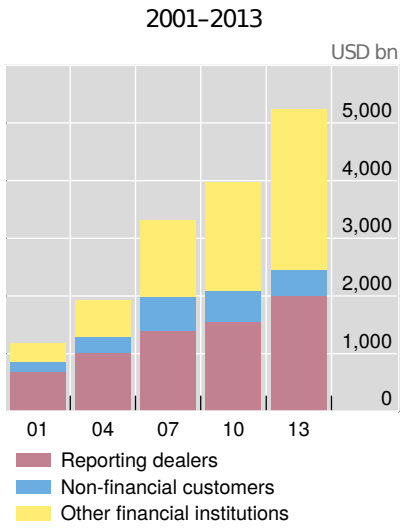

Figure 1.2: Foreign exchange turnover by counterparty. The interdealer market share is shrinking consistently throughout the 2001-2013 period. Source: Bank for international settlements [16].

the market to sell them against euros. These non-financial customers account for only 9% of the turnover. All numbers in this section come from the last triennial survey from the BIS [16]. The second broad category is constituted by the financial institutions that are not registered dealers (other financial institutions, in the BIS terminology). It includes asset managers, small banks, currency funds, pension funds, mutual funds, hedge-funds, HFT firms and central banks. They are really diverse in their trading needs, strategies and time horizons, but relative to the corporate customers they trade larger volumes and are more informed. This group accounts for 53% of the market turnover. Finally, the "official" dealers, mainly large banks, represent 39% of the total turnover. The domination of the other financial institutions in terms of trading volume is concomitant with the emergence of the electronic platforms outside the interdealer realm (see Sec. 1.1.2.2) as demonstrated by Figure 1.2. The matching of an increasing fraction of customer trades internally by large dealing banks also contributed to the decline of the interdealer market share.

1.2.2 New trends

Electronic platforms are clearly the favored trading methods nowadays with a 50% market share. This fraction rises to 64% if we look only at spot transactions.

This high fraction is partly due to the explosion of automatic trading in the recent years. The fragmented nature of this market triggers the need for liquidity aggregation algorithms that scan the many possible liquidity sources and choose one to execute orders. Different computer algorithms seek to optimize execution strategies (to minimize market impact), hedging, high-frequency strategies, etc. The proportion of automatic trading at EBS rose from 28% to 68% between 2007 and 2013. A few consequences are illustrated in chapter 2.

A second notorious trend is the emergence of retail trading. Until the beginning of the XXI century, transaction costs were excessive for small investors, but new online brokers called retail aggregators (RA in figure 1.1) started to stream FX dealers prices to their clients. These firms bundle small trades from many investors and execute them at once with a major dealing bank. The order sizes getting much larger, dealers are keen to provide liquidity indirectly to these retail customers. Some of these brokers also match some trades internally, acting as a dealer for the client. The order flow under scrutiny in chapter 4 emanates from such firms. In 2013, retail trading represented 3.8% of spot turnover.

Chapter 2

Tick size reduction and price clustering in the EBS order book

2.1 Introduction

The foreign exchange (FX) market, being the largest financial market in the world, affecting output, employment and inflation, rightly draws a lot of attention from academics. Since the 1990s, researchers have been able to access large datasets with intra-day resolutions. Starting with Wasserfallen and Zimmermann [109] and Müller et al. [87], this wealth of data led to the emergence of a growing body of empirical studies on high-frequency FX rates. A set of stylized facts, reviewed for the first time by Guillaume et al. [51] is now firmly established. The most important ones are the fat-tailed distribution of returns, the absence of linear autocorrelation of returns (except on short time scales) and the volatility clustering phenomenon. These properties are common to a wide range of assets: equities, commodities, bonds, etc. More recent and comprehensive results can be found in Refs. [40, 33].

The lack of readily available data on FX order books explains the relatively small number of empirical investigations on FX microstructure (beyond level I). Notable exceptions are Lo and Sapp [77, 78] and Kozhan and Salmon [68]. These studies use data from Reuters, while our provider is EBS (Electronic Broking Service). These are the two largest FX electronic communication networks nowadays. Typically, dealing bank traders (and also hedge funds via prime brokerage) use EBS to conduct high volume transactions in order to liquidate unwanted accumulated inventory. For a detailed description of the foreign exchange market structure, see Chapter 1. Electronic interdealer trading represents around one third of spot FX trades [10]. This market can be seen as the heart of the global

FX market; therefore, it is relevant for FX order book studies. The EBS market with high-frequency data was already considered by several authors. Berger et al. [14] investigate the relationship between order flow and exchange rates. Berger et al. [13] analyze the factors driving the volatility persistence. Interestingly, they have shown that variations in market sensitivity to information play at least as large a role as do variations in the flow of information reaching the market through the trading process. Hashimoto et al. [55] show that a "run"[1] has some predictive power on the direction of the next price move. Finally, the mechanisms behind FX rates tail events are studied by Osler and Savaser [96], who found that price-contingent trading may be a major source of extreme returns.

Here, we analyze data about two currency pairs: EUR/USD and USD/JPY that contains more quotes on each side of the book than the aforementioned studies. Crucially, they cover a period during which a major change of price resolution occurred. Indeed, in March 2011, EBS decided to reduce the tick size by a factor ten. More details about the data are provided in section 2.2.

In this chapter, we want to take advantage of these features to analyze the order book most important properties and see how they are affected by the change in tick size. We also test the validity of the stylized facts concerning returns against our FX data.Goldstein and Kavajecz [47] analyzed similar tick size reduction in the equity context. The average shape of the book is deeply modified with the appearance of peaks at round distances and the spread distribution became bi-modal in the EUR/USD case. The very high frequency sampling of our data (which makes them almost tick-by-tick in the EBS case) allows us to devise a method to infer the stream of orders (limit orders and cancellations) from the deal and quotes data. This reveals strange patterns in the order placement and volume. We show that these facts stem from the emergence of a strong price clustering, i.e. a tendency for prices to congregate around some specific values, after the tick size reduction. Surprisingly in such a liquid and mature market, the clustering is very strong and stable in time. Goodhart and Curcio [48], Sopranzetti and Datar [103] and Mitchell and Izan [84] also noticed clustering in spot FX rates but their studies concern the pre-euro era and use low-frequency indicative quotes[2]. Osler [97] reports clustering in stop-loss and take-profits orders placed at a large dealing bank (National Westminster Bank). Our study shows that the phenomenon is pervasive in the EBS market. Clustering affects transaction prices and order prices. It creates an accumulation of volume at round numbers turning them into price barriers, thus presumably hindering markets' ability to

[1] Continuous increases or decreases in deal prices for the past several ticks.

[2] Indicative quotes are non-binding quotes that have been posted by individual banks to the electronic data networks for informative purpose.

process information efficiently. Finally, we show that the clustering is mainly due to manual traders who do not use the new price resolution. Automatic traders (computer algorithms) take advantage of this behavior by submitting buy (sell) limit orders just above (below) prices ending with a zero or a five, thus easily taking price priority over large clusters.

The plan for the remainder of this chapter is as follows. Section 2 describes the data we use. Section 3 checks that the so-called stylized facts are valid for the electronic platform under scrutiny. Section 4 reports basic results on the order book, presents the orders reconstruction procedure and describes the characteristics of the resulting orders. Section 5 documents the existence of price clustering in EBS and provides evidence on its origin. Section 6 concludes.

2.2 Data

Most of spot interdealer trading occurs on two competing platforms: EBS Spot and Reuters D-3000. Due to network externalities, liquidity naturally gravitated to just one platform for each currency. EBS has long dominated interdealer trading for the EUR, JPY, and CHF, while Reuters dominates the GBP, AUD, CAD, and the Scandinavian currencies. In this paper, we study two major currency pairs for which EBS is the leader: EUR/USD and USD/JPY. The following periods of historical data were bought from EBS: from August 1, 2010 to July 31, 2011 and from January 1, 2012 to March 31, 2012. The dataset contains a *Quote Record* and a *Deal Record* on a $0.1s$ time-slice basis. The Quote Record is a snapshot of the ten best levels of the book at the end of a time-slice (if a price or a volume in the book changed within the time-slice). The Deal Record lists the highest buying deal price and the lowest selling deal price (with the dealt volumes) during the time-slice. For the second time-period, we also know the total signed volume of trades in a time-slice. This is so far the best available data from EBS in terms of frequency (almost tick by tick) and depth (10 levels). In March 2011, EBS decided to set in a tick *decimalization*. The EUR/USD tick size changed from 10^{-4} to 10^{-5} and the USD/JPY one from 10^{-2} to 10^{-3}. In FX terminology, the market went from *pip-pricing* to *decimal pricing*.

2.3 Stylized facts

Statistical regularities in the price series have been noticed in financial data in a wide range of different markets. In this section, we examine if the EBS market follows these facts.

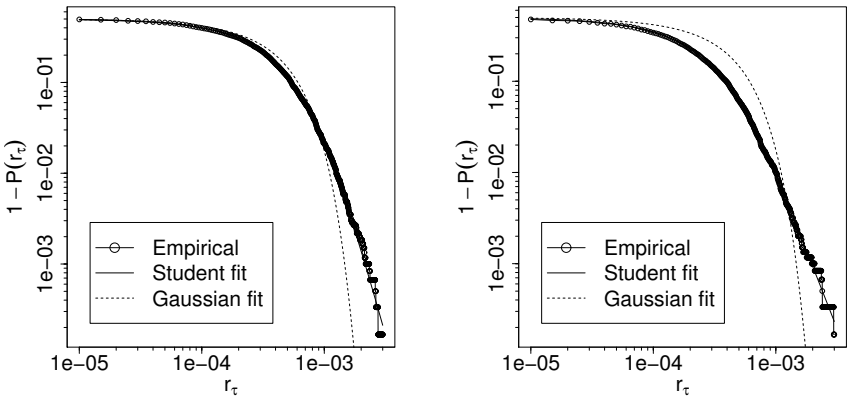

Figure 2.1: (Left) Empirical cumulative distribution of the EUR/USD 5-minute log-returns between January and March 2012. (Right) Same distribution for USD/JPY. We observe the same shape for negative returns.

2.3.1 Returns distribution

2.3.1.1 Fat-tails

Let p_t be the price of a financial asset at time t. We define its log-return over a period of time τ to be:

$$r_\tau(t) = \log(p(t+\tau)) - \log(p(t)) \tag{2.1}$$

Perhaps the most important stylized fact, first observed by Mandelbrot [79], is the fact that the empirical distributions of financial returns and log-returns are fat-tailed. There is however no consensus on the exact form of the distribution, which varies with the timescale. Figure 2.1 shows the empirical cumulative distribution of the EUR/USD and USD/JPY positive log-returns. Log-returns are useful for FX rates because they can be inverted easily ($r_\tau(USD/JPY) = -r_\tau(JPY/USD)$) and one can then directly compare different pairs and different tick size periods. We compute the log-returns by sampling (every 5 mins) our high-frequency dataset from 8 a.m. to 6 p.m. from 1 January 2012 to 31 March 2012 (except weekends and holidays). The distributions are clearly not-gaussian and fat-tailed as excepted. A Student fit is quite satisfactory. The tail exponent (obtained through maximum likelihood during the Student fit) is roughly equal to 5 for EUR/USD and to 3.5 for USD/JPY. The exponents are fairly stable through time.

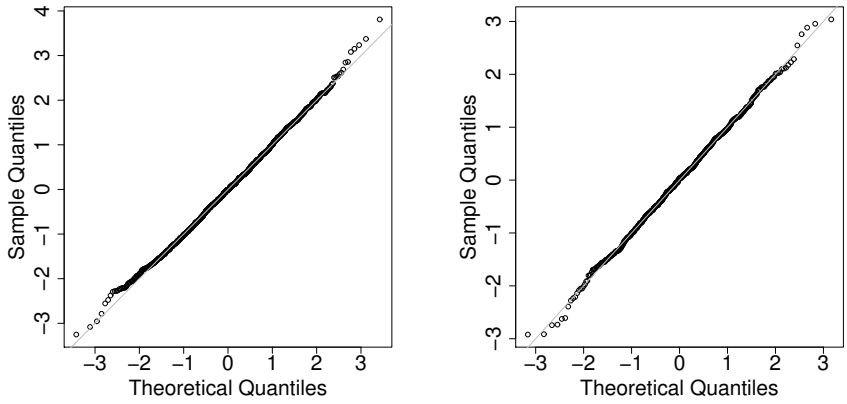

Figure 2.2: QQplots: Log-returns (sampled every 1000 trades) quantiles against normal quantiles. (Left) EUR/USD case. (Right) USD/JPY case.

2.3.1.2 Aggregational normality

It has been observed by Gopikrishnan et al. [49] that as τ increases, the distribution of returns becomes better approximated by a normal distribution. This *aggregational normality* is easier to see using trade time [59]. Once again, we compute the log-returns distribution but this time we sample our data every 1000 trades. Jarque-Bera and Shapiro normality tests were not able to reject the normal hypothesis. Figure 2.2 shows QQ-plots for visual confirmation.

2.3.2 Auto-correlation of returns

Another widely known empirical fact is that there is no evidence of a linear correlation between successive returns (except on very short timescales, see Ref. [26]). In figure 2.3, we plot the autocorrelation of log-returns. The autocorrelation function is not statistically different from zero.

2.3.3 Volatility

In this section, we present some major empirical findings about volatility.

2.3.3.1 Volatility clustering

Time series of absolute (or square) mid-price returns has been found to display positive autocorrelation that is slowly decaying. This phenomenom is called

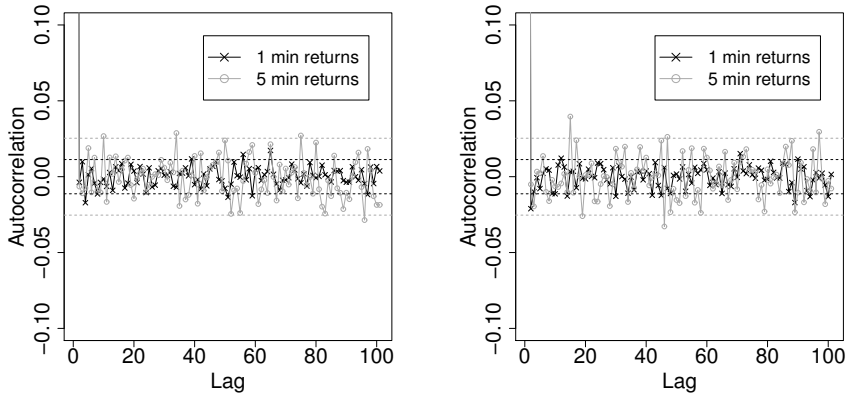

Figure 2.3: (Left) Autocorrelation function of the EUR/USD returns between January and March 2012. (Right) ACF of USD/JPY returns.

volatility clustering . Figure 2.4 shows the autocorrelation function of absolute returns. The function is statistically different from zero for a significant time horizon, demonstrating the existence of volatility clustering in our data. The effect is stronger for the USD/JPY case.

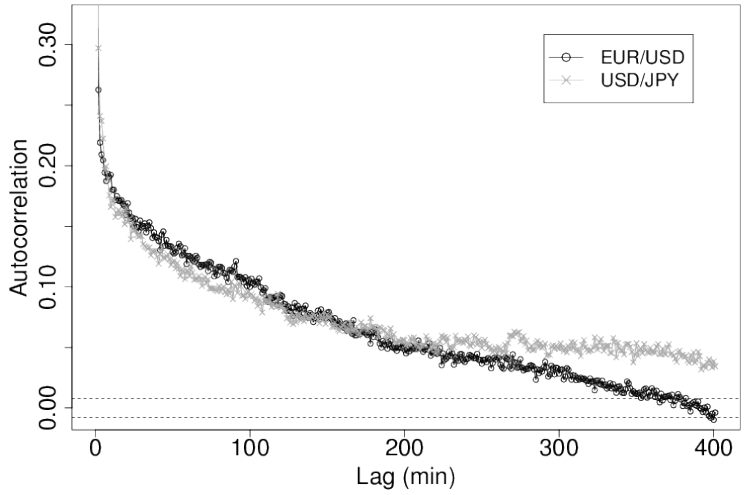

Figure 2.4: Autocorrelation function of 1-min absolute returns between January and February 2012.

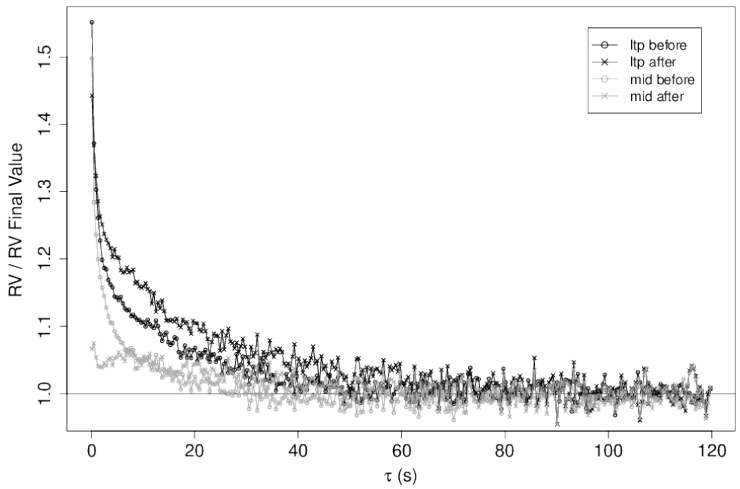

Figure 2.5: EUR/USD signature plot. Similar results for USD /JPY. Final value is reached faster before decimalization. The sampling period before decimalization goes from September to November 2010 and after decimalization from January to March 2012.

2.3.3.2 Signature plot

If returns were i.i.d., the signature plot (realized volatility as a function of τ) would be a flat line, but Andersen et al. [3] showed that financial returns realized volatility grows strongly as τ goes to zero due to microstructure noise. Figure 2.5 displays the EUR/USD signature plot with two different prices: last traded price (ltp) and mid-quotes (mid). We estimate the realized volatility using the two most active hours each day (14h-16h) and averaging over 65 trading days. Last traded price realized volatility is stronger than the mid-quotes one for small τ because of the bid-ask bounce effect. We notice two shape changes due to the tick size decimalization. First, the final value is reached faster before decimalization. Second, at high frequency before decimalization, mid-quote realized volatility explodes. This signals the passage from a large tick asset to a small tick one. These two empirical facts are contradictory regarding the choice of an optimal tick size.

2.3.4 Correlation

The two currency pairs are correlated but the correlation vanishes at very high frequency (Epps effect). Figure 2.6 is a plot of the correlation between EUR/USD

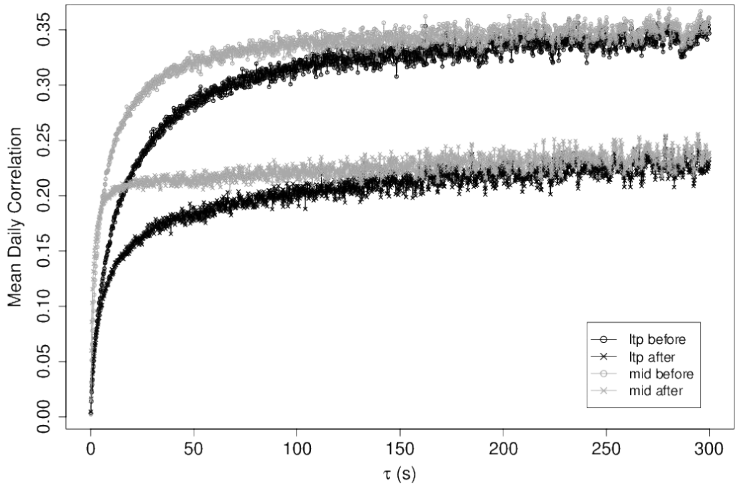

Figure 2.6: Correlation between EUR/USD and JPY/USD as a function of τ. Clear Epps effect. The sampling period before decimalization goes from September to November 2010 and after decimalization from January to March 2012.

and JPY/USD as a function of τ. After the tick change, the correlation is lower and it takes longer to reach the final state.

2.3.5 Intraday seasonality

EBS market operates 24/5 but the activity is of course not constant throughout the day. Figure 2.7 plots the average turnover and average spread in a 30-minute interval. We can clearly see a night and day pattern: from 8 a.m to 6 p.m. the volatility and turnover are much higher and the spread is lower than in the rest of the time; it corresponds to London working hours. The afternoon session shows a greater turnover (and volatility) than the morning session because it overlaps with New-York trading. Such global intra-day variation of market activity is a well-known fact on equities (see [26, 40] and references therein) and it had been studied on the EBS market by Ito and Hashimoto [62].

2.4 The limit order book

Most of today's financial markets use a limit order book mechanism to facilitate trade. Thanks to the computerization of markets, researchers can access extensive

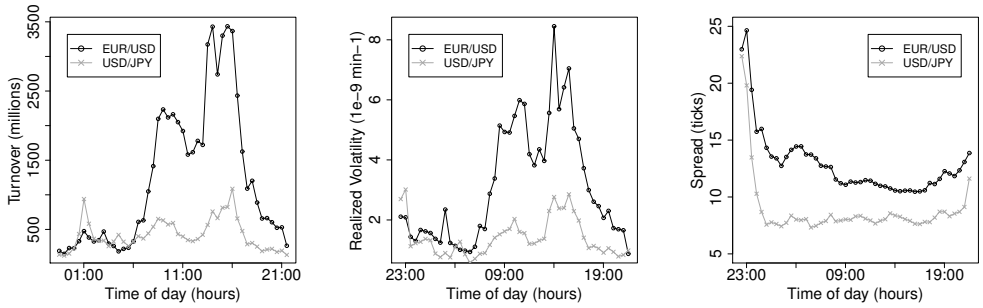

Figure 2.7: Intraday seasonalities. (Left) Average turnover in a 30-minutes interval. (Middle) Average realized volatility in a 30-minutes interval. (Right) Average spread in a 30-minutes interval.

data on order books allowing them to put the market under the microscope. The price dynamics emerges as a complex interplay between the order book and the order flow. Starting with Biais et al. [15], researchers from different fields are getting new insights to understand this complexity [82, 18, 113]. A recent review can be found in Ref. [26]. In this section, we use the EBS data to revisit the basic order book results. Our motivation is two-fold. First, as we mentioned in the introduction, few studies deal with the FX order book. Second, a tick decimalization occurred in March 2011, which is part of our dataset; therefore, we can study the reaction of the market to this change. Intraday seasonalities are remarkable in the EBS platform (see Sec. 2.3.5), therefore in the following, we will use London opening hours only (8 a.m. - 6 p.m.).

2.4.1 Average shape of the order book

A usual way to represent the shape of the book is to compute the (physical) time-averaged volume in the order book as a function of the distance from the current bid (or ask). We use one-month of data sampled every second from 8 a.m. to 6 p.m. In figure 2.8 we plot the shape before and after the decimalization. As already highlighted by other studies (see, e.g., Ref. [18] for the equity market) the maximum is not located at the best quote. Before the decimalization, we can see the well-known hump-shaped curve with a maximum at two pips. The volume after ten ticks seems very small because we only have access to the ten first levels. After decimalization, the shape is unusual. The volume decreases after the best quotes and then increases to reach a maximum at 10 (which corresponds to one pip). For EUR/USD we can also see small peaks at 5,15 and 20 ticks. We

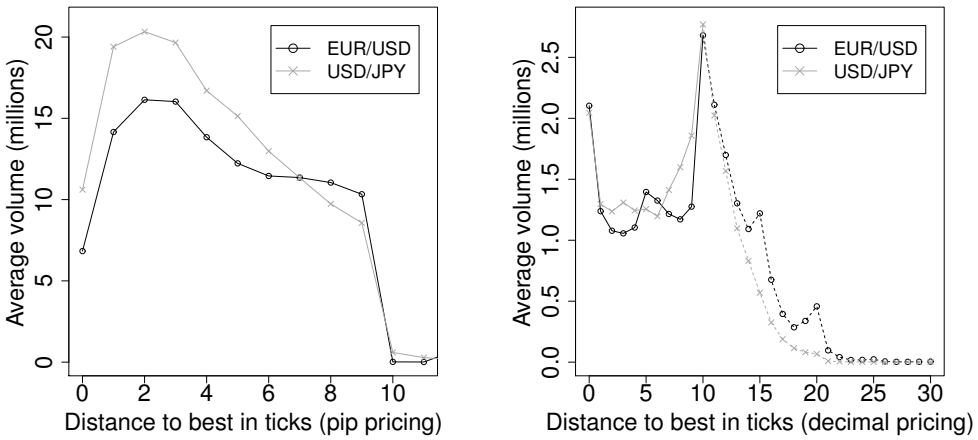

Figure 2.8: Average shape of the book ask-side. Similar shape for the bid side. (Left) Before decimalization: February 2011. (Right) After decimalization: March 2012. The shape is exact up to 10 ticks.

investigate these peculiarities in section 2.5. The results are similar for the bid side and for others months in the dataset.

Another interesting quantity related to the shape of the book is the average gap, i.e. the price distance between two levels. We plot this quantity in figure 2.9 after decimalization. Before decimalization, the gap is almost always equal to one-tick. After decimalization, it decreases with the level. The results are similar for the ask side and do not change in time.

2.4.2 Spread

One of the most important quantity for traders is the difference between best ask and best bid, called the *spread*, because it measures the cost of making a transaction immediately through a market order. Before decimalization, the spread was equal to one tick 65% of the time and to two ticks otherwise. We want to know the extent of the impact of the decimalization on the spread. For this purpose, we compute the spread distribution using one-month data sampled every second from 8 a.m. to 6 p.m. The results are presented in figure 2.9. The spread can now take many values around the previous typical spread: 10. For the EUR/USD the distribution presents a bimodality. The first mode is at 9 and the second is at 13. The second mode may change slightly depending on the considered time period, but we always have the first mode at 9 and the second one greater than 10. The

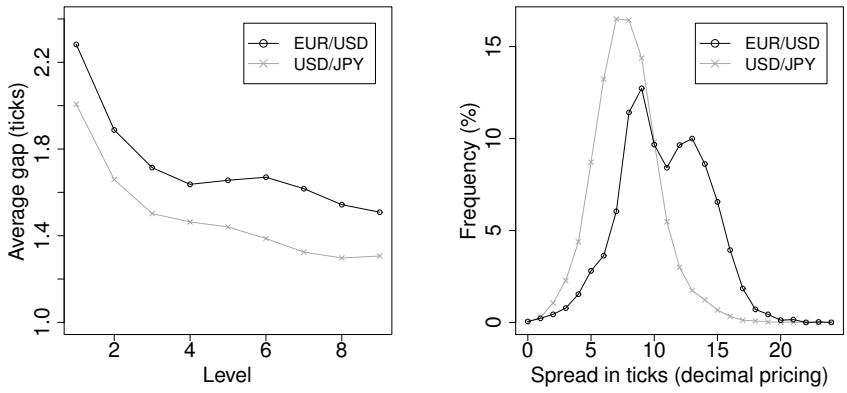

Figure 2.9: (Left) Average gap in the book bid-side. Similar results for the ask side. (Right) Spread distribution. Data from March 2012.

results are similar for different sampling frequencies. The USD/JPY distribution seems more "natural" and the USD/JPY spread is smaller than the EUR/USD one. We explain the bimodality in section 2.5.5.

The two previous subsections illustrate the change from a large tick asset to a small tick asset. The spread and the gaps switched from one tick to a few ticks and the volume at each level was reduced. In other terms, a very dense book became sparse (in ticks, not in absolute prices).

2.4.3 Order reconstruction

EBS does not publish data on the submission of limit orders and cancellations. Nevertheless, we can infer order submissions from deal and quote messages in our high-frequency data. Perfect reconstruction is impossible: the 10-levels limitation and the $0.1s$ time-slicing imply loss of information. Nevertheless, in most cases nothing happens within a time-slice. For example, between 2 p.m. and 4 p.m. (two most active hours of the day) nothing changes in the book in half of the time-slices. Moreover market participants also face this 0.1 s time-slicing so they cannot act at a much higher frequency. We therefore capture most of the order book changes. To build the order flow, we first split the data according to their side (bid or ask) and then we go sequentially through the quotes and compare two subsequent snapshots. When there is no transaction between two subsequent snapshots, all the changes between the snapshots are easily explained by limit orders and cancellations. When there is a transaction within a time-slice, we face

	Limit orders	Cancellations	Trades
EUR/USD	220	200	25
USD/JPY	75	70	7

Table 2.1: Average number of events per day (in thousands).

two cases. Case 1 (around 75 per cent of the cases): the total traded volume is equal to the reported trade[3] volume. In this case, we know that there is a unique trade in that time-slice (for the considered side), then we match the deal volume with the corresponding volume decrease between the two corresponding subsequent snapshots. The rest of the liquidity changes are explained by limit orders and cancellation. Case 2: the total traded volume is greater than the reported trade volume. We proceed like the previous case for the reported trade and we randomly distribute the remaining dealt volume among the available prices (from best price to the reported trade price). Again the remaining liquidity changes are attributed to limit orders and cancellations. Some orders of magnitude obtained with this procedure are given in Table 2.1 along with the number of deals for comparison.

2.4.4 Order characteristics

The total traded volume is only available for 2012 data so we cannot look into the order characteristics before the decimalization.

2.4.4.1 Volume

In the EBS market, order size must be a multiple of 1 million (of the base currency). Figure 2.10 shows the distribution of EUR/USD order volumes (data from March 2012). The majority (around 80%) of the order sizes is at the minimal value (1 million euros). We observe a strong representation of limit orders and cancellations around 5, 10, 15, 20, 25, etc. Although weaker, the effect is also present for deals volumes. A similar round numbers preference has been found for trade sizes in the Chinese stock market [86]. These peaks are explained in section 2.5.4. The results are stable through time and are similar for USD/JPY.

[3]By reported trade, we mean the lowest selling deal or highest buying deal depending on the considered side, see section 2.2

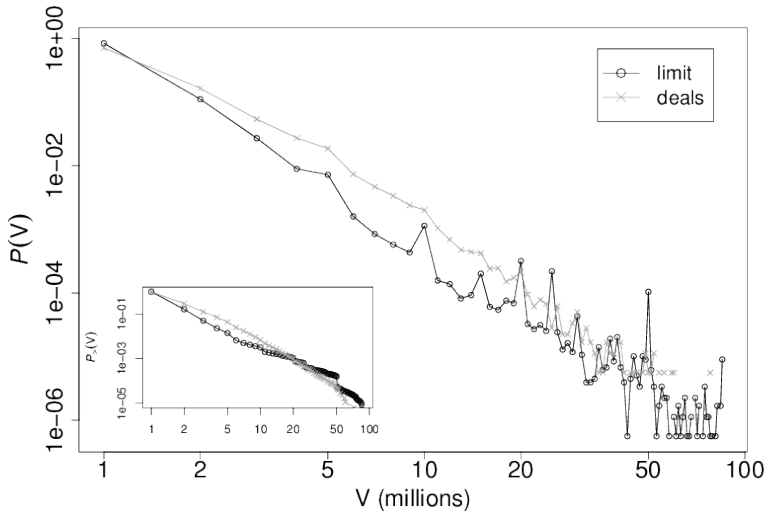

Figure 2.10: EUR/USD orders volume distribution in loglog scale. Similar results for USD/JPY. Sample period: march 2012. Inset: Cumulative distribution.

2.4.4.2 Placement

Let δ be the distance between the current price and an incoming limit order price. More precisely: $\delta = b_0(t-) - b(t)$ (resp. $a(t) - a_0(t-)$) if a bid (resp. ask) order arrives at price $b(t)$ (resp. $a(t)$), where $b_0(t-)$ (resp.$a_0(t-)$) is the best bid (resp. ask) before the arrival of this order. A classic interesting question concerns the distribution of δ. Results for EBS are plotted in figure 2.11.

These graphs being computed with incomplete data (ten best limits), we do not observe a placement as broad as in Ref. [18]. The empirical distribution is asymmetric: the left side is less broad than the right side. Since the left side represents limit orders submitted *inside* the spread, this is expected. The distribution has a maximum located at the current best price ($\delta = 0$) and a high value at -1, which is the smallest price improvement. In the EUR/USD case, we notice two clear peaks at 5 and 10, which correspond to a half-pip and a pip distances. The origin of these peaks is discussed in section 2.5.

2.4.4.3 Arrival times

It is now clearly established that the Poisson hypothesis for the arrival times of orders is not empirically verified (see [26] and references therein). The data resolution prevents us to study directly the inter-arrival times but we can compute

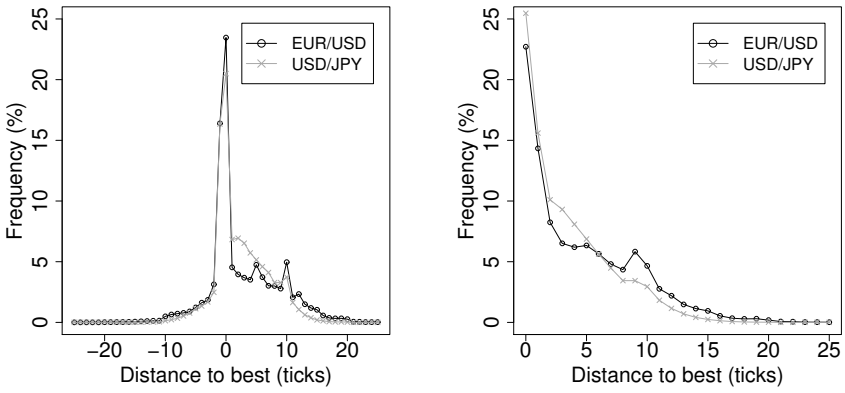

Figure 2.11: Placement of orders using the same best quote reference for February 2012. (Left) Limit orders. (Right) Cancellations. Ask side and bid side are similar.

the number of events in a 10 s window[4]. Let us consider 6 types of events: limit orders, market orders and cancellations on each side of the book and investigate clustering and inter-dependence phenomena among them. We restrict ourselves to linear dependencies with the autocorrelations and cross-correlations for the time-series of the number of events. Figure 2.12 (left) shows that the autocorrelation is statistically different from 0 at several lags. The cumulative sum of the auto-correlation coefficients, which saturates for large lags, shows that the generating process does not have long-memory. Table 2.2 is the correlation matrix for the six time-series. Significant correlations (as high as 0.84) are present, demonstrating the inter-dependence between order arrival processes. The independent Poisson processes hypothesis is clearly rejected. As suggested by recent studies, Hawkes processes are better candidates for orders time arrival modeling.

2.4.4.4 Signs memory

Lillo and Farmer [73] demonstrated that the signs of orders in the London Stock Exchange obey a long-memory process. In a similar fashion, we compute the mean sign ($+1$ for buy orders, -1 for sell orders) in a 10 s window for limit orders, cancellations and market orders and look at the autocorrelation function of these time-series. The results are plotted in figure 2.12 (right). Contrary to the equity market, the market order autocorrelation function is rapidly decaying and can be

[4]We chose the window length as a trade-off between the number of windows and the number of empty windows .

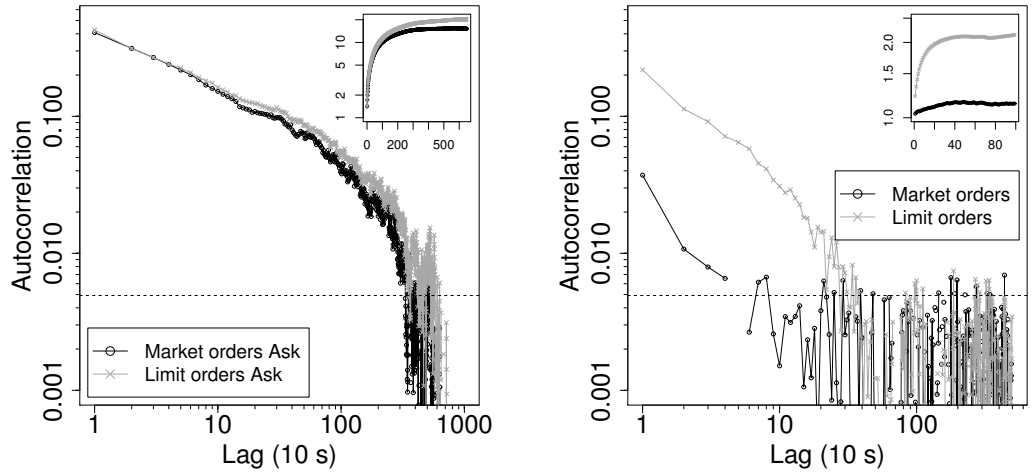

Figure 2.12: (Left) Autocorrelation function of the number of orders in a 10 s window. (Right) Autocorrelation function of the mean sign in a 10 s window. Cancellations and limit orders are similar. Insets: Cumulative sum in semi-logarithmic scale. Sampling period: January and February 2012.

	Limit Ask	Cancel Ask	Market Ask	Limit Bid	Cancel Bid	Market Bid
Limit Ask	1.000	0.837	0.561	0.563	0.750	0.716
Cancel Ask	0.837	1.000	0.602	0.751	0.716	0.571
Market Ask	0.561	0.602	1.000	0.726	0.585	0.644
Limit Bid	0.563	0.751	0.726	1.000	0.841	0.566
Cancel Bid	0.750	0.716	0.585	0.841	1.000	0.600
Market Bid	0.716	0.571	0.644	0.566	0.600	1.000

Table 2.2: Correlation matrix of the number of events time-series for EUR/USD. Similar values for USD/JPY.

considered null after about 2 minutes. This result was communicated to us by Curato et al. [36]. The autocorrelation function for limit orders (and for cancellations) displays a slower decay and becomes statistically zero after approximately 5 minutes.

2.5 Price clustering

Using previously gathered empirical facts, especially the average shape of the book and the limit order placement in section 2.4, we can anticipate a strong price clustering due to the tick decimalization. Price clustering is the tendency for prices to center around certain values. Many empirical studies have revealed that investors do not fully use the price resolution allowed by the tick size. The first statistical investigation on this phenomenon was the work by Osborne [95] followed by Niederhoffer [88, 89]. Since then, the focus has been mainly on equity markets [34, 1, 25, 60, 93], the relevant foreign exchange literature is given in the introduction. There are a number of proposed explanations regarding this clustering property. The price resolution hypothesis [9] argues that the degree of clustering varies inversely with the information about the underlying value of the asset. If the value is well known, traders will use a finer price grid. The negotiation hypothesis [54] posits that traders coordinate to restrict themselves to a smaller set of prices in order to reduce negotiations costs. The attraction hypothesis [48] states that investors have a natural attraction towards round numbers. The collusion hypothesis [31, 32] asserts that dealers avoid certain prices (odd-eights in the NASDAQ case) to maintain artificially wide spreads. In the EBS market, we show that the clustering is due to specific patterns in limit order placement. For liquid currency pairs, if the tick size is appropriate, no clustering should occur. As we are interested in clustering due to the decimalization, we are going to focus on the price last digit[5]. A spurious bid/ask asymmetry may appear if one looks at the last digit directly on both sides. For example, let us suppose that the best quotes are on prices whose last digit is 0, which we define as *integer prices*. In this situation, if a limit order improves the best quote by one tick at the bid, its last digit is 1, whereas if it is posted at the ask, its last digit is 9. In our view, it is the same situation and should give the same "decimal part". In the following, we will use the term last digit on both sides, but when it concerns a price on the ask side, it will actually designate the distance (number of ticks) to the smallest

[5]We have checked that there is no clustering on the other digits. Of course the first digit seems clustered but it is just an intrinsic value of the exchange rate. It is very unlikely that EUR/USD rises above 2 for example.

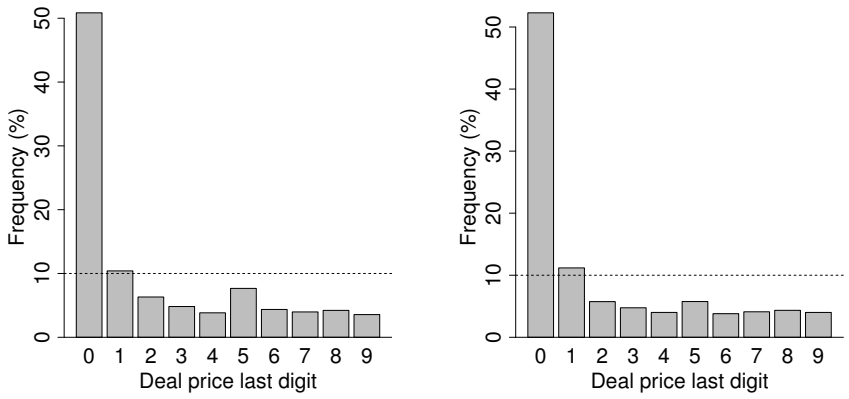

Figure 2.13: Trade prices (ask side) last digit distribution for March 2012. The dashed line represents the theoretical frequency under the uniform hypothesis. (Left) EUR/USD. (Right) USD/JPY. Around 50% of the trades occurs at integer prices. Same results for bid side.

integer price bigger than the price.

2.5.1 Trade price

During periods of active trading it is natural to assume that realized trades should not cluster at certain prices. Under this assumption, the distribution of the last digit of price should be uniform. Figure 2.13 plots the frequencies $\hat{p_i} = \frac{n_i}{n}$, where n_i is the number of trades with last digit $i \in \{0, 1, ..., 9\}$ and n is the total number of trades. The measurement uncertainties can be estimated by $1.96[\frac{\hat{p_i}(1-\hat{p_i})}{n}]^{\frac{1}{2}}$. They are all smaller than 0.0016, so we can visually assess that the frequencies are very far from uniformity. To confirm this statement, we performed χ^2 tests on different months and the uniform distribution hypothesis is always rejected at the 1% level[6]. Deal prices with a 0 as last digit (integer prices) represent about 50% of the trades. In other words, the old tick is somehow still present. We also checked that the distribution was uniform before the decimalization.

2.5.2 Limit orders

We now look at price clustering for limit orders. The last digit frequencies for limit order prices are plotted in figure 2.14. The frequency uncertainties order

[6]This is true for the two following subsections.

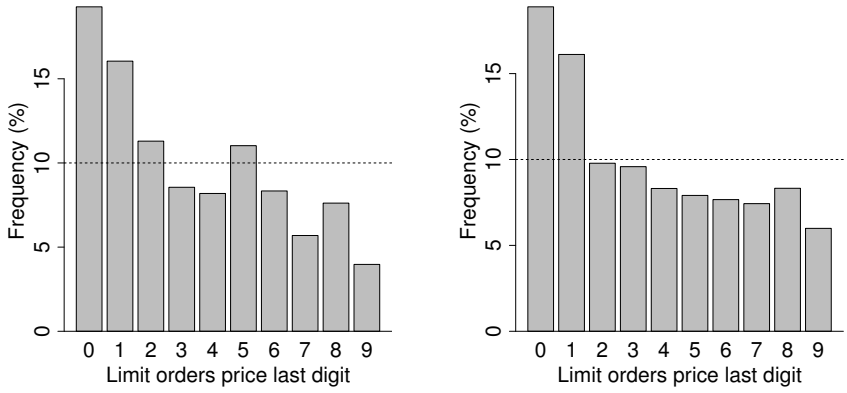

Figure 2.14: Distribution of last digits in limit order prices in March 2012. (Left) EUR/USD. (Right) USD/JPY. Around 20% of the orders are posted at integer prices.

of magnitude is 5×10^{-4}. Again, the fractions are not equally frequent. The clustering is less pronounced than in trade prices but it is still strong. Around 20% of the orders are posted at integer prices and a half-integer peak is present in the EUR/USD case. The next prominent last digit is 1, this is certainly related to the strategic behavior of some traders, who anticipate clustering tendencies and step-ahead round prices to obtain priority.

The limit order relative price distribution peaks or the average shape peaks (both at 5 and 10, see figure 2.11 and 2.8) might rise the question: is there also a round distance preference? The answer is negative, these peaks come from the price clustering. We verified this by computing the aforementioned distributions conditionally on the best quote last digit. The peaks positions change in a way that favors round prices and not round distances. The depth accumulation at 5 and 10 in the average shape comes directly from price clustering.

2.5.3 Best quote - Price Barriers

Limit orders have most of the time a size of 1 million (figure 2.10), then a clustering in terms of number of orders is also a clustering in terms of volume. Consequently, price clustering generates depth accumulation at round prices, affecting the best quote dynamics. Since more volume will be necessary to push the price through integers and halves, round best quotes may constitute "price barriers". The fact that there are more transactions at round prices may offset the depth accumula-

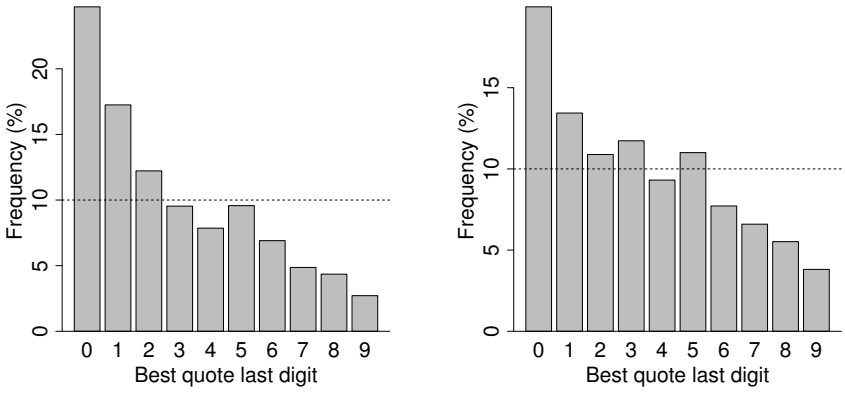

Figure 2.15: Best quote last digit distribution for January 2012 (1 s sampling). (Left) EUR/USD. (Right) USD/JPY. The best bid (or ask) spend most of the time on integers.

tion. We show that this is not the case by recording the best quote last digit every 0.1 s and plotting its distribution in figure 2.15. A congestion effect is present, since more time is spent on round prices. This is an important point since it shows that a change in a microstructure parameter can affect the price formation process.

2.5.4 Two types of traders

To shed light on the EBS price clustering we start by noting that market participants can be divided into two groups: manual traders and automatic traders (computers algorithms). Then, we analyze the reaction of each of these groups to the decimalization. According to King et al. [67] there is now a 50/50 split (in orders volume) between algorithmic traders and manual traders with a keypad. We know from discussions with traders working at major banks that manual traders do not care about price improvement to that last decimal point if they are trying to trade in large sizes. Besides, they have been used to pip-pricing for many years and are not eager to adapt to the new system. On the contrary, automatic traders adapted quickly to the new tick size (just an algorithm adjustment) and take advantage of manual trading conservatism. They anticipate clustering tendencies and easily obtain priority by posting limit orders just above the best

bid or just below the best ask[7] More precisely, some traders try to genuinely take priority in order to deal at better price (the proportion of deals with last digit 1 is above 10%) while others are simply practicing flash trading[8]. This explains the -1 strong value in the limit orders placement distribution (figure 2.11) and the prominence of 1 in the distribution of last digits in limit order prices (figure 2.14). We now present two arguments which corroborate the analysis above.

Firstly, some clustering effects are expected right after the decimalization, but they should decrease regularly as traders get used to the new tick. However, in the case of EBS the clustering is strong and stable. Therefore, part of the traders took into account the new situation whereas others did not and will not. Secondly, it is enlightening to look at the order volume depending on their price last digit. Figure 2.16 plots in log-log scale the distribution of limit order volume in which the data were previously split into two samples: orders with integer price and orders with non-integer price. For integer orders the volume is broadly distributed. A discrete power-law maximum likelihood estimator returns exponents 2.6 for EUR/USD and 2.7 for USD/JPY (smaller values were reported for NASDAQ stocks, see Ref. [82]). Moreover, we observe peaks on big round volumes (5,10,15,20... millions) which is a trace of manual trading. In banks, large volume deals with customers are usually left to human dealers, therefore they accumulate large positions and they need to submit large orders to EBS to reduce their exposure quickly. On the other hand, for decimal orders the distribution is exponentially decreasing, typical of algorithmic trading. Indeed, automated market making systems are designed to avoid the accumulation of a large inventory and even if they have to liquidate a large position, they split it into small orders to limit their market impact.

2.5.5 EUR/USD post-decimalization spread and clustering

The EUR/USD spread bimodality (section 2.4.2) is intriguing and deserves further attention. In order to qualitatively understand the shape of the spread distribution, it is important to notice that the integer preference (or, equivalently, manual traders behavior) leads to a "natural" spread value of 10, when the best bid and the best ask are on integers (it corresponds to the pre-decimalization minimal spread). Then, the "step-ahead" (section 2.5.4) strategy explains the first peak

[7]We can also add that of course traders are reluctant to post orders just behind clusters, it would imply losing price priority over the large depth available at integer prices.

[8]Flash traders send an order at the top of the book (new best price) followed by a cancellation to lure the other book observers. The goal is to make people believe that there is a bid (during the 250 ms minimum quote life) at a certain level to trigger sales at this level. Those who give themselves a small margin to be sure of making the sale by showing a lower price to bid will automatically be executed with lower bids left in the book by the flash trader.

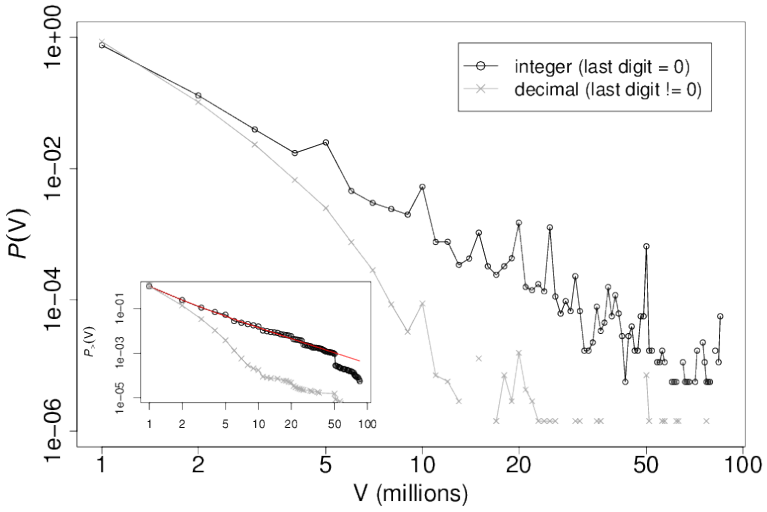

Figure 2.16: Distribution of EUR/USD limit orders volume. The distribution depends on the order price last digit. Power-law for integer orders and exponential for decimal orders. The power-law exponents are stable through time: 2.6 for EUR/USD and 2.7 for USD/JPY. Inset: Cumulative distribution with power-law fit (discrete maximum likelihood estimator).

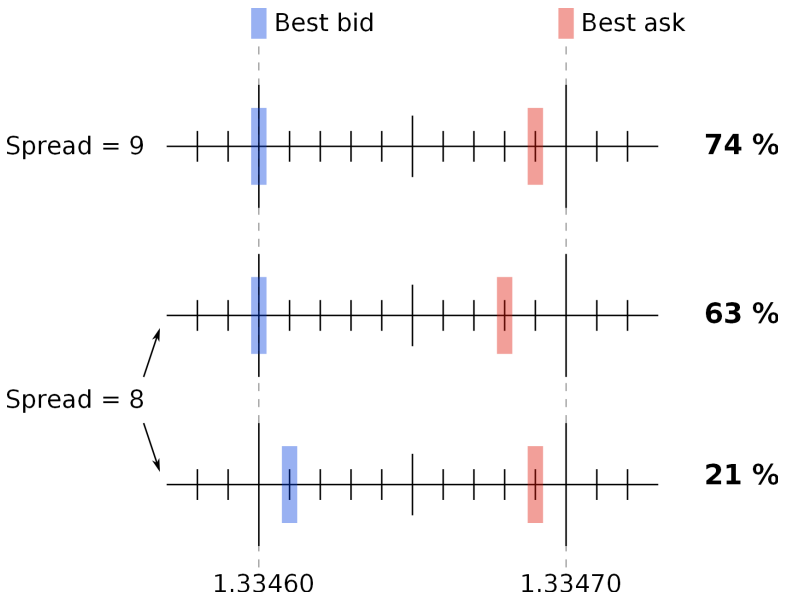

Figure 2.17: Small spreads typology. Price clustering leads to a drastic limitation of spread configurations in terms of bid and ask last digit. We do not draw symmetric configurations (ask on integer instead of bid) but we count them. If the configurations were equally likely one would expect 20% per configuration for 9 (odd spread : 5 possible configurations) and 16.67% per configuration for 8 (even spread : 6 possible configurations), instead of the observed higher percentage.

at 9 and the large value at 8. To confirm this hypothesis, we look at the quotes when the spread is 8 or 9. In theory, this can happen in a number of scenarios, depending on the last digits of bid and ask quotes. However, in our dataset we observe only one or two cases (see figure 2.17). For example, when the spread is equal to 9, one of the best quotes of the configuration is an integer price for 74% of the time.

The second hump arises from the combination of clustering at integer values for one side (bid or ask) with the smaller clustering at half-integer values for the other side. Once again, some traders take price priority by posting limit orders just above or just below price barriers, which favors spreads slightly below 15. This second peak in spread distribution is not as high as the first one because half-integer clustering is weaker than integer clustering and there is a natural tendency for the spread to revert to smaller values near 10[9].

[9]Due to the existence of market making effects (also known as order book resiliency).

2.6 Conclusion

In this paper we have reported the main outcome of a tick size change in the interdealer FX limit order book: an ubiquitous price clustering in both limit orders and actual trades. Interestingly, the uneven use of allowed price fractions can be explained by the interplay between manual traders, who stick to the old paradigm, and fast-adapting algorithmic traders. The most important consequence of this phenomenon is the appearance of support and resistance levels at integer and half-integer prices, which are not motivated by economic factors. We also note a strong distortion in the average shape of the book and in the spread distribution.

Our study led naturally to the question of what the optimal tick size for the EBS market is. On the one hand, the pip-pricing is praised by traditional dealing banks because it maintains a minimum level of profits and their traders are used to it. On the other hand, hedge-funds prefer finer grids because it reduces transaction costs, thus making it easier to use high-frequency strategies. Therefore, EBS market designers have to find a trade-off that satisfies both communities in order to maximize traded volume on their platform. From an empirical point of view, each tick size has advantages and drawbacks. When the old tick size was in place, the book exhibited high liquidity at each level and stable prices but the spread was almost always equal to one pip, suggesting that a tick size reduction was required. In contrast, the decimal pricing lowers the spread but generates a strong price clustering and allows the advent of high-frequency disruptive practices (e.g. flash trading). The USD/JPY situation is slightly better than the EUR/USD one: smaller spread and less pronounced price barriers. One possible explanation is that the USD/JPY relative tick size[10] is bigger than the EUR/USD one. Our findings suggest that the optimal tick size lies somewhere between pip and decimal and show that taking into account traders biases is essential to design efficient trading structures. Remarkably, in September 2012, during the writing of this paper, EBS decided to change again the tick size and to go for half-pip pricing. It would be relevant to see if it managed to reduce price clustering while maintaining a small spread. We leave this question for future studies.

[10] By relative tick size we mean $\delta = \frac{tick}{price}$. We find $\delta_{eurusd} \simeq 7 \times 10^{-6}$ and $\delta_{usdjpy} \simeq 1.2 \times 10^{-5}$, so $\frac{\delta_{usdjpy}}{\delta_{eurusd}} \simeq 1.7$.

Chapter 3

The limits of statistical significance of Hawkes processes fitted to FX data

3.1 Introduction

Hawkes processes are a natural extension of Poisson processes in which self-excitation causes event clustering [56, 57]. Originally applied to the modeling of earthquake occurrences [91, 92], they have proven to be useful in many fields (e.g. neuroscience, criminology and social networks modeling [30, 99, 85, 35, 111]). This is because of their tractability and the ever-increasing number of estimation methods [80, 101, 72, 6, 38, 4]. Since many types of financial market events such as mid-quote changes, extreme return occurrences or order submissions are clustered in time, Hawkes processes have become a standard tool in finance too.

In the context of market microstructure, Hawkes processes were first introduced by Bowsher [20], who simultaneously analyzed trades time and mid-quotes changes with a multivariate framework. Two others pioneer approaches are the ones by Bauwens and Hautsch [11] and Hewlett [58] who focused on the durations between transactions. Subsequently, Large [71] supplemented transaction data with limit orders and cancellations data in a ten-variate Hawkes process in order to measure the resilience of an London Stock Exchange order book. Bacry et al. [7] have recently modeled the mid-price change as the difference between two Hawkes processes and showed that the resulting price exhibits microstructure noise and the Epps effect. Jaisson and Rosenbaum [63] established that under a suitable rescaling a nearly unstable Hawkes process converges to a Heston model. Bacry and Muzy [5] used an enhanced version of the model to account for market

impact. Finally, Jedidi and Abergel [64] modeled the full order book with a multivariate Hawkes setup and proved that the resulting price diffuses at large time scales. Remarkably, Hawkes processes are also applied to other financial topics such as VaR estimation [29, 28], trade-through modeling [105], portfolio credit risk [42], or financial contagion across regions [2] and across assets [17].

It is widely accepted among researchers that only a small fraction of price movements is directly explained by external news releases (e.g. Cutler et al. [37], Joulin et al. [66]). Thus, the price dynamics is mostly driven by internal feedback mechanisms, which corresponds to what Soros calls "market reflexivity" [104]. In the framework of Hawkes processes, endogeneity comes from self-excitation while the baseline activity rate is deemed exogenous (see Sec. 3.2 for a mathematical definition). In other words, these processes provide a straightforward way to measure the importance of endogeneity, for example in the E-mini S&P futures [44, 53]. Filimonov and Sornette [44] argued that the level of endogeneity has increased steadily in the last decade due to the advent of high-frequency and algorithmic trading. Hardiman et al. [53] showed that it is only the *short-term* endogeneity (linked to increases of computer power and speed, and, indeed, HFT) that has increased over the years, while the endogeneity factor has been very stable and close to 1, the special value at which the process becomes totally self-referential and unstable. Fitting Hawkes processes to financial data requires some care: one should not use a single exponential self-excitation kernel [53], while many other biases may affect fits with long-tailed kernels on long time periods [45].

Nobody claims that Hawkes process are the exact description of the whole dynamics of financial markets. However, testing the significance of the fits is not a current priority in the literature. Given the fact that the fits are usually visually satisfactory, it seems obvious that statistical significance may be obtained in some cases. Here, we wish to assess the extent (and the limits) of the explanatory power of Hawkes processes with several possibly types of parametric kernels, according to three statistical tests. One of the difficulties in obtaining significant fits come from jumps in trading activity such as those occurring when markets open and close. This is why we work on data from FX markets which have the advantage of operating continuously for longer periods. There may still be discontinuities, either implicit (e.g. fixing time) or explicit (e.g. week-end closures) in our FX data, but at least one day of FX data spans many more hours than one day of equity market data and is thus more suitable to our aim. Hence, *a minori*, one may extrapolate most of our failures to fit correct Hawkes processes to other types of data with more significant activity discontinuities.

The two other papers on FX data and Hawkes processes have a different focus

than ours: Hewlett [58] deals with the relatively illiquid EUR/PLN currency pair and uses a single-exponential kernel. Rambaldi et al. [100] also use EBS data (with the same time resolution as ours) and studies the dynamics of best quotes around important news. Because our data set consists of order book snapshots every 0.1 s (see Sec. 3.3 for more details), we can trace most trades but not mid price changes. This is why we fit a univariate Hawkes process to EUR/USD trade arrivals. The endogeneity parameter is then the average number of trades triggered by a single trade.

The structure of the chapter is as follows: we first define Hawkes processes, the fitting method, the parametric kernels and the statistical tests that we will use. We first show that Hawkes processes excel at fitting one hour of FX data, are fairly good for a single day, and fail when used for two consecutive days. We also present a non-parametric confirmation of the aforementioned results.

3.2 Hawkes process

3.2.1 Definition

An univariate Hawkes process is a linear self-exciting point process with an intensity given by

$$\begin{aligned}\lambda_t &= \mu(t) + \int_0^t \phi(t-s)dN_s \\ &= \mu(t) + \sum_{t_i<t} \phi(t-t_i),\end{aligned} \tag{3.1}$$

where $\mu(t)$ is a (deterministic) baseline intensity describing the arrival of exogenous events and the second term is a weighted sum over past events. The kernel $\phi(t-t_i)$ describes the impact on the current intensity of a previous event that took place at time t_i.

A Hawkes process can be mapped to a branching process, where exogenous "mother" events occurring with intensity μ_t can trigger one or more "child" events. In turn, each of these children, can trigger multiple child events (or "grand-child" respectively to the original event), and so on. The quantity $n \equiv \int_0^\infty \phi(s)ds$ controls the size of the endogenously generated families. Indeed, n is the *branching ratio* of the process, which is defined as the average number of children for any event. Therefore, n quantifies market reflexivity in an elegant way. Three regimes exist depending on the branching ratio value:

- a sub-critical regime ($n < 1$) where families dies out almost surely,
- the critical regime ($n = 1$), where one family lives indefinitely without exploding. In the language of Hawkes process, this requires $\mu = 0$ to be properly defined and it is equivalent to Hawkes process without ancestors studied by Brémaud and Massoulié [22],
- the explosive regime ($n > 1$), where a single event triggers an infinite family with a strictly positive probability.

Evaluating n gives a simple measure of the market "distance" to criticality. For $n \leq 1$, the process is stationary if μ_t is constant. In this case, the branching ratio is also equal to the average proportion of endogenously generated events among all events.

3.2.2 Kernel Parametrization

We compare the performance of the following kernels, each labeled by its own index.

- Sum of exponentials:

$$\phi_M(t) = \sum_{i=1}^{M} \alpha_i e^{-t/\tau_i},$$

  where M is the number of exponentials. The amplitudes α_i and timescales τ_i of the exponentials are the estimated parameters. The branching ratio is then given by: $n = \sum_{i=1}^{M} \alpha_i \tau_i = \sum_{i=1}^{M} n_i$.

- Approximations of power-laws have the advantage of needing a few parameters only. As a consequence, fitting them to data is much easier. Approximate power-law kernel is given by

$$\phi_M^{\mathrm{PL}}(t) = \frac{n}{Z} \sum_{i=0}^{M-1} a_i^{-(1+\epsilon)} e^{-\frac{t}{a_i}},$$

  where

$$a_i = \tau_0 m^i.$$

  M controls the range of the approximation and m its precision. Z is defined such that $\int_0^\infty \phi_{PL}(t)dt = n$. The parameters are the branching ratio n, the tail exponent ϵ and the smallest timescale τ_0.

- Approximate power-law with a short lags cut-off [53]:

$$\phi_M^{\text{HBB}}(t) = \frac{n}{Z}\left(\sum_{i=0}^{M-1} a_i^{-(1+\epsilon)} e^{-\frac{t}{a_i}} - S e^{-\frac{t}{a_{-1}}}\right),$$

  the definition is the same as ϕ_M^{PL} with the addition of a smooth exponential drop for lags shorter than τ_0. S is defined such that $\phi_M^{\text{HBB}}(0) = 0$.

- We propose a new type of kernels, made up of an approximate power-law ϕ_M^{PL} and one exponential with free parameters. This is to allow for a greater freedom in the structure of time scales. The kernel is then defined as

$$\phi_M^{\text{PLx}}(t) = \frac{n}{Z}\left(\sum_{i=0}^{M-1} a_i^{-(1+\epsilon)} e^{-\frac{t}{a_i}} + b e^{-\frac{t}{\tau}}\right),$$

  where the exponential term adds two parameters b and τ. The other variables have the same meaning as above.

When a kernel is a sum of exponentials, one can exploit a recursive relation for the log-likelihood calculation that reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ (see Ozaki [98]). It provides reasonable computation time on a single workstation since N is $\mathcal{O}(10^4)$. The first form is the most flexible and can approximate virtually any continuous function, at the cost of extra-parameters and more sloppiness [110]. The second and third ones aim to reproduce the long memory observed in many market but are less flexible; their effective support may span well beyond the fitting period. The last one tries to combine the best of both worlds.

Once a kernel form is specified, we use the L-BFGS-B algorithm [24] to estimate the parameters that maximize the log-likehood. For each fit we try different starting points to avoid local maxima.

Using multivariate Hawkes process to fit the arrival and the reciprocal influence of buy and sell trades systematically yields null cross-terms. Both buy and sell trades yield indistinguishable results; we therefore focus on buy trades.

3.2.3 Mathematical results

We will now present the mathematical results required for the following sections. We state them with the general kernel: $\phi_M(t) = \sum_{i=1}^{M} \alpha_i e^{-t/\tau_i} = \sum_{i=1}^{M} \alpha_i e^{-\beta_i t}$, but they are valid for all the proposed kernels in Sec. 3.2.2, since they all possess this structure up to trivial terms rewriting.

3.2.3.1 Maximum Likelihood Estimation

The standard way to calibrate a Hawkes process is to perform a maximum likelihood estimation. The log-likelihood of a simple point process with intensity λ_t and counting process N_t, is written:

$$\log \mathcal{L}\left((N_t)_{t\epsilon[0,T]}\right) = \int_0^T (1-\lambda_s)ds + \int_0^T \log \lambda_s dN_s, \tag{3.2}$$

where log designates the natural logarithm. Let $\{t_i\}_{i=1,\dots,n}$ be the point process's jump times and let us assume that $T = t_n$and $t_0 = 0$. In the case of a Hawkes model (eq. 3.1) with a kernel given by ϕ_M, Eq. 3.2 can be explicitly written as:

$$\begin{aligned}\log \mathcal{L}\left(\{t_i\}_{i=1,\dots,n}\right) &= t_n - \int_0^{t_n} \lambda_s ds + \sum_{i=1}^{n} \log \lambda_{t_i} \\ &= t_n - \int_0^{t_n} \lambda_s ds + \sum_{i=1}^{n} \log \left[\mu(t_i) + \sum_{j=1}^{M}\sum_{k=1}^{i-1} \alpha_j e^{-\beta_j(t_i - t_k)}\right]\end{aligned} \tag{3.3}$$

Following Ogata [90], we can simplify this expression with a recursive formula. We note that:

$$\begin{aligned}R_j(i) &\equiv \sum_{k=1}^{i-1} e^{-\beta_j(t_i - t_k)} \\ &= e^{-\beta_j(t_i - t_{i-1})} \sum_{k=1}^{i-1} e^{-\beta_j(t_{i-1} - t_k)} \\ &= e^{-\beta_j(t_i - t_{i-1})} \left(1 + \sum_{k=1}^{i-2} e^{-\beta_j(t_{i-1} - t_k)}\right) \\ &= e^{-\beta_j(t_i - t_{i-1})} \left(1 + R_j(i-1)\right).\end{aligned} \tag{3.4}$$

Then equation 3.3 can be written:

$$\log \mathcal{L}\left(\{t_i\}_{i=1,\dots,n}\right) = t_n - \int_0^{t_n} \lambda_s ds + \sum_{i=1}^{n} \log \left[\mu(t_i) + \sum_{j=1}^{M} \alpha_j R_j(i)\right], \tag{3.5}$$

where $\forall j \epsilon [1, M]$, the$\{R_j(i)\}_{i>1}$ are given by equation 3.4 and $R_j(i) = 0$. This recursive relation reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. The term $\int_0^{t_n} \lambda_s ds$ is obtained through direct computation:

$$\int_0^{t_n} \lambda_s ds = \int_0^{t_n} \mu(s) ds + \sum_{i=2}^{n} \int_{t_{i-1}}^{t_i} \sum_{t_k < s} \phi_M(s - t_k) ds. \tag{3.6}$$

Let us explicit the term in the summation:

$$\begin{aligned}
\int_{t_{i-1}}^{t_i} \sum_{t_k<s} \phi_M(s-t_k) ds &= \int_{t_{i-1}}^{t_i} \sum_{k=1}^{i-1} \phi_M(s-t_k) ds \\
&= \sum_{k=1}^{i-1} \int_{t_{i-1}}^{t_i} \phi_M(s-t_k) ds \\
&= \sum_{k=1}^{i-1} \int_{t_{i-1}}^{t_i} \sum_{j=1}^{M} \alpha_j e^{-\beta_j (s-t_k)} ds \\
&= \sum_{k=1}^{i-1} \sum_{j=1}^{M} \frac{\alpha_j}{\beta_j} \left[e^{-\beta_j (t_{i-1}-t_k)} - e^{-\beta_j (t_i - t_k)} \right].
\end{aligned} \tag{3.7}$$

Plugin 3.7 in equation 3.6 gives a telescopic sum and after simplification:

$$\begin{aligned}
\int_0^{t_n} \lambda_s ds &= \int_0^{t_n} \mu(s) ds + \sum_{j=1}^{M} \frac{\alpha_j}{\beta_j} \left[\sum_{i=2}^{n} 1 - \sum_{k=1}^{n-1} e^{-\beta_j (t_n - t_k)} \right] \\
&= \int_0^{t_n} \mu(s) ds + \sum_{i=1}^{n-1} \sum_{j=1}^{M} \frac{\alpha_j}{\beta_j} \left(1 - e^{-\beta_j (t_n - t_k)} \right).
\end{aligned} \tag{3.8}$$

Finally, combining equations 3.5 and 3.8 gives:

$$\begin{aligned}
\log \mathcal{L} \left(\{t_i\}_{i=1,\dots,n} \right) &= t_n - \int_0^{t_n} \mu(s) ds \\
&\quad + \sum_{i=1}^{n-1} \sum_{j=1}^{M} \frac{\alpha_j}{\beta_j} \left(1 - e^{-\beta_j (t_n - t_k)} \right) \\
&\quad + \sum_{i=1}^{n} \log \left[\mu(t_i) + \sum_{j=1}^{M} \alpha_j R_j(i) \right].
\end{aligned} \tag{3.9}$$

The term $\int_0^{t_n} \mu(s)ds$, depends on the chosen specification of $\mu(t)$, we will explicit it in the specific applications (hourly fits, daily fits). Expression 3.9 is then numerically maximized to obtain the maximum-likelihood estimators of the parameters.

3.2.3.2 Goodness of fits tests

After a parameter estimation, it is natural to assess the quality of the fits. The goal is to answer the question: is the Hawkes process a good statistical model for the data? Any test is built on the time-rescaling theorem. Let us state it:

Theorem 1. *Let N be a point process on $\mathbb{R}_+$ with a strictly positive intensity λ_s. Let t_τ be the stopping time defined by:*

$$\int_0^{t_\tau} \lambda_s ds = \tau. \tag{3.10}$$

Then the process $\tilde{N}(\tau) = N(t_\tau)$ is an homogeneous Poisson process with a constant intensity equal to one.

Proof. See Ref. [23]. □

From the time-rescaling theorem , we can build a goodness-of-fit procedure:

1. Compute the time-deformed series of durations[1] $\{\theta_i\}_{i=1,\dots,n}$, defined by $\theta_i = \int_{t_{i-1}}^{t_i} \hat{\lambda}_t dt$, where $\hat{\lambda}$ is the estimated intensity and $\{t_i\}$ are the empirical timestamps.

2. Statistically test that the θ_i's are (i) independent and (ii) exponentially distributed with rate equal to 1.

Step 1. To compute θ_i, let us first denote $A_j(i-1) \equiv \sum_{t_k \leq t_{i-1}} e^{-\beta_j(t_{i-1}-t_k)}$, with $A_j(0) = 0$, thus from equation 3.7, we get

$$\theta_i = \int_{t_{i-1}}^{t_i} \mu(s)ds + \sum_{j=1}^{M} \left(1 - e^{-\beta_j(t_i - t_{i-1})}\right) A_j(i-1),$$

[1] Also called: integrated intensities, compensators or residuals.

and again we use a recursive relation from Ogata [90] to speed up numerical computations:

$$\begin{aligned} A_j(i-1) &= 1 + e^{-\beta_j(t_{i-1}-t_{i-2})} \sum_{t_k \le t_{i-2}} e^{-\beta_j(t_{i-2}-t_k)} \\ &= 1 + e^{-\beta_j(t_{i-1}-t_{i-2})} A_j(i-2). \end{aligned}$$

Step 2. We start by inspecting the distribution of θ visually in QQ-plots, then we use more rigorous tools.

Property (i) can be tested by the Ljung-Box test, which examines the null hypothesis of absence of auto-correlation in a given time-series. We use here a slight modification of the original test statistic from Ljung and Box [76], defined as

$$Q = N(N+2) \sum_{k=2}^{h+1} \frac{\hat{\rho}_k^2}{n-k},$$

where N is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under the null, Q follows a χ^2 with h degrees of freedom. Note that we start the sum at $k = 2$ (instead of 1). This is because of the systematic small one-step anti-correlation introduced by the data cleaning procedure (Sec. 3.3.2). In other words, we wish to test the absence of auto-correlation at lags that are unaffected by this procedure.

Property (ii) is assessed by two tests

1. Kolmogorov-Smirnov test (KS henceforth), based on the maximal discrepancy between the empirical cumulative distribution and the exponential cumulative distribution. The asymptotic distribution under the null is the Kolmogorov distribution. It is known to be a very (even excessively) demanding test.

2. Engle and Russell [41] Excess Dispersion test (ED henceforth), which verifies the lack of excess dispersion in the residuals. The test statistic reads:

$$S = \sqrt{N} \frac{\hat{\sigma}^2 - 1}{\sqrt{8}},$$

where $\hat{\sigma}^2$ is the sample variance of θ which should be equal to 1. Under the null, S has a limiting normal distribution.

All these three tests check basic but essential properties of the θs.

3.2.3.3 Simulations

Simulations of Hawkes processes can be useful to test the fitting procedure or to assess the impact of various data treatments on a pure Hawkes process (See Sec. 3.3.2). We recall here (algorithm 3.1) the most common method to efficiently simulate an univariate Hawkes process on $[0, T]$, due to Ogata [90].

Algorithme 3.1 Ogata's thinning procedure.

---

1. **Initialization:** Set $\lambda^* \leftarrow \mu$, $n \leftarrow 1$.

2. **First event:** Generate $U \rightsquigarrow \mathcal{U}_{[0,1]}$, where $\mathcal{U}_{[0,1]}$is the uniform distribution on the interval $[0, 1]$, and set $s \leftarrow -\frac{1}{\lambda^*} \log U$.

   If $s \leq T$,

   Then $t_1 \leftarrow s$,

   Else go to step 4.

3. **General routine:** Set $n \leftarrow n + 1$.

   (a) *Update maximum intensity:* Set $\lambda^* \leftarrow \lambda(t_{n-1}) + \alpha$.

   (b) *New event:* Generate $U \rightsquigarrow \mathcal{U}_{[0,1]}$ and set $s \leftarrow -\frac{1}{\lambda^*} \log U$.
   If $s \geq T$,
   Then go to step 4.

   (c) *Rejection Test:* Generate $D \rightsquigarrow \mathcal{U}_{[0,1]}$.
   If $D \leq \frac{\lambda(s)}{\lambda^*}$,
   Then $t_n \leftarrow s$ and go through the general routine again,
   Else update $\lambda^* \leftarrow \lambda(s)$ and try a new time at step 3(b).

4. **Output:** The times (t_n) forms the simulated process.

3.2.4 Caveat

Let us now illustrate all the previous notions (simulation, estimation and quality of fit assessment) and highlight one source of difficulties with Hawkes process modeling. We generate a very simple process with a constant baseline intensity $\mu = 1.2$ and an exponential kernel $\phi_1(t) = \alpha e^{-\beta t}$, with $\alpha = 0.35$, $\beta = 0.5$ and $T = 10^4$. The branching ratio is then equal to 0.7. Fig. 3.1 shows a realization

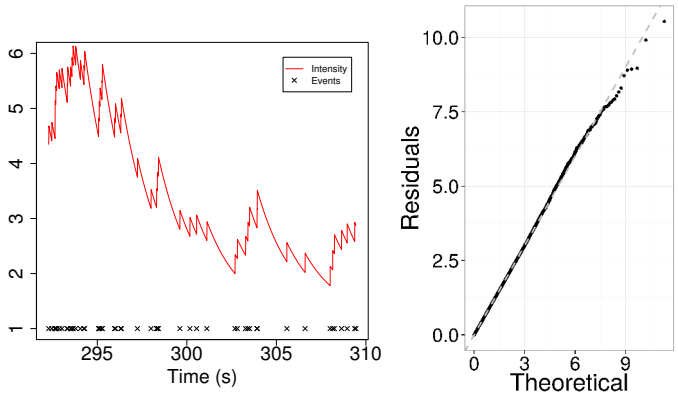

Figure 3.1: Left: Realization of a univariate exponential Hawkes process. Parameters: $\mu = 1.2$, $\alpha = 0.35$ and $\beta = 0.5$. Right: QQ-plot, residual v.s. exponential distribution.

over a 50 s window.

The fitting procedure quickly converges close to the true parameters: $\hat{\mu} = 1.25$, $\hat{\alpha} = 0.36$, $\hat{\beta} = 0.52$. The p-values of the tests presented in Sec.3.2.3.2 are $p_{KS} = 0.69$, $p_{ED} = 0.46$, and $p_{LB} = 0.82$, well above any standard threshold. To underline the importance of a good model specification, we will now concatenate two simulated Hawkes process with the previous kernel. The first one with $\mu_1 = 0.8$ and the second one $\mu_2 = 1.2$. We voluntarily fit it with a wrong model: μ constant over the whole period. The estimated parameters are the following: $\hat{\mu} = 0.69$, $\hat{\alpha} = 0.33$ and $\hat{\beta} = 0.39$. Interestingly, the procedure does not find a μ between μ_1 and μ_2 as we might conjecture but a value below both of them. The branching ratio is overestimated $\hat{n} = 0.86$. The true difficulty comes from the fact that it is almost impossible to detect the miss-specification from a statistical analysis of the residuals. Indeed, the QQ-plot is very satisfactory (Fig.3.1 right) and the p-values above standard thresholds ($p_{KS} = 0.46$, $p_{ED} = 0.19$, $p_{LB} = 0.62$), despite the high number of points: 100151. This should warn us about the bold use of a constant baseline intensity, especially with seasonal financial data.

3.3 Data

3.3.1 Description

We study EUR/USD inter-dealer trading from January 1, 2012 to March 31, 2012. The data comes from EBS, the leading electronic trading platform for this

currency pair. A message is recorded every 0.1 s. It contains the highest buying deal price and the lowest selling deal price with the dealt volumes, as well as the total signed volume of trades in the time-slice. Orders on EBS must have a volume multiple of 1 million of the base currency, which is therefore the natural volume unit. This is, to our knowledge, the best data available from EBS in terms of frequency (almost tick by tick) and, above all, has the invaluable advantage of containing information about traded volumes.

3.3.2 Treatment

The data must be filtered to improve the accuracy of fits. The coarse time resolution introduces a spurious discretization of the duration data, as illustrated in Fig. 3.2 (left plot). To overcome this issue, we added a time shift, uniformly distributed between 0 and 0.1, to trade occurrence times (Fig. 3.2, middle plot).

The number of transactions on one side during a time-slice can be determined from the total signed volumes in 92% of cases. Indeed, when the total signed traded volume (V_{total}) is equal to the reported trade volume (V_{report}), only one trade occurred and the only uncertainty is about the exact time of the event. However, when $V_{total} > V_{report}$, one knows that more than one trade occurred. If $V_{total} - V_{report} = 1$, exactly two trades occurred, one with volume V_{report} and one with volume 1; their respective event time are randomly uniformly drawn during the time slice. Finally, the case $V_{total} - V_{report} > 1$ (about 8% of the non-empty time-slices) is ambiguous because the extra volume may come from more than one trade and hence may be split in different ways. We tried different schemes: not adding any trade, adding one trade, adding a trade per extra million, adding a uniform random number of trades between 1 and $V_{total} - V_{report}$ and a self-consistent correction that uses the most probable partition according to the distribution of the volume of unambiguously determined trades. All of them give similar estimated fitting parameters for all kernels. However, statistical significance is best improved by adding one trade irrespective of the kernel choice . We therefore apply this procedure in this paper; as a consequence, all statistical results closely depend on this choice. The distribution of resulting durations are plotted in Fig 3.2 (right plot).

This simple correction procedure introduces a weak, short-term memory effect. Figure 3.3 (left) plots the linear auto-correlation function of the sequence $\{\theta_i\}$, for a particular day (March 3rd 2012) (other days yield similar results). All the coefficients are almost statistically equal to zero except at the first lag (which is why we apply Ljung-Box test starting from the second lag). This negative value is induced by the correction procedure (see Sec. 3.3.2) since the same

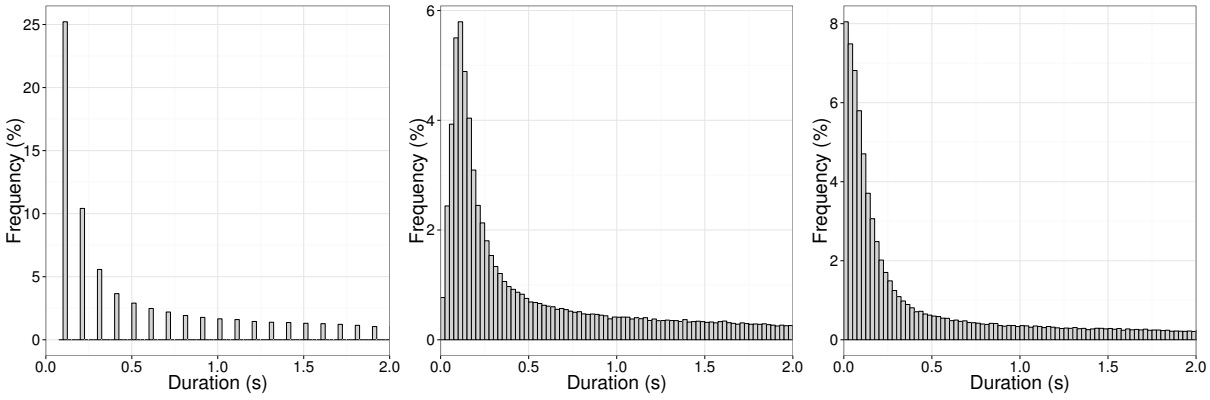

Figure 3.2: Duration distribution. Left: Raw times. Middle: Randomized times. Right: Corrected times. Three months of data, restricted to London active hours (9am-5pm).

measure performed in raw displays no memory at all (Fig. 3.3 (right)). The auto-correlation of the$\{(\theta_i)^2\}$ series is however null with the correction procedure. This test therefore shows that the time stamp correction procedure, without which no fit ever passes a Kolmogorov-Smirnov test, is not entirely satisfactory from this point of view. Nevertheless, the side effects are small and most of the auto-correlation of the corrected timestamps is well explained by a Hawkes model.

There may be other unwanted side effects caused by limited time resolution and by the randomization of timestamps within a given interval. In particular, one may wonder if limited time resolution introduces a spurious small time scale in the fits. Appendix A reports extensive numerical simulations that assess the effect of limited time resolution and time stamp shuffling and shows first that this is not the case when time stamps are shuffled in an interval. In addition, the smallest fitted time scale is influenced by the limited time resolution, but to a limited extent.

3.4 Results

3.4.1 Hourly fits

Hourly intervals are long enough to obtain reliable calibrations, at least on active hours during which 1500 events take place on average. In such short intervals, the endogenous activity μ_t in Eq. (3.1) can be approximated by a constant. We choose $m = 2$ and $M = 15$ for the power-law types of kernel. At the hourly scale,

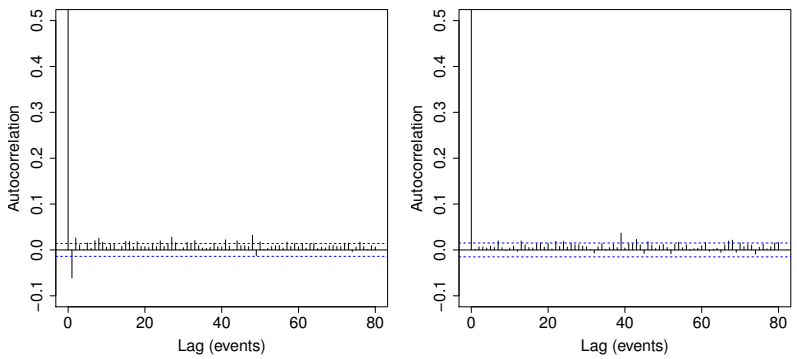

Figure 3.3: Time-adjusted durations autocorrelation function for March 3rd 2012. Left: with correction. Right: without correction.

the results are fairly insensitive to changes in these parameters.

3.4.1.1 Kernel comparisons

Table 3.1 summarizes the results of the 8 types of kernels for the three tests. The mono-exponential kernel ϕ_1 is clearly much worse than all the other specifications and we can safely rule it out as a possible description of the data. Taking more than two exponentials only marginally improves the fits of hourly activity. QQ-plots (Fig. 3.4) illustrate the inadequacy of ϕ_1 and show indeed that ϕ_2 is a good kernel: for this time length, two time scales are enough to describe a whole hour of the arrival of FX trades. We judge the trade-off between log-likelihood and the number of parameters with Akaike criterion, denoted by AIC_p , Akaike weights w_i of kernel i, and N_{max}, the number of intervals in which kernel i was the best. Both Akaike criteria are averaged over all the intervals. In the end, both w_i and N_{max} convey (almost) the same information because most of the time only one kernel has a weight almost equal to one. Power-law types of kernels also achieve good results, in particular $\phi_{15}^{\mathrm{PL}_\times}$, but all indicate a larger endogeneity factor n than kernels with free exponentials. Akaike weights strongly suggest that ϕ_2 is the best model at an hourly time scale. In addition we note that the means and medians of the fitted parameters of ϕ_n $(n = 1, 2, 3)$ kernels are very similar, while those of kernels that approximate power laws are significantly different, which points to the fact that this type of kernel is prone to fitting difficulties at an hourly time scale. Finally, the free exponential of $\phi_{15}^{\mathrm{PL}_\times}$ has a timescale of 0.06 s.

	ϕ_1	ϕ_2	ϕ_3	ϕ_{15}^{HBB}	ϕ_{15}^{PL}	ϕ_{30}^{HBB}	ϕ_{30}^{PL}	ϕ_{15}^{PLx}
μ	0.13	0.08	0.08	0.07	0.07	0.07	0.07	0.06
n	0.41	0.64	0.64	0.67	0.67	0.77	0.75	0.72
ϵ	NA	NA	NA	0.23	0.38	0.26	0.40	0.28
pKS	0.16	0.69	0.68	0.56	0.52	0.56	0.52	0.56
pED	0.03	0.57	0.55	0.63	0.60	0.58	0.57	0.62
pLB	0.11	0.38	0.38	0.34	0.31	0.29	0.28	0.33
$\log \mathcal{L}_p$	4022.9	4069.5	4069.9	4055.6	4062.9	4045.8	4060.3	4064.2
AIC_p	−8035.9	−8122.7	−8117.0	−8098.2	−8112.7	−8078.5	−8107.6	−8105.8
w	0.01	0.55	0.14	0.05	0.06	0.03	0.04	0.11
N_{max}	21	692	84	65	70	21	21	116

Table 3.1: Comparison the ability of various kernels to fit Hawkes processes on hourly time windows. pKS, pED and pLB are respectively the Komogorov-Smirnov, Excess-Dispersion and Ljung-Box test average p-values. $\log \mathcal{L}_p$ is the log likelihood per point for the fit of each intervals, averaged over all intervals, and multiplied by the average number of points per interval. Idem for the Akaike information criterion AIC_p. The Akaike normalized weights $w[\phi] = \frac{1}{W} \exp\left(-\frac{AIC[\phi]-AIC_{min}}{2}\right)$, are the probabilites that kernel ϕ is the best according to Kullback–Leibler discrepancy [108]. $N_{max}[\phi]$ is the number of intervals in which the Akaike weight of kernel ϕ is the largest one. Values averaged over the fits on 1090 non-overlapping windows with more than 200 trades.

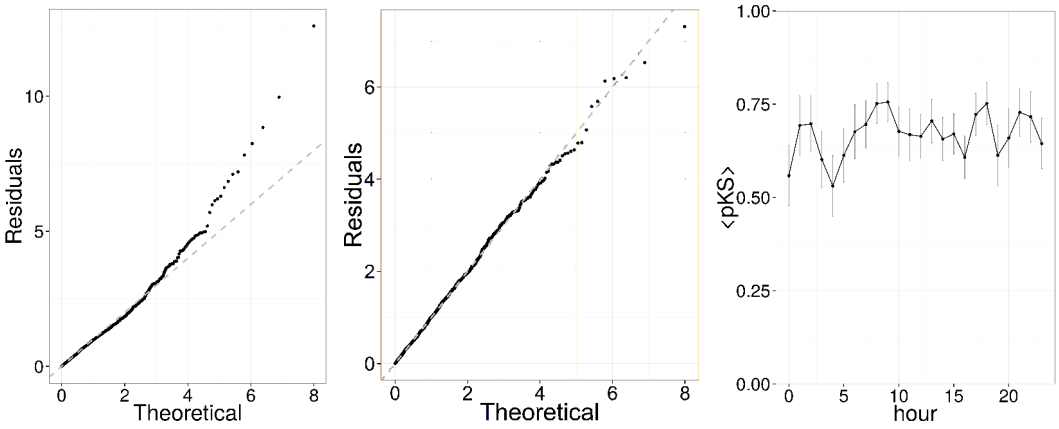

Figure 3.4: Goodness of Fit tests under the null hypothesis of exponentially distributed time-deformed durations. Left: A typical QQ-plot (February 1st 2012, 3-4pm) for ϕ_1. Middle: Same for ϕ_2. Right: Kolmogorov-Smirnov test average p-value. Error bars set at two standard deviations.

3.4.1.2 Detailed results for ϕ_2

Given its simplicity and good performance, it is interesting to look further into the results for the double exponential case. We note that Rambaldi et al. [100] also suggest that this kernel is a good candidate for the modeling of mid-quotes changes in EBS data (without signed volumes). We characterize each hourly time-window by averaging the fits over three months.

First, let us have a look at goodness of fits results. Fig. 3.4 (left plot) reports the quantiles of $\{\theta_i\}$ for a particular day and hourly window against the exponential theoretical quantiles. The fit is visually very satisfactory. Other time windows of all days yield similar results. Fig. 3.4 (right plot) demonstrates that all hours of the day pass Kolmogorov-Smirnov test by a large margin.

In Fig. 3.5 (left), the number of trades displays the well-known intraday pattern of activity in the FX market [39, 62]. The average fitted exogenous part $\langle \mu \rangle$ perfectly reproduces this activity pattern (Fig. 3.5, right plot).

Remarkably, the endogeneity level n is relatively stable (within statistical uncertainty) for all hours (Fig. 3.6) given the fact that the typical trading activity is 10 times smaller at nights (Fig. 3.5). This is particularly striking for the endogeneity associated to largest time scale, n_2. Endogeneity associated with the smallest time scale, n_1, follows, albeit with a much smaller relative change, the daily average activity, except for the lunch time lull, which comes from the largest time scale. This suggests that while automated algorithmic trading takes

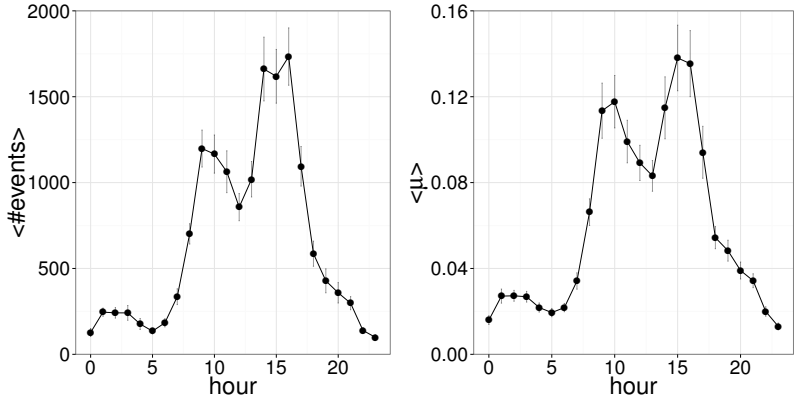

Figure 3.5: Left: Average number of trades (left) and average baseline intensity (right) throughout the day. Error bars set at two standard deviations.

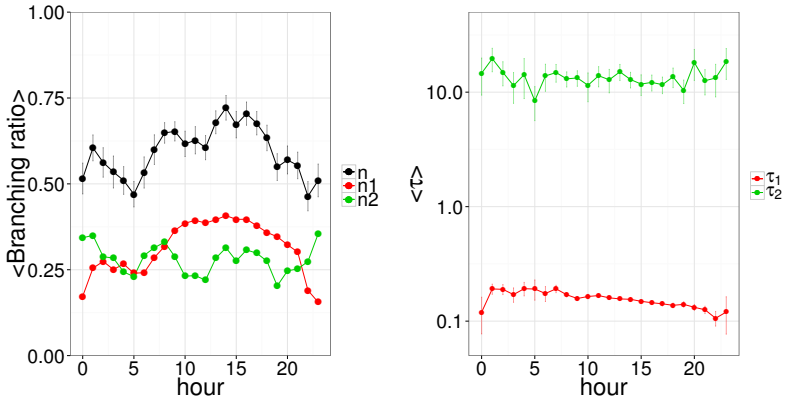

Figure 3.6: Average branching ratio throughout the day (left); black symbols: total ratio; green symbols: branching ratio of the largest time-scale; blue symbols: branching ratio of the smallest time-scale. Average associated times-scales on the right. Error bars set at two standard deviations.

no pause, human traders do have a break. In turn, this means that at this scale, most of the endogeneity at the smallest time scale comes from algorithmic trading, and that a sizable part of the endogeneity at longer times scales is caused by human trading.

3.4.2 Whole-day fits

The relative stability of the branching ratio and the high p-values of e.g. KS tests encourages us to fit longer time windows. As we will see, this is possible for a full day at a time. In this case, μ cannot be considered constant anymore (see Fig. 3.5). As suggested by Bacry and Muzy [5], a time-of-the-day dependent background intensity is a good way to account for the intraday variation of activity. This method has the advantage of not mixing data from other days like classic detrending methods do. We thus approximate, for each day, μ_t by a piecewise linear function with knots at 0 am (when the series begin), 5 am, 9 am, 12 pm, 4 pm and at the end of the series. The 6 knots values are additional fitting parameters.

3.4.2.1 Kernel comparison

The results are synthesized in table 3.2.

Only ϕ_2 and ϕ_3 pass the Ljung-Box test. This time ϕ_3 is the favored model according to the Akaike weights and performs well with respect to the three tests. We note that $\phi_{15}^{\mathrm{PL_X}}$, whose free exponential has a timescale equal to 0.11 s, is also a strong contender. We can gain a global insight across days from QQplots. Indeed, under the null hypothesis, the residuals possess the same distribution independently of the considered day. We therefore merge all the residuals from all the daily fits and construct the QQplot against the exponential distribution. Fig. 3.7 reports the performance of four families of kernel and bring a visual confirmation of the results in Table 3.2. In addition, it allows one to understand where each kernel performs best and worst. For example, ϕ_{30}^{PL} is better in the extreme tails than in the bulk of the distribution. One also sees the problems of ϕ_3 in this region, solved by adding a fourth exponential (see ϕ_4).

3.4.2.2 Detailed results for ϕ_3

Let us investigate in details the fits of ϕ_3, the overall best kernel for whole days. We also show some results for ϕ_2 for sake of comparison. The background intensity fitted values are summarized in Fig. 3.8 and are in line with the average intraday activity pattern.

Table 3.2: Kernel comparison. Full day fits. 59 points.

Kernel	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_{15}^{HBB}	ϕ_{15}^{PL}	ϕ_{30}^{HBB}	ϕ_{30}^{PL}	$\phi_{15}^{\mathrm{PL_x}}$
n	0.48	0.79	0.83	0.85	0.81	0.83	0.98	0.97	0.88
pKS	$7e-13$	0.09	0.13	0.16	4×10^{-6}	2×10^{-7}	6×10^{-4}	$4e-6$	0.04
pED	0	0.10	0.31	0.45	0.61	0.51	0.52	0.49	0.66
pLB	0	0.058	0.056	0.012	$9e-5$	0.001	1×10^{-4}	$2e-4$	0.006
$\log\mathcal{L}_p$	60271.0	61559.3	61596.2	61575.5	61279.6	61340.2	61286.3	61340.6	61468
AIC_p	-120525.4	-123097.7	-123167.4	-123121.8	-122540.4	-122661.7	-122553.9	-122662.5	-122912.0
ϵ	NA	NA	NA	NA	0.090	0.115	0.13	0.14	0.08
w	0	0.13	0.38	0.24	0	0	0	0	0.24
N_{max}	0	9	22	14	0	0	0	0	14

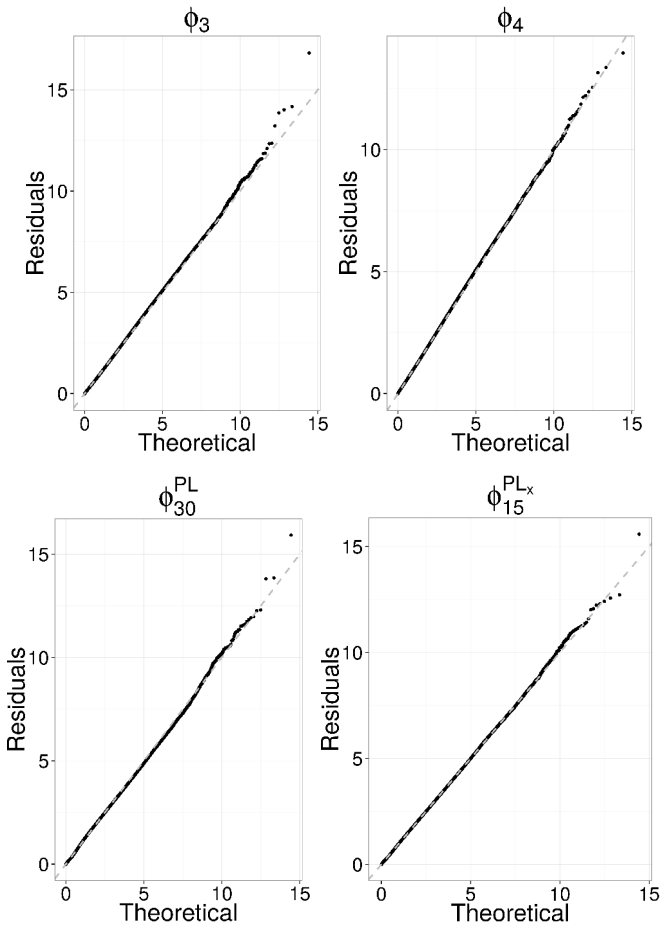

Figure 3.7: QQ-plot of the residuals merged from all intervals (one-day fits).

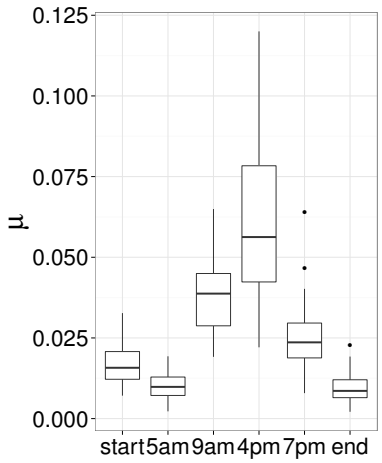

Figure 3.8: Tukey boxplot of baseline intensity knots values.

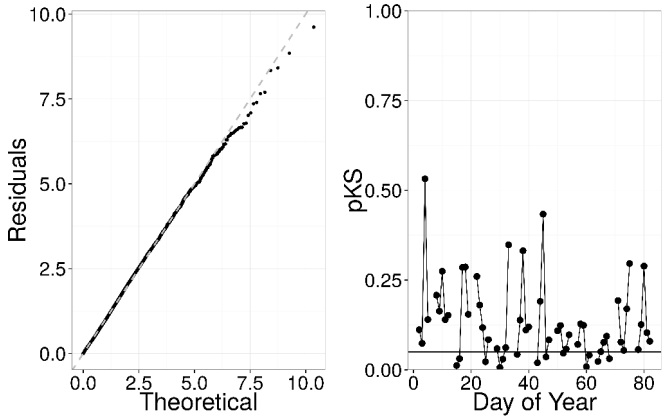

Figure 3.9: Goodness of Fit tests under the null hypothesis of exponentially distributed time-deformed durations. Left: A typical QQ-plot (March 3rd 2012). Right: Kolmogorov-Smirnov test p-value. The continuous line is the 0.05 significance level.

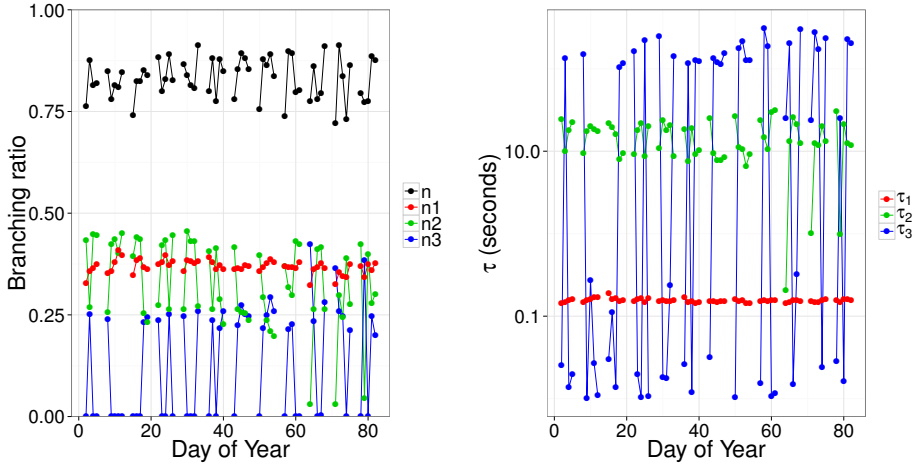

Figure 3.10: Daily branching ratio (left) and associated time-scales (right). The shortest characteristic timescale is very stable; the model captures 1 or 2 longer time scales depending on the day.

	2 timescales	3 timescales
$\langle\tau_1\rangle$	0.16 s	0.15 s
$\langle\tau_2\rangle$	21.9 s	9.3 s
$\langle\tau_3\rangle$	NA	161 s

Table 3.3: Average timescales when two or three timescales are found by fitting ϕ_3 to whole days.

Figure 3.9 reports the Kolmogorov-Smirnov p-value for each fitted day. Again, the null hypothesis of exponentially distributed $\{\theta_i\}$, i.e., good fits, cannot be rejected. Fits are however less impressively significant that those of hourly fits case because of additional non-stationarities. On this plot and on all the remaining plots of the section, line breaks correspond to weekends. The QQ-plot (left plot of Fig. 3.9) visually confirms the accuracy of the fit.

While the total branching ratio oscillates around 0.8 (Fig. 3.6), the parameters associated to each exponential make it clear that three timescales are only found on some days. This, once again, may either be because some days do not require three timescales, or because of the sloppiness of sums of exponentials. As reported by Table 3.3, the shortest timescale $\langle\tau_1\rangle$ does not depend on the effective number of timescales, while the second indeed does.

Kernel	ϕ_2	ϕ_3	ϕ_4	ϕ_{15}^{HBB}	ϕ_{15}^{PL}	ϕ_{30}^{HBB}	ϕ_{30}^{PL}	ϕ_{15}^{PLx}
n	0.80	0.87	0.88	0.82	0.84	0.98	0.97	0.91
pKS	0.02	0.04	0.06	1×10^{-12}	3×10^{-15}	1×10^{-6}	3×10^{-10}	0.04
pED	0.04	0.46	0.54	0.50	0.35	0.59	0.44	0.58
pLB	0.010	0.008	0.011	2×10^{-6}	4×10^{-6}	3×10^{-8}	4×10^{-8}	0.001
$\log \mathcal{L}_p$	119666.0	119819.2	119814.5	119094.1	119221.6	119086.8	119207.4	119656.4
AIC_p	−239303.5	−239605.7	−239592.4	−238161.7	−238416.7	−238147.1	−238388.2	−239282.2
ϵ	NA	NA	NA	0.08	0.11	0.13	0.15	0.10
w	0	0.34	0.40	0	0	0	0	0.26
N_{max}	0	9	10	0	0	0	0	7

Table 3.4: Kernel comparison of two-days fits. 26 points.

3.4.3 Multi-day fits

Extending fits to two days requires to account for weekly seasonality. First and most importantly, EBS order book does not operate at week-ends, which implies that Mondays and Fridays most likely have a dynamics distinctly different from the other days. Thus we fit all pairs Tuesdays-Wednesdays, and Wednesdays-Thursdays, which amounts to 26 fits (2 points per week, 13 weeks). Before proceeding, it is important to keep in mind that Figure 3.8 forewarns that the daily variations of activity at various times of the day are ample, particularly at about 4pm, the time of the daily fixing. This may also prevent a single kernel to hold for several days in a row, the composition of the reaction times of the population of traders being potentially subject to similar fluctuations between two days.

Table 3.4 compares the performance of all kernels. Kernel ϕ_2 performs poorly, while ϕ_3, ϕ_4 and ϕ_{15}^{PLx} are the best ones according to AIC_p criterion. No kernel can pass the three tests at the same time (ϕ_3 does for a single pair of days). The timescales of ϕ_3 are stable and similar to those of single-day fits ($\langle \tau_1 \rangle \simeq 0.15\,\text{s}$, $\langle \tau_2 \rangle \simeq 10.6\,\text{s}$, $\langle \tau_3 \rangle \simeq 178\,\text{s}$), while ϕ_4 sometimes manages to find a fourth timescale. For the record, we tried to use 5 exponentials, but never found a fifth timescale. It is noteworthy that ϕ_4 has an acceptable average pKS. The free exponential of ϕ_{15}^{PLx} has a timescale of 0.13 s.

3.5 A non-parametric investigation

This section presents a complementary approach where the kernel is estimated with non-parametric means. The statistical relevance is not the focus anymore

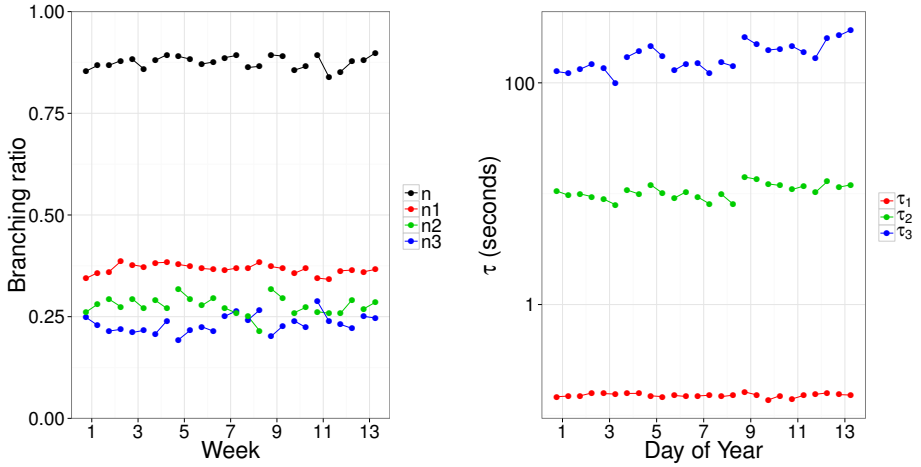

Figure 3.11: Endogeneity factors (left plot) and associated timescales (right plot) for fits of ϕ_3 to two consecutive days

	3 timescales	4 timescales
$\langle\tau_1\rangle$	0.15 s	0.15 s
$\langle\tau_2\rangle$	13.5 s	7.1 s
$\langle\tau_3\rangle$	226 s	33 s
$\langle\tau_4\rangle$	NA	295 s

Table 3.5: Average timescales when three or four timescales are found by fitting ϕ_4 to two consecutive days.

but we would rather like to cross-check the kernel shapes and investigate the kernel estimation when calibrated on a much longer time period.

3.5.1 Method

Here, we present only the key practical steps of the method developed by Bacry et al. [6]. This method is valid for a stationary Hawkes process, meaning that μ is constant and $n < 1$.

As before, let N_t and λ_t be respectively the counting process and intensity of a univariate Hawkes process. We define the normalized auto-covariance of the process at scale h and lag τ by

$$v_\tau^{(h)} = \frac{1}{h}\mathrm{Cov}\left(N_{t+h} - N_t, N_{t+h+\tau} - N_{t+\tau}\right).$$

Since we consider a stationary process, $v_\tau^{(h)}$ does not depend on the time, thus the quantity is easily estimated empirically.

Key steps:

1. Set the sampling period Δ and fix $h = \Delta$. We will then use h in the following to designate both the sampling period and the bin size.
2. Estimate the unconditional intensity $\bar{\lambda} = E(\lambda_t)$.
3. Estimate $v_\tau^{(h)}$ and compute its Fourier transform $\hat{v}_{i\omega}^{(h)}$.
4. Compute A defined by $A^2 = \hat{v}_{i\omega}^{(h)}/\bar{\lambda}$. A is real because the auto-covariance is an even function.
5. Compute the Fourier transform of the kernel thanks to: $\hat{\phi}_{i\omega} = 1 - e^{-\log A + i\mathcal{H}(\log A)}$, where $\mathcal{H}(.)$ designates the Hilbert transform operator [94].
6. Inverse Fourier transform to obtain $\phi(t)$ (more precisely $\phi(n\Delta)$, $n \epsilon \mathbb{N}$).

3.5.2 One numerical test

We simulate a Hawkes process close to the parametric ones from Sec. 3.4.2 and apply the methodology to the synthetic data. Figure 3.12, show that the procedure is able to retrieve the true kernel.

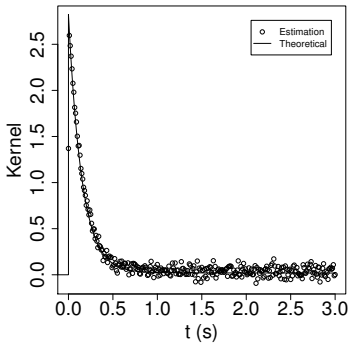

Figure 3.12: Non parametric estimation of a simulated Hawkes process with $300,000$ events, h =0.1 s. Kernel parameters: ϕ_2, $\mu = 0.1$, $\alpha_1 = 0.022$, $\alpha_2 = 2.8$, $\tau_1 = 20$, $\tau_2 = 20$.

3.5.3 Results

This calibration method needs more points than the MLE because of the auto-covariance estimation. We estimate the auto-covariance every day and average all these estimates. The kernel is then deducted thanks to the method described in Sec. 3.5.1.

3.5.3.1 Cross-check

To back up the previous parametric results, we calibrate the kernel to the 2012 EUR/USD dataset (3 months). We set h =0.1 s, smaller h are not considered because of the temporal resolution of the data. Two methods are used to reduce biases associated with the seasonality of the data. (1) Each day we choose a relatively small time window (2 p.m. to 4 p.m), with high activity where we can consider μ constant . (2) Detrending the data: the number of events in a h-bin is divided by its average computed over the whole three months with bins of width h_d.

In figure 3.13, we plot the empirical kernel alongside an estimated level of noise in blue. The noise level is determined by the application of the above methodology to an ensemble of simulated Poisson processes (one per day) with a rate equal to $\bar{\lambda}$. The blue kernel should be zero (no clustering in a Poisson process) so the blue points represent a level of significance for the kernel amplitude. We conclude from this analysis that the empirical kernel is statistically null after about 50 s for (1) and about 100 s for (2). The solid red line is a nonlinear least square curve fitting

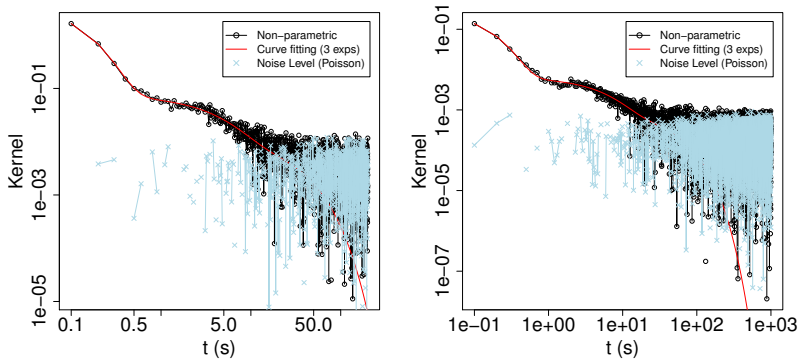

Figure 3.13: Non-parametric kernel estimates. Blue crosses represents the noise level. The red line is a curve fit with a sum of three exponentials. Left: Method (1). Daily time window: 2 p.m.-4 p.m. and no detrending. Right: Method (2). Daily time window: 6 a.m.-9 p.m. and $h_d = 5\,\text{min}$

with a ϕ_3 function, on the area above the noise[2]. It confirms that ϕ_3 is a good fit to the data. Using ϕ_2 also gives visually good results.

3.5.3.2 Branching ratio

In principle, We can also extract an estimate of the branching ratio. Let Φ be the integral of the kernel: $\Phi(t) = \int_0^t \phi(s)ds$, by definition we have $\lim_{t\to+\infty} \Phi(t) = n$. We can approximate $\Phi(t)$ by the cumulative sum of the previously estimated kernel times h. Fig. 3.14 left shows that Φ saturates at large lags, giving $\hat{n} \simeq 0.93$. In Fig. 3.14 middle, we see that method (2) completely change Φ's shape. The branching ratio estimate is higher: 0.98. The distortion is mainly to the detrending procedure as shown by Fig. 3.14 right that reproduces the analysis without it. Actually, we noticed that $\hat{n}$ also depends on h. This may be a sign that this method is not reliable for this purpose or at least that a systematic study of $\hat{n}$ as a function of h and h_d is needed before drawing any conclusion.

[2]The fit parameters are: $\alpha_1 = 0.064$, $\alpha_2 = 0.010$, $\alpha_3 = 4.101$, $\tau_1 = 3.78$, $\tau_2 = 27.03$, $\tau_3 = 0.10$ for method 1

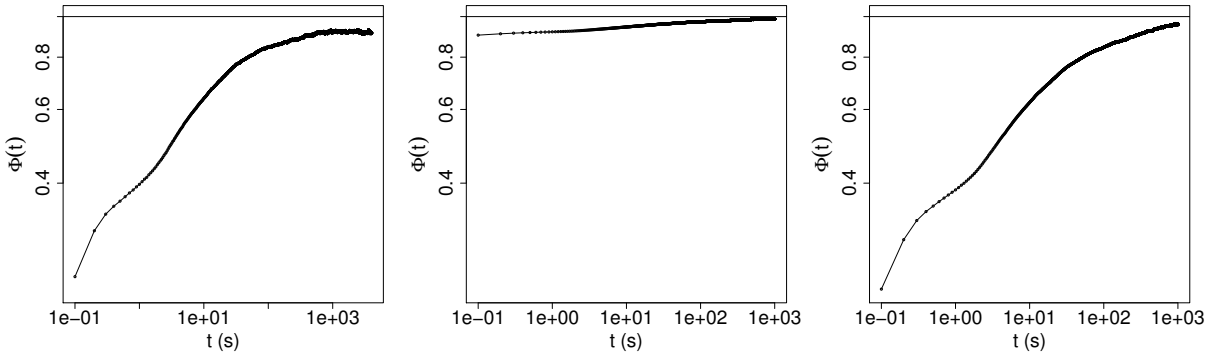

Figure 3.14: Branching ratio analysis thanks to$\Phi(t)$. Left: Method (1). Middle: Method (2). Right: Same conditions than method (2) without the detrending procedure.

3.6 Discussion and conclusions

Our results are mostly positive: Hawkes processes can indeed be fitted in a statistically significant way according to three tests to a whole day of data. This means that they describe very precisely a large number of events (around on average 15000). This is all the more remarkable because the fitted timescales are quite small. This shows that the endogenous part, which account for about 80% of the events, is limited to short time self-reactions in FX markets. This also means that at these time horizons, the instantaneous distribution of reaction time scales of the traders influences much the fitted kernels, as shown by the lunch lull in endogeneity. This is one reason why fitting more than one day with the same kernel is very hard since nothing guarantees that the composition of the trader population will be the same for several days in a row.

Fitting longer and longer time periods requires more and more exponentials. Fitting sums of exponentials with free parameters yields successive timescales whose ratios are not constant, which contrasts with the assumption of kernels that approximate power-laws. This is why the kernel $\phi_{15}^{\mathrm{PL}_\times}$, which adds one free exponential to the latter, has an overall better performance than pure approximations of power-laws. Longer time periods also leads to larger endogeneity factors, which makes sense since measuring long memory by definition requires long time series. As it clearly appears in all the tables, the use of power law-like kernels mechanically increases the apparent endogeneity factor, some of them being dangerously close to 1 (e.g ϕ_{30}^{HBB} and ϕ_{30}^{PL}). That said, and quite importantly, the best

kernels are never those with the largest endogeneity factors.

One may wonder if significance could be much improved by using data with a much better time resolution. It would certainly help, but only to a limited extent. As shown in Appendix A, only the KS test is affected by introduction of limited time resolution. Since the fits also fail to pass the the LB test for two consecutive days that is not affected by a limited time resolution, it is safe to assume that this failure has deeper reasons. The main problem resides in the difficulties caused by the non-stationarities of both exogeneity and endogeneity. The example of the lunch lull is striking: assuming a constant kernel shape for all times of the day, while a good approximation, cannot lead to statistical significance of fits over many days. In this precise case, one could add a daily seasonality on some weights.

Our results may well be specific to FX markets. In particular, the endogeneity is never close to 1, in contrast with studies on futures on equity indices. However given the nightly closure of equities markets (for example) and their short opening times, and given the difficulties encountered for FX data, it seems difficult to envisage a statistically satisfying comparison.

Chapter 4

Implicit communication networks of traders

4.1 Introduction

In this chapter, we leave the interdealer market to focus on the other side of the foreign exchange trading: the customer market. Specifically, we study data from two different segments: a large dealing bank, providing liquidity to financial institutions and an online broker-dealer whose clients are mainly individuals. Despite the fact that their leading position as the main source of FX liquidity as been challenged in recent years, large banks still handle enormous transaction volumes and are a key component that aggregates customer information into prices. Retail trading volume has increased dramatically volume increase thanks to the development of electronic platforms and is no longer a small phenomena that can be neglected. Hence, they both worth academic interest. In the rest of the chapter, we use indifferently the terms traders or agents to designate these two types of customers.

Understanding the behavior of thousands of agents is a daunting task. A promising solution from complex system sciences is to cluster the agents with similar behavior into groups and study their interaction. Switching from individual units to groups reduces the dimensionality of the problem and allows one to explore phenomena beyond the individual activity of each agent. It is also appealing for market-makers that could adjust their quotes according to the category of each client. Traders communicate all the time, explicitly (phone, Bloomberg station, Internet, Facebook, Twitter) but more interestingly also implicitly. Indeed, every trader observes the *same price* at the same time and takes his decision based on more or less complicated strategies, most of them built with the *same blocks*

(moving averages and combinations thereof...). It is thus reasonable to expect a certain degree of synchronization between the trades times when the strategies are similar enough. Suitable statistical methods might be then able to detect groups of traders that coordinate their actions.

Community detection is an active area with a vast literature, however applications to financial agents are complicated due to confidentiality reasons that prevents researchers to access trader resolved data. Nevertheless, a growing number of studies tries to identify groups of investors. A first approach took advantage of correlations between agents decisions over time: typically, one would build a time-series of buy or sell signals for each agent, according to the volume imbalance in each time-step. Then the correlation matrix of these time-series is analyzed thanks to random matrix theory and a structure is extracted from the eigenvectors associated to the eigenvalues outside the theoretical bands. [52, 61, 112, 74, 114]. The most interesting result obtained with this method, irrespective of the type of agents (banks, institutions, firms, individuals) is the presence of a large eigenvalue much larger than the other ones and way above predicted threshold for structureless matrices. This finding is reminiscent of the "market mode" found in stocks correlations [69]. However, correlation matrices are hard to estimate, especially in presence of a large number of very heterogeneous agents. Moreover the bursty nature of human trading may introduce spurious synchronization, converted into falsely large correlation coefficients. To overcome these issues Tumminello et al. [106] developed a new method coined statistically validated networks (SVN.). A pair of traders whose synchronicity cannot be explained by pure randomness, is linked; one then identifies the communities in the resulting network with standard cluster detection methods. Tumminello et al. [107] later applied this methodology to the clustering of Finnish investors. An other direction based on network theory and information on both counter-parties of a deal is the one by Jiang and Zhou [65]. They have constructed a network where a transaction is interpreted as a directed link from the seller to the buyer, the trade size being its weight. They systematically find a giant component and a power-law degree distribution in the resulting networks.

In this work, we show that trader clustering at an hourly timescale leads to more complex networks and communities that by using daily data. We extend previous literature results on investors clustering to a new asset class (FX) and with several types of agents (firms, institutions, individuals). More interestingly, we also consider the possible lead-lag connexions between the detected communities. A particular strategy used by an agent might activate a trading decision with a more or less stable lag with respect to the strategy used by an other agent. Considering groups instead of agents reinforces this effect. Indeed, the aggregation

Platform	Timespan	Resolution	Instruments	Clients	Trades
SQ	2012/01/31 → 2012/08/10	1 s	68	∼ 4000	∼ 1400000
LB	2013/01/01 → 2014/09/17	1 ms	12	11849	1440042

Table 4.1: Datasets description.

of the actions of agents belonging to the same cluster reduces the fluctuations in the timing of trading, stabilizing the lag with other groups actions. There exists then the possibility that a particular group buying seems to be an effective signal to sell later for an other group. Using a reformulation of SVNs, we obtain such links in a statistically valid way. We thus unravel an implicit lead-lag interaction network between investors.

Robust lead-lag relationships are an indicator of predictability in traders activity. We prove that accurate forecast of the aggregated trading direction (buy or sell) is possible. To do so, we leverage the clustering procedure by using random forests, a type of machine learning.

4.2 Marketplaces and data

As sketched in the introduction, the proprietary customer data in our possession come from two sources, a large dealing bank (LB hereafter), and a considerable broker-dealer: Swissquote (SQ hereafter). LB electronic market-marking desk offers liquidity to large clients such as commercial companies, financial institutions, pension funds, hedge-funds. SQ act as a broker (bundling many small orders that are executed with a large dealer) and also as a counterparty (internal matching of the clients orders) and provides an electronic platform targeted mainly at retail investors (see Sec. 1.2.2). The reader might refer to Chapter 1 for details about the FX market organization.

For both market participants, we have the same data structure. All the customers orders are recorded with all the relevant information: the side (buy or sell) of the transaction, the price, the amount, the currency pair and the client identifier (a unique anonymous number). The resolution is a millisecond for LB and a second for SQ. More details about the data are in Table. 4.1.

	N_{SQ}(CurrencyPair)	N_{SQ}(Trader)	N_{LB}(Trader)
N_{min}	454	534	702
α	1.49	2.32	2.02

Table 4.2: Discrete power-law fitting results.

4.3 FX Customers descriptive statistics

4.3.1 Heterogeneity in activity

The activity level greatly varies from one trader to the other and from one instrument to the other. It is clear from the broad distribution of the number of trades per currency pair and per trader in Fig. 4.1. A discrete power-law model is fitted using a maximum likelihood procedure giving exponent α. The value N_{min} after which the power-law is a good description of the data is also estimated by minimizing the Kolmogorov- Smirnoff statistic, using R package `poweRlaw` [46]. Results are in Table. 4.2. For LB, we could not look at the distribution for the number of trades per currency pair, since there are only 12 for this dataset.

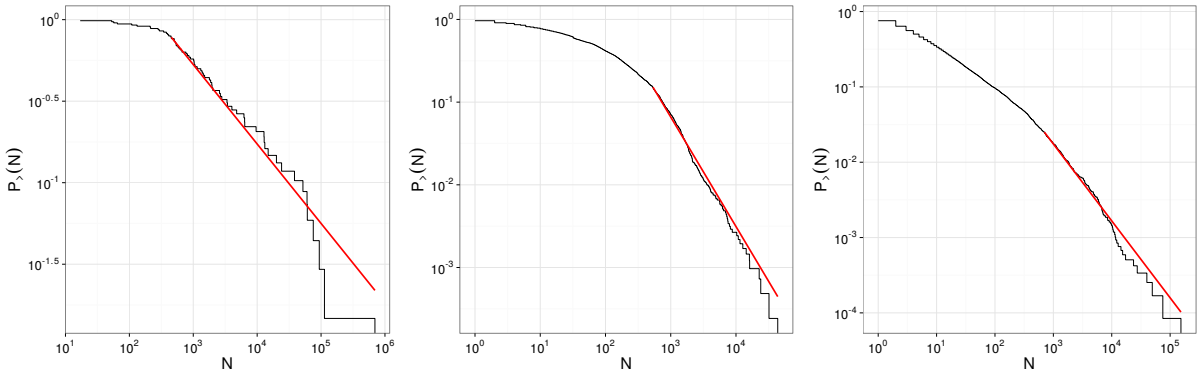

Figure 4.1: Heterogeneity in the trading activity. SQ cumulative distribution of the number of trades per currency pair (left) and per trader (middle). LB cumulative distribution of the number of trades per trader (right) in log-log scale. They both display power-law tails.

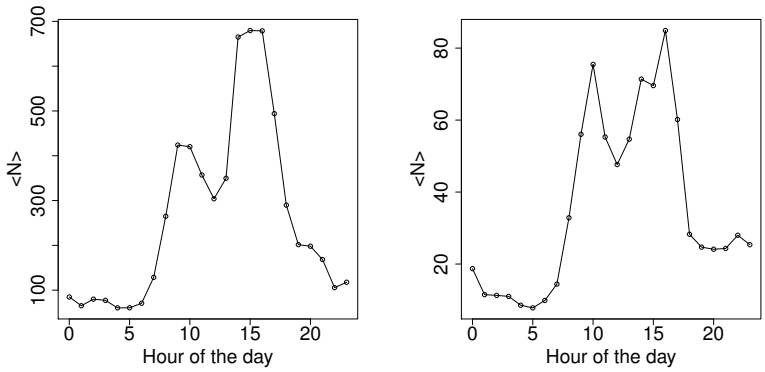

Figure 4.2: SQ seasonalities. Left: EUR/USD. Right: GBP/USD.

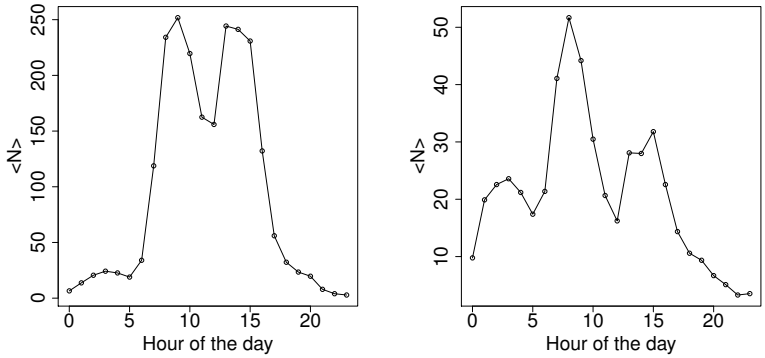

Figure 4.3: LB seasonalities. Left: EUR/USD. Right: USD/JPY.

4.3.2 Intraday pattern

We have seen previously (Sec. 2.3.5) a clear seasonality in activity for the inter-bank market. Do we retrieve a similar pattern for the dealer-customer market? We plot in Figs. 4.2 and 4.3, the average number of trades per hour for the two most traded currency pairs (for each platform). We observe qualitatively the same behavior than on EBS. The well-know pattern for the EUR/USD, which roughly follows London working hours, with a first peak corresponding to the Asian market and a drop at lunch time. Interestingly, the higher activity period (2 p.m. to 4 p.m.) is only present for the SQ data. The GBP/USD follows a similar pattern whereas the USD/JPY has of course more pronounced peaks during the Asian session.

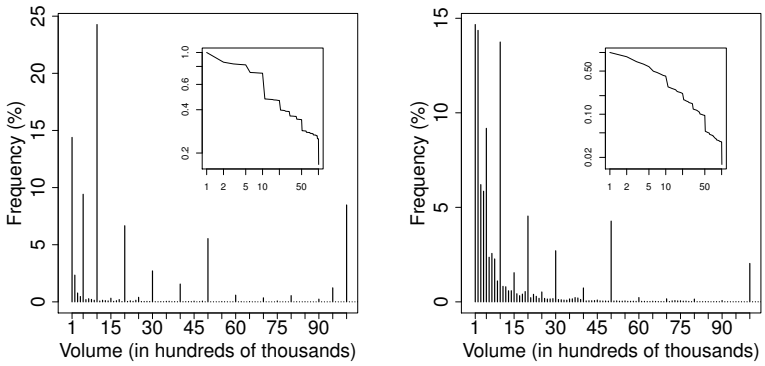

Figure 4.4: Distribution of EUR/USD transaction sizes. Left: SQ. Right: LB. Strong peaks are present at round volumes (5,10,15,...). Inset: Empirical cumulative distribution.

4.3.3 Amount distribution

The typical trade size of the two dealers' clients are very different. We thus present the results with a different unit. For SQ, in multiple of 1000 of the base currency and for LB in multiple of 100000. In Fig. 4.4, we plot the distribution of the transaction sizes (Fig. 4.4), for the EUR/USD. Similar results are obtained for other pairs. We obverse peaks at round numbers (5,10,15,...), like in the EBS market. The effect is stronger for SQ, consistent with the fact that its clients are less experienced than the LB ones, so more prone to be affected by psychological biases.

4.4 Clustering: the classical approach

Our attempt to unravel communities of FX traders starts with the standard approach based on correlations of the trading activity presented in the introduction. Details about the differences between Refs. [114, 74, 112] and our work are available in table 4.3.

Thanks to random matrix theory (Sec. 4.4.3), the three studies finds meaningful structure in the correlation matrices where the first factor variations are essentially explained by returns. Based on the correlation between the strategy and the returns, investors can be categorized in three groups: trend-followers (positive correlation), mean-reverters (negative correlation) and mixed or uncategorized traders (no significant correlation).

We replicate this study as a benchmark before exploring a more recent method. It is also interesting to see if the aforementioned results can be extended to the FX market for different kind of investors, mainly individual traders for SQ and institutional ones for LB. We find similar results with daily data and a richer structure with hourly data.

4.4.1 Strategies: Definition and examples

We define agent i strategy s_i during interval t as the net volume sign:

$$s_i(t) = \text{sign}\left[V_{buy}(t) - V_{sell}(t)\right],$$

where $V_{buy}(V_{sell})$, is the total amount bought (sold) by the agent during a unit time interval (one day or one hour in the following). s_i can only take three values: $+1(-1)$ when the agent is a net buyer (seller) during the interval or 0 when the agent is inactive or finishes the period flat. In fig. 4.5, we show some sample cumulative strategies at the daily scale. These plots reinforce the intuition that some communities should exist, discriminated by trading styles.

4.4.2 Correlation matrices

We construct the strategy matrix for both datasets. Each column is an agent's strategy over the whole time-steps. Because one needs at the very least a number of time-steps T greater than the number of agents N, we only keep a subset of the traders. The parameters retained for the following are stored in table. 4.4. We focus on the EUR/USD currency pair as it is the most traded, results are qualitatively similar for the G10 rates, although less clear-cut.

The correlation matrices of the strategies are represented at the daily scale on Fig. 4.6. A visual inspection suggests the existence of large moderately correlated (red) and anti-correlated (blue) groups. For the SQ case, we can also see two small groups of extremely correlated traders. A closer look at the data shows that in each group, agents trade always within the same second which cannot be due to an indirect link. Indeed, discussions with people working at SQ confirms that it is the same person managing several accounts and using exactly the same strategy on each of them. These "multiple-accounts" will also appear in the following method (Sec. 4.5).

Study	Signal	Timescale	Agents	Market
Zovko and Farmer [114]	Net volume sign	Hour	Institutions	London Stock Exchange
Lillo et al. [74]	Inventory variation	Day	Institutions	Spanish Stock Market
Zhou et al. [112]	Inventory variation	Day	Individuals and Institutions	Chinese Stock Market
LB	Net volume sign	Hour + Day	Institutions	FX Dealer
SQ	Net volume sign	Day/Hour	Individuals+Institutions	FX Broker-Dealer

Table 4.3: Review of the literature about strategies clustering and position of our work within.

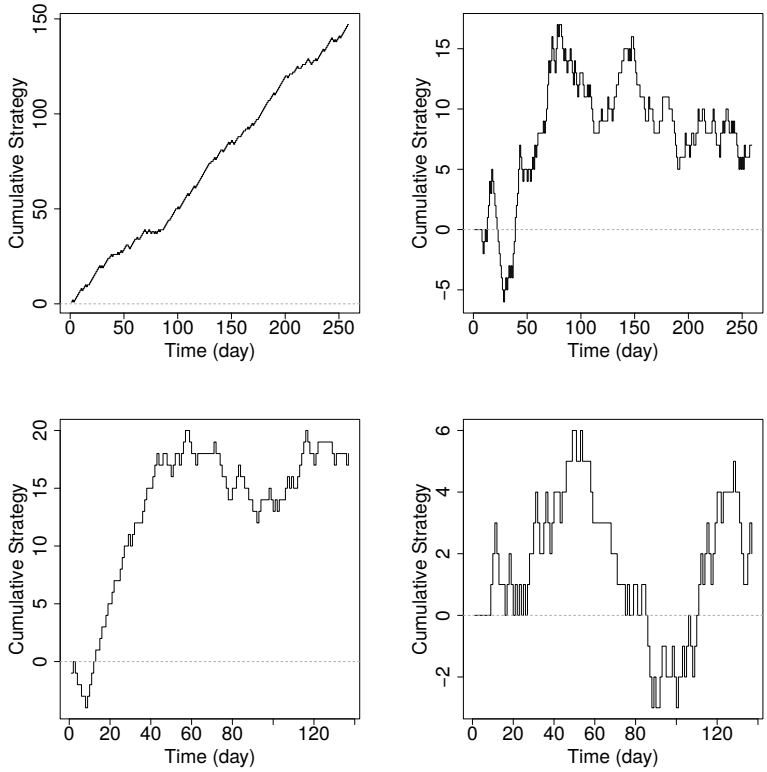

Figure 4.5: Cumulative strategies: cumsum($s_i(t)$) v.s. t. Different trading styles are noticeable. Top left trader seems to build a position whereas the top and bottom right ones seems to use mean-reverting strategies. The bottom left trader displays a mix of the two styles.

Dataset	Timescale	N	T
SQ	Day	100	137
LB	Day	100	259
SQ	Hour	400	1644
LB	Hour	400	3108

Table 4.4: Parameters used for the correlations matrices estimation.

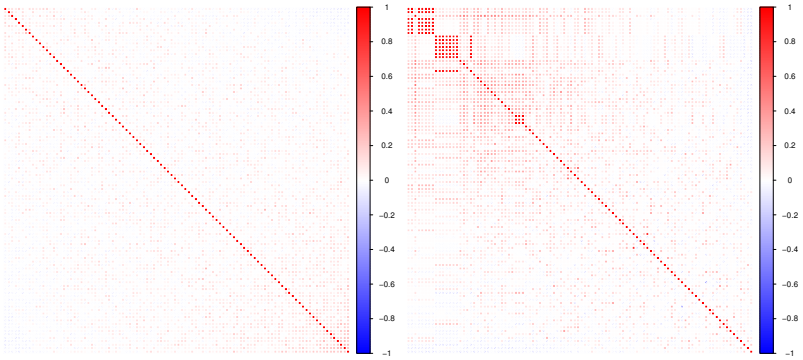

Figure 4.6: Correlations matrices of trading strategies. Left: LB data. Right: SQ data.

4.4.3 Eigenvalues analysis

A elegant way to see if there is structure in the correlation matrix is to examine its eigenvalues. Marčenko and Pastur [81] gave a result for the eigenvalues density for N independent and identically distributed variables of length T. In the limit $N \to \infty$ and $T \to \infty$ while keeping the ratio $Q = \frac{T}{N} \geq 1$, fixed, the density of eigenvalues $p_{MP}(\lambda)$ is given by:

$$p_{MP}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda},$$

where $\lambda \in [\lambda_{min}, \lambda_{max}]$, with:

$$\lambda_{min}^{max} = \left(1 \pm \frac{1}{\sqrt{Q}}\right)^2.$$

Fig. 4.7 plots the empirical eigenvalues distribution (in black) at the daily scale and fig. 4.8 at the hourly scale. In both figures, the red line is the Marcenko-Pastur density (p_{MP}) and the blue line is the empirical distribution when the strategy matrix is first randomly shuffled by column. The left plot is for LB, the middle one for SQ and the right one also for SQ but the multiple-accounts (described in Sec. 4.4.2) are removed from the strategy matrix. In the three plots of Fig. 4.7, one eigenvalue is clearly significant and out of the random bulk. The two other large eigenvalues for the middle plot are a consequence of the two multiple-accounts as demonstrated by their absence in the right plot. The picture is strikingly similar to the one found for correlation matrices of stock returns where one common factor drive all the stocks, the so-called market mode.

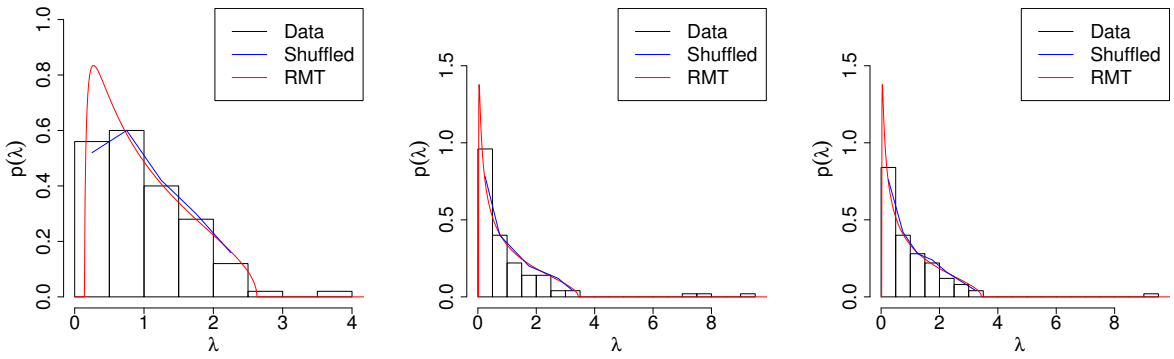

Figure 4.7: Daily scale distribution of eigenvalues (black) alongside the Marcenko-Pastur density (red line) and distribution from shuffled agents' time-series (blue line). Left: LB. Middle: SQ. Right: filtered SQ. The largest eigenvalue is clearly out of the bulk predicted by random matrix theory.

At the hourly scale the structure is even richer as demonstrated by Fig.4.8. At least four eigenvalues are meaningful in each dataset. Again, removing the multiple-accounts removes two large eigenvalues.

4.4.4 Eigenvectors analysis

The deviating eigenvalues must bear some information about the collective behavior of some agents. First, we check that the eigenvalues between λ_{min} and λ_{max} do not bring any information. We aggregate the normalized components of all the corresponding eigenvectors and inspect its distribution. The QQ-plots with the normal distribution as a reference (Fig. 4.9) show no deviation at both scales, in agreement with random matrix theory, meaning that there is no communities among traders delimited by those directions. The results are similar for LB.

At the opposite, the distribution of the eingenvectors components associated to deviating eigenvalues show strong divergences from the normal distribution (especially for the SQ case, see Fig. 4.10).

4.4.5 Information content of the first eigenvalue

Here we focus on extracting information from the first eigenvalue, because it is the only one present at the daily scale and it is much larger than the other one at the hourly scale (hence more reliable). In the case of a stock returns correlation

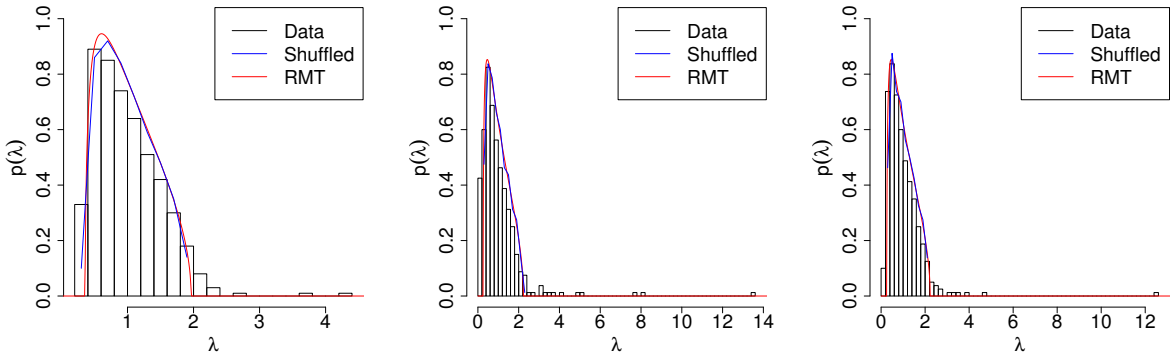

Figure 4.8: Hourly scale distribution of eigenvalues (black) alongside the Marcenko-Pastur density (red line) and distribution from shuffled agents' time-series (blue line). Left: LB. Middle: SQ. Right: filtered SQ. Several eigenvalues are clearly out of the bulk predicted by random matrix theory.

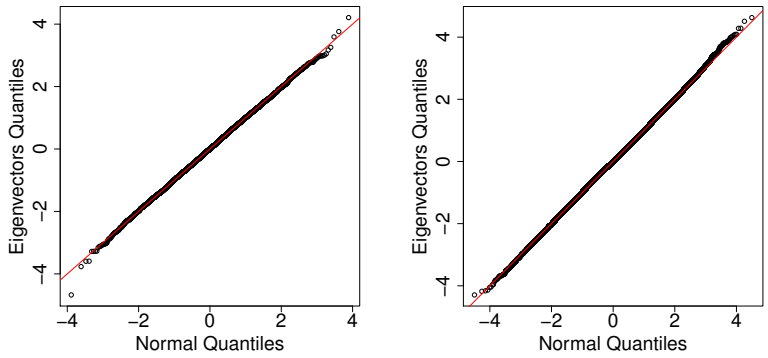

Figure 4.9: Normalized bulk eigenvectors components quantiles versus Normal quantiles. The good agreement between the two distribution demonstrates the absence of information in the eigenvectors corresponding to bulk eigenvalues. Left: Daily scale. Right: Hourly scale.

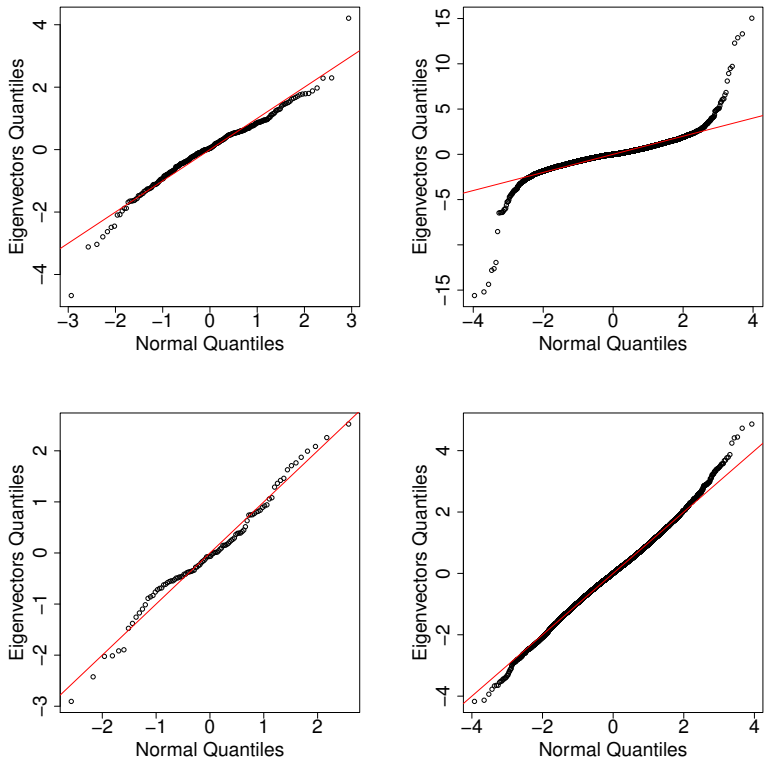

Figure 4.10: Normalized deviating eigenvectors components quantiles versus Normal quantiles. Top row: SQ data at daily and hourly scale. Bottom row: LB data at daily and hourly scale. The distribution of the components clearly posses discrepancies with the normal distribution.

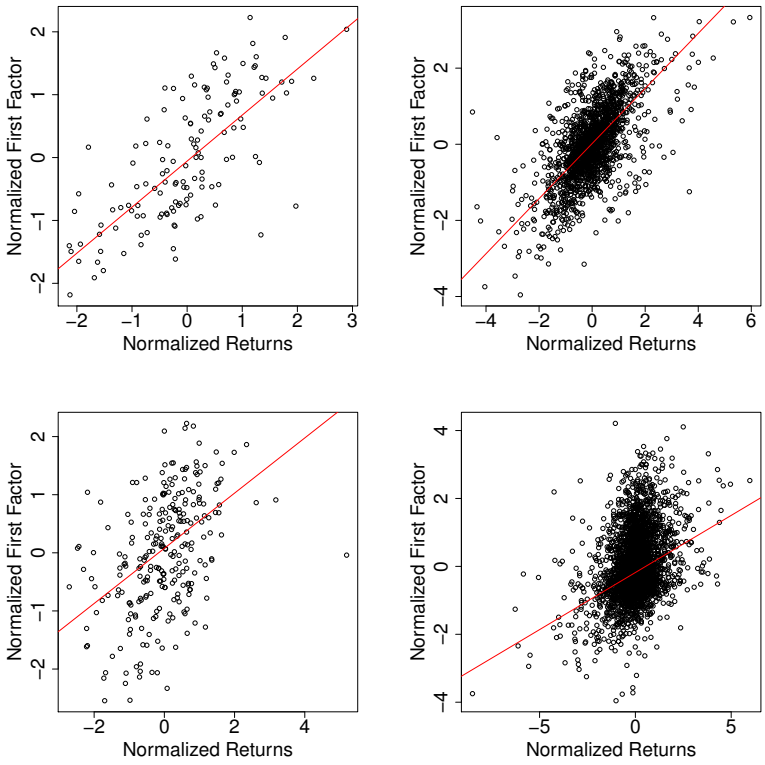

Figure 4.11: Scatter plot of the first factor versus the returns. The solid red line is a robust linear regression. The coefficients are given within the following brackets. Top row: SQ data at daily (0.73 ± 0.06) and hourly (0.72 ± 0.02) scale. Bottom row: LB data at daily (0.47 ± 0.07) and hourly (0.34 ± 0.01) scale.

matrix, it is known ([69]) that the first eigenvector represents a collective mode to which all stocks contribute. Similarly, with our data there is a part of the agents activity that can be attributed to a global quantity. We can anticipate that the driving force of this collective mode is the currency pair returns, as it is the most closely monitored market quantity by the traders. Following refs. [74, 112], we project the s_i variables onto the first eigenvector to obtain the first factor $F(t)$. A scatter plot of $F(t)$ versus the returns $r(t)$ in Fig. 4.11, reveals a strong linear dependence between the two variables. The phenomena is much clearer for SQ than for LB.

This observation suggests that the behavior of agents with respect to the returns is a good classifier. Following Lillo et al. [74], we compute for each agent i the correlation $\rho_{s_i r}$ between its strategy s_i and the returns r. The agents are

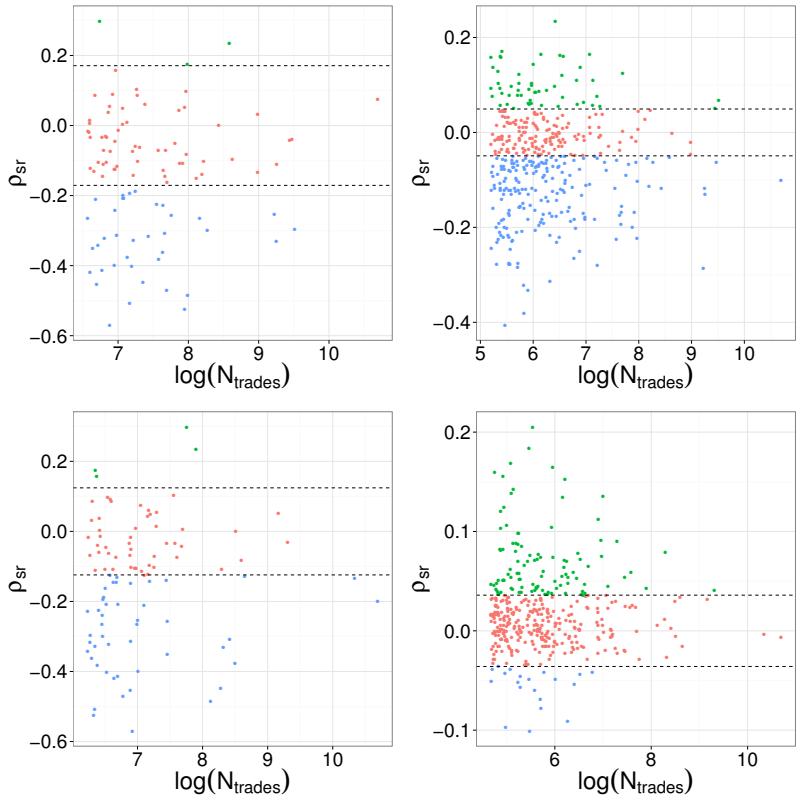

Figure 4.12: Scatter plot of $\rho_{sr}(t)$ versus the $r(t)$. The dashed lines are significance thresholds. Green points represents trend-followers, blue points represents mean-reverters and red points are uncategorized agents. Top row: SQ data at daily and hourly scale. Bottom row: LB data at daily and hourly scale.

categorized according to the value of the correlation compared to the significance threshold $\pm 2/\sqrt{N}$. If $\rho_{s_i r}$ is above (below) the upper (lower) threshold, it means that the agent is positively (negatively) correlated with the returns, thus it is assigned to the trending (reversing) category. In-between the thresholds, the agents are uncategorized, they probably use more complicated mixed-strategies. Figure 4.12, displays the results. At the daily scale (left column), we notice that many agents are in the mean-reverting category whereas we see only 3 trend followers for SQ (top) and 4 for LB (bottom). The situation is a slightly more balanced for SQ at the hourly scale (top right) with 13.5% of trend-following agents and 50% of mean-reverting ones. For LB (bottom right) the repartition is inverted compared to the daily scale: 69% of trend-followers against only 5% of negatively correlated traders.

At the hourly scale, we have seen many deviating eigenvalues besides the largest one, their information content is not clear. There is no easy interpretation in terms of economic sectors as in the stock returns case. They probably separate the traders according to their behavior, but how? We were not able to link these groups to clear patterns of trading. We shall now turn to a more powerful clustering method, to deepen further our understanding of traders communities.

4.5 Clustering: Statistically validated networks approach

4.5.1 Method description

The method is based on the statistically validated networks (SVN) technique recently developed by Tumminello et al. [106] and applied to the clustering of Finish investors in Ref. [107]. The goal is to identify groups of traders that act in a very similar way. We find a surprisingly strong synchronization in agents' decisions. The main edge of this method is its robustness with respect to the large heterogeneity in traders activity (see Fig. 4.1 right). It can therefore reliably detect statistically valid links (not purely random) between two traders. The aim is to be able to make statement of the type: when agent i buys, agent j sells, for instance.

4.5.1.1 Signal

We introduce a signal variable $s_i(t)$ close to the "strategies" of section 4.4.1. For each agent i, we compute an imbalance ratio:

$$\rho_i(t) = \frac{V_i^b(t) - V_i^s(t)}{V_i^b(t) + V_i^s(t)},$$

where $V_i^b(t)$ is the volume of base currency bought by the agent during time-step t and $V_i^s(t)$ the volume sold. Let ρ_0, be a real number with $0 < \rho_0 < 1$, defining an imbalance threshold. If $\rho_i(t) > \rho_0$, the agent is in a buying state and we set $s_i(t) = 1$. At the opposite, the agent is in a selling state when $\rho_i(t) < -\rho_0$ and we set $s_i(t) = -1$. When, $-\rho_0 < \rho_i(t) < \rho_0$, the agent is in a buying-selling state and we set $s_i(t) = 2$ (n.b. this is only a classification). When an agent did not trade in a time-step, we call it the inactive state and we set $s_i(t) = 0$. Note that in this setting an exact compensation $V_i^b(t) = V_i^s(t)$, causing a state 2 for the agent, is different from inactivity (state 0). The results are fairly similar for $\rho_0 \in [0.01, 0.1]$; in the following $\rho_0 = 0.01$.

4.5.1.2 Links validation

If a pair of agents has the same signal for several time-steps, we want to know if there is a genuine synchronization between them and not only a random effect. If we neglect the time-step heterogeneity, i.e. the number of states of a particular kind per interval, we can formulate a null hypothesis and compute a p-value that takes into account the activity level of each agent. In the following, we do not consider the synchronization in inactivity (state 0). The procedure to validate links goes as follows:

- Pick each pair of agents (i, j).
- Compute N_P, the number of time-steps t such as $s_i(t) = P$, with $P \in \{-1, 1, 2\}$.
- Compute N_Q, the number of time-steps t such as $s_j(t) = Q$, with $Q \in \{-1, 1, 2\}$.
- Compute N_{PQ}, the number of time-steps t such as $s_i(t) = P$ AND $s_j(t) = Q$.

Under the null, the N_P and N_Q trading signals under consideration are allocated randomly to the T possible time-steps (days or hours). The probability of having k common time-steps is equal to the probability of getting k white balls in N_Q draws from an urn containing N_P white balls and $T - N_P$ black balls. Hence, it is given by the hypergeometric distribution,

$$\Pr(k) = H(k|T, N_P, N_Q) = \frac{\binom{N_P}{k}\binom{T-N_P}{N_Q-k}}{\binom{T}{N_Q}}.$$

- Compute the statistical test p-value which is the probability of getting at least N_{PQ} random co-occurrence.

$$p(N_{PQ}) = 1 - \sum_{k=0}^{N_{PQ}-1} H(k|T, N_P, N_Q).$$

- Compare the p-value to a level of significance p_c to decide whether there is link of type PQ between i and j.

Some attention is needed when it comes to set the significance level. It cannot be an arbitrary number like such as the usual 0.01 or 0.05 because we have to take into account simultaneously the 9 types of co-occurrence (buy-buy, buy-sell,

etc...) and all the pairs of agents. Therefore, the level needs a multiple hypothesis test correction. We use the powerful false discovery rate (FDR) method [12]. The FDR is designed to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries", false links in our case). The correction is obtained in two steps. First the p-values from all the performed tests are sorted in increasing order. Then, the FDR threshold p_c is the p-value corresponding to the largest index l such as $p_l < l\frac{p_0}{N_{tests}}$, where N_{tests} is the number of performed statistical tests and p_0 is the chosen significance level for a single test. Here, N_{tests} is equal to the number of different types of links times the number of pairs of traders: $N_{tests} = 9\left(\frac{N(N-1)}{2}\right)$. We set a strict threshold, $p_0 = 0.01$ in the following.

With this tool, we are now able to statistically acknowledge the existence of links between a pair of traders while limiting the number of false links. We thus unravel the implicit communication network between traders with the distinctive feature of having multi-links of different nature.

4.5.1.3 Finding communities

The last step is to extract communities from this network. This will enable us to find groups with similar trading patterns, making possible a better understanding of traders behavior and strategies. It is also a good way to perform a dimensionality reduction by replacing the full set of agents by a few meaningful and representative entities, which turns out to be useful for flux prediction (see Sec. 4.6).

Community detection in networks is a rich field [70], Tumminello et al. [107] suggest the use of one of the most powerful technique: the InfoMap method [102]. Unfortunately, the method is not suited for multi-links networks. An easy workaround is to convert multi-links into weighted links by assigning a weight equal to the number of validated links between two traders. Prior to the community detection, we exclude links between opposite action (buy-sell), as we are primarily interested in finding groups with similar trading strategies.

4.5.2 Results on simultaneous actions

Within this section, we keep only the 500 most active clients, to have at least 100 trades per client per year. We exclude data from the weekends as they do not reflect the “normal” behavior of the market.

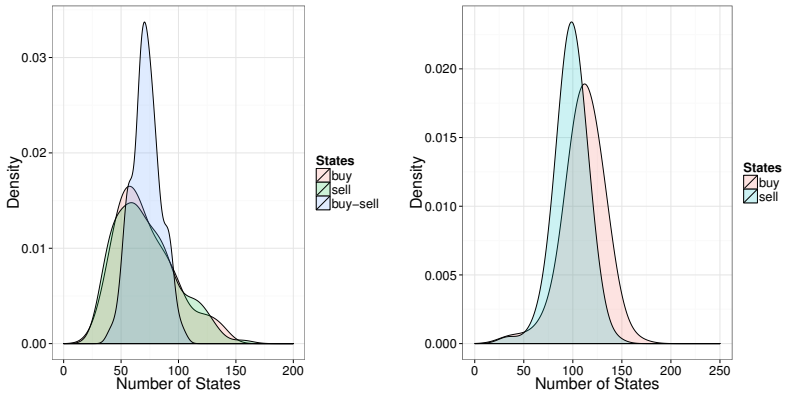

Figure 4.13: Distribution of the number of states per day. Left: SQ. Right: LB. EUR/USD case, similar results for other pairs. We don't display the distribution of the number of buy-sell states for LB as it is highly peaked around 7, hiding all the other information.

4.5.2.1 Daily

We first apply the method at the daily scale, to test it and compare the results to the RMT analysis ones in Ref. [107]. We keep only the data from 9a.m. to 7p.m. (London time). First, we control that we can neglect the variability in the number of states per day. Fig. 4.13 shows that it is a reasonable hypothesis.

The procedure gives the following networks (Fig. 4.14) on the EUR/USD currency pair. We plot only the links in the same direction, for better visibility and also because it is the ones used for the cluster detection. There is about 6 times more links in the SQ case. Let us denote the number of validated links in the same (opposite) direction by N_{same} (N_{opp}). We obtain for SQ: $N_{same} = 642$ and $N_{opp} = 147$, whereas for LB: $N_{same} = 68$ and $N_{opp} = 67$. There is a clear imbalance towards N_{same} for SQ, hinting at a strong herding behavior for individual investors.

For SQ, we retrieve the "multiple-accounts" identified in the activity correlation matrices (see Sec. 4.4.2) in the form of complete sub-graphs with all type of links present. For both data-sets, we see a large component surrounded by many satellite small groups.

4.5.2.2 Hourly

Hourly data increases the number of points, consequently we need fewer days to reliably validate links and detect communities than in the daily case. Therefore, we will be able to track the evolution of the network during time. It will also

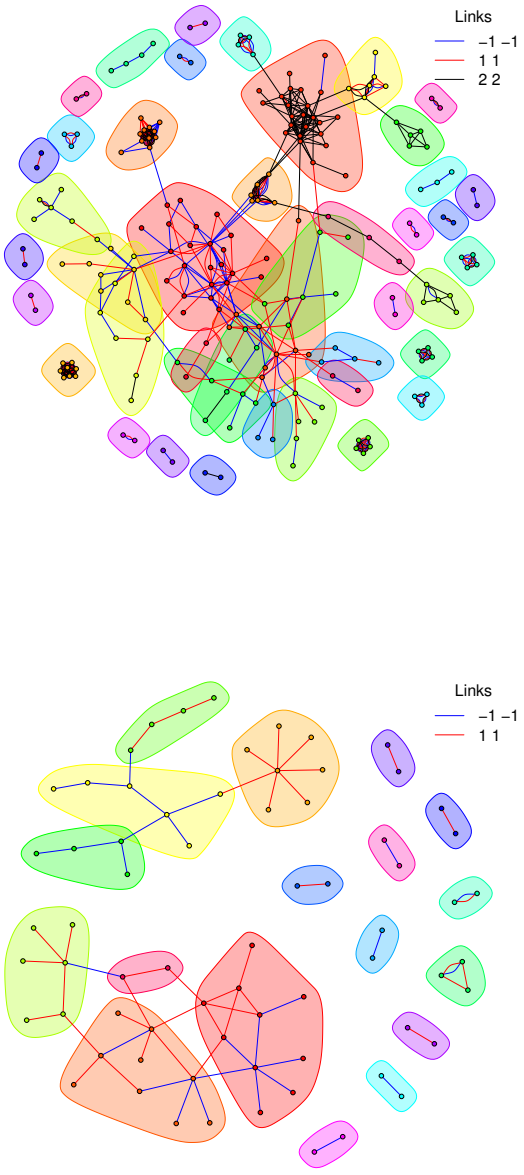

Figure 4.14: Examples of obtained networks at the daily scale. Top: SQ EUR/USD. Bottom: LB EUR/USD.

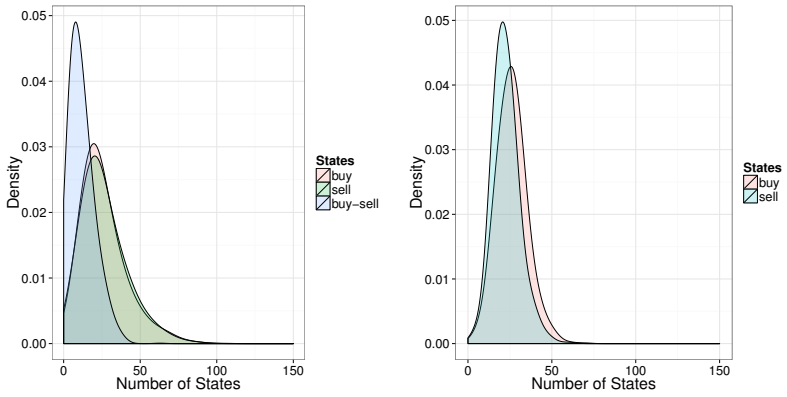

Figure 4.15: Distribution of the number of states per hour. Left: SQ. Right: LB. EUR/USD case, similar results for other pairs. We don't display the distribution of the number of buy-sell states for LB as it is highly peaked at 0, hiding all the other information.

allow us to perform a prediction task in Sec. 4.6. We keep only the data from 9a.m. to 4p.m. (London time). Again it is acceptable to neglect the variability in the number of states per timesteps, because its distribution presents a bell-shape with a moderate range of fluctuations (Fig. 4.15).

We compute the network of traders over 4 months overlapping periods. The shifting time between each network determination is one day. Fig. 4.16 shows a sample of such networks. The network characteristics displays mild variations over time for SQ whereas for LB an interesting phenomenon occur in January 2014. In Fig. 4.17, we see the number of detected groups decreasing drastically meanwhile the number of links didn't change much. This shows the emergence of super group containing almost all the classified traders and thus the existence at this point in time of a macroscopic collective strategy. We were not able to find a simple explanation for this to have happened.

This section clearly demonstrates the existence of synchronized trading behavior that can be summarized in a few number of meaningful groups.

4.5.3 Lead-lag relationships

4.5.3.1 Modified method

In this section, we extend the framework devised by Tumminello et al. [106] to try to unravel lead-lag relationships between the groups obtained in the previous section. Is the buying or selling status of group g_i at time t can give us information

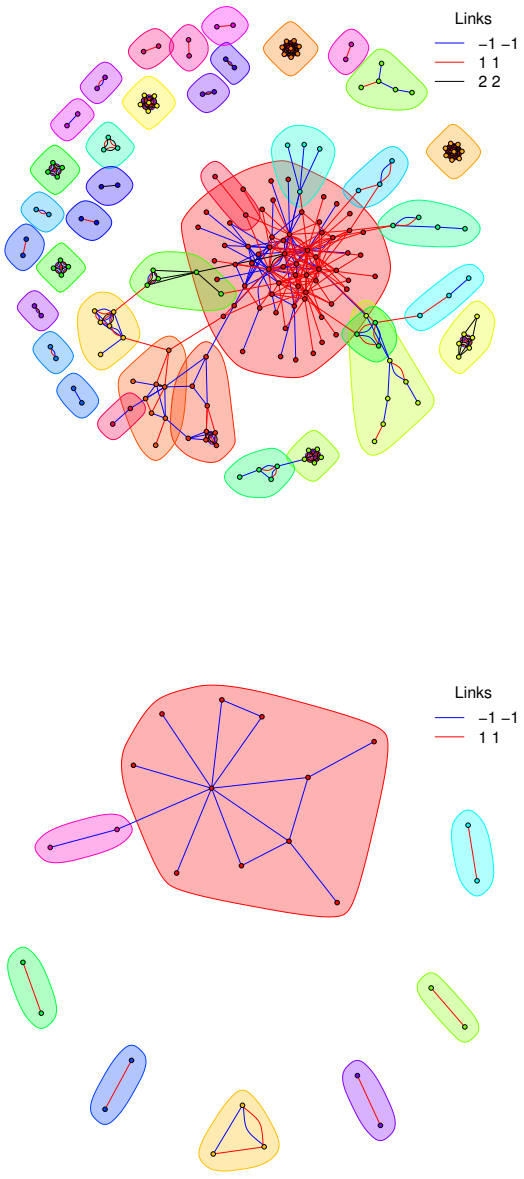

Figure 4.16: Examples of obtained networks at the hourly scale. Top: SQ EUR/USD. Bottom: LB EUR/USD.

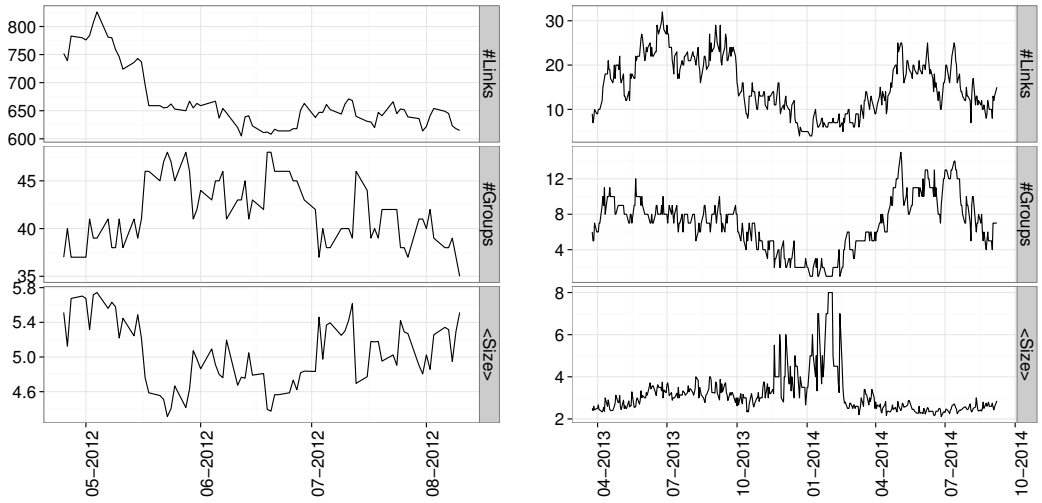

Figure 4.17: Network of clients: characteristics over time. Left: SQ. Right: LB.

on the buying or selling status of group g_j at time $t+1$? Finding such predictability in traders behavior could be of great interest for brokers and market-makers in order to match trades internally more efficiently.

To address this question, we use the same ideas that in Sec. 4.5.1.1 but with different variables. Instead of looking each agent individually, we analyze the trading decision at the group level. More precisely, we create the strategy time-series $s_{g_i}(t)$ for each group $g_i \in g_1, g_2, ..., g_{N_g}$, with N_g being the number of detected groups. To do so, the ratio ρ_{g_i} is computed by aggregating the volume of all the agents in the group g_i and $s_{g_i}(t)$ is deducted from $\rho_{g_i}(t)$ in a similar manner as Sec. 4.5.1.1. The same methodology is applied but one of the group strategy is *lagged* with respect to the other, breaking the $i \leftrightarrow j$ symmetry and allowing for self-links. Step-by-step, the procedure goes as follows:

- Find the groups with the method of Sec. 4.5.1.
- Pick each permutation with repetition (Lg_i, g_j), where L is the lag operator.
- Compute N_P, the number of time-steps t such as $s_{g_i}(t-1) = P$, with $P \in \{-1, 1, 2\}$.
- Compute N_Q, the number of time-steps t such as $s_{g_j}(t) = Q$, with $Q \in \{-1, 1, 2\}$.

- Compute N_{PQ}, the number of time-steps t such as $s_{g_i}(t-1) = P$ AND $s_{g_j}(t) = Q$.
- Compute the statistical test p-value which is the probability of getting at least N_{PQ} random co-occurrence.

$$p(N_{PQ}) = 1 - \sum_{k=0}^{N_{PQ}-1} H(k|T, N_P, N_Q).$$

- Compare the p-value to a level of significance p_{fdr} to decide whether there is link of type PQ between Lg_i and g_j. Computation of p_{fdr} is modified because N_{tests} is equal to $9N_g^2$ in this setting.

4.5.3.2 Results

The same parameters than Sec. 4.5.2.2 are kept for inter-groups links detection. We discuss only the hourly results. Several links are found between the groups and we can construct statistically validated lead-lag networks. We display two representative networks in Fig. 4.18. More links are present for SQ compared to LB, especially self-links. Interestingly, more links that validate opposite directions are present in the lead-lag case compared to the contemporaneous one. The evolution of the number of links over time is shown in Fig. 4.19. There is a drop in the number of links around January 2014 for LB due to the appearance of the large unique group (thus no other group to link to, see Fig. 4.17).

The presence of links, valid under severe statistical inspection, clearly demonstrate the existence of predictability in the investors trading directions. We try to exploit this information in the following section to predict the investors' flux.

4.6 Flux prediction

The purpose of this section is to demonstrate the feasibility of accurate traders' flux forecasts. We simplify the task and ambition to predict only the sign of the clients future flux Φ_{t+1}. The flux is defined by:

$$\Phi_t = \sum_{i \in groups} V_i(t),$$

where $V_i(t)$ is the signed volume of agent i in hour t. The summation is to be taken only on the agents assigned to a group. It is known, in the machine learning

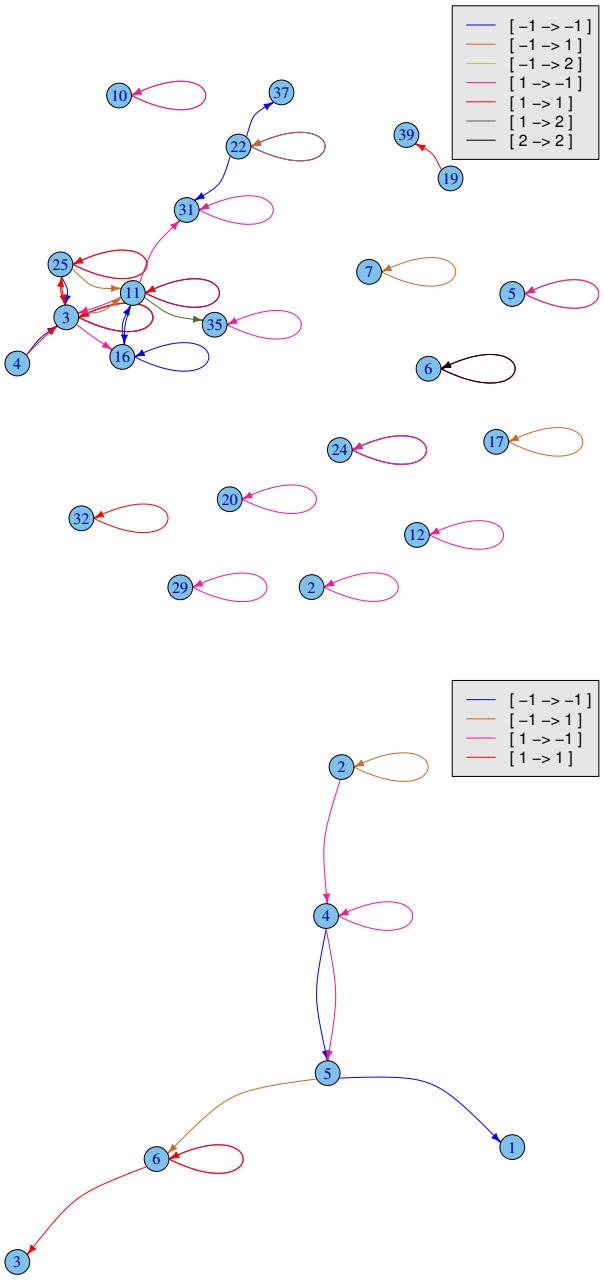

Figure 4.18: Examples of statistically validated lead-lag networks at the hourly scale. Top: SQ EUR/USD. Bottom: LB EUR/USD.

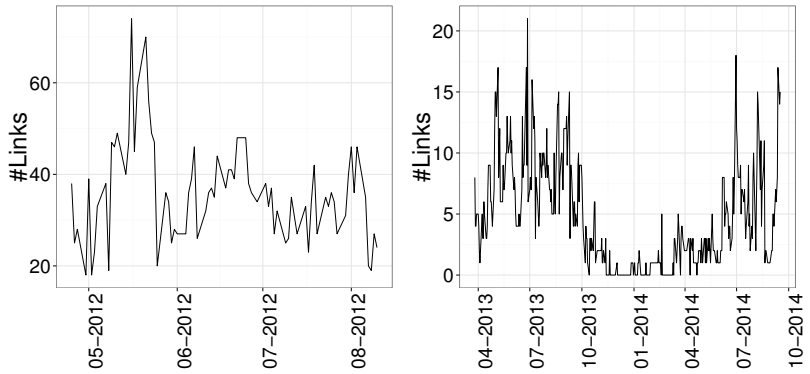

Figure 4.19: Lead-Lag network of clients: characteristics over time. Left: SQ. Right: LB.

literature [27] that clustering the features before applying classification algorithms improve their accuracy, it also reduces the computational burden. We therefore use the groups detected in the previous sections to perform the task. As no theory hints at a possible dependence structure between the future flux and current traders' actions, we use a very generic data mining method, namely random forests (RF). RF are a fully non-parametric statistical method developed by Breiman [21]. In short, RF are a bagging of decision trees, stabilizing model estimates and avoiding one of single decision tree curse: over-fitting. RF are robust, completely non-linear and posses a very good predictive accuracy on many "standard" datasets [43]. It is to the best of our knowledge the first attempt to forecast investors imbalance with this kind of techniques. It is also a novel approach to use SVN as a clustering layer before a classification task. The standard way being the use of PCA or K-means clustering.

4.6.1 Setup

In the section, we describe the "experimental" setup. The procedure is applied on rolling windows of 4 months (T_{in}) to predict the client flux over each hour of the next day. On each training period, the 500 most active traders are selected and the procedure described in Sec. 4.5.1 is applied at the hourly scale. We construct the groups strategy matrix: each column is a time-series $s_{g_i}(t)$ (see Sec. 4.5.3.1). This matrix constitutes the basis of our predictors. Several columns are added: lagged versions of those predictors (up to 10), in order to take into account effects at timescales larger than one hour and the hour of the day, because it is an important

driver of the total flux (see intraday patterns in Sec. 4.3.2). Each line of this large matrix represents the value of all the predictors for a particular hour t in the training set. The response function is the vector: $\text{sign}\left(\Phi_{t+1}\right)_{t\in T_{in}} \in \{-1, 0, 1\}$, an hour after the corresponding line in the predictors matrix. The predictors matrix and the vector $\left(\Phi_{t+1}\right)_{t\in T_{in}}$ are *in-sample* data. The RF learns how to predict the response function from the predictors, a step named "growing the forest" in RF's terminology. Finally, we apply the grown forest on the day following the training period, and predict $\text{sign}\left(\Phi_{t+1}\right)$ each hour. The model is updated only on a daily basis to save computation time and because we do not expect the model estimates to change much when only one hour of data out of 4 months is modified. It is important to note that the predictions are entirely made *out-of-sample*.

4.6.2 Benchmarks

To evaluate the above procedure, we merge the predictions from all the out-of-sample days and plot the cumulative sum of a success/failure variable (x_t). x_t equals $+1$ if the predicted sign is equal to the realized sign ($s_t = \text{sign}\Phi_t \in \{-1, 0, 1\}$), -1 otherwise. If is also interesting to weight x_t by the realized volume ($x_t = \pm \mid \Phi_t \mid$) to assess if the performance is higher when the realized volumes are large. It is crucial for a market-maker because large volumes potentially bear more benefits and more risks.

Let us denote x_t's cumulative sum by X_t,

$$X_t = \sum_{\tau=1}^{t} x_\tau.$$

It is imperative to compare X_t to a value obtained by randomness. A naive null hypothesis ($H_0^{(1)}$) is a prediction made by choosing one of the three outcomes with equal probabilities, meaning $p_{-1} = p_0 = p_1 = \frac{1}{3}$. A more sophisticated one ($H_0^{(2)}$), is to use different probabilities to take into account the unconditional statistics of Φ_t. The probability of each outcome is estimated in-sample and then used to forecast the following day:

$$\hat{p}_{-1} = \frac{1}{T_{in}} \sum_{t \in T_{in}} \mathbf{1}_{\{\Phi_t < 0\}},$$
$$\hat{p}_0 = \frac{1}{T_{in}} \sum_{t \in T_{in}} \mathbf{1}_{\{\Phi_t = 0\}},$$
$$\hat{p}_1 = \frac{1}{T_{in}} \sum_{t \in T_{in}} \mathbf{1}_{\{\Phi_t > 0\}}.$$

A final refinement ($H_0^{(3)}$), is to estimate the previous probabilities conditionally on the hour of the day. The forecast is then performed according to the foreseen hour. We compare the RF's forecasts to the following theoretical bounds: $\mathbb{E}[X_t] \pm \sigma_{X_t}$, for these three scenarios.

Under $H_0^{(1)}$, the expected value of the increment x_t is given by:

$$\mathbb{E}[x_t] = \frac{1}{3}(+1) + \frac{2}{3}(-1) = -\frac{1}{3}$$

and the variance:

$$\sigma_{x_t}^2 = \frac{1}{3}(+1)^2 + \frac{2}{3}(-1)^2 - \frac{1}{9} = \frac{8}{9}$$

The bounds are therefore: $X_{\pm}^{(1)} = -\frac{t}{3} \pm \frac{1}{3}\sqrt{8t}$. With the weighted metric, we easily obtain: $X_{\pm}^{(1)} = -\frac{1}{3}\sum_{\tau=1}^{t} \mid \Phi_\tau \mid \pm \frac{1}{3}\sqrt{8 \sum_{t=\tau}^{t} \Phi_\tau^2}$.

Under $H_0^{(2)}$, the expected value of the increment x_t depends explicitly on the realized sign s_t and is given by:

$$\mathbb{E}[x_t] = (\hat{p}_{s_t}(+1) + (1 - \hat{p}_{s_t})(-1) = 2\hat{p}_{s_t} - 1,$$

where the probabilities depends on the time because they are updated every day.The variance is given by:

$$\sigma_{x_t}^2 = \hat{p}_{s_t}(+1)^2 + (1 - \hat{p}_{s_t})(-1)^2 - (2\hat{p}_{s_t} - 1)^2 = 1 - (2\hat{p}_{s_t} - 1)^2 = 4\hat{p}_{s_t}(1 - \hat{p}_{s_t}).$$

The bounds are therefore:

$$X_{\pm}^{(2)} = \sum_{\tau=1}^{t}(2\hat{p}_{s_\tau} - 1) \pm \sqrt{\sum_{\tau=1}^{t}(4\hat{p}_{s_\tau}(1 - \hat{p}_{s_\tau}))},$$

and with the weights,

$$X_{\pm}^{(2)} = \sum_{\tau=1}^{t}(2\hat{p}_{s_\tau} - 1) \mid \Phi_\tau \mid \pm\sqrt{\sum_{\tau=1}^{t}\left(4\hat{p}_{s_\tau}(1-\hat{p}_{s_\tau})\right)\Phi_\tau^2}.$$

For the third scheme $X_{\pm}^{(3)}$, the formula are exactly the same as $X_{\pm}^{(2)}$ but with $\hat{p}_s$ estimated conditionally on the hour.

4.6.3 Results

We run the procedure on each dataset. In every figure of this section, the predictions are in black and the theoretical bounds are the dashed lines in red (H_0^1), blue (H_0^2) and green (H_0^3). In Fig. 4.20, we show the results for the EUR/USD pair. The performance for SQ is not significant whereas for LB, the accuracy is satisfactory and above the theoretical bounds. Interestingly for SQ, $X_{\pm}^{(2)}$ and $X_{\pm}^{(3)}$ are almost superimposed and symmetric about the x-axis, meaning an almost 50-50 repartition between buy and sell in each hour of the training set. This lack of trend might explain the forecasting difficulties. An other possible explanation is that institutional investors are more predictable than erratic retail traders.

Though not stunning which is understandable due to the noisy nature of financial data, it is nevertheless valuable for the broker-dealer because of the enormous volume at stakes. For LB, it means a net 10 billions of euros anticipated in the right direction over a year and a half. Also it might improve the internal matching of customer orders. For example a predicted -1 after a realized $+1$ gives an incentive to keep the inventory in the hope of offloading it with future customers, at the opposite, a predicted $+1$ or 0 might give a signal to hit the interdealer market depending on the defined level of acceptable inventory.

We cross validate the results with different currency pairs for LB. Unfortunately for SQ, we have required data only for EUR/USD. Fig. 4.21 indicates good performances.

A different way to look at the results is to analyses the performance conditioned on the hour of the day. In Fig. 4.22, we compute the t-stat of x_t conditional on the hour of the day. We notice a tendency to perform well in periods of high activity. An easy way to improve the forecast accuracy is then to predict only the "good" hours.

We highlight two remarkable features of the random forest prediction to end this section. Adding the returns to the feature set did not improve the forecast, probably because the returns effect are already embedded in groups actions, in line with the factor analysis of Sec. 4.4.5. Remarkably, multiple runs of the procedure revealed that there is no need to update the model every day: an

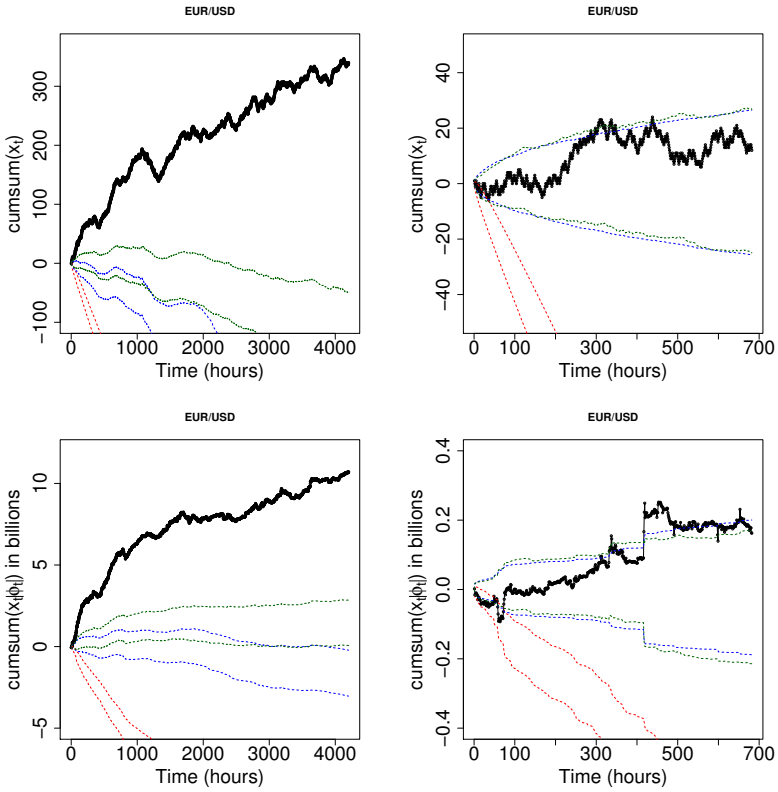

Figure 4.20: EUR/USD customer flux predictions successes. Left: LB dataset. Right: SQ dataset. The bottom row presents the volume weighted metric. The random forest (black) outperforms the benchmarks (dashed lines: red (H_0^1), blue (H_0^2) and green (H_0^3)) for LB and achieve a remarkable performance whereas forecasts for SQ are within the boundaries.

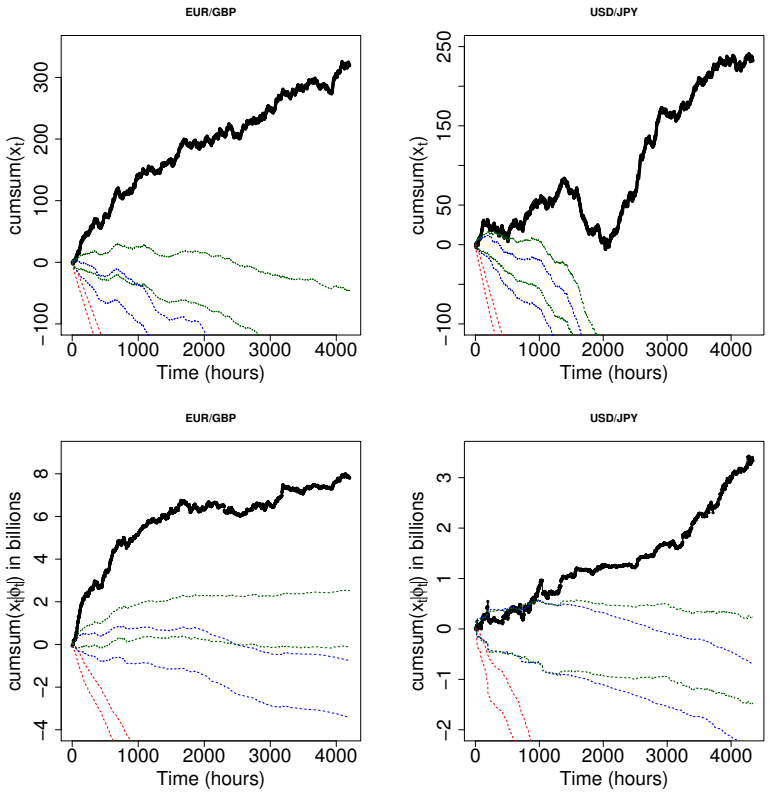

Figure 4.21: LB customer flux predictions. Left: EUR/GBP. Right: USD/JPY. The random forest (black) outperforms the benchmarks (dashed lines: red (H_0^1), blue (H_0^2) and green (H_0^3)) for both pairs.

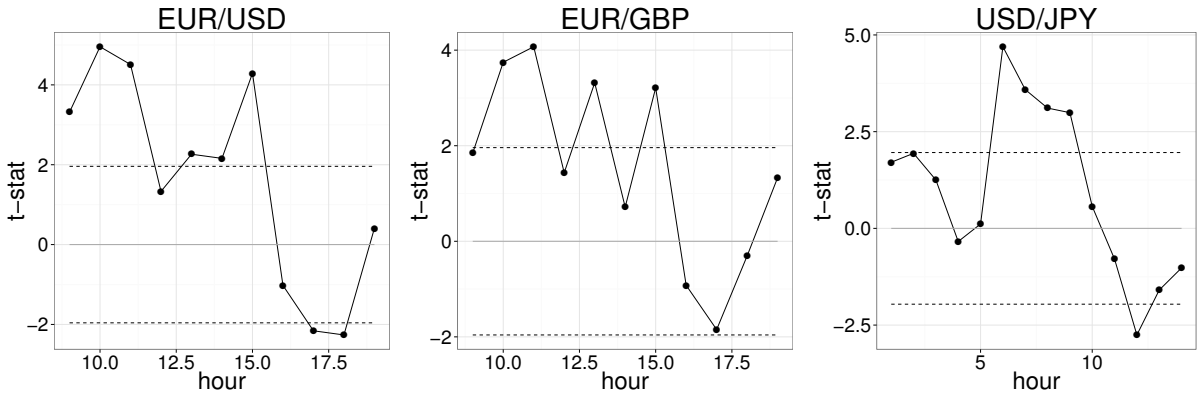

Figure 4.22: x_t's t-stat per hour for LB.

update every 5 days gave approximately the same results. It is an important point for the practitioner in order reduce computation time opening the way for real-time applications.

4.6.4 Future improvements

In the previous section, we tried to give a "proof of principle", it demonstrates from an academic point of view that an exploitable structure exists in investors data. Needless to say, any application in banking needs more work to improve forecasts performances and robustness. Possible directions are: a fine tuning of the parameters (training period, number of considered lags, update period, statistical threshold in the groups construction,...) and variable selection trough variable importance score.

One may also consider "oblique" random forests (oRF). Its principle is identical of the RF one, but the decision trees are grown differently. While in the standard RF, decision trees separate the feature space by hyperplanes that are orthogonal to single feature axes, oRF trees use arbitrary split directions. More details can be found in [83]. Preliminary results with oRF shows small improvements.

4.7 Conclusion and perspectives

Despite the high variability in FX traders behavior, we have shown in this work that trading co-occurrences are more frequent than expected from pure statistical fluctuations. Thanks to SVNs, we have been able to uncover communities for two types of FX traders: institutions and individuals. Moreover, the existence

of a lead-lag structure in the form of a inter-temporal network between these communities has been established. There is no obvious reason for this result to be specific to the FX market, its extrapolation to other asset classes is therefore tempting. Nevertheless, confirmation is needed from future research.

The origin of the groups lie into explicit communication (exchange of information, analyst recommendations) but more importantly in the implicit one. Through the use of similar strategies on the same price, a high degree of synchronization appears between investors. Finding the strategies corresponding to each detected group is the future challenging step of this work. Besides, a thorough analysis of the groups stability over time is needed.

The application of random forests explicitly confirmed that the groups' buying or selling states bear information about future customer's flux. The natural next step is to consider the flux amplitude on top of its sign turning the forecast procedure into a regression task.

It would be relevant to compare these results to recent methods from social networks studies aimed at unveiling latent network structure thanks to large multidimensional Hawkes processes [75, 8]. These methods are hard to apply in the present form to heterogeneous individual traders because of the inherent difficulties involved in Hawkes process estimation (see chapter 3 and Ref [45]). However, it is an appealing direction as it would be a bridge between several topics addressed in the thesis. It would also raise the very interesting question of individual traders self and cross kernels. Do they have any meaning at all? What are there shape, exponential, power-law? And more importantly, how are they translated into to the market global kernel. Answering these questions would shed new lights on the connections between the customer and the interdealer market.

Appendix A

Simulations

We simulate a Hawkes process with a ϕ_2 kernel with parameters similar to those of hourly fits on real data: we set $\mu = 0.05$, $n_1 = 0.37$, $n_2 = 0.42$, τ_2 =21 s and vary τ_1 from 0.05 s to 1.5 s. For each value of τ_1 we perform 50 simulations of 22 hours. Then, on each simulated time-series, we artificially reduce the data resolution to 0.1 s, introducing time slicing as in our data set, and then randomize the timestamps within each time slice in order to mimic the procedure applied on empirical data (see Sec. 3.3.2). We fit each resulting time-series and average the results over the 50 runs with two- and three-exponential kernels.

Figure A.1 reports the fitted smallest time scale as a function of the original time scale and shows that the shuffling of time stamps within an interval leads artifically increases the apparent smallest time scale, particularly (and quite expectedly) for small τ_1^{sim}. Nevertheless, this increase is small, of the order of 15%. In addition, shuffling does not introduce a spurious third time scale, as fits with kernels with three exponentials did not yield any third time scale.

Finally, Fig. A.2 shows that only the Kolmogorov-Smirnov p-values are affected by the time slicing and time stamp shuffling within a time slice. Nevertheless, at $\tau_1^{fit} = 0.15$s , pKS is still larger than 0.05. This is consistent with fits on real data: significance is possible, but limited time resolution does not help.

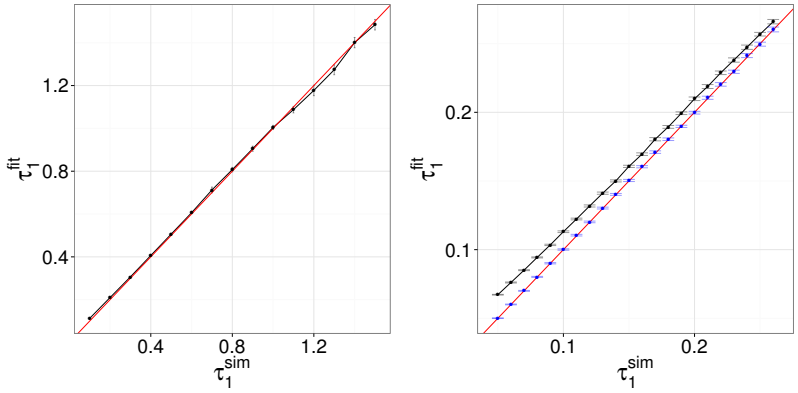

Figure A.1: Fitted short timescale (black points) versus simulated short timescale. In red, the $y = x$ line. Right plot is a zoom on the critical region (close to 0.1 s). Blue points are the fitted values without the slicing procedure. Small distortion in the short timescale determination. Error bars set at two standard deviations..

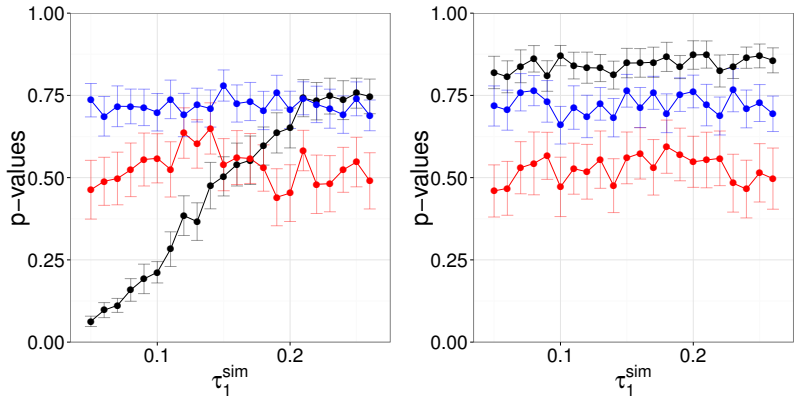

Figure A.2: Fits p-values for Kolmogorov-Smirnov test (black), Ljung-Box test (red) and Excess-Dispersion test (blue). Left plot: with time stamp shuffling within a time slice. Right plot: without shuffling. Only the Kolmogorov-Smirnov p-value is affected by the data bundling. Error bars set at two standard deviations.

Bibliography

[1] H.-J. Ahn, J. Cai, and Y. L. Cheung. Price clustering on the limit-order book: Evidence from the Stock Exchange of Hong Kong. *Journal of Financial Markets*, 8(4):421–451, Nov. 2005.

[2] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. A. Laeven. Modeling Financial Contagion Using Mutually Exciting Jump Processes. *National Bureau of Economic Research Working Paper Series*, No. 15850, 2010.

[3] T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Realized Volatility and Correlation. *Manuscript, Northwestern University, Duke University and University of Pennsylvania. Published in revised form as "Great Realizations," Risk, March 2000, 105-108.*, (March), 2000.

[4] E. Bacry and J.-F. Muzy. Second order statistics characterization of Hawkes processes and non-parametric estimation. *arXiv:1401.0903*, Jan. 2014.

[5] E. Bacry and J.-F. Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, pages 1–20, Apr. 2014.

[6] E. Bacry, K. Dayri, and J. F. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85(5):157, May 2012.

[7] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13 (1):65–77, Jan. 2012.

[8] E. Bacry, S. Gaïffas, and J.-F. Muzy. A generalization error bound for sparse and low-rank multivariate Hawkes processes. *arXiv:1501.00725*, Jan. 2015.

[9] C. A. Ball, W. N. Torous, and A. E. Tschoegl. The degree of price resolution: The case of the gold market. *Journal of Futures Markets*, 5(1):29–43, Mar. 1985.

[10] Bank for International Settlements. Triennial-Central Bank Survey-Report on global foreign exchange market activity in 2010. Technical Report December, 2010.

[11] L. Bauwens and N. Hautsch. Dynamic Latent Factor Models for Intensity Processes. *CORE Discussion Paper*, 103, Feb. 2003.

[12] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, Jan. 1995.

[13] D. Berger, A. Chaboud, and E. Hjalmarsson. What drives volatility persistence in the foreign exchange market? *Journal of Financial Economics*, 94 (2):192–213, Nov. 2009.

[14] D. W. Berger, A. P. Chaboud, S. V. Chernenko, E. Howorka, and J. H. Wright. Order flow and exchange rate dynamics in electronic brokerage system data. *Journal of International Economics*, 75(1):93–109, May 2008.

[15] B. Biais, P. Hillion, and C. Spatt. An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse. *The Journal of Finance*, 50 (5):1655–1689, 1995.

[16] BIS. Triennial Central Bank Survey. 2013.

[17] G. Bormetti, L. M. Calcagnile, M. Treccani, F. Corsi, S. Marmi, and F. Lillo. Modelling systemic price cojumps with Hawkes factor models. *Quantitative Finance*, 15(7):1137–1156, Mar. 2015.

[18] J.-P. Bouchaud, M. Mézard, and M. Potters. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256, Aug. 2002.

[19] J.-P. Bouchaud, J. D. Farmer, and F. Lillo. CHAPTER 2 - How Markets Slowly Digest Changes in Supply and Demand. In T. H. R. B. T. H. o. F. M. D. Schenk-Hoppé and Evolution, editors, *Handbooks in Finance*, pages 57–160. North-Holland, San Diego, 2009. ISBN 15684997.

[20] C. G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, Dec. 2007.

[21] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[22] P. Brémaud and L. Massoulié. Hawkes Branching Point Processes without Ancestors. *Journal of Applied Probability*, 38(1):122–135, Mar. 2001.

[23] E. N. Brown, R. Barbieri, V. Ventura, R. E. Kass, and L. M. Frank. The Time-Rescaling Theorem and Its Application to Neural Spike Train Data Analysis. *Neural Computation*, 14(2):325–346, Feb. 2002.

[24] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, Sept. 1995.

[25] A. Cellier and D. Bourghelle. Limit Order Clustering and Price Barriers on Financial Markets: Empirical Evidence from Euronext. In *AFFI, Conférence internationale Brest*, Feb. 2009.

[26] A. Chakraborti, I. M. Toke, M. Patriarca, and F. Abergel. Econophysics review: I. Empirical facts. *Quantitative Finance*, 11(7):991–1012, June 2011.

[27] M. Chavent, R. Genuer, and J. Saracco. Combining clustering of variables and random forests for high-dimensional supervised classification. In *COMPSTAT*, 2012.

[28] V. Chavez-Demoulin and J. A. McGill. High-frequency financial data modeling using Hawkes processes. *Journal of Banking & Finance*, 36(12):3415–3426, Dec. 2012.

[29] V. Chavez-Demoulin, A. C. Davison, and A. J. McNeil. Estimating value-at-risk: a point process approach. *Quantitative Finance*, 5(2):227–234, Apr. 2005.

[30] E. S. Chornoboy, L. P. Schramm, and A. F. Karr. Maximum likelihood identification of neural point process systems. *Biological Cybernetics*, 59 (4-5):265–275, 1988.

[31] W. G. Christie and P. H. Schultz. Why do NASDAQ Market Makers Avoid Odd-Eighth Quotes? *The Journal of Finance*, 49(5):1813–1840, Dec. 1994.

[32] W. G. Christie, J. H. Harris, and P. H. Schultz. Why did NASDAQ Market Makers Stop Avoiding Odd-Eighth Quotes? *The Journal of Finance*, 49(5): 1841–1860, Dec. 1994.

[33] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, Feb. 2001.

[34] J. W. Cooney Jr., B. Van Ness, and R. Van Ness. Do investors prefer even-eighth prices? Evidence from NYSE limit orders. *Journal of Banking & Finance*, 27(4):719–748, Apr. 2003.

[35] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, Oct. 2008.

[36] G. Curato, F. Lillo, and P. Pennesi. Order flow and exchange rate dynamics in interdealer FX market. *Unpublished*, 2014.

[37] D. M. Cutler, J. M. Poterba, and L. H. Summers. What Moves Stock Prices? Working Paper 2538, National Bureau of Economic Research, 1988.

[38] J. Da Fonseca and R. Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 2013.

[39] M. M. Dacorogna, U. A. Müller, R. J. Nagler, R. B. Olsen, and O. V. Pictet. A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance*, 12(4):413–438, Aug. 1993.

[40] M. M. Dacorogna, R. Gençay, U. A. Müller, R. B. Olsen, and O. V. Pictet. *An Introduction to High-Frequency Finance*. Academic Press, San Diego, 2001. ISBN 978-0-12-279671-5.

[41] R. F. Engle and J. R. Russell. Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, 66(5): 1127–1162, 1998.

[42] E. Errais, K. Giesecke, and L. Goldberg. Affine Point Processes and Portfolio Credit Risk. *SIAM Journal on Financial Mathematics*, 1(1):642–665, Jan. 2010.

[43] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

[44] V. Filimonov and D. Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, May 2012.

[45] V. Filimonov and D. Sornette. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *arxiv:1308.6756*, Aug. 2013.

[46] C. S. Gillespie. *Fitting heavy tailed distributions: the poweRlaw package*, 2014.

[47] M. A. Goldstein and K. A. Kavajecz. Eighths, sixteenths, and market depth: changes in tick size and liquidity provision on the NYSE. *Journal of Financial Economics*, 56(1):125–149, Apr. 2000.

[48] C. Goodhart and R. Curcio. The clustering of bid-ask prices and the spread in the foreign exchange market. *LSE Financial Market Group Discussion Paper*, 110, 1991.

[49] P. Gopikrishnan, V. Plerou, L. A. Nunes Amaral, M. Meyer, and H. E. Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60(5):5305–5316, Nov. 1999.

[50] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, Aug. 2013.

[51] D. M. Guillaume, M. M. Dacorogna, R. R. Davé, U. A. Müller, R. B. Olsen, and O. V. Pictet. From the bird's eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets. *Finance and Stochastics*, 1(2):95–129, 1997.

[52] M. Gutiérrez-Roig and J. Perelló. Volatility polarization of non-specialized investors' heterogeneous activity. *arXiv:1302.3169*, Feb. 2013.

[53] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud. Critical reflexivity in financial markets: a Hawkes process analysis. *Eur. Phys. J. B*, 86(10), Oct. 2013.

[54] L. Harris. Stock Price Clustering and Discreteness. *The Review of Financial Studies*, 4(3):389–415, Jan. 1991.

[55] Y. Hashimoto, T. Ito, T. Ohnishi, M. Takayasu, H. Takayasu, and T. Watanabe. Random walk or a run. Market microstructure analysis of foreign exchange rate movements based on conditional probability. *Quantitative Finance*, 12(6):893–905, Dec. 2010.

[56] A. G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1):83–90, Apr. 1971.

[57] A. G. Hawkes. Point Spectra of Some Mutually Exciting Point Processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3): 438–443, Jan. 1971.

[58] P. Hewlett. Clustering of order arrivals, price impact and trade path optimisation. In *Workshop on Financial Modeling with Jump Processes*, 2006.

[59] N. Huth and F. Abergel. The times change: multivariate subordination. Empirical facts. *Quantitative Finance*, 12(1):1–10, Jan. 2012.

[60] D. L. Ikenberry and J. P. Weston. Clustering in US Stock Prices after Decimalisation. *European Financial Management*, 14(1):30–54, Jan. 2008.

[61] G. Iori, R. Renò, G. De Masi, and G. Caldarelli. Trading strategies in the Italian interbank market. *Physica A: Statistical Mechanics and its Applications*, 376:467–479, Mar. 2007.

[62] T. Ito and Y. Hashimoto. Intraday seasonality in activities of the foreign exchange markets: Evidence from the electronic broking system. *Journal of the Japanese and International Economies*, 20(4):637–664, Dec. 2006.

[63] T. Jaisson and M. Rosenbaum. Limit theorems for nearly unstable Hawkes processes. *arXiv:1310.2033*, Oct. 2013.

[64] A. Jedidi and F. Abergel. On the Stability and Price Scaling Limit of a Hawkes Process-Based Order Book Model. *SSRN Electronic Journal*, May 2013.

[65] Z.-Q. Jiang and W.-X. Zhou. Complex stock trading network among investors. *Physica A: Statistical Mechanics and its Applications*, 389(21): 4929–4941, Nov. 2010.

[66] A. Joulin, A. Lefevre, D. Grunberg, and J.-P. Bouchaud. Stock price jumps: news and volume play a minor role. *arXiv:0803.1769*, Mar. 2008.

[67] M. R. King, C. Osler, and D. Rime. Foreign Exchange Market Structure, Players, and Evolution. In *Handbook of Exchange Rates*, pages 1–44. John Wiley & Sons, Inc., 2012. ISBN 9781118445785.

[68] R. Kozhan and M. Salmon. The information content of a limit order book: The case of an FX market. *Journal of Financial Markets*, 15(1):1–28, Feb. 2012.

[69] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise Dressing of Financial Correlation Matrices. *Phys. Rev. Lett.*, 83(7):1467–1470, 1999.

[70] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80(5):56117, Nov. 2009.

[71] J. Large. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1):1–25, Feb. 2007.

[72] E. Lewis and G. Mohler. A Nonparametric EM algorithm for Multiscale Hawkes Processes. *Submitted*, 2011.

[73] F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics Econometrics*, 8(3), 2004.

[74] F. Lillo, E. Moro, G. Vaglica, and R. N. Mantegna. Specialization and herding behavior of trading firms in a financial market. *New Journal of Physics*, 10(4):43019, 2008.

[75] S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. *arXiv:1402.0914*, Feb. 2014.

[76] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, Aug. 1978.

[77] I. Lo and S. G. Sapp. The submission of limit orders or market orders: The role of timing and information in the Reuters D2000-2 system. *Journal of International Money and Finance*, 27(7):1056–1073, Nov. 2008.

[78] I. Lo and S. G. Sapp. Order aggressiveness and quantity: How are they determined in a limit order market? *Journal of International Financial Markets, Institutions and Money*, 20(3):213–237, July 2010.

[79] B. Mandelbrot. The Variation of Certain Speculative Prices. *The Journal of Business*, 36(4):394–419, Oct. 1963.

[80] D. Marsan and O. Lengliné. Extending Earthquakes' Reach Through Cascading. *Science*, 319(5866):1076–1079, Feb. 2008.

[81] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[82] S. Maslov and M. Mills. Price fluctuations from the order book perspective, empirical facts and a simple model. *Physica A: Statistical Mechanics and its Applications*, 299(1-2):234–246, Oct. 2001.

[83] B. Menze, B. Kelm, D. Splitthoff, U. Koethe, and F. Hamprecht. On Oblique Random Forests. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases SE - 29*, volume 6912 of *Lecture Notes in Computer Science*, pages 453–469. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23782-9.

[84] J. Mitchell and H. Y. Izan. Clustering and psychological barriers in exchange rates. *Journal of International Financial Markets, Institutions and Money*, 16(4):318–344, Oct. 2006.

[85] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, 106(493):100–108, Mar. 2011.

[86] G.-H. Mu, W. Chen, J. Kertész, and W.-X. Zhou. Preferred numbers and the distributions of trade sizes and trading volumes in the Chinese stock market. *The European Physical Journal B*, 68(1):145–152, 2009.

[87] U. A. Müller, M. M. Dacorogna, R. B. Olsen, O. V. Pictet, M. Schwarz, and C. Morgenegg. Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis. *Journal of Banking & Finance*, 14(6):1189–1208, Dec. 1990.

[88] V. Niederhoffer. Clustering of Stock Prices. *Operations Research*, 13(2): 258–265, Mar. 1965.

[89] V. Niederhoffer. A New Look at Clustering of Stock Prices. *The Journal of Business*, 39(2):309–313, Apr. 1966.

[90] Y. Ogata. On Lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31, 1981.

[91] Y. Ogata. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association*, 83(401):9–27, Mar. 1988.

[92] Y. Ogata. Seismicity Analysis through Point-process Modeling: A Review. *pure and applied geophysics*, 155(2-4):471–507, 1999.

[93] J.-P. Onnela, J. Töyli, and K. Kaski. Tick size and stock returns. *Physica A: Statistical Mechanics and its Applications*, 388(4):441–454, Feb. 2009.

[94] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, USA, 1999. ISBN 0-13-754920-2.

[95] M. F. M. Osborne. Periodic Structure in the Brownian Motion of Stock Prices. *Operations Research*, 10(3):345–379, May 1962.

[96] C. Osler and T. Savaser. Extreme returns: The case of currencies. *Journal of Banking & Finance*, 35(11):2868–2880, Nov. 2011.

[97] C. L. Osler. Currency Orders and Exchange Rate Dynamics: An Explanation for the Predictive Success of Technical Analysis. *The Journal of Finance*, 58(5):1791–1819, Oct. 2003.

[98] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

[99] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter. Recurrent interactions in spiking networks with arbitrary topology. *Physical Review E*, 85(3):31916, Mar. 2012.

[100] M. Rambaldi, P. Pennesi, and F. Lillo. Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. *Physical Review E*, 91(1):12819, Jan. 2015.

[101] P. Reynaud-Bouret and S. Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5): 2781–2822, 2010.

[102] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23, Jan. 2008.

[103] B. J. Sopranzetti and V. Datar. Price clustering in foreign exchange spot markets. *Journal of Financial Markets*, 5(4):411–417, Oct. 2002.

[104] G. Soros. *The Alchemy of Finance: Reding the Mind of the Market*. 1987.

[105] I. M. Toke and F. Pomponio. Modelling Trades-Through in a Limited Order Book Using Hawkes Processes Trades-through. *Economics E-journal*, 32, 2011.

[106] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, and R. N. Mantegna. Statistically Validated Networks in Bipartite Complex Systems. *PLoS ONE*, 6 (3):e17994, Mar. 2011.

[107] M. Tumminello, F. Lillo, J. Piilo, and R. N. Mantegna. Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, 14, 2012.

[108] E.-J. Wagenmakers and S. Farrell. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004.

[109] W. Wasserfallen and H. Zimmermann. The behavior of intra-daily exchange rates. *Journal of Banking & Finance*, 9(1):55–72, Mar. 1985.

[110] J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, and J. P. Sethna. Sloppy-Model Universality Class and the Vandermonde Matrix. *Physical Review Letters*, 97(15):150601, Oct. 2006.

[111] S. Yang and H. Zha. Mixture of Mutually Exciting Processes for Viral Diffusion. *Journal of Machine Learning Research*, 28(2):1–9, 2013.

[112] W.-X. Zhou, G.-H. Mu, and J. Kertesz. Random matrix approach to the dynamics of stock inventory variations. *New Journal of Physics*, 14(9): 93025, 2012.

[113] I. Zovko and J. D. Farmer. The power of patience: a behavioural regularity in limit-order placement. *Quantitative Finance*, 2(5):387–392, Oct. 2002.

[114] I. Zovko and J. D. Farmer. Correlations and clustering in the trading of members of the London Stock Exchange. In *AIP Conference Proceedings*, volume 965, pages 287–299. AIP, 2007.