CentraleSupélec

Thèse de doctorat

Spécialité :
Mathématiques Appliquées

Laboratoire d'accueil :
Mathématiques Appliquées aux Systèmes

présentée par

Marouane Anane

pour l'obtention du
GRADE DE DOCTEUR

Une approche mathématique de l'investissement boursier

A Mathematical Approach To Stock Investing

Dirigée par Frédéric Abergel

Soutenue publiquement le 10 Février 2015 devant le jury composé de :

Frédéric ABERGEL	CentraleSupélec	*Directeur de Thèse*
Anirban CHAKRABORTI	Jawaharlal Nehru University	*Rapporteur*
Nicolas VAYATIS	ENS Cachan	*Rapporteur*
Damien CHALLET	CentraleSupélec	*Examinateur*
Charles-Albert LEHALLE	Capital Fund Management	*Examinateur*
Éric MOULINES	Télécom ParisTech	*Examinateur*

À mes parents, à ma soeur, à ma femme.

Remerciements

En écrivant cette page, je me rends compte que pendant ces trois années de thèse j'ai eu la chance et le plaisir de côtoyer des dizaines d'amis et de collègues. Chaque personne, a eu une contribution non négligeable dans ma formation. Je tiens donc à les remercier nominativement dans les lignes qui suivent.

Je tiens, d'abord, à remercier mon directeur de thèse Frédéric ABERGEL sans qui ce manuscrit n'aurait jamais abouti. Je suis ravi d'avoir eu la chance de travailler avec lui et d'apprendre de son expérience très riche. En plus de son encadrement et sa disponibilité pour mes travaux de thèse, il m'a introduit dans d'autres domaines, qui m'étaient jusqu'à lors inconnus, tels que l'enseignement, la participation aux conférences internationales et la publication dans les revues scientifiques. C'est grâce à lui que cette thèse fut une expérience très enrichissante autant scientifiquement que humainement.

Je tiens également à remercier Anirban CHAKRABORTI et Nicolas VAYATIS d'avoir aimablement accepté d'être rapporteurs de cette thèse. Leurs remarques m'étaient très constructives et m'ont permis de finaliser au mieux ce travail. Je remercie aussi, très chaleureusement, Damien CHALLET, Charles-Albert LEHALLE et Éric MOULINES d'avoir aimablement accepté d'honorer ma soutenance.

Je remercie vivement mon responsable à BNP Paribas Sébastien LEFORT de m'avoir accueilli pendant quatre ans au sein de son équipe. Au cours de cette très agréable expérience, j'ai énormément appris de sa vaste connaissance des marchés financiers. J'ai particulièrement apprécié son esprit de partage de connaissances ainsi que la confiance qu'il accorde aux nouvelles idées. Cette confiance est sans doute l'un des principaux moteurs de l'évolution de l'équipe.

Dans ce même cadre, je tiens à remercier mon responsable à BNP Paribas Franck MICHEL de m'avoir donné l'occasion de travailler sur des sujets clefs pour l'activité de Market Making. J'ai particulièrement apprécié sa grande maîtrise des enjeux du monde de la haute fréquence, son ouverture d'esprit et sa capacité de mettre toutes les ressources humaines (stagiaires, support..) et matérielles (système informatique à point..) au service de la recherche et de l'innovation.

Je tiens à remercier également, tous les responsables de BNP Paribas qui encouragent vivement les travaux de recherche. Les thésards ont à leur disposition d'énormes ressources matérielles et logistiques (données financières, grille de calcul, plusieurs ordinateurs par personne, support informatique réactif..) assurant, au quotidien, les meilleures conditions de travail. Par ailleurs, je tiens à saluer la très bonne perception qu'accorde la BNP aux thésards autant sur le plan humain que financier. Dans ce cadre je remercie nominativement; Stéphane ANDRE, Alexandre BENECH, Thibaut DELAHAYE, Olivier OSTY et Geoffrey RODRIGUE.

Je pense également à Axel BREUER qui m'a encadré au cours de mon stage de fin d'études et au début de cette thèse, ainsi qu'à d'autres collègues, avec lesquels j'ai travaillé, qui sont partis pour d'autres aventures, notamment Olivier CLEMENTIN, Jaques Olivier MOUSSAFIR et Martin PLANES. Dans ce même cadre, je pense à Laurent GIORELLO et à Mehdi Laurent AKKAR avec lesquels j'ai eu des discussions particulièrement enrichissantes qui ont sans doute marqué ma manière de voir les choses.

Je souhaite remercier chaleureusement mes collègues de l'équipe *Automatic Market Making* avec lesquels j'ai eu le plaisir de travailler sur différents projets, notamment: Vincent BAZINETTE, Khalil BENATIYA, Jeff BJORAKER, Guillaume BIOCHE, Carole BOUGEANT, Nicolas BOUISSET, Benoit COLEDAN, Alexandre DAVROUX, Christian DIDION, Kamal FAIK, Jonathan GIARMON, Pierre GIREAU, Jimmy KONIETZKO, Olivier RICAUD, Alexandre SAINTVILLE, Thomas SENNEVILLE, Laurent VACCA et Thanh-Niem VU.

Je n'oublie pas les amis des équipes de l'informatique qui m'ont aidé à m'améliorer techniquement: Gabriel AH-TUNE, Eva ATTAL, Clement CUNIN, Karim GARDABOU, Bernard HELMSTETTER, Bruno HESS, David JOBET, Jerome JOUVIE, Joachim JOYAUX, Benoit JUIN, François LEIBER, Cédric MABILLE, Julien MALDANT-SAVARY, Juan-Sébastien PENA-RODRIGUEZ, Hervé POUSSINEAU, Daniel TEPLY et Julien VIVENOT.

Je remercie aussi l'équipe de la recherche de BNP Paribas pour les différentes discussions intéressantes: Cédric JOULAIN, Grégoire LOEPER et Jean-Jacques RABEYRIN.

Du coté de l'école Centrale je souhaite remercier Emmanuelle COPLO, Sylvie DERVIN, Annie GLOMERON et Catherine LHOPITAL de m'avoir facilité les démarches administratives, Dany KOUOH-ETAME, Laurent SERIES et Laurent GUERBY pour leur support technique, Anirban CHAKRABORTI, Damien CHALLET, Dalia IBRAHIM et Sophie LARUELLE pour leurs discussions intéressantes. Je remercie aussi, mes compagnons thésards et ex-thésards, de la bonne ambiance: Ahmed BEL HADJ AYED, Rémy CHICHEPORTICHE, Joao DE GAMA BATISTA, Nicolas HUTH, Mehdi LALLOUACHE, Fabrizio POMPONIO et Ban ZHENG.

Je remercie profondément mes amis Sofiene EL AOUD, Aymen JEDIDI et Riadh ZAATOUR de leurs relectures de mes différents papiers ainsi que de leurs remarques judicieuses.

Je souhaite aussi remercier mes ex-responsables à Natixis qui m'ont introduit au monde de la finance: Adel BEN HAJ YEDDER et Adil REGHAI ainsi que mon professeur de théorie financière Olivier TARAMASCO.

Ce travail est dédié à ma mère qui m'a encouragé le plus à faire une thèse, à mon père qui m'a toujours poussé et soutenu tout au long de mes études, à ma sœur et à ma femme.

Résumé

Le but de cette thèse est de répondre au vrai besoin de prédire les fluctuations futures des prix d'actions. En effet, l'aléatoire régissant ces fluctuations constitue pour des acteurs de la finance, tels que les *Market Maker*, une des plus grandes sources de risque. Tout au long de cette étude, nous mettons en évidence la possibilité de réduire l'incertitude sur les prix futurs par l'usage des modèles mathématiques appropriés. Cette étude est rendue possible grâce à une grande base de données financières et une puissante grille de calcul mises à notre disposition par l'équipe *Automatic Market Making* de BNP Paribas. Dans ce document, nous présentons uniquement les résultats de la recherche concernant le trading haute fréquence. Les résultats concernant la partie basse fréquence présentent un intérêt scientifique moindre pour le monde académique et rentrent par ailleurs dans le cadre des résultats confidentiels. Ces résultats seront donc volontairement omis.

Dans le premier chapitre, nous présentons le contexte et les objectifs de cette étude. Nous présentons, également, les différentes méthodes utilisées, ainsi que les principaux résultats obtenus.

Dans le chapitre 2, nous nous intéressons à l'apport de la supériorité technologique en trading haute fréquence. Dans ce but, nous simulons un trader ultra rapide, omniscient, et agressif, puis nous calculons son gain total sur 3 ans. Les gains obtenus sont très modestes et reflètent l'apport limité de la technologie en trading haute fréquence. Ce résultat souligne l'intérêt primordial de la recherche et de la modélisation dans ce domaine.

Dans le chapitre 3, nous étudions la prédictibilité des prix à partir des indicateurs de carnet d'ordre. Nous présentons, à l'aide des espérances conditionnelles, des preuves empiriques de dépendances statistiques entre les prix et les différents indicateurs. L'importance de ces dépendances résulte de la simplicité de la méthode, éliminant tout risque de surapprentissage des données. Nous nous intéressons, ensuite, à la combinaison des différents indicateurs par une régression linéaire et nous analysons les différents problèmes numériques et statistiques liés à cette méthode. Enfin, nous concluons que les prix sont prédictibles pour un horizon de quelques minutes et nous mettons en question l'hypothèse de l'efficience du marché.

Dans le chapitre 4, nous nous intéressons au mécanisme de formation du prix à partir des arrivés des évènements dans le carnet d'ordre. Nous classifions les ordres en douze types dont nous analysons les propriétés statistiques. Nous étudions par la suite les dépendances entre ces différents types d'ordres et nous proposons un modèle de carnet d'ordre en ligne avec les observations empiriques. Enfin, nous utilisons ce modèle pour prédire les prix et nous appuyons l'hypothèse de la non-efficience des marchés, suggérée au chapitre 3.

Mots-clés: Trading haute fréquence, Microstructure, Apprentissage statistique, Régression linéaire, Processus de Hawkes, Backtest, Stratégies de trading.

Abstract

The aim of this thesis is to address the real need of predicting the prices of stocks. In fact, the randomness governing the evolution of prices is, for financial players like market makers, one of the largest sources of risk. In this context, we highlight the possibility of reducing the uncertainty of the future prices using appropriate mathematical models. This study was made possible by a large base of high frequency data and a powerful computational grid provided by the *Automatic Market Making* team at BNP Paribas. In this paper, we present only the results of high frequency tests. tests are of less scientific interest in the academic world and are confidential. Therefore, these results will be deliberately omitted.

In the first chapter, the background and the objectives of this study are presented along with the different methods used and the main results obtained.

The focus of chapter 2 is on the contribution of technological superiority in high frequency trading. In order to do this, an omniscient trader is simulated and the total gain over three years is calculated. The obtained gain is very modest and reflects the limited contribution of technology in high frequency trading. This result underlines the primary role of research and modeling in this field.

In Chapter 3, the predictability of prices using some order book indicators is studied. Using conditional expectations, the empirical evidence of the statistical dependencies between the prices and indicators is presented. The importance of these dependencies results from the simplicity of the method, eliminating any risk of over fitting the data. Then the combination of the various indicators is tested using a linear regression and the various numerical and statistical problems associated with this method are analyzed. Finally, it can be concluded that the prices are predictable for a period of a few minutes and the assumption of market efficiency is questioned.

In Chapter 4, the mechanism of price formation from the arrival of events in the order book is investigated. The orders are classified in twelve types and their statistical properties are analyzed. The dependencies between these different types of orders are studied and a model of order book in line with the empirical observations is proposed. Finally, this model is used to predict prices and confirm the assumption of market inefficiency suggested in Chapter 3.

Keywords: High frequency trading, Market microstructure, Statistical learning, Linear regression, Hawkes process, Backtest, Trading strategies.

Contents

1 Contexte, Méthodes et Résultats (In French) 13
1.1 Contexte et objectifs . . . 13
1.2 Limitation empirique du trading haute fréquence . . . 15
1.2.1 Préliminaires . . . 15
1.2.2 Borne supérieure de gain d'un trader nuisible . . . 16
1.2.3 Fréquence optimale de trading . . . 17
1.2.4 Conclusion . . . 18
1.3 Preuves empirique de l'inefficience du marché : Prédictions des prix d'actions . . . 19
1.3.1 Préliminaires . . . 19
1.3.2 Preuve empirique de la prédictibilité des prix d'actions . . . 20
1.3.3 Prédiction par un modèle linéaire . . . 22
1.3.4 Conclusion . . . 24
1.4 Modélisation mathématique du carnet d'ordres: Nouvelle approche de prédiction des prix d'actions . . . 25
1.4.1 Préliminaires . . . 25
1.4.2 Propriétés empiriques de la dynamique du carnet d'ordre . . . 26
1.4.3 Prédiction par un processus de Hawkes multivarié . . . 27
1.4.4 Conclusion . . . 29

2 Optimal High Frequency Strategy in Omniscient Order Book 33
Introduction . . . 34
2.1 Preliminaries . . . 35
2.1.1 Aggressive HFT . . . 35
2.1.2 Data and framework . . . 35
2.2 Omniscient optimal HFT strategy . . . 36
2.2.1 Problem formulation . . . 36
2.2.2 Resolution . . . 38
2.3 Upper bound for HFT strategy and optimal holding period . . . 42
2.3.1 Omniscient order book trading - one step . . . 42
2.3.2 Omniscient order book trading - N steps . . . 48
2.4 Conclusions . . . 50

3 Empirical Evidence of Market Inefficiency: Predicting Single-Stock Returns 51
Introduction . . . 52
3.1 Data, methodology and performances measures . . . 53
3.1.1 Data . . . 53
3.1.2 Methodology . . . 54
3.1.3 Performance measures . . . 54
3.2 Conditional probability matrices . . . 55
3.2.1 Binary method . . . 56
3.2.2 Four-class method . . . 60
3.3 Linear regression . . . 62

3.3.1 Ordinary least squares (OLS) . . . 62
3.3.2 Ridge regression . . . 64
3.3.3 Least Absolute Shrinkage and Selection Operator (LASSO) . . . 70
3.3.4 ElASTIC NET (EN) . . . 72

4 Mathematical Modeling of the Order Book: New Approach of Predicting Single-Stock Returns 75
Introduction . . . 76
4.1 Empirical study of the order book events . . . 77
4.1.1 Data and Framework . . . 77
4.1.2 Introduction to the order book mechanism . . . 77
4.1.3 Statistical properties of the order book events . . . 79
4.1.4 Statistical dependencies between the different order book events . . . 83
4.2 Modeling framework . . . 88
4.2.1 Introduction to point process . . . 88
4.2.2 Introduction to Hawkes process . . . 89
4.2.3 Simulation of Hawkes process . . . 91
4.2.4 Goodness of fit . . . 92
4.2.5 Maximum likelihood estimation of Hawkes process parameters . . . 94
4.3 Mathematical modeling of the order book . . . 97
4.3.1 Poisson Model . . . 97
4.3.2 Univariate Hawkes Model . . . 103
4.3.3 Multivariate Hawkes Model . . . 105

Bibliography 116

5 Appendix 117

Chapter 1

Contexte, Méthodes et Résultats (In French)

1.1 Contexte et objectifs

C'est en 1250 à Toulouse qu'est née la première société par actions connue au monde: "Les moulins du Bazacle" [84] . Les particuliers pouvaient alors acheter des parts de la société et partager ainsi les risques et les gains. Ce système très pratique, permettant à de petits investisseurs de participer à de très gros projets, s'est propagé de ville en ville donnant naissance aux premières bourses européennes.

Victime de son succès, le système boursier se voit dériver de son objectif principal -i.e. mettre en relation des investisseurs et des entrepreneurs- vers la spéculation et l'avidité de l'argent facile. C'est ainsi que s'est produit le premier krach boursier célèbre de l'histoire, "la crise des tulipes" au 17ème siècle à Amsterdam [40] . Malgré l'évolution des marchés financiers, un krach semblable s'est reproduit aux Etats-Unis le jeudi 24 Octobre 1929 [2] . Nommé le jeudi noir, ce jour a vu la bourse de New York chuter de 30% en une journée, signalant le début d'une longue et douloureuse crise économique mondiale. Depuis, de nombreux krachs violents ont continué a bouleverser le monde de la finance. Du lundi noir le 19 octobre 1987 [21] , à la bulle immobilière en 2008 [13] en passant par la bulle Internet en 2000 [68] , les crises changent de noms mais la raison profonde reste la même: la déconnexion entre la bourse et l'économie réelle.

En parallèle avec l'évolution de la finance, la fin du 20ème siècle est marquée par la révolution digitale. Les échanges classiques, entre des agents, dans une salle physique, laissent progressivement la place aux échanges virtuels, sur le réseau, entre des humains ou des automates [56] . Dès lors, un particulier peut investir depuis son PC, dans une société cotée en bourse. Il peut même changer son avis après quelques minutes voire même quelques secondes et se retirer de cet investissement. Aussi bouleversant que ceci pourrait paraître, ce mécanisme assure une grande liquidité pour les sociétés cotées en bourse, qui voient une proportion non négligeable de leurs capitaux s'échanger chaque jour, sans que cela n'affecte vraiment leurs axes de développement.

C'est dans ce cadre que de grands acteurs de la finance proposent un service de liquidité en continu nommé "Market Making". Ces acteurs proposent, à tout instant, un prix d'achat et un prix de vente, pour toutes les actions, prenant le risque d'échanger contre des agents plus informés et de se retrouver ainsi avec des investissements perdants [39] . La difficulté majeure de ce service est de déterminer, à chaque instant, le juste prix de chaque actif, relativement à un horizon d'investissement donné. En effet, le prix instantané représente le consensus des acheteurs et des vendeurs à l'instant même; l'actif vaut ce qu'il vaut car il existe autant d'acheteurs qui croient à sa hausse que de vendeurs qui croient à sa baisse.

Par ailleurs, l'hypothèse de l'efficience des marchés [65] suggère que nul ne peut avoir une meilleure valorisation d'un actif -i.e. une valeur plus juste- que le marché. Ceci se justifie par l'hypothèse de l'absence d'opportunité d'arbitrage (AOA). L'AOA peut être résumée dans l'exemple suivant: "Si on peut prédire que le prix d'un actif va augmenter, les agents achèteront l'actif et entraîneront l'augmentation de son prix -par le mécanisme de l'offre et de la demande-. Le prix s'établira donc à un niveau qui annule le pouvoir prédictif." Aussi cohérent que ce raisonnement puisse paraître, il nous conduit au paradoxe suivant: Le prix est à chaque instant à sa juste valeur, parce que dès qu'il en s'éloigne les participants agissent dans le sens qui le ramène à cette valeur. Ceci n'est pas évident à admettre pour au moins les deux raisons qui suivent. Premièrement, en supposant l'existence d'agents qui corrigent le prix en continu, on admet l'existence d'instants -i.e. juste avant chaque correction- pour lesquels le prix n'est pas à sa juste valeur. Deuxièmement, l'hypothèse de l'efficience des marchés sous-entend qu'à chaque instant, il existe dans le marché suffisamment d'agents informés -i.e. qui ont un bon pouvoir prédictif- pour établir le juste consensus de prix. Cette dernière hypothèse n'a jamais été prouvée.

Par ailleurs, deux observations factuelles nous incitent à prendre avec beaucoup de précautions l'hypothèse des marchés efficients. La première observation est que les investisseurs n'ont pas les mêmes horizons d'investissement. Un investisseur qui croit que le prix de l'action augmentera de 10% sur l'année, n'aura pas de regret s'il l'achète au début d'une journée pendant laquelle son prix baisse de 0.1%. En face de lui, un vendeur à découvert, à un horizon d'investissement d'une journée, serait content d'empocher 0.1% de performance pour une journée de trading. Ce raisonnement s'applique, aussi bien, à ce même vendeur qui n'aura pas de regret s'il rachète le lendemain son action 0.0001% plus cher, par rapport à son nouveau prix, face à un trader haute fréquence. Cette observation met en cause l'existence même d'un juste prix absolu -i.e. indépendant des fonctions d'utilité des différents agents [41] -. La deuxième observation est que les acteurs professionnels de la finance gèrent des portefeuilles qui surperforment le marché significativement chaque année. Ayant vécu moi-même cette expérience, je crois fortement à la limitation du rôle du hasard dans ces résultats.

En s'intéressant à la détermination du juste prix des actifs financiers on s'intéresse automatiquement aux facteurs susceptibles d'agir sur ce prix. Dans ce cadre, il est important de ramener chaque facteur à son échelle temporelle. Les facteurs fondamentaux [80] [35] , tel que l'avancement des projets menés par l'entreprise, déterminent le prix de l'action à une échelle macro-économique. Ces facteurs influent la tendance principale sur plusieurs semaines, voire plusieurs mois, mais n'expliquent pas les fortes fluctuations des prix autour de cette tendance. En raccourcissant l'échelle temporelle à quelques jours, on retrouve les facteurs techniques [11] [66] , liés à la psychologie des investisseurs. On observe souvent des effets de réversion [33] -i.e. retour à la moyenne-, des gros mouvements des prix, expliqués par des récupérations de bénéfices. On observe aussi à cette échelle, des momentum [86] -i.e. des petits rendements successifs de même signe- expliqués par une tendance des investisseurs à acheter les actions qui semblent surperformer le marché et à délaisser celles qui semblent le sous-performer. Enfin, à l'échelle de la haute fréquence [45] , on observe, en direct, le mécanisme de la formation de prix, régi par la loi de l'offre et la demande.

L'objectif de cette thèse est de répondre au vrai besoin de déterminer le juste prix, relativement à un horizon, des actifs financiers. Nous proposons différentes méthodes mathématiques pour estimer ce prix. Nous démontrons, par ailleurs, que grâce à ce genre de méthodes, les professionnels du trading algorithmique font des bénéfices tout en étant au service des autres acteurs du marché.

Les différents modèles mathématiques proposés dans ce travail ont déjà été étudiés avec plus de détails dans des papiers académiques. Cette thèse ne constitue qu'une tentative modeste de renforcer le lien entre le monde académique régi par la beauté de la science et le monde professionnel régi par l'obligation des résultats.

Dans la suite de ce chapitre, nous présentons les différentes parties de ce travail en résumant les méthodes et les résultats.

1.2 Limitation empirique du trading haute fréquence

L'image du trading haute fréquence (THF) est très négative dans l'opinion publique et politique [61] [60] [64] [77] [59] . Malheureusement, le THF est généralement mal compris par ceux qui en parlent le plus dans les medias. Ceci entraîne des débats médiatiques, sans fin, qui amplifient les dérives de la finance quantitative pour expliquer les échecs économiques des sociétés qui travaillent de moins en moins.

Nous admettons que, comme tout autre domaine, le THF comporte des risques opérationnels (tels que les bugs informatiques, les interférences entre des algorithmes non compatibles..) qui provoquent occasionnellement des krachs éclair (flash crash) [79] . Cependant, ces krachs sont aussi rares que les krachs d'avions et ne constituent donc pas une preuve de la nuisance du THF.

Dans le chapitre 2, nous proposons des réponses chiffrées à quelques idées reçues sur le THF.

1.2.1 Préliminaires

Dans un marché gouverné par les ordres [70] , chaque participant peut poster publiquement ses intérêts dans l'objectif d'échanger avec les autres participants du marché. Les intérêts postés par tous les agents constituent à chaque instant le carnet d'ordre. **La Figure 1.1** représente un exemple de carnet d'ordre. À gauche, sont postés les ordres d'achat; un agent est prêt à acheter 100 actions à 45.5 euros et un autre est prêt à acheter 70 à 45.4 euros. À droite, sont postés les ordres de vente; 80 actions sont à vendre à 45.7 euros et 90 actions sont à vendre à 45.8 euros. Dans l'état actuel du carnet, il n y a pas d'intérêts compatibles entre les acheteurs et les vendeurs. Aucune transaction n'est donc exécutée.

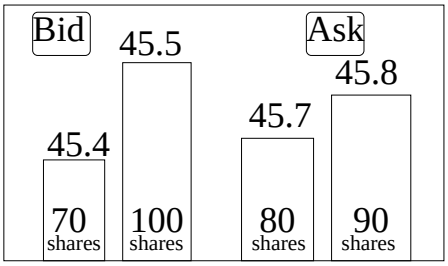


Figure 1.1: Exemple de carnet d'ordre

Par exemple, si un trader veut acheter 10 actions, il peut passer un ordre agressif, nommé un ordre au marché, pour les acheter au prix du marché. Ce prix est donné par la meilleure offre disponible au côté opposé; soit 45.7 euros. Dans ce cas le trader est un consommateur de liquidité (désigné par *liquidity taker*) et considéré, par conséquence, comme un trader agressif. Dans le cas contraire, le trader peut apporter la liquidité au marché, en postant un nouvel ordre (appelé ordre limite) d'achat de 10 actions à un prix inférieur à 45.7. Dans ce cas, le trader est désigné par *liquidity provider* et est considéré comme un trader passif.

le THF est largement perçu comme un abus de la supériorité technique [74] . Les entreprises de THF sont accusées d'utiliser des moyens mathématiques et informatiques démesurés pour profiter des investisseurs les moins équipés. Les traders HF sont soupçonnés d'avoir un accès plus rapide aux informations [73] . Ils peuvent donc en tirer profit en agressant le marché pour prendre les bonnes positions avant la propagation de l'information. Les investisseurs en face perdent systématiquement l'argent à cause de cette asymétrie informationnelle. L'objectif du chapitre 2 est de relativiser cette hypothèse et de quantifier empiriquement ses limitations.

1.2.2 Borne supérieure de gain d'un trader nuisible

Nous considérons qu'un trader est nuisible s'il agit exclusivement par des ordres au marché. Celui qui agit par des ordres limite ne fait qu'améliorer la liquidité disponible et n'induit aucun risque pour les autres participants [102] [48] [69] [85] . Par ailleurs, nous considérons qu'un trader profite de la technologie haute fréquence s'il garde ses positions pour des périodes très courtes (de l'ordre de la seconde).

Nous nous intéressons, dans cette partie, à la borne supérieure de gain d'un trader HF nuisible. Dans ce cadre, nous introduisons un trader omniscient -i.e. qui connaît parfaitement le futur. En particulier, chaque 10 millisecondes, il connaît parfaitement l'état du carnet d'ordre à l'instant même t, ainsi que son état après une période h. Il peut donc acheter ou vendre toute la quantité disponible à t et faire l'opération inverse à $t+h$. Par définition, ce trader ne prend donc que des positions gagnantes. **La Figure 1.2** schématise la stratégie de ce trader [55] .

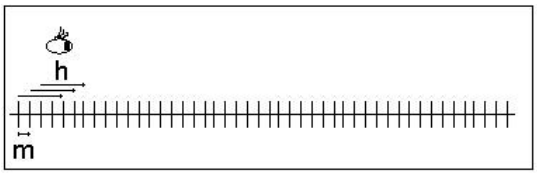


Figure 1.2: Chaque (m=10) millisecondes, le trader peut voir l'état du carnet d'ordre, ainsi que son état à t+h, il peut donc prendre toutes les positions profitables à t et les solder à t+h.

Nous avons calculé le gain total de ce trader omniscient sur les 50 actions européennes de l'indice Eurostoxx 50 pour la période de 2011 à 2013. Le **Tableau 1.1** résume les résultats obtenus pour des périodes de portage de 10 millisecondes à 10 secondes. La première colonne est, sans doute, la plus surprenante. Un trader qui prend toutes les décisions gagnantes, à un horizon de 10 millisecondes, et qui traite sans frais de transaction, ne gagne que 4.4 millions sur 3 ans. Ce gain est très modeste par rapport au frais de fonctionnement d'une entreprise haute fréquence. Cette stratégie est donc sans intérêt. Pour des périodes de portage plus longues, la profitabilité de la stratégie s'améliore, cependant, l'hypothèse de l'omniscience est de moins en moins valide.

	10 ms	100 ms	500 ms	1 sec	10 sec
Gain total [millions d'euros]	4.4	97	974	2,634	84,948
Gain moyen [euros]	136	3,051	30,562	82,631	2,658,734
Nombre moyen de trades	34	842	6,873	16,573	279,914
Rendement moyen [points de base]	2.8	3.2	3.4	3.5	4.3
Gain moyen par trade [euros]	6.7	8.5	11	12	18

Table 1.1: Profitabilité maximale d'une stratégie HF nuisible

Cette partie nous permet de conclure qu'une stratégie nuisible ne peut pas être profitable en très haute fréquence. Dans la partie suivante nous étudions la fréquence optimale d'une stratégie nuisible.

1.2.3 Fréquence optimale de trading

Dans le paragraphe précédent, nous avons imposé au trader d'avoir une période de portage, h, fixe et égale à sa période d'omniscience. Nous avons, ensuite, calculé ses gains pour différentes valeurs de h et nous avons conclu que le THF n'est pas rentable quand la période h est très courte. Dans ce paragraphe, nous autorisons au trader de changer indéfiniment ses positions pendant la période d'omniscience. Les transactions ne sont plus forcément équi-espacées. Par conséquence, les périodes de portage engendrées sont différentes et inférieures en moyenne à la période d'omniscience. Le but de cette partie est de calculer la période moyenne de portage.

Désormais, à l'instant t, pour une période d'omniscience h, le trader connaît les états du carnet d'ordre pour tout instant t_i, $t \leq t_i \leq t + h$. Il peut ainsi prendre à l'instant t les différentes décisions à exécuter pour tous les instants t_i. Le trader est uniquement soumis à deux contraintes; il ne peut pas exécuter plus que la quantité totale disponible dans le marché, et il doit être capable de solder toutes ses positions à l'instant $t + h$.

Pour obtenir la période de portage moyenne du trader, nous cherchons à déterminer sa stratégie de trading à partir des états du carnet d'ordre. Nous assumons que le trader applique à chaque instant la stratégie qui maximise son gain.

La **Figure 1.3** représente un exemple d'une stratégie de trading définie par les positions (v_i). Nous introduisons $\boldsymbol{\delta v}$ ($\delta v_i = v_i - v_{i-1}$ pour $i > 0$), le vecteur de toutes les transactions à effectuer pour appliquer la stratégie $\boldsymbol{v}$.

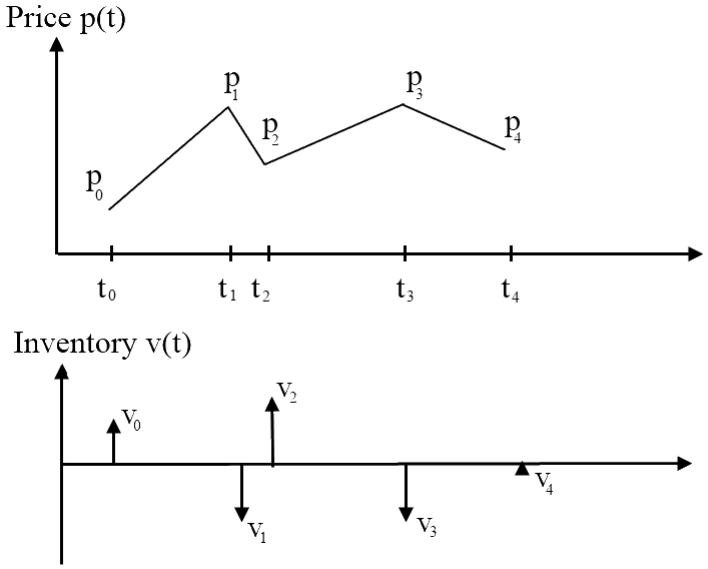


Figure 1.3: À chaque instant t_i, le trader décide d'avoir un nombre d'actions v_i.

Nous supposons que les coûts de transaction sont linéaires et définis par un facteur λ. Nous nous intéressons à U_T la fortune finale du trader qui applique la stratégie $\boldsymbol{v}$.

U_T peut-être exprimée, en fonction des transactions et des prix, comme suit:

$$U_T(\boldsymbol{\delta v}) = \sum_{i=0}^{T} -\delta v_i p_i + p_T \sum_{i=0}^{T} \delta v_i - \lambda \sum_{i=0}^{T} |\delta v_i p_i|$$

L'objectif du trader est de maximiser sa fortune finale. Sa stratégie s'obtient donc par la maximisation de U_T en respectant les contraintes de liquidité à chaque instant. Le problème

ainsi obtenu est linéaire et peut donc être résolu facilement. Nous avons ainsi tous les éléments nécessaires pour calculer la période moyenne de portage.

Pour chaque journée de trading, nous extrayons les données avec une résolution de 10 millisecondes puis nous divisons la journée en intervalles de 10 secondes. Sur chaque intervalle, nous calculons la stratégie optimale comme définie précédemment. Nous obtenons ainsi toutes les transactions effectuées par le trader omniscient.

La liste des transactions effectuées par le trader, nous permet de calculer sa fréquence moyenne de trading. Rappelons que la fréquence maximale possible correspond à la résolution des données, soit à une période de portage de 10 millisecondes. D'autre part, la fréquence minimale correspond à une stratégie constante sur chaque intervalle, soit à une période de portage de 10 secondes. Intuitivement, plus la fréquence optimale est proche de la fréquence maximale, plus le rôle de la technologie est important en THF. Gagner l'argent serait, dans ce cas, une simple conséquence de l'avantage informationnel.

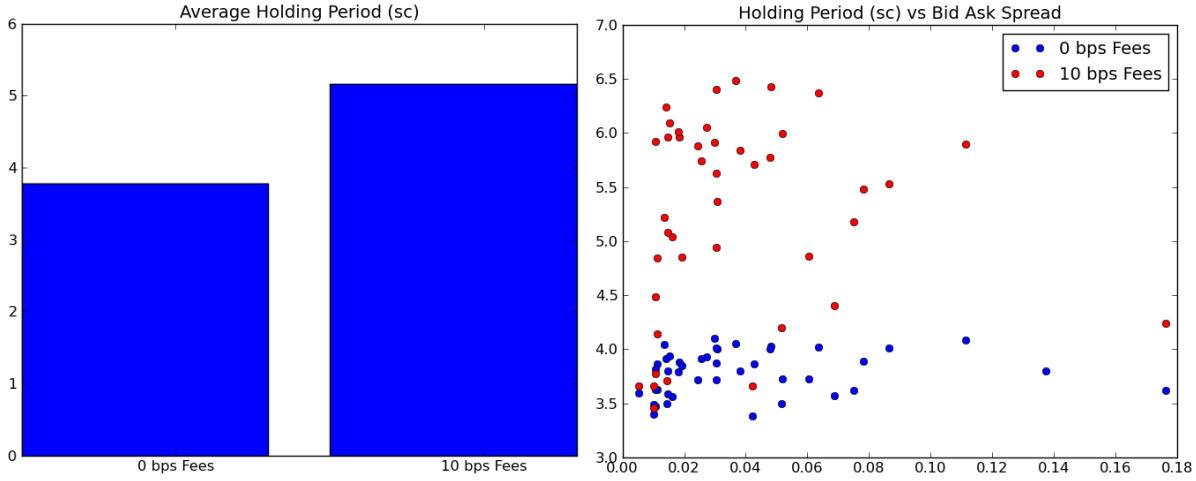


Figure 1.4: Période moyenne de portage

La **Figure 1.4** résume les résultats de cette partie. La période optimale de portage, pour des frais de transaction nuls, est de 3.8 secondes, soit 380 fois supérieure à la période minimale. Autrement dit, la fréquence de trading optimale est 380 fois inférieure à la fréquence maximale.

Ce résultat souligne l'apport limité de la vitesse en THF. En effet, Les ordres au marché coûtent cher et ne sont pas rentables à très court terme. Un trader nuisible est donc obligé de réduire sa fréquence de trading à des valeurs raisonnables, et perd par la suite une grande partie de son avantage informationnel.

1.2.4 Conclusion

Dans ce chapitre nous avons montré que le THF agressif n'est pas rentable à cause du bid-ask spread. Ces résultats sont en ligne avec d'autres études effectuées sur différents instruments (Forex, US Equities..) [55] [30] [3] [17] [10] [1] et modèrent les propos sur la nuisibilité du THF. Nous avons aussi montré que les plus gros profits en HF se réalisent plutôt en plusieurs secondes qu'en quelques fractions de seconde. L'enjeu en THF n'est donc pas uniquement technologique. Ainsi, pour réduire son risque de sélection adverse, un market maker ne peut pas se contenter d'annuler rapidement ses ordres, sur des signaux informationnels, mais doit aussi avoir des modèles de prédiction sur quelques secondes, voire quelques minutes. Ces modèles sont développés dans les prochains chapitres.

1.3 Preuves empirique de l'inefficience du marché : Prédictions des prix d'actions

Prédire l'évolution des actifs financiers intrigue les esprits des investisseurs et des scientifiques depuis des siècles. La meilleure formulation de la problématique est probablement celle énoncée par L. Bachelier, en 1900, dans sa thèse intitulée *Théorie de la spéculation* [4] : "Le Calcul des probabilités ne pourra sans doute jamais s'appliquer aux mouvements de la cote et la dynamique de la Bourse ne sera jamais une science exacte. Mais il est possible d'étudier mathématiquement l'état statique du marché à un instant donné, c'est-à-dire d'établir la loi de probabilité des variations de cours qu'admet à cet instant le marché."

Dans le chapitre 3, nous étudions la prédictibilité des prix sur des horizons fixes de 5, 10 et 30 minutes.

1.3.1 Préliminaires

Nous définissons, ci-dessous, des indicateurs du carnet d'ordre susceptibles de contenir de l'information sur les prix futurs.

Le rendement passé: L'utilisation du rendement passé, est justifiée par deux observations empiriques; la réversion et le momentum. La réversion désigne la correction d'une déviation non justifiée du prix. Plus précisément, des réactions exagérées des investisseurs, ou des anomalies ponctuelles de l'équilibre entre l'offre et la demande, peuvent dévier, brusquement et fortement, le prix de son niveau habituel. En moyenne, cette déviation est suivie par un mouvement opposé -i. e. une réversion- ramenant le prix à son niveau de référence. Au contraire, si la déviation du prix se réalise progressivement et lentement, elle peut indiquer un vrai signal sur l'action. En moyenne, d'autres participants adhèrent au mouvement et l'accentuent encore davantage. Cet effet de boule de neige est appelé momentum.

Le déséquilibre du carnet d'ordre: La liquidité à l'achat (respectivement à la vente) peut être définie comme la quantité d'actions demandée par les acheteurs (respectivement proposée par les vendeurs). Le déséquilibre du carnet est obtenu par le rapport entre la liquidité à l'achat et la liquidité à la vente. Un niveau élevé de cet indicateur indique une pression des acheteurs et constitue souvent un signal d'un mouvement haussier. De même, un niveau faible de l'indicateur permet de prédire un mouvement baissier.

Le flux de quantité: Cet indicateur est obtenu simplement par le rapport entre le nombre d'actions achetées et le nombre d'actions vendues pendant un intervalle de temps. Une action est dite achetée (respectivement vendue) si l'ordre au marché initiant la transaction est un ordre à l'achat (respectivement à la vente). Le flux est connu pour son autocorrélation positive [14] [36] . L'idée de l'utiliser pour prédire les rendements est de vérifier si la persistance des flux engendre une persistance de rendements.

Dans la suite, nous testons différentes méthodes de prédiction de prix et nous étudions la profitabilité des stratégies de trading correspondantes. Chaque méthode est donc qualifiée statistiquement par son taux de réussite et financièrement par le gain de la stratégie associée. Nous estimons que la prédiction est satisfaisante si le gain permet de payer un coût de transaction de 0.005% (0.5 point de base) -i.e. le coût approximatif d'exécution des banques et des fonds-.

Dans le paragraphe suivant, X désigne un indicateur observable à partir de l'état courant du carnet d'ordre et Y désigne la variable à prédire -i.e. le rendement décalé d'une période dans le temps-.

1.3.2 Preuve empirique de la prédictibilité des prix d'actions

Pour vérifier la pertinence d'un indicateur X, nous essayons, à partir d'un échantillon d'observations $(X_n, Y_n)_{n \leq N}$ et de la valeur X_{n+1}, de prédire Y_{n+1}. Nous considérons que l'indicateur est pertinent si la prédiction est fiable. Nous définissons, dans ce paragraphe, un estimateur de Y_{n+1} basé sur le principe des espérances conditionnelles. Nous avons choisi cet estimateur très simple pour éviter tout risque de sur optimisation des paramètres.

Pour obtenir une estimation de Y_{n+1}, nous commençons par classifier les X_n dans un petit nombre d'états, par exemple en 2 classes $C_1^X = \{X_n < \overline{X}\}$ et $C_2^X = \{X_n > \overline{X}\}$. Nous définissons ensuite $\widehat{Y}_1$ (respectivement $\widehat{Y}_2$) comme la moyenne des Y_n sur l'ensemble $(X_n, Y_n)_{n \leq N \cap X_n \in C_1^X}$ (respectivement $(X_n, Y_n)_{n \leq N \cap X_n \in C_2^X}$). L'estimation de Y_{n+1} peut être donnée par l'espérance de Y conditionnellement à la classe de X_{n+1}.

Formellement : $\widehat{Y}_{n+1} = \widehat{Y}_1 \mathbb{1}_{X_{n+1} \in C_1^X} + \widehat{Y}_2 \mathbb{1}_{X_{n+1} \in C_2^X}$.

Nous avons appliqué cette méthode, avec une fenêtre d'apprentissage de 10 jours glissants, à notre échantillon de données. La **Figure 1.5** résume la qualité statistique de la prédiction du signe du rendement 1-minute.

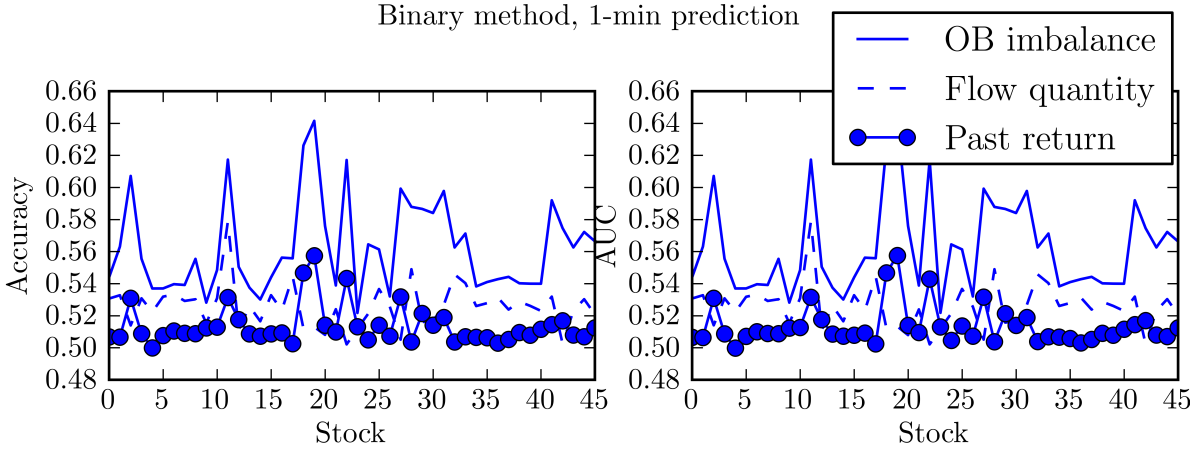


Figure 1.5: Qualité de la prédiction par classification binaire: La fréquence de réussite est supérieure à 50% pour toutes les actions. Les 3 indicateurs testés semblent être informatifs.

Les graphiques précédents montrent que tous les indicateurs sont statistiquement pertinents. Par ailleurs, l'indicateur de déséquilibre du carnet semble être le plus informatif, alors que le rendement passé semble être le moins informatif.

Pour mesurer l'intérêt pratique des prédictions, nous associons à chaque estimateur $\widehat{Y}$ une stratégie de trading qui achète/vend 100,000 euros de l'action, à l'instant n, si $\widehat{Y}_{n+1}$ est positif/négatif.

La **Figure 1.6** montre, que les stratégies associées aux trois indicateurs sont profitables en absence de coûts de transaction. Cependant, l'ajout d'un coût de 0.5 bp dégrade considérablement les performances (voire **Figure 1.7**).

Enfin, nous avons appliqué la même démarche en utilisant une classification en quatre états. La **Figure 1.8** montre que cette nouvelle méthode performe mieux que la méthode binaire.

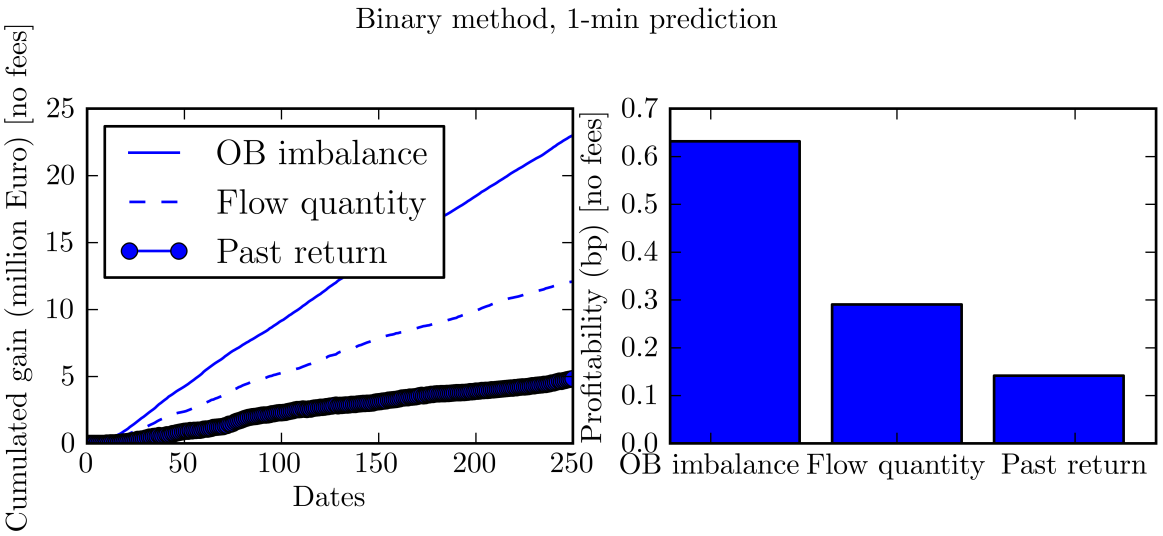


Figure 1.6: Qualité de la prédiction par classification binaire.

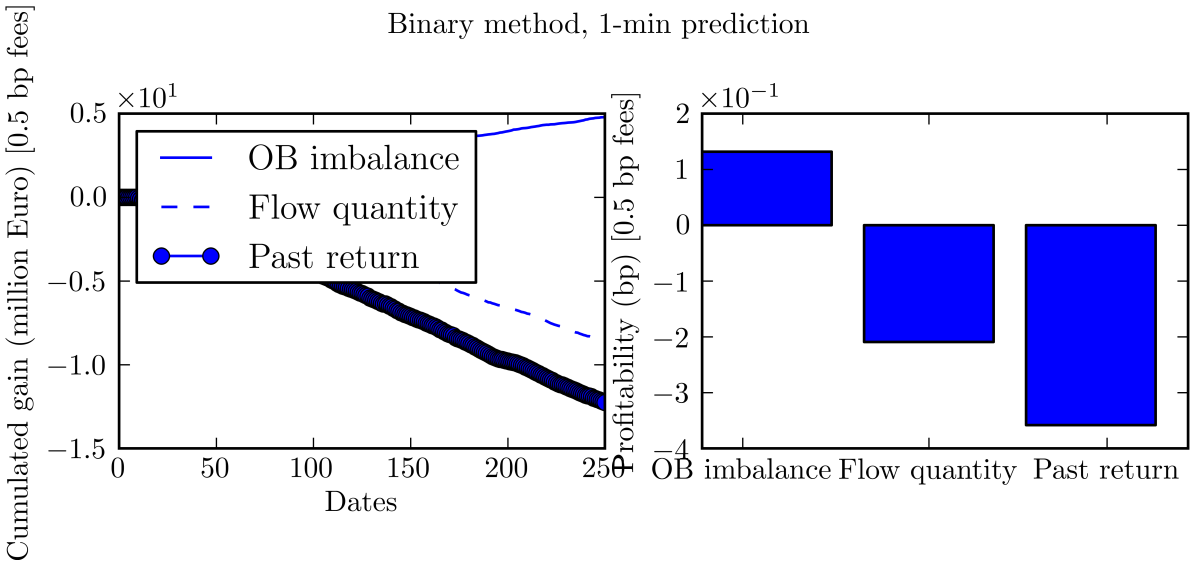


Figure 1.7: Qualité de la prédiction par classification binaire.

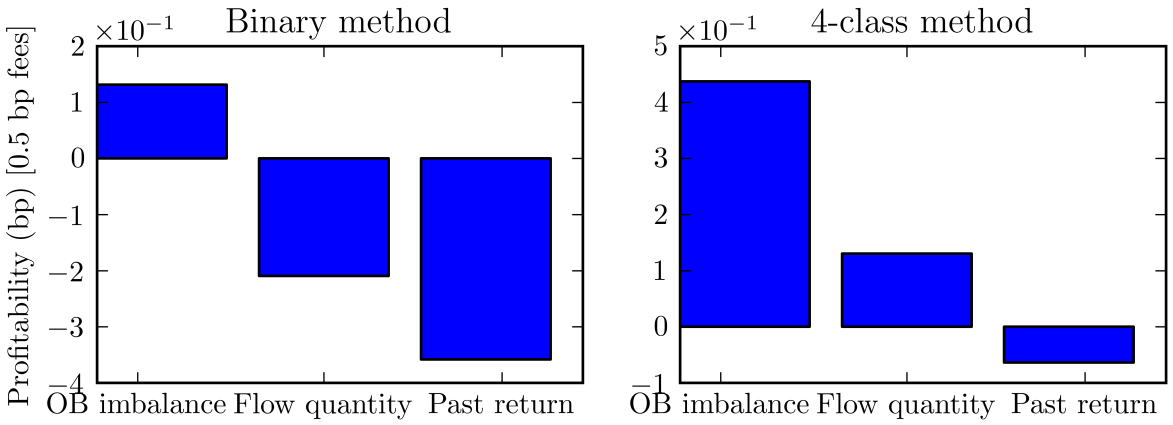


Figure 1.8: Qualité de la prédiction par classification en 4 classes.

Bien que les méthodes utilisées soient très basiques, les résultats sont relativement satisfaisants. Ceci suggère que les prix ne sont pas totalement imprévisibles.

1.3.3 Prédiction par un modèle linéaire

Dans cette partie, nous combinons tous les indicateurs, ainsi que leurs moyennes mobiles de différentes fréquences, dans une même matrice X. Cette matrice contient par construction plus d'information que chaque indicateur pris individuellement. Nous modélisons les dépendances entre les rendements futurs et les indicateurs par le modèle multilinéaire [96] suivant:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Nous calculons le paramètre de dépendance β à partir des données, par l'estimateur des moindres carrés classique (OLS pour Ordinary Least Squares) défini comme suit:

$$\widehat{\beta} = argmin_{\beta}(||Y - X\beta||_2^2)$$

Enfin, nous appliquons la même stratégie de trading, définie dans la première partie, aux prédictions du modèle linéaire. Intuitivement, nous nous attendons à une amélioration significative des résultats due à l'ajout d'information. La **Figure 1.9** compare les résultats du modèle linéaire à ceux du modèle binaire.

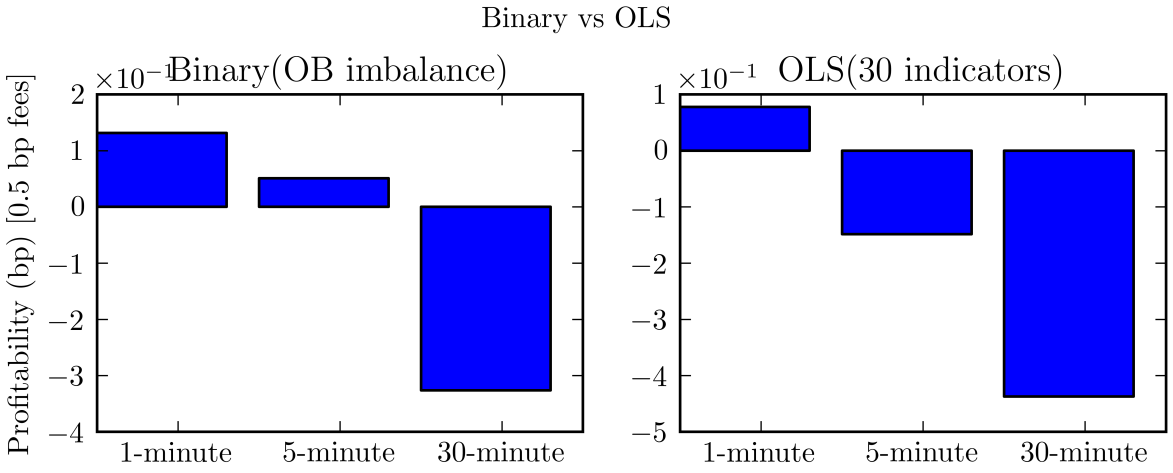


Figure 1.9: La qualité de la prédiction OLS: Les résultats de la méthode OLS ne sont pas meilleurs que ceux de la méthode binaire.

Étonnamment, les nouveaux résultats ne sont pas meilleurs que les précédents. Faire une régression linéaire avec 30 indicateurs ne surperforme pas la simple décision basée uniquement sur la classe instantanée du déséquilibre du carnet d'ordre. Ceci nous pousse à examiner la qualité de la calibration du modèle linéaire.

La Figure 1.10 montre 2 anomalies de cette calibration. Nous observons que le coefficient de régression associé au déséquilibre du carnet est négatif sur certaines périodes. Nous observons aussi des coefficients de régressions très différents pour des indicateurs très proches. Ces deux résultats sont contre l'intuition financière et soulignent un problème numérique ou statistique de calibration.

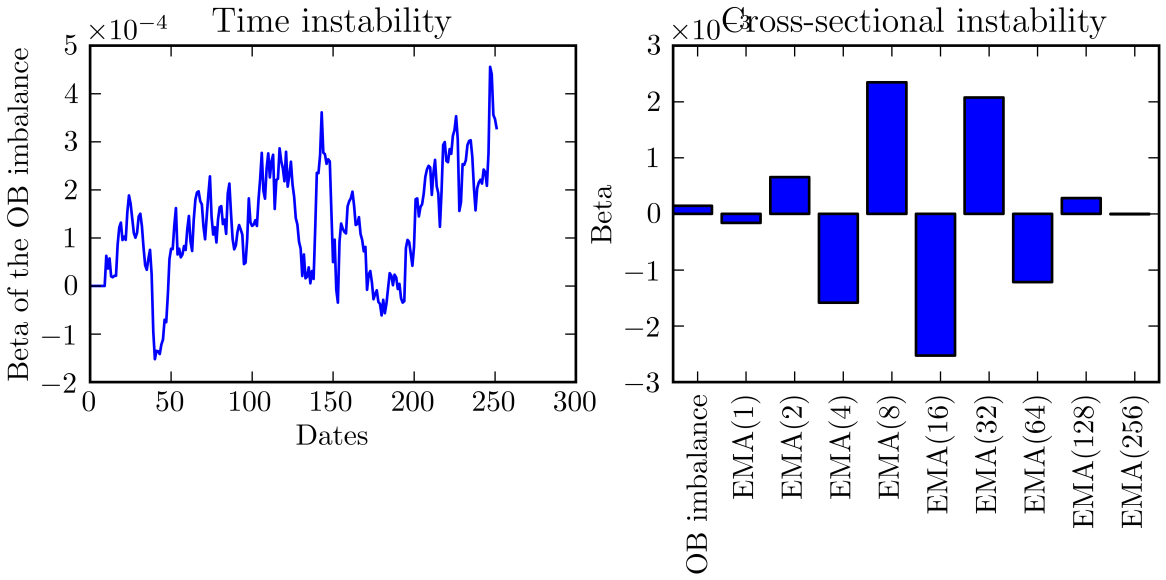


Figure 1.10: La qualité de la prédiction OLS: Le graphique à gauche montre l'instabilité du coefficient de la régression pour l'indicateur du déséquilibre du carnet (Dans cet exemple, la courbe est tracée à partir des données de l'action Deutsche Telekom pour l'année 2013). Le graphique à droite montre, pour une journée aléatoire, des coefficients très différents pour des indicateurs très proches; l'indicateur de déséquilibre du carnet et ses moyennes mobiles exponentielles (EMA pour Exponential Moving Average).

Pour calculer le paramètre β la méthode OLS passe par l'inversion de la matrice ${}^tXX$. Dans le cas de variables fortement corrélées, l'inversion de cette matrice peut conduire à des résultats non fiables numériquement et statistiquement [43] .

Pour remédier à ce problème, une fonction de régularisation peut être ajoutée à la fonction coût des moindres carrés. Cette régularisation favorise une forme particulière du paramètre β (petite norme, contient des zéros...) et stabilise significativement son estimation.

Dans cette étude, nous avons testé 3 méthodes de régularisation: La régression Ridge [51] ; la régression LASSO [90] et la régression Elastic Net [101] . Ci-dessous les définitions des estimateurs β associés aux 3 méthodes:

$$\widehat{\beta}_{Ridge} = argmin_{\beta}(||Y - X\beta||_2^2 + \lambda_{Ridge}||\beta||_2^2)$$

$$\widehat{\beta}_{Lasso} = argmin_{\beta}(||Y - X\beta||_2^2 + \lambda_{Lasso}||\beta||_1)$$

$$\widehat{\beta}_{EN} = argmin_{\beta}(||Y - X\beta||_2^2 + \lambda_{EN_1}||\beta||_1 + \lambda_{EN_2}||\beta||_2^2)$$

Nous avons utilisé les différents modèles pour tester la stratégie de trading basée sur la prédiction des prix. **La Figure 1.11** compare les performances de toutes les méthodes de régression. Nous observons que la méthode Elastic Net donne les meilleurs résultats alors que la méthode OLS donne les résultats les moins performants. Nous observons aussi que la simple méthode de classification performe presque aussi bien que la méthode Elastic Net.

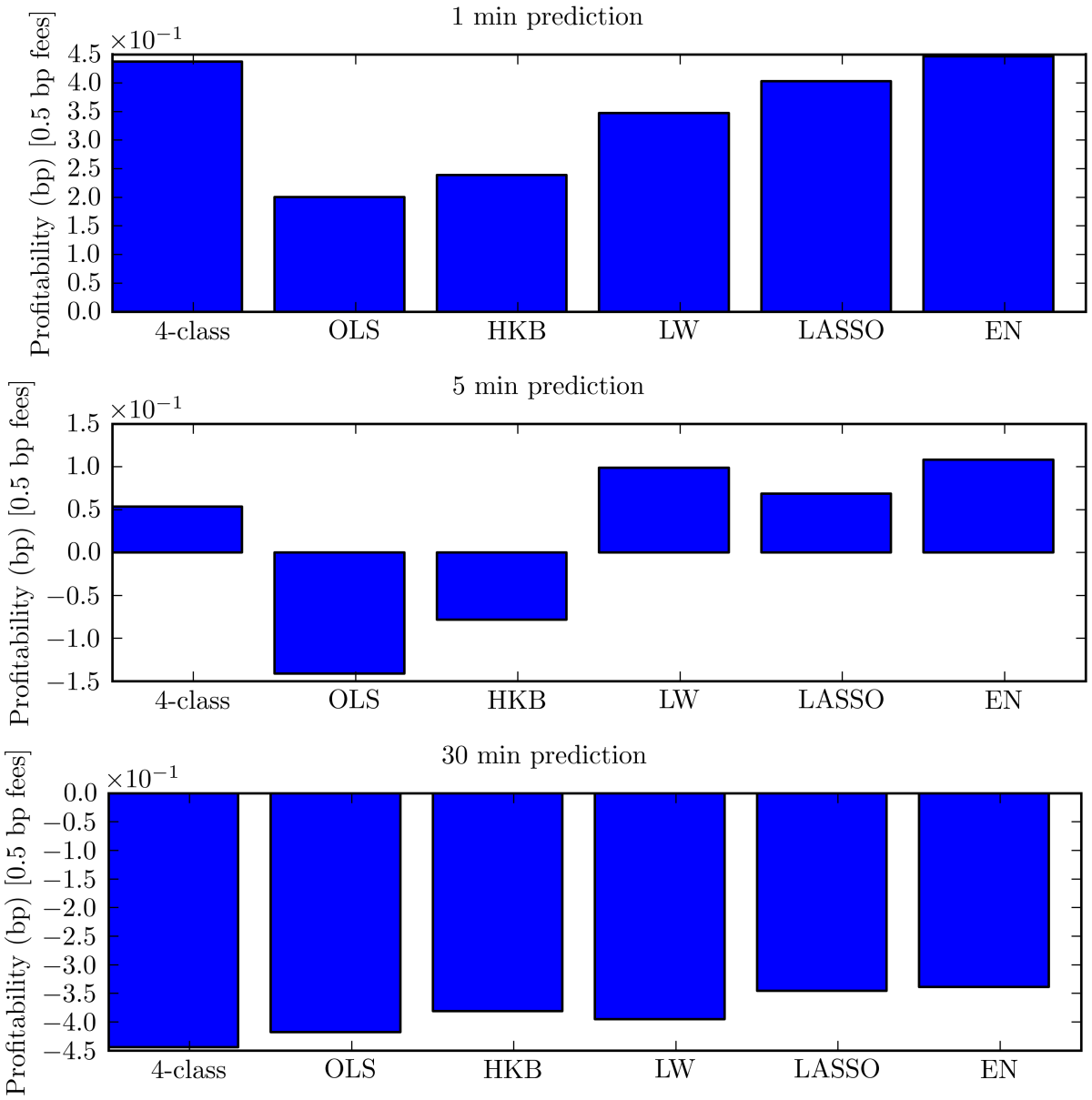


Figure 1.11: La qualité des prédictions: Le graphique compare les performances des stratégies de trading associées aux différentes méthodes de prédictions. La régression EN est la méthode la plus performante alors que la régression OLS est celle la moins performante. La méthode simple des espérances conditionnelles donne à son tour des résultats proches de la méthode EN.

1.3.4 Conclusion

Dans ce chapitre nous avons montré qu'à l'horizon de la minute, les prix des actions ne sont pas complètement imprévisibles. En particulier, dans le cadre de frais de transaction réaliste, une stratégie de trading simple, basée sur le déséquilibre du carnet d'ordre, est significativement profitable. Nous concluons aussi que la combinaison de plusieurs indicateurs, à l'aide d'une régression linéaire, nécessite une attention particulière aux problèmes numériques et statistiques liés à cette méthode.

Dans cette partie, les prédictions sont calculées sur des grilles de temps physique (toutes les minutes, toutes les 5 minutes..). En pratique, les évènements du carnet d'ordre peuvent donner de forts signaux prédictifs qui ne sont visibles qu'autour des arrivées des évènements. L'étude de ces signaux fait l'objet de la dernière partie de la thèse.

1.4 Modélisation mathématique du carnet d'ordres: Nouvelle approche de prédiction des prix d'actions

À l'échelle de la microstructure, le prix se forme, en continu, par l'arrivée des ordres émis par les différents participants [23] . Ce mécanisme, régi par l'offre et la demande, assure la cohérence des prix et la stabilité des marchés. Chaque ordre reflète une conviction du participant qui l'a envoyé et aura par conséquence, selon ses caractéristiques, un impact plus ou moins important sur l'évolution du prix. Une bonne compréhension du processus des arrivées des ordres aidera donc à réduire l'incertitude sur les prix futurs.

Dans le chapitre 4, les ordres, répartis en 12 types, sont modélisés par un processus ponctuel multivarié [29] . L'intensité instantanée de ce processus est utilisée pour prédire le sens de l'évolution du prix. Similairement au chapitre 4, la pertinence de la modélisation est testée par la mise en place d'une stratégie de trading basée sur les prédictions obtenues.

1.4.1 Préliminaires

Comme explicité dans le paragraphe **1.2.1**, un trader peut agir par des ordres agressifs (ordres au marché) ou par des ordres passifs (ordres limite). Par ailleurs, un ordre limite peut être annulé à tout instant avant son exécution. Nous avons ainsi 3 types d'ordres; les ordres au marché (*Market Order*), les ordres limite (*Limit Order*) et les annulations (*Cancellation*). Ces 3 types d'ordres sont répartis, selon leurs impacts instantanés sur le prix, et selon leurs sens (achat ou vente), en 12 types d'ordres élémentaires. Les notations correspondantes aux différents types d'ordres sont résumées dans la **Table 1.2**.

Notation	Définition
M, L, C, O	*market order*, *limit order*, *cancellation*, tout ordre.
M_{buy}, M_{sell}	*market order* à l'achat/à la vente.
M^0_{buy}, M^0_{sell}	*market order* à l'achat/à la vente qui ne change pas le prix.
M^1_{buy}, M^1_{sell}	*market order* à l'achat/à la vente qui change le prix.
L_{buy}, L_{sell}	*limit order* à l'achat/à la vente.
L^0_{buy}, L^0_{sell}	*limit order* à l'achat/à la vente qui ne change pas le prix.
L^1_{buy}, L^1_{sell}	*limit order* à l'achat/à la vente qui change le prix.
C_{buy}, C_{sell}	*cancellation* à l'achat/à la vente.
C^0_{buy}, C^0_{sell}	*cancellation* à l'achat/à la vente qui ne change pas le prix.
C^1_{buy}, C^1_{sell}	*cancellation* à l'achat/à la vente qui change le prix.
M^0, L^0, C^0, O^0	*market order*, *limit order*, *cancellation*, tout ordre qui ne change pas le prix.
M^1, L^1, C^1, O^1	*market order*, *limit order*, *cancellation*, tout ordre qui change le prix.

Table 1.2: Notations des différents types d'ordres

Pour répartir les évènements entre les classes O^0 et O^1, nous avons considéré uniquement le changement de prix instantané causé par l'évènement. Les changements de prix décalés dans le temps rentrent plutôt dans le cadre des études du *market impact* [71] [88] [98] et ne font pas partie de nos critères de classification.

Avec les notations de la **Table 1.2**, nous définissons le sous-ensemble d'évènements qui engendrent une hausse (respectivement baisse) immédiate du prix: $E_{up} = \{L^1_{buy}, C^1_{sell}, M^1_{buy}\}$ (respectivement $E_{down} = \{L^1_{sell}, C^1_{buy}, M^1_{sell}\}$).

L'objectif du chapitre 4 est de déterminer à chaque instant les probabilités d'occurrence des évènements de types E_{up} et E_{down}. Ceci permettrait de déduire facilement le sens d'évolution du prix. Pour calculer ces probabilités, nous modélisons le carnet d'ordre par un processus ponctuel multivarié d'intensité $\lambda(t) = (\lambda_1(t), \ldots, \lambda_{12}(t))$. Dans le paragraphe suivant, nous étudions les propriétés empiriques des différents évènements, afin de déterminer un modèle adéquat pour λ.

1.4.2 Propriétés empiriques de la dynamique du carnet d'ordre

La première ligne de la **Table 1.3** donne les probabilités historiques d'occurrence par type d'évènement. Nous observons que les ordres qui changent instantanément le prix représentent moins de 10% de la totalité des évènements. Nous observons aussi que les ordres limite et les annulations sont significativement plus récurrents que les trades. La deuxième ligne représente la répartition des ordres qui changent le prix. Nous observons, en particulier, une répartition équilibrée entre les différents types d'ordres. Ceci souligne l'importance de tenir compte de tous les évènements dans le modèle.

	L^0_{buy}	L^0_{sell}	C^0_{buy}	C^0_{sell}	M^0_{buy}	M^0_{sell}	L^1_{buy}	L^1_{sell}	C^1_{buy}	C^1_{sell}	M^1_{buy}	M^1_{sell}
$\|O$	22.82	22.93	19.80	20.03	2.99	3.00	2.07	2.12	0.85	0.88	1.27	1.26
$\|O^1$							24.52	25.12	9.71	10.06	15.36	15.23

Table 1.3: Probabilities (in %) of occurrences per event type

Nous avons aussi étudié les interactions entre les évènements en calculant les différentes probabilités conditionnelles d'occurrence. Pour simplifier la représentation des résultats, nous divisons ces probabilités par les probabilités inconditionnelles et nous arrondissons au plus proche entier. La **Table 1.4** résume les résultats obtenus.

	L^0_{buy}	L^0_{sell}	C^0_{buy}	C^0_{sell}	M^0_{buy}	M^0_{sell}	L^1_{buy}	L^1_{sell}	C^1_{buy}	C^1_{sell}	M^1_{buy}	M^1_{sell}
$\|L^0_{buy}$	2	0	1	1	1	1	1	1	0	1	1	0
$\|L^0_{sell}$	0	2	1	1	1	1	0	1	1	0	0	1
$\|C^0_{buy}$	1	1	2	0	0	1	1	1	2	0	0	0
$\|C^0_{sell}$	1	1	0	2	1	0	1	1	0	2	0	0
$\|M^0_{buy}$	1	0	0	0	**12**	0	**4**	0	1	1	**9**	0
$\|M^0_{sell}$	0	1	1	0	0	**11**	0	**3**	2	1	0	**9**
$\|L^1_{buy}$	1	0	0	1	2	2	1	1	**7**	2	2	**8**
$\|L^1_{sell}$	0	1	1	0	2	2	1	1	2	**6**	**10**	2
$\|C^1_{buy}$	1	1	2	0	0	0	**4**	2	1	0	0	0
$\|C^1_{sell}$	1	1	0	2	0	0	2	**4**	1	1	0	0
$\|M^1_{buy}$	1	0	0	1	1	0	**6**	**4**	1	1	1	0
$\|M^1_{sell}$	1	1	2	0	0	1	**3**	**4**	1	2	1	1

Table 1.4: Probabilités conditionnelles relatives

Les résultats montrent que les ordres M^0 augmentent très fortement la probabilité d'avoir des ordres au marché de même sens. Ceci est un résultat classique [63] expliqué par *l'order splitting* et le momentum. Par ailleurs, la table montre des interactions surprenantes entre les ajouts

d'ordres dans le spread (L^1) et les annulations totales à la première limite dans le même sens (C^1). Ces interactions peuvent être une conséquence d'algorithmes de manipulation de marché. À notre connaissance, ce résultat n'a pas été étudié dans d'autres papiers. Il mérite ainsi d'être approfondi dans de prochaines études. Nous observons aussi que les ordres M^1 augmentent la probabilité des ordres L^1 de sens opposé. Ceci correspond à des *liquidity provider* qui remplacent la liquidité consommée. Les ordres M^1 augmentent aussi la probabilité des ordres L^1 de même sens. Ceci représente une nouvelle limite qui se crée à un nouvel niveau confirmant le mouvement initié par l'ordre M^1. Ces observations montrent l'existence de plusieurs dépendances fortes, en temps évènementiel, entre les différents types d'ordres. Dans le paragraphe suivant, nous analysons les dépendances observées en temps physique.

Pour un processus de comptage $(N(t))_{t\in\mathbb{R}_+} = (N_1(t), \dots, N_M(t))$, une duration h et un lag τ, nous définissons la matrice de corrélation infinitésimale $Cr^h_\tau(i,j)_{1\leq i,j\leq M}$ du processus par :

$$Cr^h_\tau(i,j) = Correlation(N_i(t+h+\tau) - N_i(t+\tau), N_j(t+h) - N_j(t))$$

Dans la suite, nous choisissons h à 0.1 seconde et $\tau \in \{0.1, 0.2, \dots, 0.9\}$ et nous estimons les corrélations à partir des données historiques. Pour chaque type d'évènement i, la fonction $Cr^h_{i,j}(\tau)$ représente la décroissance de l'impact de l'arrivée d'un évènement j sur l'intensité d'arrivée de l'évènement i (voir par exemple 1.12).

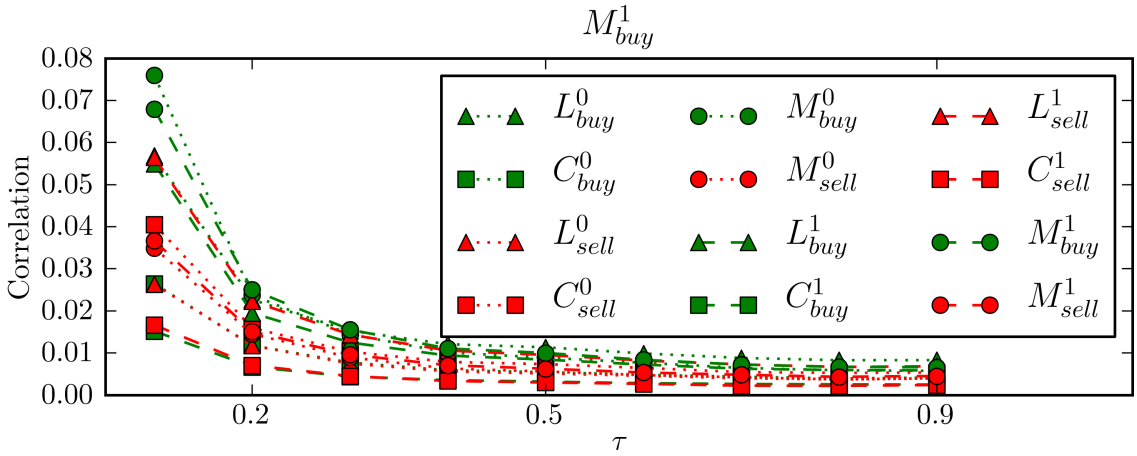


Figure 1.12: Les fonctions d'impact associées à M^1_{buy}: Le graphique confirme que les évènements qui impactent le plus l'intensité d'arrivée de M^1_{buy} sont M^0_{buy} et M^1_{buy}.

Pour les différents évènements, nous avons croisé ce résultat graphique avec les résultats des probabilités conditionnelles pour arriver à un modèle de Hawkes 12-variate avec un grand nombres de coefficients forcés à zéro. Ce modèle ainsi que son application sont détaillés dans le paragraphe suivant.

1.4.3 Prédiction par un processus de Hawkes multivarié

Nous modélisons le carnet d'ordre par un processus de Hawkes 12-variate avec un kernel mono-exponentiel. Rappelons que pour $m \in 1, .., 12$ l'intensité est donnée par:

$$\lambda_m(t) = \mu_m + \sum_{n=1}^{M} \sum_{T_i<t} \alpha_{mn} e^{-\beta_{mn}(t-T_i)} \mathbb{1}_{\{X_i=n\}}$$

Compte tenu des résultats de la partie empirique, nous nous intéressons exclusivement aux interactions les plus importantes résumées dans la **Table 1.5**. Les coefficients associés aux autres interactions sont forcés à zéro.

Évènements	Évènements influentes
L^1_{buy}	$\{M^0_{buy}, L^1_{buy}, M^1_{buy}, M^1_{sell}\}$
L^1_{sell}	$\{M^0_{sell}, L^1_{sell}, M^1_{buy}, M^1_{sell}\}$
C^1_{buy}	$\{L^1_{buy}\}$
C^1_{sell}	$\{L^1_{sell}\}$
M^1_{buy}	$\{M^0_{buy}, M^1_{buy}\}$
M^1_{sell}	$\{M^0_{sell}, M^1_{sell}\}$

Table 1.5: La matrice de dépendances

Nous calibrons le modèle retenu par un maximum de vraisemblance [78] et nous calculons λ_{up} (respectivement λ_{down}) comme la somme des intensités associées aux évènements de l'ensemble E_{up} (respectivement E_{down}).

Enfin, nous testons la stratégie qui achète (respectivement vend) 100,000 euros de l'action si $\lambda_{up} > \lambda_{down}$ (respectivement $\lambda_{up} < \lambda_{down}$). Les performances sont résumées dans la **Table 1.6**.

Ticker	Fréquence de réussite	Gain [Euros]	Profitabilité [Bps]	Période de portage
ADS	0.72	28428	0.08	1.46
ALV	0.70	33436	0.07	1.13
BAS	0.73	43995	0.08	1.00
BAYN	0.73	38894	0.08	1.38
BEI	0.73	14665	0.10	3.70
BMW	0.72	41168	0.09	1.25
CBK	0.69	48038	0.17	1.88
CON	0.74	37682	0.12	1.68
DAI	0.71	48337	0.08	0.88
DB1	0.73	22699	0.13	3.14
DBK	0.70	53172	0.08	0.88
DPW	0.72	33775	0.08	1.34
DTE	0.70	29932	0.09	1.59
EOAN	0.71	34662	0.09	1.48
FME	0.71	18334	0.10	2.94
FRE	0.69	17525	0.12	3.84
HEI	0.73	28147	0.12	2.59
HEN3	0.73	24911	0.09	2.11
IFX	0.73	30362	0.11	1.99
LHA	0.70	33421	0.15	2.53
LIN	0.72	21490	0.08	2.09
LXS	0.71	23976	0.16	3.67
MRK	0.69	15869	0.12	4.25
MUV2	0.71	24105	0.08	1.86
RWE	0.72	37955	0.11	1.52
SAP	0.72	32530	0.06	1.06
SDF	0.70	26084	0.17	3.64
SIE	0.72	39092	0.07	0.94
TKA	0.71	26506	0.13	2.82
VOW3	0.72	38411	0.08	1.16
Average	0.71	31,587	0.10	2.06
Min	0.69	14,665	0.06	0.88
Max	0.74	53,172	0.17	4.25

Table 1.6: Performances de la stratégie basée sur le modèle de Hawkes.

Les résultats obtenus sont relativement bons (71% de bonne prédiction en moyenne). Ceci suggère l'adéquation du modèle aux données. Cependant, la courte période de portage (2 secondes en moyenne) conduit à une faible profitabilité (0.1 bp en moyenne).

Pour améliorer la profitabilité de la stratégie, nous augmentons sa période de portage en appliquant une moyenne mobile exponentielle aux intensités. Ceci ralentit le signal du trading et réduit le bruit.

La nouvelle stratégie obtenue gagne moins, mais elle a une meilleure profitabilité. La **Figure 1.13** représente le gain des 2 stratégies sur 4 mois. En particulier, nous observons que la stratégie ralentie reste profitable après 0.5 bp de coût de transactions.

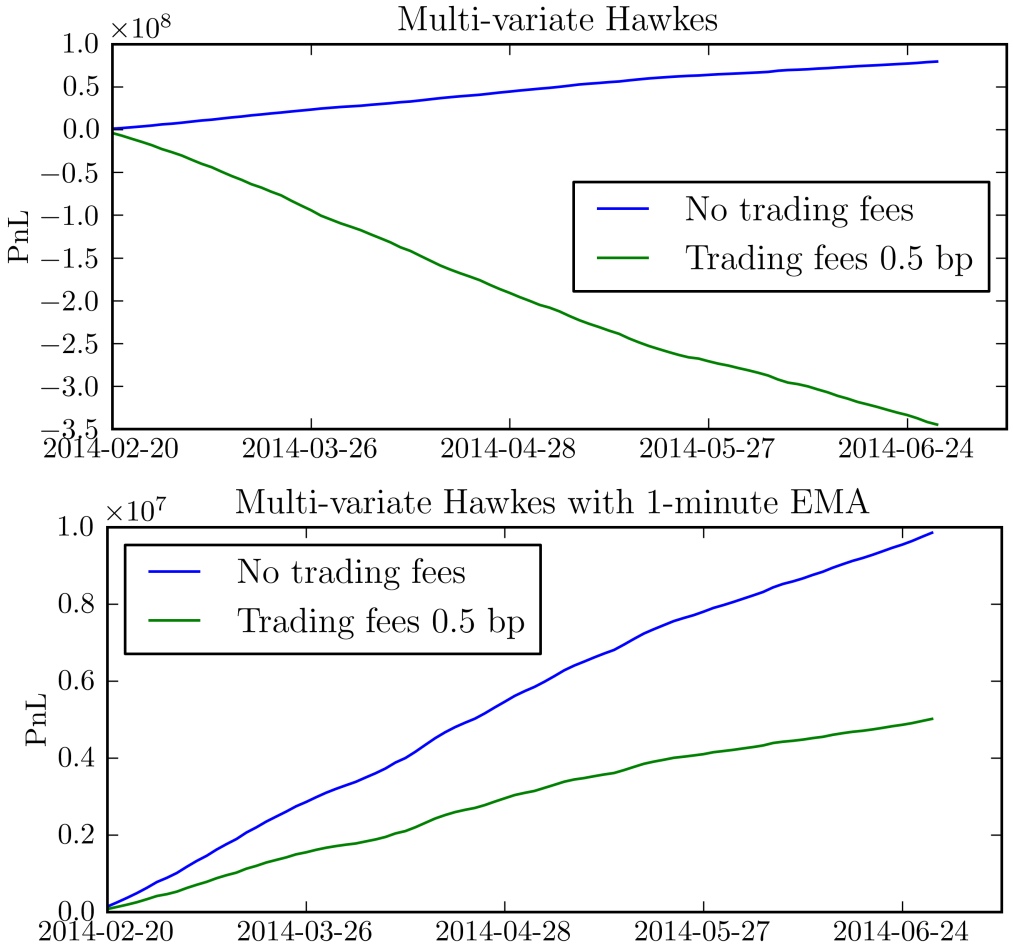


Figure 1.13: Gain cumulé sur 4 mois

1.4.4 Conclusion

Dans ce chapitre, nous avons étudié les dépendances entre les différents types d'évènements du carnet d'ordre pour modéliser mathématiquement leurs temps d'arrivées. En particulier, nous avons montré que le modèle le mieux adapté aux observations empiriques est le processus de Hawkes multivarié. Ce modèle nous a permis d'avoir des prédictions fiables de l'évolution des prix.

Note

Chaque chapitre de cette thèse a été publié séparément. Nous avons gardé, volontairement, les versions originales des papiers. Ceci permet à chaque lecteur de comprendre parfaitement la partie qui l'intéresse sans besoin de lire les parties précédentes.

Each chapter of this thesis was published separately. We kept, deliberately, the original versions of the papers. This allows each reader to fully understand the part that interests him without need to read the previous parts

Chapter 2

Optimal High Frequency Strategy in Omniscient Order Book

Note:
- This chapter is submitted to "The Journal Of Empirical Finance".
- This chapter is presented in the forum "Big Data in Finance and Insurance", Institut Louis Bachelier, Paris, March 2014.
- This chapter is presented in the conference "Finance, Risk and Accounting Management Perspectives Conference", University of Oxford, London, September 2014.

Abstract

The aim of this study is to quantify the low latency advantage of High Frequency Trading (HFT) and to compute, empirically, an optimal holding period of a HF trader. Critics claim that low latency leads to information asymmetry, victimizing retail investors. However, objective studies measuring the gain due to this asymmetry are rare. In order to perform the study, new methods are introduced in this paper, in particular, the optimal strategy problem is formulated and ideas are given to compute it in a reasonable amount of time. A new measure, the weighted mean holding period, is introduced and an algorithm to compute it is suggested. Using the previous concepts, a large empirical study based on the optimal omniscient strategy is presented and evidence of the low latency advantage limitation is provided. In particular, it is shown that the bid ask spread and the transaction costs lead to a trading frequency much lower than the information renewal frequency.

Contents

Introduction **34**
2.1 Preliminaries **35**
2.1.1 Aggressive HFT 35
2.1.2 Data and framework 35
2.2 Omniscient optimal HFT strategy **36**
2.2.1 Problem formulation 36
2.2.2 Resolution 38
2.3 Upper bound for HFT strategy and optimal holding period **42**
2.3.1 Omniscient order book trading - one step 42
2.3.2 Omniscient order book trading - N steps 48
2.4 Conclusions **50**

Introduction

Since the last financial crisis, proprietary trading, especially High Frequency Trading, has been widely criticized and assumed to be one of the main causes of market instability. In 2010, President Obama's adviser argued [26] that such speculative activity played a key role in the financial crisis of 2007-2010. Many regulation ideas have been suggested. Tobin Tax [92] is a well-known example.

The rationale behind penalizing HFT agents is to protect investors from such professional speculators. HFT firms are widely assumed to be armed with sophisticated mathematical algorithms and a strong software framework [57] allowing them to make large profits by rapidly making the best decisions. Due to the short holding periods, HFT seems to be a risk-free activity [31] providing huge profits, victimizing less sophisticated investors. HFT is also assumed to cause flash crashes, artificial volatility, and to increase market adverse selection by hitting the order book systematically at each arbitrage opportunity [60] .

Despite all these assumptions, empirical papers published by various authors studying the US market claim modest upper bounds on profit. Kearns, et al. [55] demonstrated that HFT profits are modest compared to the traded volume. In particular, their study found an upper bound of HFT profit on US stock market equal to 21 billion dollars/year for a 10-second holding period and only 21 million for a 10-millisecond holding period. Duhigg [30] suggested the same 21 billion dollar upper bound, Arnul et al. [3] suggested 1.5 to 3 billion dollar upper bound while Brogaard [17] suggested 3 billion dollars. Baron et al. [10] studied the E-mini S&P 500 futures contract from August 2010 to August 2012 and found an estimation of HFT profits equal to 100 million. Aldridge [1] studied the HFT profit on the forex market and concluded that the upper bound on returns is 4 basis point.

As far as is known, there is no equivalent study dealing with recent data on the European Market. In addition, no paper was found studying the HFT holding period.

The main goals of this study are to define a theoretical optimal strategy for a HF Trader, to analyze the factors that might explain HFT profit, and to find the optimal holding period according to the bid-ask spread trading cost. This optimal holding period quantifies the low latency advantage effect and helps in understanding the behavior of HF traders. The focus of this paper is on aggressive strategies based on market orders. Limit orders do not increase the adverse selection risk for other participants and are thus widely considered to be harmless[18] .

This work is organized as follows: The first section presents some general but insightful concepts. In the second section, the optimal strategy is formulated as a solution of a linear problem. The computation time problem is addressed and some ideas are proposed to enhance the computing performances. In the third section, a one-step omniscient trader method is developed and used to analyze the HFT profit. Results confirm the modest upper bound, discussed above and show a strong dependence of HFT profit on the volatility. Finally, the one-step assumption is relaxed and the methodology, formulated in the second section, is applied to compute the optimal holding period. Results of this section are surprising and show that the optimal trading frequency is not as high as widely assumed.

Notation

Bold, lowercase characters represent vectors, and bold capital characters represent matrices.
In particular, the following denote :

- $\boldsymbol{v}$: A column vector.

- $\boldsymbol{v}^T$: A row vector equal to the transpose of $\boldsymbol{v}$.
- $\boldsymbol{O}$: A matrix which all elements are equal to zero.
- $\boldsymbol{o}$: A vector which all elements are equal to zero.
- $\boldsymbol{I}$: The identity matrix.
- $\boldsymbol{i}$: A vector which all elements are equal to one.
- $\boldsymbol{L}$: A lower full triangular matrix with all non-zero elements equals to one.

2.1 Preliminaries

2.1.1 Aggressive HFT

In order to buy/sell a number of shares on an order book driven market [70] , the trader can either match other participants' interests or provide a new offer to the market. For example, **Figure 2.1** represents an order book with two limits. Some participants are currently willing to buy (Bid side) 100 shares and 70 shares, at 45.5 and 45.4 euros respectively. Other participants are willing to sell (Ask side) 80 shares and 90 shares, at 45.7 and 45.8 euros respectively. At the current state of the order book there are no matching interests. Thus, no transaction is executed.

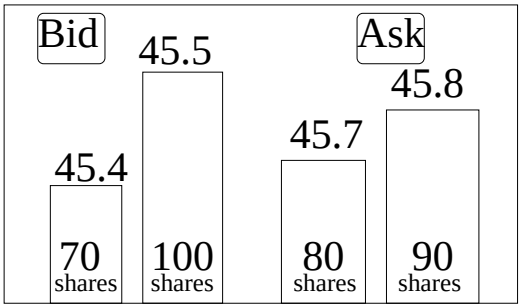


Figure 2.1: Example of an order book with 2 limits on the ask side (participant willing to sell) and 2 limits on the bid side (participant willing to buy)

Suppose a trader wants to buy 50 shares, he can either "hit the order boo" and "consume liquidity" by buying 50 shares at 45.7 euros, or post a "buy order" at a price below 45.7 euros. In the first case, the order is called a "market order" and the participant is a "liquidity taker." In the second case, the order is called a "limit order" and the participant is a "liquidity provider."

This paper deals exclusively with a liquidity taker trader, i.e. one who uses exclusively market orders. HF traders acting through limit orders can be viewed as liquidity providers to the market, and there seem to be a consensus that providing more liquidity to market participants is harmless, see [102] [48] [69]. This study also focuses on profit made when running a strategy based on short holding periods. Lower frequency strategies can be run with any framework and thus, are not specific to HFT.

2.1.2 Data and framework

This study focuses on the EURO STOXX 50 stocks. Three years of full daily order book data provided by the "Chair of Quantitative Finance" at Ecole Centrale Paris are used. Snapshots are extracted every 10 milliseconds. Auction phases are ignored since traders can not hit the order book during those phases. Thanks to the Mesocentre of the Ecole Centrale Paris, millions of calculations were computed in a reasonable amount of time.

2.2 Omniscient optimal HFT strategy

2.2.1 Problem formulation

This section aims to mathematically define an optimal strategy relative to some criteria. Knowing the price time series, the available Bid and Ask quantities, and the transaction fees, the following question is answered, "What strategy would have maximized a given utility function?". To achieve this work, the final wealth U_T is considered as the utility function.

A strategy is defined as the vector $\boldsymbol{v}$ such that the i^{th} coordinate v_i is the signed number of shares to hold between the time t_i and the time t_{i+1} (see example in Figure 2.2). Given the price time series, $\boldsymbol{p}$, and the chosen strategy, $\boldsymbol{v}$, the final wealth, U_T is to be calculated.

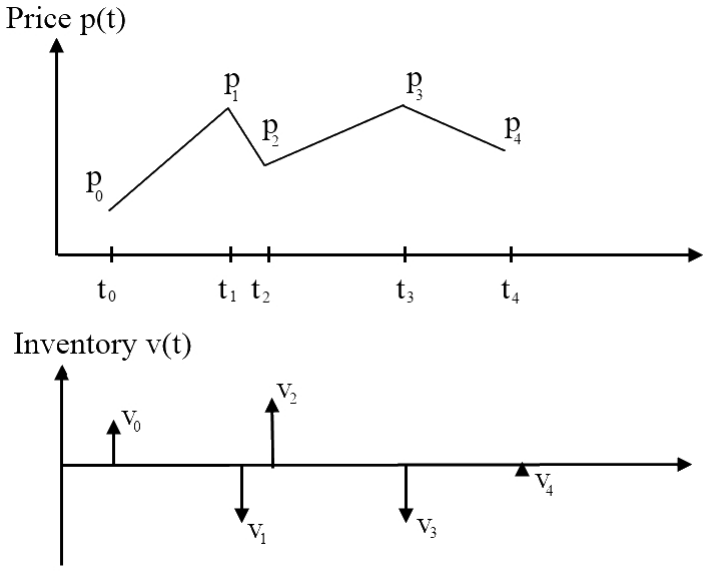


Figure 2.2: At each time t_i, the trader decides to have v_i shares on his portfolio.

We define $\boldsymbol{\delta v}$ ($\delta v_i = v_i - v_{i-1}$ for $i > 0$) as the vector of all the transactions to execute in order to apply the strategy $\boldsymbol{v}$. The initial condition $\delta v_0 = v_0$ is chosen (before time 0, the portfolio is empty). Assuming that transaction fees can be assimilated to a proportional cost, λ, $\boldsymbol{U}$ can be calculated easily, for example, at the time t_1:

$$U_1 = v_0(p_1 - p_0) - \lambda|v_0|p_0 - \lambda|v_1 - v_0|p_1$$

$$U_1 = -\delta v_0 p_0 - \delta v_1 p_1 + \delta v_1 p_1 + \delta v_0 p_1 - \lambda|\delta v_0 p_0| - \lambda|\delta v_1 p_1|$$

More generally, the wealth U_T obtained by applying a strategy $\boldsymbol{v}$ over T periods is as follows:

$$U_T(\boldsymbol{\delta v}) = \sum_{i=0}^{T} -\delta v_i p_i + p_T \sum_{i=0}^{T} \delta v_i - \lambda \sum_{i=0}^{T} |\delta v_i p_i|$$

Due to the initial condition, a strategy is perfectly defined by giving indifferently $\boldsymbol{v}$ or $\boldsymbol{\delta} v$.

The focus of this study is HFT, thus it is assumed that the portfolio is empty at the end of the period T; $\sum_{i=0}^{T} \delta v_i = 0$.

When dealing only with the best limits of the order book, all notations can be simplified. Considering liquidity and trading constraints, the optimal strategy is determined by solving the following problem:

Minimize

$$J_\lambda(\boldsymbol{\delta v}) = \sum_{i=0}^{T}(\delta v_i^+ p_{ask_i} + \delta v_i^- p_{bid_i}) + \lambda \sum_{i=0}^{T}(\delta v_i^+ p_{ask_i} - \delta v_i^- p_{bid_i})$$

Subject to

- $-bidQ_i \leq \delta v_i^- \leq 0$ (Liquidity constraints)
- $0 \leq \delta v_i^+ \leq askQ_i$ (Liquidity constraints)
- $\sum_{i=0}^{T} \delta v_i = 0$ (Empty portfolio at the end of the period)
- $\delta v_i = \delta v_i^- + \delta v_i^+$ (Definition)
- Min inventory $\leq v_i \leq$ Max inventory (Trading constraints)

The simplified notations above are used in the mathematical formulations for the rest of the paper. However, the tests on the real data were computed using the multi limits formulations. Denotes K as the number of limits available and x_i^j as the value of x relative to the limit j at the time i, the optimal strategy problem is given by:

Minimize

$$J_\lambda(\boldsymbol{\delta v}) = \sum_{i=0}^{T}\sum_{j=0}^{K-1}(\delta v_i^{j+} p_{ask_i{}^j} + \delta v_i^{j-} p_{bid_i{}^j}) + \lambda \sum_{i=0}^{T}\sum_{j=0}^{K-1}(\delta v_i^{j+} p_{ask_i{}^j} - \delta v_i^{j-} p_{bid_i{}^j})$$

Subject to

- $-bidQ_i{}^j \leq \delta v_i^{j-} \leq 0$ (For each j - Liquidity constraints)
- $0 \leq \delta v_i^{j+} \leq askQ_i{}^j$ (Liquidity constraints)
- $\sum_{i=0}^{T-1} \delta v_i = 0$ (No overnight position constraint)
- $\delta v_i^+ = \sum_{j=0}^{K-1} \delta v_i^{j+}$ (Definition)
- $\delta v_i^- = \sum_{j=0}^{K-1} \delta v_i^{j-}$ (Definition)
- Min inventory $\leq v_i \leq$ Max inventory (Trading constraints)

2.2.2 Resolution

Solving the previous optimization problem might seem easy from a mathematical perspective [38] , however, when dealing with high dimensional problems, the simplest linear system might become costly in computation time [27] . This section compares different methods to solve the problem. In particular, the importance of the sparsity when dealing with big data is shown. The key to HFT is to process large amounts of data rapidly. Solving a problem becomes useless if the calculation time is long enough for input data to significantly change. In the next paragraphs, the results obtained using the CVXOPT package and those obtained using the MOSEK solver are compared. For each solver, both dense and sparse formulations of the problem are used.

2.2.2.1 Framework

Sparse matrices: A sparse matrix [89] is a matrix populated mainly by zeros. The fraction of zero elements is called the sparsity of the matrix. In programming, such particularity leads to an important gain of storage space. Instead of storing all the n^2 values of the matrix, only the p non-zero values and their coordinates in the original matrix are stored. Without any loss of the initial information, an important proportion of the storage space is economized. In numerical analysis, most of the powerful solvers [97] [44] correctly handle sparse matrices and take advantage of the sparse structure to economize time when solving numerical problems.

CVXOPT package: CVXOPT is a free software package for convex optimization based on the Python programming language [94] . The package provides solvers for linear and quadratic problems. It handles sparse matrices' implementations and it is easy to use in any external program.

MOSEK package: MOSEK is a large-scale optimization software providing solvers for linear, quadratic, general convex and mixed integer optimization problems [28] . MOSEK handles sparse matrices' implementations. The software is not free but provides free academic licenses for research and educational purposes.

Matricial formulation (dense formulation): For classic programming languages the problem is described in matricial form as follows:

Minimize

- $\boldsymbol{c}^T\boldsymbol{x}$

Subject to

- $\boldsymbol{G}\boldsymbol{x} \leq \boldsymbol{h}$

Where

- $\boldsymbol{c}^T = \left[p_{ask_0}(1+\lambda), ..., p_{ask_{T-1}}(1+\lambda), p_{bid_0}(1-\lambda), ..., p_{bid_{T-1}}(1-\lambda)\right]$
- $\boldsymbol{x}^T = \left[\boldsymbol{\delta v}^{+T}, \boldsymbol{\delta v}^{-T}\right]$
- $\boldsymbol{G} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} \\ -\boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \\ \boldsymbol{O} & -\boldsymbol{I} \\ \boldsymbol{L} & \boldsymbol{L} \\ -\boldsymbol{L} & -\boldsymbol{L} \end{bmatrix}$
- $\boldsymbol{h}^T = \left[\boldsymbol{askQ}^T, \boldsymbol{o}^T, \boldsymbol{o}^T, \boldsymbol{bidQ}^T, v_{max}, ..., v_{max}, 0, -v_{min}, ..., -v_{min}, 0\right]$

Dimension

- $\boldsymbol{x} \in \mathbb{R}^{2T}$
- $\boldsymbol{G} \in \mathbb{R}^{6T} * \mathbb{R}^{2T}$
- Number of non-zero elements in $\boldsymbol{G}$: $2T^2 + 6T$

2.2.2.2 Computation times

MOSEK and CVXOPT computation times for several dimensions are compared in **Table 2.1**.

Dimension T	MOSEK	CVXOPT
100	0.05	0.04
1000	9.60	584.00
2000	72.30	4156.40

Table 2.1: Computation times (in seconds) for several dimensions

MOSEK is 60 times faster than CVXOPT, however both solvers are slow compared to the latency needed for a HF strategy.

In order, to enhance the computation time, a new formulation of the problem is given in the next paragraph.

2.2.2.3 Variable duplication

Matricial formulation (sparse formulation): In order to reduce the number of non-zero elements, a redundant variable, $\boldsymbol{v}$, is introduced. This variable is unnecessary since v_i is perfectly defined knowing $(\delta v_j)_{0 \leq j \leq i}$. The new formulation is:

Minimize

- $\boldsymbol{c}^T \boldsymbol{x}$

Subject to

- $\boldsymbol{G}\boldsymbol{x} \leq \boldsymbol{h}$
- $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$

Where

- $\boldsymbol{c}^T = \left[p_{ask_0}(1+\lambda), ..., p_{ask_{T-1}}(1+\lambda), p_{bid_0}(1-\lambda), ..., p_{bid_{T-1}}(1-\lambda), \boldsymbol{o}\right]$
- $\boldsymbol{x}^T = \left[\boldsymbol{\delta v}^{+T}, \boldsymbol{\delta v}^{-T}, \boldsymbol{v}^T\right]$
- $\boldsymbol{G} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ -\boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \\ \boldsymbol{O} & \boldsymbol{O} & -\boldsymbol{I} \end{bmatrix}$
- $\boldsymbol{h}^T = \left[\boldsymbol{askQ}^T, \boldsymbol{o}^T, \boldsymbol{o}^T, \boldsymbol{bidQ}^T, v_{max}, ..., v_{max}, 0, -v_{min}, ..., -v_{min}, 0\right]$
- $\boldsymbol{A} = [\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\Lambda}]$
- $\boldsymbol{\Lambda} = \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$

Dimension

- $\boldsymbol{x} \in \mathbb{R}^{3T}$
- $\boldsymbol{G} \in \mathbb{R}^{6T} * \mathbb{R}^{3T}$
- $\boldsymbol{A} \in \mathbb{R}^{T} * \mathbb{R}^{3T}$
- Number of non-zero elements in $\boldsymbol{G}$ and $\boldsymbol{A}$: $10T - 1$

Remarks

- In the second formulation, the dimension of the problem is increased by 50%.
- When introducing the redundant variable the number of non-zero elements is reduced from $O(T^2)$ to $O(T)$.

Previous computation times are compared with the new ones in **Table 2.2**.

Dimension	MOSEK (dense)	MOSEK(sparse)	CVXOPT(dense)	CVXOPT(sparse)
100	0.05	0.02	0.40	0.04
1000	9.60	0.07	584.00	2.80
2000	72.30	0.12	4156.40	11.00
4000	596.00	0.24	33000.00	47.00

Table 2.2: Computation times (in seconds) for dense and sparse formulations

When using the sparse formulation, the computation time decreases spectacularly. In **Figure 2.3**, for both formulations, MOSEK is used to compute the solution and computation times are plotted for several dimensions. It can be concluded, in this case, that rewriting the problem in a sparse form, using a redundant variable, decreases the calculation cost from $O(T^3)$ to $O(T)$.

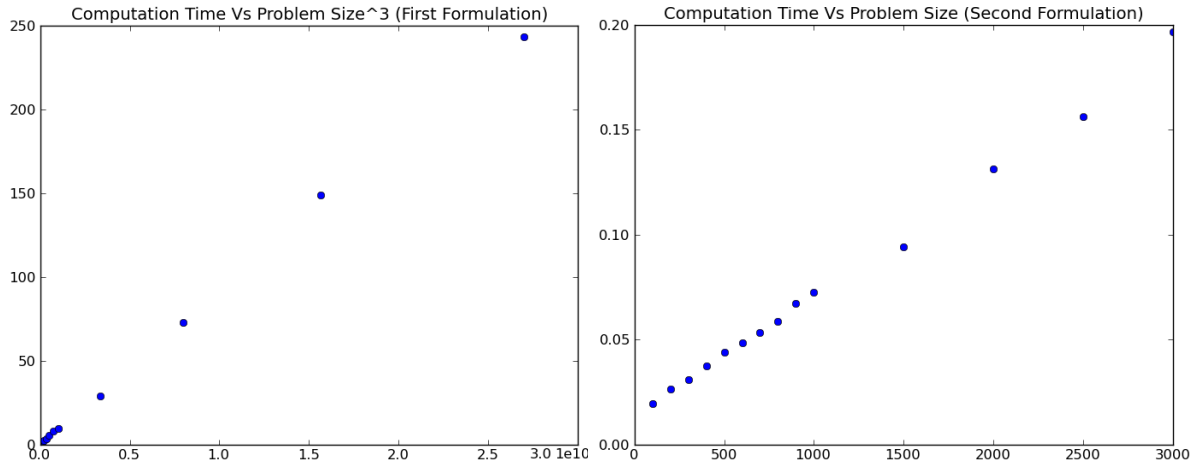


Figure 2.3: The graph on the left shows a linear dependency of computation time on T^3 when using the first (dense) formulation, and the graph on the right shows a linear dependency on T when using the second (sparse) formulation.

2.2.2.4 Importance of computation time

In high frequency the computation time is so important. If the market state changes while computing an algorithm, the computation results are less relevant. More generally, when dealing with big data the study cannot be done if the unitary calculation time is not sufficiently small.

In this paper, to study the HFT profitability and the optimal holding period, the unitary algorithm computes the optimal strategy on a bucket of 10 second data sampled with a 10 millisecond resolution. This corresponds to a problem size of $T = 1000$. Each day contains more than 2,500 buckets, and the study deals with 3 years of data of 50 stocks. This leads to approximately 90 million calculations.

The dense formulation problem can be computed in 9.6 seconds, and thus a total computation time longer than 200,000 hours of calculation. Using 200 processors, in parallel run, the results would have been computed in 1000 hours. Thanks to the sparse formulation, computation time is divided by more than 100, thus, using 200 processors, results were computed in a few hours.

2.3 Upper bound for HFT strategy and optimal holding period

This section aims to compute an upper bound for HFT profits, to analyze the main factors that explain HFT profitability and to compute an optimal holding period for HF strategy. To this end, an omniscient trader who can observe the future and act accordingly to realize benefits is simulated.

This assumption is not realistic, since the best a trader can do is predict the future with a small error. However, such results give an idea about the maximum possible HFT profit realized by executing all the profitable trades over 50 stocks for three years.

In the first part of this section, the method presented by Kearns [55] is developed and the HFT profits are explained using different market indicators. In the second part, the one-step method is generalized in the n-steps case using the previous results to compute the optimal strategy and find the HFT optimal holding period.

2.3.1 Omniscient order book trading - one step

2.3.1.1 Methodology

The experiment consists of a trader observing the present and the future state of the order book at a given frequency, and taking all profitable positions (see **Figure 2.4**). Two key time quantities are involved. The first one is the holding period, h, of any taken position. This period has to be long enough for the order book to undergo sufficiently large changes enabling the realization of profits that offset the trading costs due to bid-ask spread crossing, but short enough in order to remain in a high frequency setting.

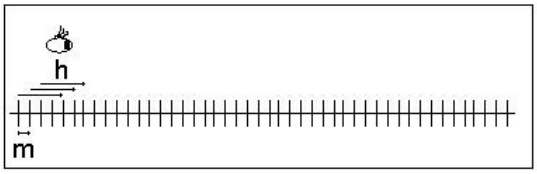


Figure 2.4: Each m second, the omniscient trader can see the current state of the order book, and its state at the time t+h, he takes all possible profitable positions at t and unwinds them at t+h.

In fact, a holding period of one millisecond is too short to observe a favorable movement in the order book. A holding period of one minute is too long, and therefore offsets the advantage of rapid exchange access, making the opportunity of profit available to non-high frequency traders.

The second key time quantity is the acting period, m. This quantity is important since it is assumed that the trader does not impact the market. Indeed, the liquidity taken by the trader when he acts at time t, is returned to the order book when he re-observes it at time $t + m$ to decide to take a new position. It is then clear that a profitable position taken at time t will be available (and then also taken) at time $t + m$ if the order book does not move. This is in accordance with the aim to estimate an upper bound, even if this upper bound can be made arbitrarily high by taking m to be arbitrarily small.

Thus, m has to be small enough in order to realize this large bound for the benefits, and large enough in order to avoid the pathological case of taking one profitable position infinitely many times. In addition, to avoid counting artificial profits, the omniscient trader is forbidden from taking positions impossible to be unwound during the next 15 seconds. The order book can show important moves after a "long period" (15 seconds or more) without any change. Thus, the omniscient profitable trade cannot be counted as a HF trade.

This step m is chosen to be $m = 10$ milliseconds. This is still very short to have a large overestimation of the profitability, as a winning position can be taken 100 times within a second if the order book does not move enough within that second. This is in accordance with the aim to overestimate the benefits, and avoids the pitfall of very large overestimation.
Another key hypothesis is that the trader is omniscient and thus always makes the good decision.

2.3.1.2 Results

The different results obtained when running the omniscient strategy over three years of data are analyzed. It was found that HFT profits are modest and negligible compared to traded volumes. It was also shown that profitable trades are very rare for short holding periods.

Global results: Results of running the omniscient strategy over 50 stocks between 2011 and 2013 are summarized in this paragraph. **Figure 2.5** shows that profits decrease rapidly with a decreased holding period. The maximum total profit possible for a holding period of 10 seconds is 85 billion euros and for a holding period of 10 milliseconds is only 4.4 million euros. As discussed in the next paragraph, these sums are modest compared to the traded volume. It can also be noted that the profit in 2011 was significantly higher than 2012 and 2013. This might be explained by a fall in volume and volatility during the last 2 years.

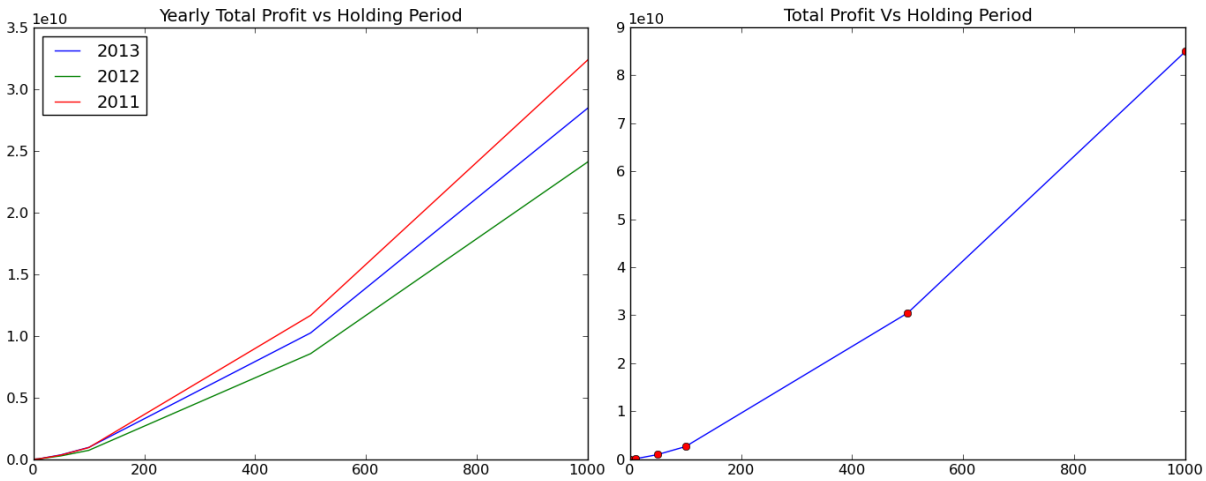


Figure 2.5: Graphs show modest total profits for short holding period

To have more familiar numbers, the average profit per stock per day is plotted in **Figure 2.6**. For a holding period of 10 milliseconds, an omniscient trader, trading aggressively, without transaction fees, taking all profitable decisions at least once, makes on average 136 euros per stock per day! The profit rises up to 2.7 million euros per stock per day for a 10-second holding period. However, it is impossible to be omniscient for 10 seconds.

Previous results also give an approximation of the possible profit of a non-omniscient trader. Let $U_T(p)$ be the wealth realized by a trader making predictions with a success probability $p < 100\%$. A simple approximation gives $U_T(p) = p * U_T(100\%) - (1-p) * U_T(100\%)$. To verify this formula, a trader with a 80% prediction success rate (**Figure 2.6**) is simulated. The average profit is approximately equal to 60% (A linear regression gives $\beta = 0.599$) of the omniscient average profit, which is coherent with the previous approximation.

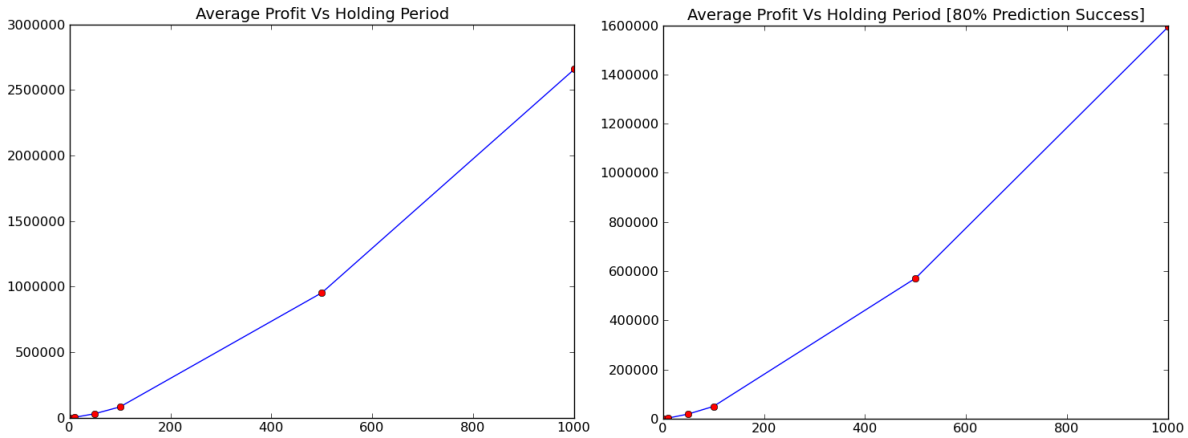


Figure 2.6: The graph on the left shows that the profit per stock per day is less than 100,000 euros for a 1-second holding period. The graph on the right shows that a 20% failure rate in prediction leads to a 40% lost in profit.

To understand the causes of small profit for short holding periods, the average number of trades and of traded shares vs holding period are plotted in **Figure 2.7**. For the 10-millisecond holding period the average number of trades is 34 and the average number of shares is 27,918. Profitable positions become rare when the holding period is short. This is mainly caused by the bid ask spread that becomes non-negligible for small moves of the order book.

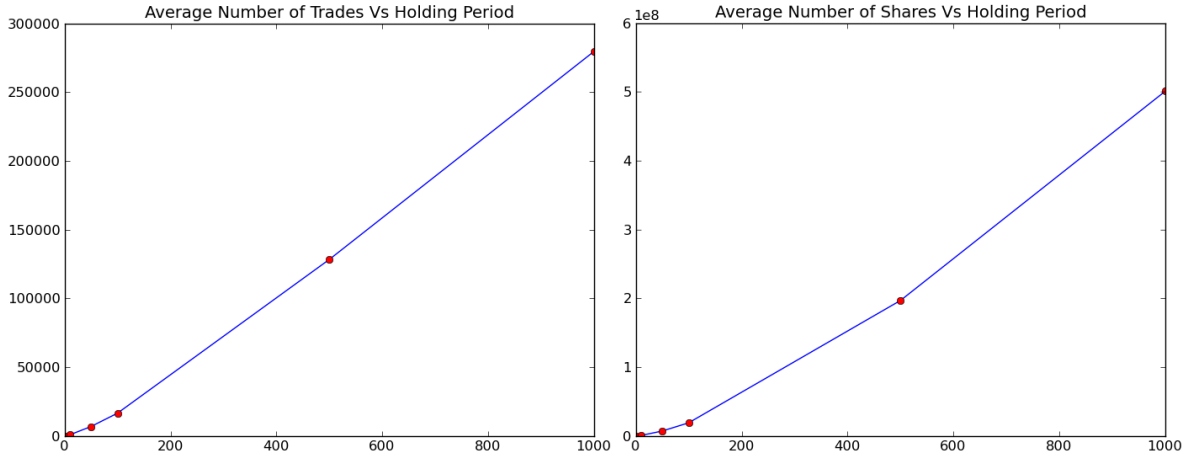


Figure 2.7: The graph on the left shows that profitable trades are rare for short holding period. The graph on the right shows that the number of traded shares decreases rapidly with decreasing holding period.

Besides the fact that profitable positions are rare for short holding periods, **Figure 2.8** establishes that they are also less profitable. For the shortest holding period, the average profit by trade is 6.7 euros and the average return is 2.8 bps (bps : basis point $= 1\% * 1\%$).

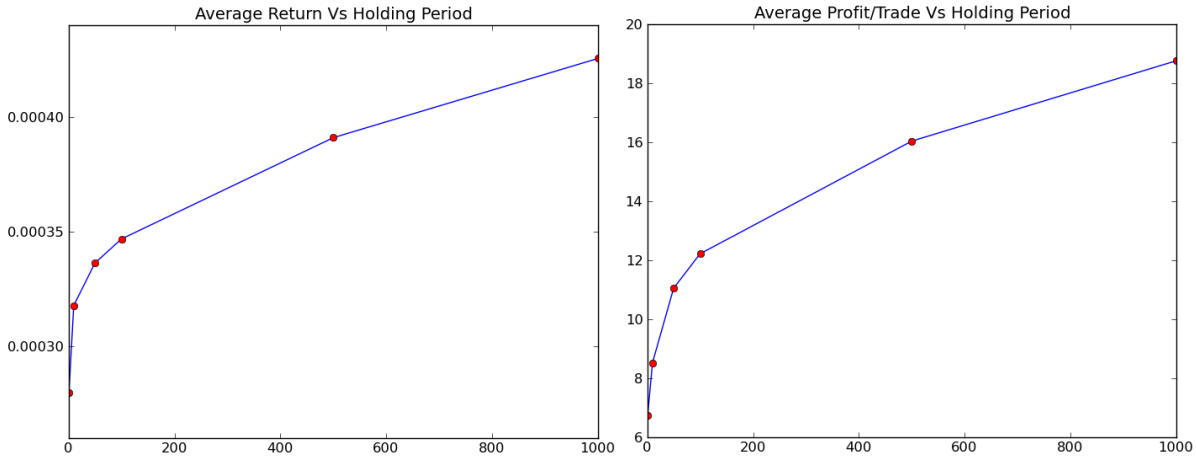


Figure 2.8: The graph on the left shows the limitation of aggressive strategies' profitability, even for a 5-second omniscience period the profitability is less than 5 basis points. The graph on the right shows that for the shortest holding periods, the average profit per trade is less than 10 euros.

The main data used to plot **Figure 2.5**, **Figure 2.6**, **Figure 2.7** and **Figure 2.8** are summarized in **Table 2.3**:

	10 ms	100 ms	500 ms	1 sec	10 sec
Total Profit (2013) [million euros]	1.2	35	359	962	28,468
Total Profit (2012) [million euro]	1.5	30	275	725	24,105
Total Profit (2011) [million euro]	1.6	31	339	947	32,375
Total Profit (All) [million euro]	4.4	97	974	2,634	84,948
Average Profit [euros]	136	3,051	30,562	82,631	2,658,734
Average Number of Trades	34	842	6,873	16,573	279,914
Average Number of Shares	27,918	702,589	7,114,299	19,050,229	501,433,780
Average Return [basis points]	2.8	3.2	3.4	3.5	4.3
Average Profit per Trade [euros]	6.7	8.5	11	12	18

Table 2.3: Global results

Detailed results: In this part HFT profitability is studied in more detail with the focus on the shortest holding period. **Figure 2.9** represents the average (single stock) daily profit during the entire studied period, and the density of daily profits.

In order to understand the main factors driving HFT profits, the daily average profit is plotted vs some features of the EURO STOXX 50. **Figure 2.10** examines the relationship between HFT profitability and the Future instrument returns ($\frac{ClosePrice - OpenPrice}{OpenPrice}$).

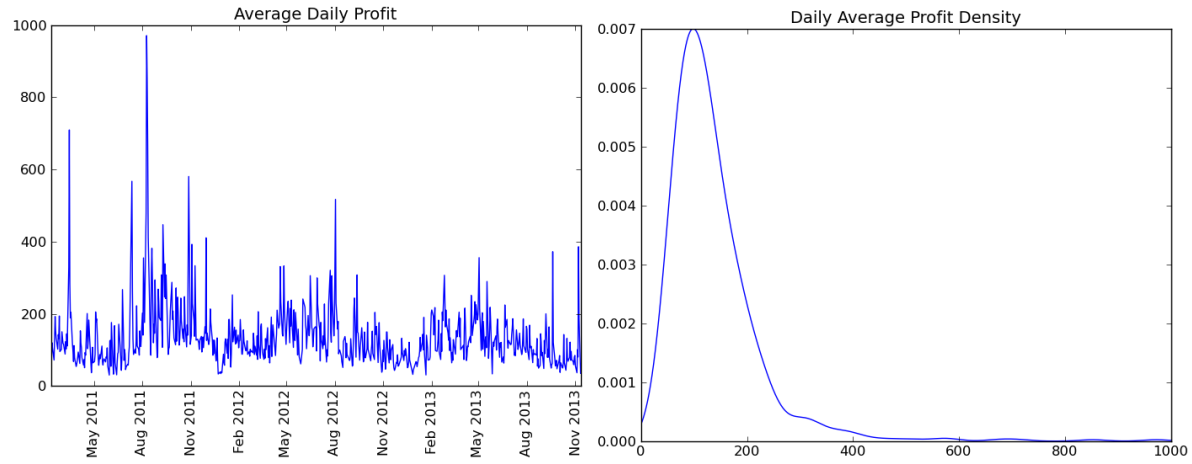


Figure 2.9: The graph on the left shows some clustering phenomenon. The periods HFT works better (summer 2011 for example) correspond to a volatile market. The graph on the left shows that for the shortest holding period, the average profit is generally less than 300 euros.

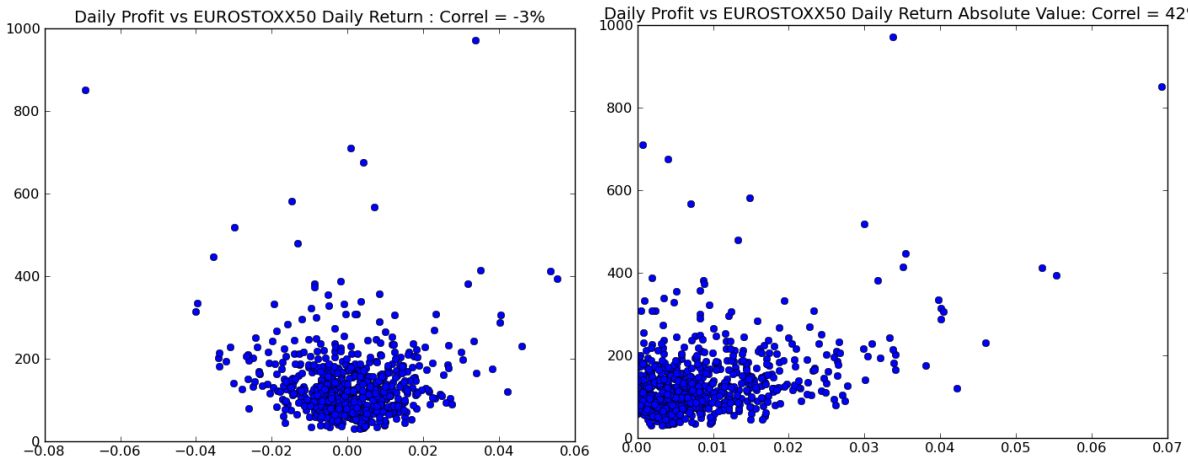


Figure 2.10: Average profit vs EURO STOXX 50 returns

The graph establishes that profits can be better explained by the returns' absolute values than by the returns themselves. A negative correlation (-3%) is observed in the first case, and a positive, more significant, correlation (42%) is observed in the second case. The first result might be explained by the fact that down moves are more brutal (because of agents' panic), thus more profitable for HFT traders. The second result is quite intuitive, since an omniscient aggressive trader makes more money when the order book shows big moves.

Since obtained results show that HFT profits are better explained by the volatility than by the returns, a better intraday volatility indicator should give results that are more significant. In **Figure 2.11** the daily range indicator (Daily High - Daily Low) is computed as a proxy of intraday volatility and HFT profits are plotted vs this indicator. The correlation rises up to 64%. In order to keep in mind the relative value of HFT profits, the average daily profit is plotted vs the Future EURO STOXX 50 total traded volume. The correlation is high (56%) which shows that to make more profit, a HFT needs big volumes. Another interesting result is that the best trading day (out of three years) of the omniscient aggressive HF trader (10-millisecond holding period) ended with less than 50,000 euros of profit. In that same day, 100 billion euros were traded on the Future EURO STOXX 50.

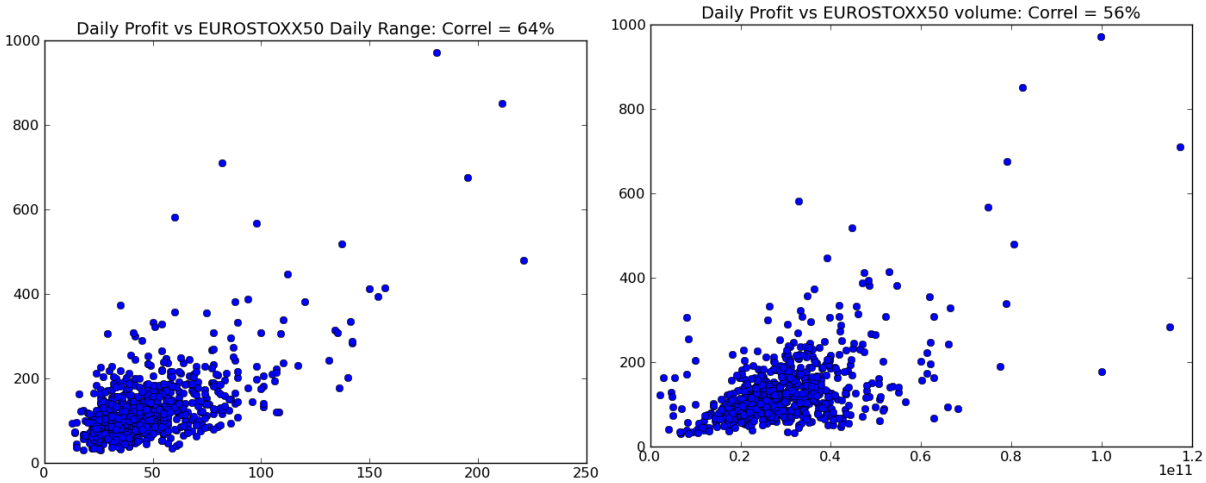


Figure 2.11: The graph on the left shows a high correlation between HFT profitability and the EURO STOXX 50 daily range.

Similar observed effects seen on temporal analysis are present on cross sectional analysis. HFT performs better on volatile and liquid stocks. In particular, a 30% correlation between the stock volatility and the profit made over the stock is observed.

This section concludes with performance comparisons over the main European markets. **Figure 2.12** establishes that in the Italian market, profitable trades are rare. This can be explained by the enormous quantities in the best bid and the best ask. It is rare to observe a big move that consumes all the best limit quantity, however when it happens, the HFT trader can make an important profit per trade due to the big available liquidity. It is also observed that the German market presents more profitable trades due to the big liquidity and small ticks.

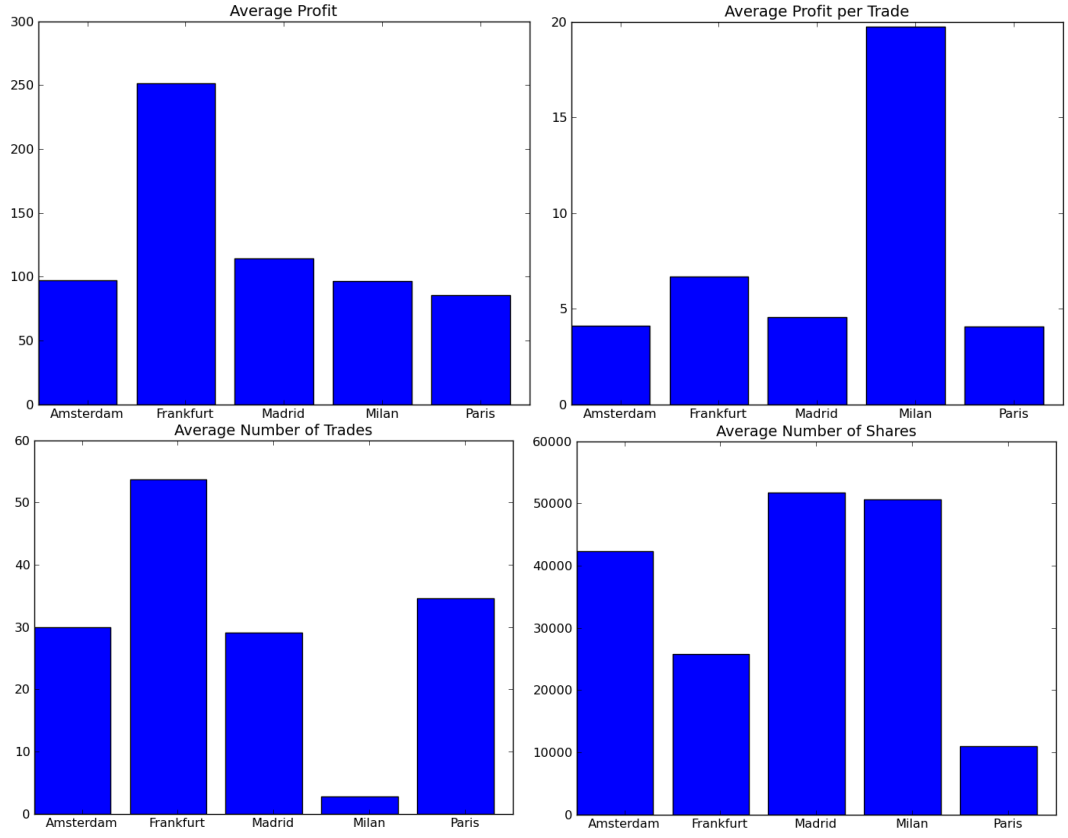


Figure 2.12: HFT Profitability in main European markets

2.3.2 Omniscient order book trading - N steps

In the previous section, empirical results prove that HFT profits are modest for short holding periods. The strategy presented supposes that the trader knows two states of the order book each time; the current state and the next state. The goal of this section is to analyze the optimal strategy in a more general case, and to understand the behavior of a trader who can perfectly predict all the changes in the order book during some omniscience period.

2.3.2.1 Methodology

Similar to the previous section, the experiment consists of a trader observing the present and future states of the order book at a given frequency, and taking all profitable positions. The new element here is that the trader knows not only the state of the order book at time t and time $t+h$, but also knows all intermediary states. The trader can switch positions indefinitely under the constraint of having an empty portfolio at the end of each omniscience period. As usual, the trader can buy or sell all the available quantities on the order book without any impact.

The aim of this section is to understand the behavior of a HF trader able to trade at any frequency relative to a 10-millisecond sampled order book and a 10-second omniscience period. If low latency advantage is important, the trader would rapidly switch his positions (every 10 milliseconds in the extreme case). On the other hand, if profit is made on slower moves, the trader would hold his positions for longer periods (10 seconds in the extreme case).

For each opened and closed position, the holding period T is computed as the difference between the closing position time and the opening position time. If the trader opens many successive positions without closing the previously opened positions, the assumption is made that positions are closed in the chronological order (first opened, first closed).

Finally, the weighted mean holding period is defined as a weighted (by the quantities) mean of all holding periods. The use of weights is very important; with equal weights, a trader holding 1000 shares for 10 seconds and 1 share for 10 milliseconds, would have a holding period of 5 seconds! For the example of **Figure 2.13** the mean holding period is given by $T = \frac{Q_1 T_1 + Q_2 T_2}{Q_1 + Q_2}$.

This measure gives a precise idea about the added value of HFT low latency. If HFT traders make the biggest part of their profits on fast trades, the mean holding period should be significantly smaller than the omniscience period.

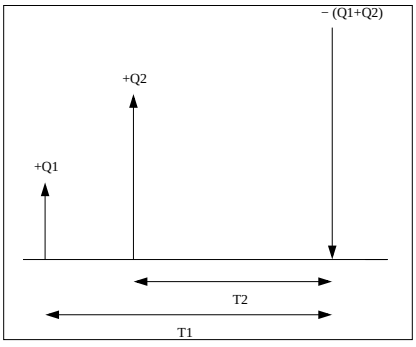


Figure 2.13: Each position is defined by a quantity and a holding period

2.3.2.2 Example

To illustrate the methodology, one stock's mid price evolution over 10 seconds and the corresponding optimal omniscient strategy according to the order book liquidity constraints are plotted in **Figure 2.14**. In this example, the omniscience period is 10 seconds.

The **Table 2.4** shows the detailed evolution of the trader's portfolio over this 10-second period. When a new trade is executed, if the new quantity has the same sign as the existing position, the quantity is added to the list of previous quantities. If the new quantity has an opposite sign, it is used to close the oldest opened position. This rule is used to compute the mean holding period following the formula given in the previous paragraph.

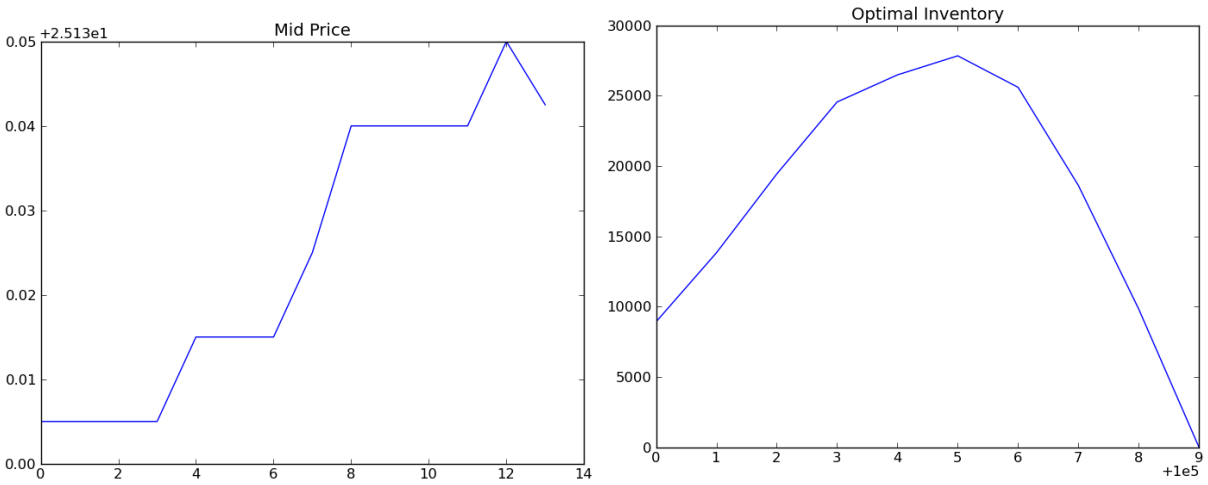


Figure 2.14: Example of mid price evolution (graph on the left) and the corresponding optimal strategy for 10-second omniscience (graph on the right).

Timer	Trade	Opening Times and Held Quantities	Mean Holding Period (seconds)
00:00	+8928	[00:00] [8928]	-
00:01	+4905	[00:00, 00:01] [8928, 4905]	-
00:02	+5603	[00:00, 00:01, 00:02] [8928, 4905, 5603]	-
00:03	+5121	[00:00, 00:01, 00:02, 00:03] [8928, 4905, 5603, 5121]	-
00:04	+1927	[00:00, 00:01, 00:02, 00:03, 00:04] [8928, 4905, 5603, 5121, 1927]	-
00:05	+1357	[00:00, 00:01, 00:02, 00:03, 00:04, 00:05] [8928, 4905, 5603, 5121, 1927, 1357]	-
00:06	-2239	[00:00, 00:01, 00:02, 00:03, 00:04, 00:05] [6689, 4905, 5603, 5121, 1927, 1357]	6.00
00:07	-6980	[00:01, 00:02, 00:03, 00:04, 00:05] [4614, 5603, 5121, 1927, 1357]	6.73
00:08	-8786	[00:02, 00:03, 00:04, 00:05] [1431, 5121, 1927, 1357]	6.63
00:09	-9836	[] []	6.29

Table 2.4: Portfolio evolution and mean weighted holding period computation.

2.3.2.3 Results

The first graph of **Figure 15** shows the main results of this section. A trader who knows the order book evolution perfectly for 10 seconds with 10-millisecond sampling, and trades with 0 costs, would have an average holding period of 3.8 seconds. This holding period is 380 times greater than the smallest possible holding period; 10 milliseconds. Such result mitigates the claim that low latency advantage is the main key of HFT profit. Making money when hitting the order book and paying the bid ask cross is very difficult. When the trader is subject to 10-bps trading costs, the holding period increases to 5.1 seconds. The number of trades decreases from 106,000 trades to only 10,000 trades per stock per day.

In the second graph of **Figure 2.15** the holding period is plotted vs the bid ask spread. It can be seen that the holding period depends strongly on trading fees. When trading becomes costly, only very profitable trades are executed. Those trades should provide a return higher than the fees. Such high returns are more likely observed on long holding periods.

The dependence of the holding period on Bid Ask spread is less clear. However, a positive correlation of 17% can be seen. The Bid Ask spread represents the average crossing cost. A positive correlation is consistent with the fact that holding periods increase with trading costs.

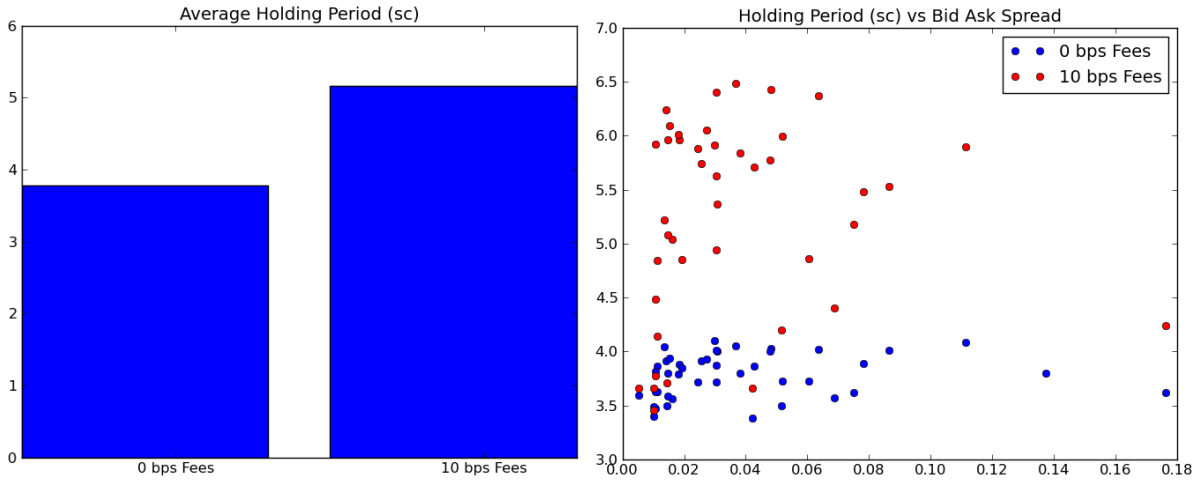


Figure 2.15: Average holding period (in seconds)

2.4 Conclusions

This paper provides a large empirical study dealing with 50 European liquid stocks over three years (2011-2013). To compute an objective upper bound of aggressive HFT profits, a one-step omniscient strategy is applied. The results confirm studies from other papers dealing with other markets (Forex, US Equities..) [55] [30] [3] [17] [10] [1] . Profits are rather modest and even negligible for the shortest holding periods.

To get rid of the fixed holding period hypothesis, a new method to compute an optimal HFT strategy is introduced: the n-steps omniscient strategy. This method is used to compute a new measure: the weighted mean holding period. Results show that this period is 400 times greater than the smallest possible period. In other words, an omniscient trader is trading on average with a frequency 400 times slower than the highest available frequency, which shows that hitting the order book rapidly in order to take advantage of low latency information asymmetry is not that profitable.

Chapter 3

Empirical Evidence of Market Inefficiency: Predicting Single-Stock Returns

Note:
- This chapter is published in the proceedings of the conference "Econophysics-Kolkata VIII: Econophysics and data driven modelling of market dynamics".
- This chapter is submitted to the forum "Scenarios, Stress and Forecasts In Finance", Institut Louis Bachelier, Paris, March 2015.

Abstract

Although it is widely assumed that the stock market is efficient, some empirical studies have already tried to address the issue of forecasting stock returns. As far as is known, it is hard to find a paper involving not only the forecasting statistics but also the forecasting profitability. This paper aims to provide an empirical evidence of the market inefficiency and to present some simple realistic strategies based on forecasting stocks returns. In order to achieve this study, some linear and non linear algorithms are used to prove the predictability of returns. Many regularization methods are introduced to enhance the linear regression model. In particular, the RIDGE method is used to address the colinearity problem and the LASSO method is used to perform variable selection. The different obtained results show that the stock market is inefficient and that profitable strategies can be computed based on forcasting returns. Empirical tests also show that simple forecasting methods perform almost as well as more complicated methods.

Contents

Introduction **52**
3.1 Data, methodology and performances measures **53**
3.1.1 Data 53
3.1.2 Methodology 54
3.1.3 Performance measures 54
3.2 Conditional probability matrices **55**
3.2.1 Binary method 56
3.2.2 Four-class method 60
3.3 Linear regression **62**
3.3.1 Ordinary least squares (OLS) 62
3.3.2 Ridge regression 64
3.3.3 Least Absolute Shrinkage and Selection Operator (LASSO) 70
3.3.4 ElASTIC NET (EN) 72

Introduction

Forecasting the market has been one of the most exciting financial subjects for over a century. In 1900, L. Bachelier [4] admitted, "Undoubtedly, the Theory of Probability will never be applicable to the movements of quoted prices and the dynamics of the Stock Exchange will never be an exact science. However, it is possible to study mathematically the static state of the market at a given instant to establish the probability law for the price fluctuations that the market admits at this instant." 70 years later, Fama [34] proposed some formal definitions of the market efficiency; "A market in which prices always fully reflect available information is called efficient." Opinions have been always divergent about the market efficiency. B Malkiel [67] concluded that most investors trying to predict stocks' returns always ended up with profits inferior to passive strategies. In his famous book, *Fooled by Randomness*, N. Taleb [87] argued that even the best performances can be explained by luck and randomness. On the other hand, finance professionals demonstrated, in real life, that they can always make money beating the market; see Warren Buffett's response to efficient market claims [20].

The recent rise in electronic markets lead to big available financial data. The attempt to discover some predictable, and hopefully profitable, signal in the middle of those millions of numbers has never been as high as today.

In the academic world, the order book empirical properties were studied in many papers (see for example [12] , [52] , [15] , [22] and [63]). In particular, A. Chakraborti et al [22] studied in detail the statistical properties of the intraday returns, and came to the conclusion that there is no evidence of correlation between successive returns. Similarly, Lillo and Farmer [63] concluded that stock returns contain negligible temporal autocorrelation. Fortunately, B. Zheng, E. Moulines and F. Abergel [100] found some promising results, in particular the liquidity imbalance on the best bid/ask seems to be informative to predict the next trade sign.

In the professional world, many books present hundred of strategies predicting the market and always earning money; see [72], [95] for example. When testing those strategies in other samples, results are so different and the strategies are no longer profitable. It is possible that the overfit of such methods played a key role in the good performances published in those books.

This study was performed from both an academic and a professional perspective. For each prediction method, not only are statistical results presented, but also presented are the performances of the correspondant strategies. The aim is to give another point of view of a good prediction and of an efficient market.

This work is organized as follows: In the first section, the data and the test methodology are presented. In the second section a non linear method, based on conditional probability matrices, is used to test the predictive power of each indicator. In the last section, the linear regression is introduced to combine the different indicators and many regularization ideas are tested in order to enhance the performances of the strategies.

3.1 Data, methodology and performances measures

3.1.1 Data

This paper focuses on the EURO STOXX 50 European liquid stocks. One year (2013) of full daily order book data provided by BNP Paribas are used to achieve the study. For a stock with a mid price S_t at time t, the return to be predicted over a period dt is $Ln(\frac{S_{t+dt}}{S_t})$. At the time t, one can use all the available data for any time $s \leq t$ to perform the prediction.

In section 2 and section 3, the focus is on predicting the stocks' returns over a fixed period dt using some order book indicators. Once the returns and the indicators are computed, the data are sampled on a fixed time grid from 10h to 17h with a resolution dt. Three different resolutions are tested; 1, 5 and 30 minutes.

Below are the definitions of the studied indicators and the rationale behind using them to predict the returns:

Past return: The past return is defined as $Ln(\frac{S_t}{S_{t-dt}})$. Two effects justify the use of the past return indicator to predict the next return; the mean-reversion effect and the momentum effect. If a stock, suddenly, shows an abnormal return that, significantly, deviates the stock's price from its historical mean value, the mean reversion effect is observed when an opposite return occurs rapidly to put the stock back in its usual average price range. On the other hand, if the stock shows, progressively, an important and continuous deviation; the momentum effect occurs when more market participants are convinced of the move and trade in the same sense increasing the deviation even more.

Order book imbalance: The liquidity on the bid (respectively ask) side is defined as $Liq_{bid} = \sum_{i=1}^{5} w_i b_i bq_i$ (respectively $Liq_{ask} = \sum_{i=1}^{5} w_i a_i aq_i$), where b_i (respectively a_i) is the price at the limit i on the bid (respectively ask) side, bq_i (respectively aq_i) is the corresponding available quantity, and w_i is a decreasing function on i used to give more importance to the best limits. Those indicators give an idea about the instantaneous money available for trading on each side of the order book. Finally, the order book imbalance is defined as $Ln(\frac{Liq_{bid}}{Liq_{ask}})$. This indicator summarizes the order book static state and gives an idea about the buy-sell instantaneous equilibrium. When this indicator is significantly higher (respectively lower) than 0, the available quantity at the bid side is significantly higher (respectively lower) than the one at the ask side; only few participants are willing to sell (respectively buy) the stock, which might reflect a market consensus that the stock will move up (respectively down).

Flow quantity: This indicator summarizes the order book dynamic over the last period dt. Q_b (respectively Q_s) is denoted as the sum of the bought (respectively sold) quantities, over the last period dt and the flow quantity is defined as $Ln(\frac{Q_b}{Q_s})$. This indicator is close to the order flow and shows a high positive autocorrelation. The rationale behind using the flow quantity is to verify if the persistence of the flow is informative about the next return.

EMA: For a process $(X)_{t_i}$ observed on discrete times (t_i), the Exponential Moving Average $EMA(d, X)$ of delay d is defined as $EMA(d, X)_{t_0} = X_{t_0}$ and for $t_{1 \leq i}$, $EMA(d, X)_{t_i} = \omega X_{t_i} + (1 - \omega) EMA(d, X)_{t_{i-1}}$, where $\omega = min(1, \frac{t_i - t_{i-1}}{d})$. The EMA is a weighted average of the process with an exponential decay. The smaller d is, the shorter the EMA memory is.

3.1.2 Methodology

The aim of this study is to prove, empirically, the market inefficiency by predicting the stocks' returns for three different periods: 1, 5 and 30 minutes.

In section 2, the used indicators are the past returns, the order book imbalance and the flow quantity. A simple method based on historical conditional probabilities is used to prove, separately, the informative effect of each indicator.

In section 3, the three indicators and their $EMA(X, d)$ for $d \in (1, 2, 4, 8, 16, 32, 64, 128, 256)$ are combined in order to perform a better prediction than the mono indicator case. Different methods, based on the linear regression, are tested. In particular, the statistical and the numerical stability problems of the linear regression are addressed.

In the different sections, the predictions are tested statistically, then used to design a simple trading strategy. The goal is to verify, whether or not one can find a profitable strategy covering 0.5 basis point trading costs. This trading cost is realistic and corresponds to many funds, brokers, and banks trading costs. The possibility of computing, if it exists, a strategy, profitable, after paying the costs, would be an empirical argument of the market inefficiency.

Notice that, in all the sections, the learning samples are sliding windows containing sufficient number of days, and the testing samples are the next days. The models parameters are fitted on the learning sample (called in-sample) and the strategies are tested on the testing sample (called out of sample). The sliding training avoids any overfit problem since performances are only computed out of sample.

3.1.3 Performance measures

In the most of the studies addressing the market efficiency, the results are summarized in the linear correlation. However, this measure is not enough to conclude about the returns predictability or the market efficiency. Results interpretation should depend on the predicted signal and the trading strategy. A 1% correlation is high if the signal is supposed to be totally random, and 99% correlation is insuffissant if the signal is supposed to be perfectly predictable.

Moreover, a trader making 1 euro each time trading a stock with 50.01% probability and losing 1 euro with 49.99% probability, might be considered as a noise trader. However, if this strategy can be run, over 500 stocks, one time a second, for 8 hours a day, at the end of the day the gain will be the sum S_n of $n = 14.4$ million realisations. Using the central limit theorem, $\frac{S_n}{n}$ has a normal law $N(E, \frac{\sigma}{\sqrt{n}})$ (with the classic notations). Thus the probability of having a negative trading day is $\Phi(\frac{-E\sqrt{n}}{\sigma}) = \Phi(-0.62) = 26.5\%$, so much lower than the one of a noise trader.

In this paper, returns are considered predictable and thus the market is considered inefficient, if one can run a profitable strategy covering the trading costs.

3.2 Conditional probability matrices

The conditional probability matrices method uses observed frequencies as an estimation of the conditional probability law. To apply this method, data need to be descritized in a small number of classes. Denote the explanatory variable as X, the return as Y and the frequencies matrix as M. Denote the classes of X (respectively Y) as $C^X = \{C_i^X : i \in \mathbb{N}_+ \cap \{i \leq S_X\}\}$ (respectively $C^Y = \{C_i^Y : i \in \mathbb{N}_+ \cap \{i \leq S_Y\}\}$). S_X (respectively S_Y) denotes the total number of classes for X (respectively Y). For a given learning period $[0, T]$ containing N observations, the frequencies matrix at the time T is constructed as:

$$M_T^{i,j} = card(\{(X_{t_n} \in C_i^X, Y_{t_n} \in C_j^Y)\})$$

where $n \in \mathbb{N}_+ \cap \{n \leq N\}$, and X_{t_n} (respectively Y_{t_n}) is the n^{th} observed value of X (respectively Y), observed at the time t_n. Note that the return Y_{t_n} is backshifted for one instant (namely $Y_{t_n} = Ln(\frac{S_{t_{n+1}}}{S_{t_n}})$). Finally, the prediction of the next Y conditional to the last observed X_T can be computed using the matrix M_T.

The idea of this method is a simple application of the statistical independence test. If some events A="$X_{t_n} \in C_i^X$" and B="$Y_{t_n} \in C_j^Y$" are statistically independent then $P(A|B) = P(A)$. For example, to check if the past returns (denoted X in this example) can help predicting the future returns (denoted Y in this example), the returns are classified into 2 classes, then the empirical historical frequencies matrix is computed. **Table 3.1** shows the results for the 1-minute returns of Deutsh Telecom over the year 2013.

	$A =$ "$Y < 0$"	$B =$ "$Y > 0$"
$A =$ "$X < 0$"	19,950	21,597
$B =$ "$X > 0$"	21,597	20,448

Table 3.1: Historical frequencies matrix for Deutsh Telecom over 2013

In probabilistic terms, the historical probability to observe a negative return is $P(A) = 49.70\%$ and to observe a positive return is $P(B) = 50.30\%$. Thus a trader always buying the stock would have a success rate of 50.30%. Notice that: $P(A/A) = 48.02\%$, $P(B/A) = 51.98\%$, $P(A/B) = 51.37\%$, $P(B/B) = 48.63\%$. Thus, a trader playing the mean-reversion (buy when the past return is negative and sell when the past return is positive), would have a success rate of 51.67%. Notice that the same approach as **1.3** gives a success rate, when trading the strategy over 500 stocks, of 54.38% for the buy strategy and of 72.91% for the mean reversion strategy.

This simple test shows that the smallest statistical bias can be profitable and useful for designing a trading strategy. However the previous strategy is not realistic; the conditional probabilities are computed in sample and the full sample data of Deutsh Telecom was used for the computation. In reality, predictions have to be computed using only the past data. It is, thus, important to have stationary probabilities. **Table 3.2** shows that the monthly observed frequencies are quite stable, and thus can be used to estimate out of sample probabilities. Each month, one can use the observed frequencies of the previous month as an estimator of current month probabilities.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
P(A/A)	0.49	0.48	0.47	0.48	0.47	0.50	0.48	0.51	0.51	0.47	0.46	0.44
P(B/A)	0.51	0.52	0.53	0.52	0.53	0.50	0.52	0.49	0.49	0.53	0.54	0.56
P(A/B)	0.50	0.52	0.51	0.53	0.52	0.49	0.51	0.50	0.51	0.50	0.51	0.55
P(B/B)	0.50	0.48	0.49	0.47	0.48	0.51	0.49	0.50	0.49	0.50	0.49	0.45

Table 3.2: Monthly historical conditional probabilities: In the most cases, $P(A/A)$ and $P(B/B)$ are lower than 50% where $P(B/A)$ and $P(A/B)$ are higher than 50%.

In the following paragraphs, frequencies matrices are computed on sliding windows for the different indicators. Several classification and prediction methods are presented.

3.2.1 Binary method

In the binary case, returns are classified into positive and negative as the previous example and explanatory variables are classified relatively to their historical mean. A typical constructed matrix is shown in **Table 3.1**. Denote, in the **Table 3.1** example, $C_1^X = \{X < \overline{X} = 0\}$, $C_2^X = \{X > \overline{X} = 0\}$, $C_1^Y = \{Y < 0\}$, $C_2^Y = \{Y > 0\}$. Y can be predicted using different formula based on the frequency matrix. Below some estimators examples:

$\widehat{Y_1}$: The sign of the most likely next return conditionally to the current state.
$\widehat{Y_2}$: The expectation of the most likely next return conditionally to the current state.
$\widehat{Y_3}$: The expectation of next return conditionally to the current state.

$$\widehat{Y_1} = \begin{cases} +1 & \text{if } X_T \in C_1^X \\ -1 & \text{if } X_T \in C_2^X \end{cases}$$

$$\widehat{Y_2} = \begin{cases} E(Y|Y \in C_2^Y \cap X \in C_1^X) & \text{if } X_T \in C_1^X \\ E(Y|Y \in C_1^Y \cap X \in C_2^X) & \text{if } X_T \in C_2^X \end{cases}$$

$$\widehat{Y_3} = \begin{cases} E(Y|X \in C_1^X) & \text{if } X_T \in C_1^X \\ E(Y|X \in C_2^X) & \text{if } X_T \in C_2^X \end{cases}$$

In this study, only results based on the estimator $\widehat{Y_3}$ (denoted $\widehat{Y}$ in the rest of the paper) are presented. Results computed using different other estimators are equivalent and the differences do not impact the conclusions. To measure the quality of the prediction, four tests are applied:

AUC: (Area under the curve) [37] combines the true positive rate and the false positive rate to give an idea about the classification quality.

Accuracy: defined as the ratio of the correct predictions (Y and $\widehat{Y}$ have the same sign).

Gain: computed on a simple strategy to measure the prediction performance. Predictions are used to run a strategy that buys when the predicted return is positive and sells when it is negative. At each time, for each stock the strategy's position is in $\{-100,000, 0\ , +100,000\}$.

Profitability: defined as the gain divided by the traded notional of the strategy presented above. This measure is useful to estimate the gain with different transaction costs.

Figure 3.1 summarizes the statistical results of predicting the 1-minute returns using the three indicators. For each predictor, the AUC and the accuracy are computed over all the stocks. Notice that for each stock, results are computed over more than 100,000 observations and the amplitude of the 95% confidence interval is around 0.6%. For the three indicators, the accuracy and the AUC are significantly higher than the 50% random guessing threshold. The graph shows also that the order book imbalance gives the best results and that the past resturn is the least successful predictor. Detailed results per stock are given in **Table 5.1** of **Appendix 2**.

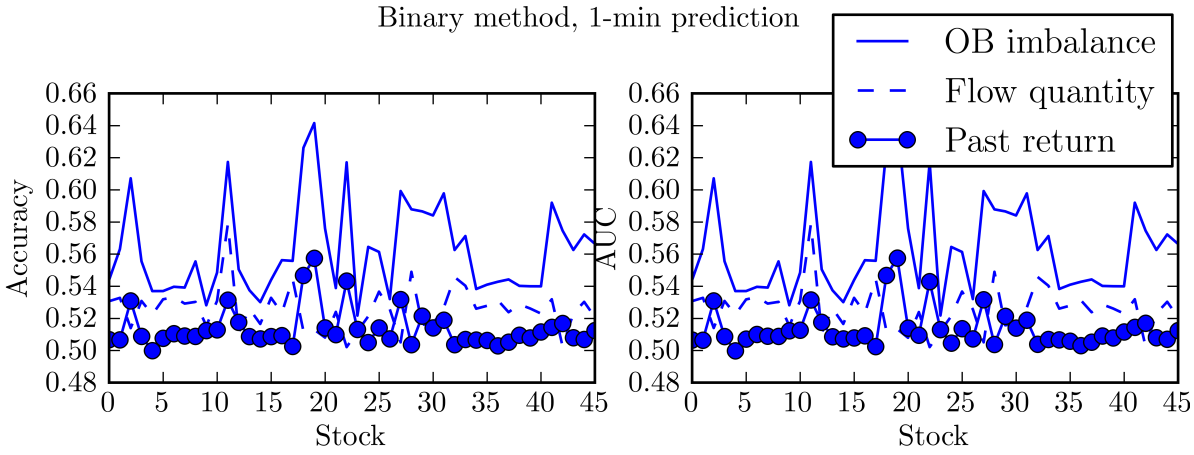


Figure 3.1: The quality of the binary prediction: The AUC and the Accuracy are higher than 50%. The three predictors are better than random guessing and are significatly informative.

In **Figure 3.2**, the performances of the trading strategies based on the prediction of the 1-minute returns are presented. The strategies are profitable and the results confirm the predictability of the returns (see the details in **Table 5.2** of **Appendix 2**).

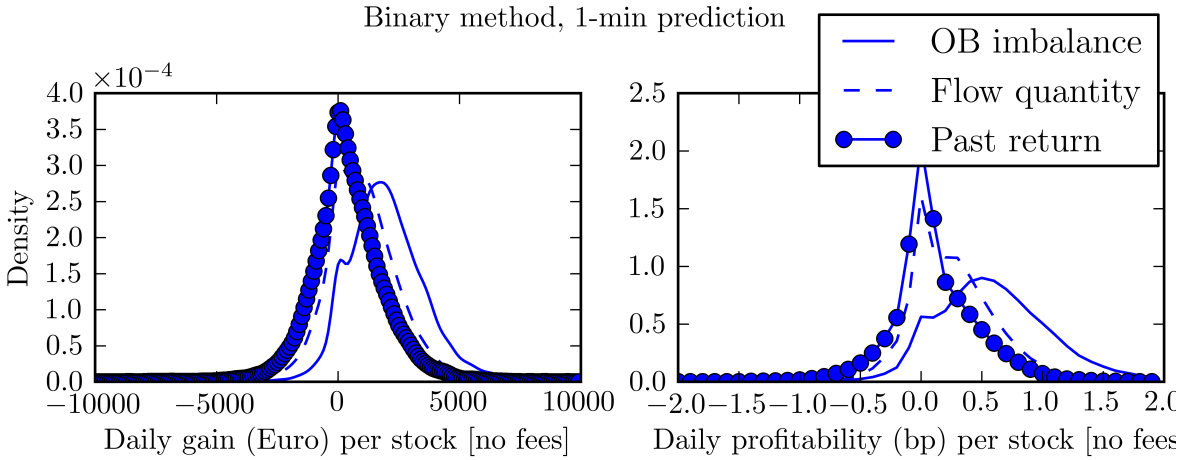


Figure 3.2: The quality of the binary prediction: For the 3 predictors, the densities of the gain and the profitability are positively biased, confirming the predictability of the returns.

In **Figure 3.3**, the cumulative gains of the strategies based on the 3 indicators over 2013 are represented. When trading without costs, predicting the 1-minute return using the past return and betting 100,000 euros at each time, would make a 5-million Euro profit. Even better, predicting using the order book imbalance would make more than 20 million Euros profit. The results confirm the predictability of the returns, but not the inefficiency of the market. In fact, **Figure 3.4** shows that, when adding the 0.5 bp trading costs, only the strategy based on the order book imbalance remains (marginally) positive. Thus, no conclusion, about the market efficiency, can be made (see more details in **Table 5.3** of **Appendix 2**).

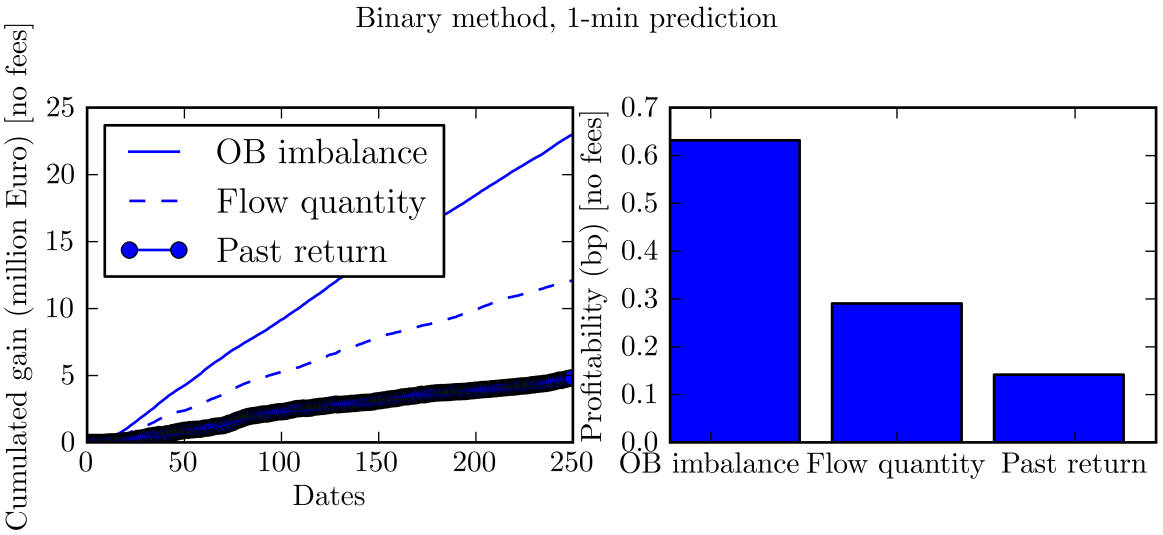


Figure 3.3: The quality of the binary prediction: The graphs confirm that the 3 indicators are informative and that the order book imbalance indicator is the most profitable.

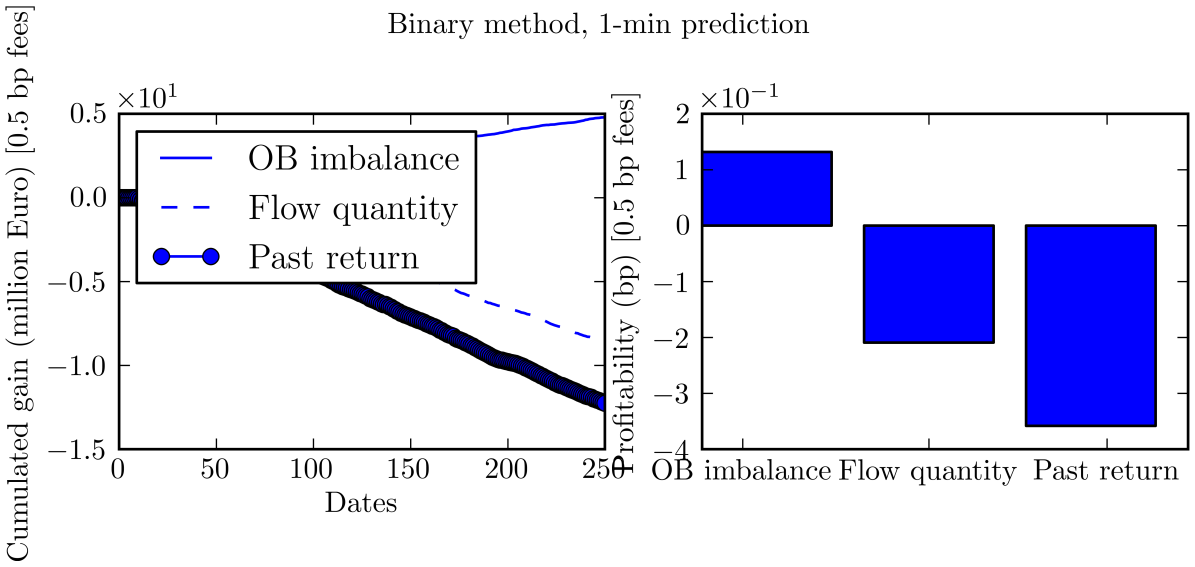


Figure 3.4: The quality of the binary prediction: When adding the 0.5 bp trading costs, the strategies are no longer very profitable.

Figure 3.5 represents the cumulative gain and the profitability for the 5-minute and the 30-minute strategies (with the trading costs). The strategies are not profitable. Moreover, the predictive power decreases with an increasing horizon. Similar as the 1-minute prediction,

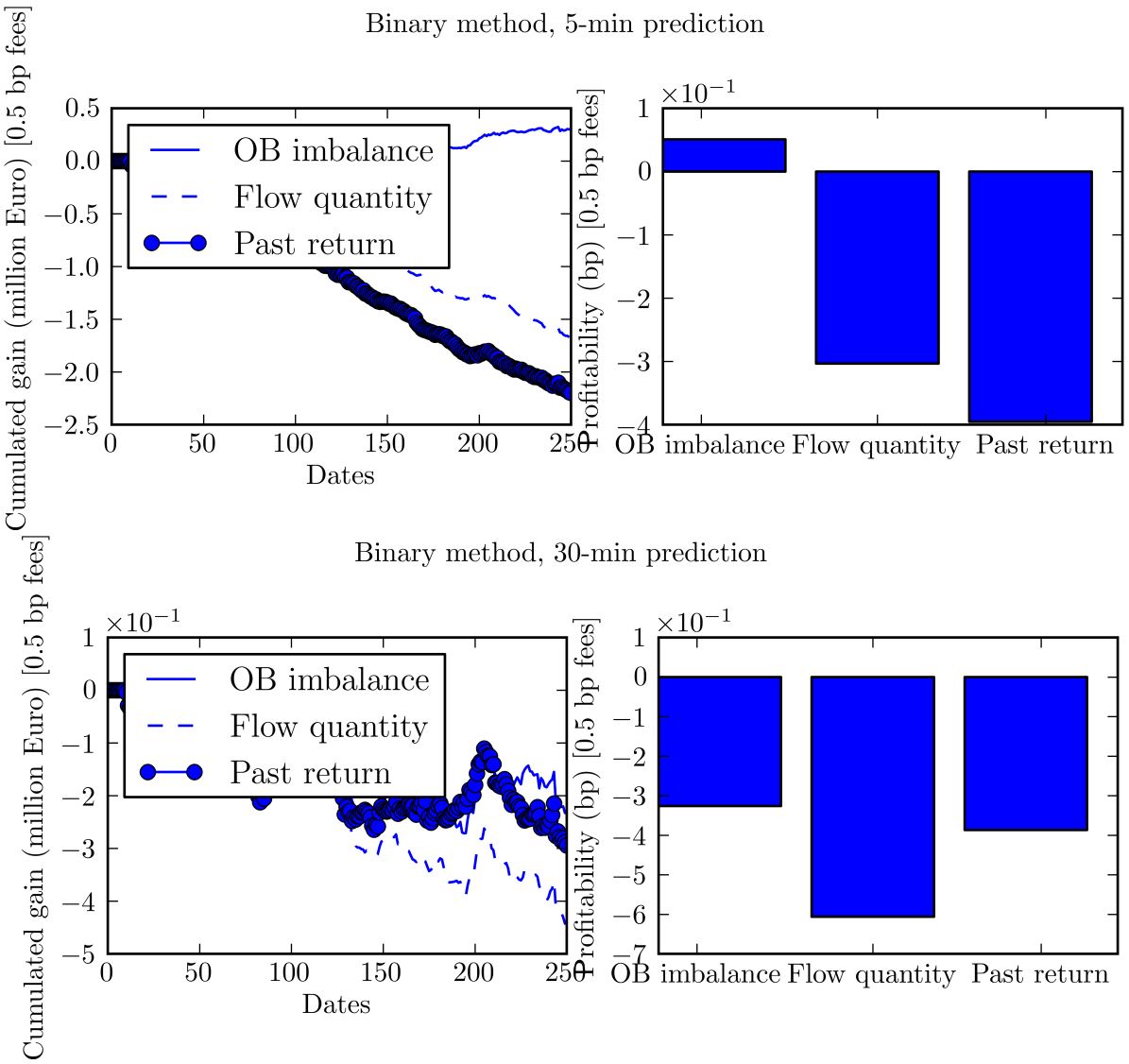


Figure 3.5: The quality of the binary prediction: The strategies are not profitable. Moreover, the performances decreases significantly compared to the 1-minute horizon.

the detailed results of the 5-minute prediction can be found in **Tables 5.4, 5.5 and 5.6** of **Appendix 2**. Those of the 30-minute prediction can be found in **Tables 5.7, 5.8 and 5.9** of the same Appendix.

The results of the binary method show that the returns are significantly predictable. Nevertheless, the strategies based on those predictions are not sufficiently profitable to cover the trading costs. In order to enhance the predictions, the same idea is applied to the four-class case. Moreover, a new strategy based on a minimum threshold of the expected return is tested.

3.2.2 Four-class method

The indicator X is now classified into 4 classes; "very low values" C_1^X, "low values" C_2^X, "high values" C_3^X and "very high values" C_4^X. At each time t_n, Y is predicted as $\widehat{Y} = E(Y|X \in C_i^X)$, where C_i^X is the class of the current observation X_{t_n}. As the previous case, the expectation is estimated from the historical frequencies matrix. Finally, a new trading strategy is tested. The strategy is to buy (respectively sell) 100,000 euros when $\widehat{Y}$ is positive (respectively negative) and $|\widehat{Y}| > \theta$, where θ is a minimum threshold (1 bp in this paper). Notice that the case $\theta = 0$ corresponds to the strategy tested in the binary case.

The idea of choosing $\theta > 0$ aims to avoid trading the stock when the signal is noisy. In particular, when analyzing the expectations of Y relative to the different classes of X, it is always observed that the absolute value of the expectation is high when X is in one of its extreme classes (C_1^X or C_4^X). On the other hand, when X is in one of the intermediary classes (C_2^X or C_3^X) the expectation of Y is close to 0 reflecting a noisy signal.

For each indicator X, the classes are defined as $C_1^X =]-\infty, X_a[$, $C_2^X =]X_a, X_b[$, $C_3^X =]X_b, X_c[$ and $C_4^X =]X_c, +\infty[$. To compute X_a, X_b and X_c, the 3 following classifications were tested:

Quartile classification: In the in-sample period, the quartile Q_1, Q_2 and Q_3 are computed for each day then averaged over the days. X_a, X_b and X_c corresponds, respectively, to $\overline{Q_1}$, $\overline{Q_2}$ and $\overline{Q_3}$.

K-means classification: The K-means algorithm [46], applied to the in-sample data with $k = 4$, gives the centers $G_1, G2, G_3$ and G_4 of the optimal (in the sense of the minimum within-cluster sum of squares) clusters. X_a, X_b and X_c are given respectively by $\frac{G_1+G_2}{2}$, $\frac{G_2+G_3}{2}$ and $\frac{G_3+G_4}{2}$.

Mean-variance classification: The average $\overline{X}$ and the standard deviation $\sigma(X)$ are computed in the learning period. Then, X_a, X_b and X_c correspond, respectively, to $\overline{X} - \sigma(X)$, $\overline{X}$ and $\overline{X} + \sigma(X)$.

In this paper, only the results based on the mean-variance classification are presented. The results computed using the two other classifications are equivalent and the differences do not impact the conclusions.

Figure 3.6 compares the profitabilities of the binary and the 4-class methods. For the 1-minute prediction, the results of the 4-class method are significantly better. For the longer horizons, the results of the both methods are equivalent. Notice also that, using the best indicator, in the 4-class case, one could obtain a significantly positive performance after paying the trading costs. The detailed results per stock are given in **Tables 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17**, and **5.18** of **Appendix 3**.

The interesting result of this first section is that even when using the simplest statistical learning method, the used indicators are informative and provide a better prediction than random guessing. However, in most cases, the obtained performances are too low to conclude about the market inefficiency.

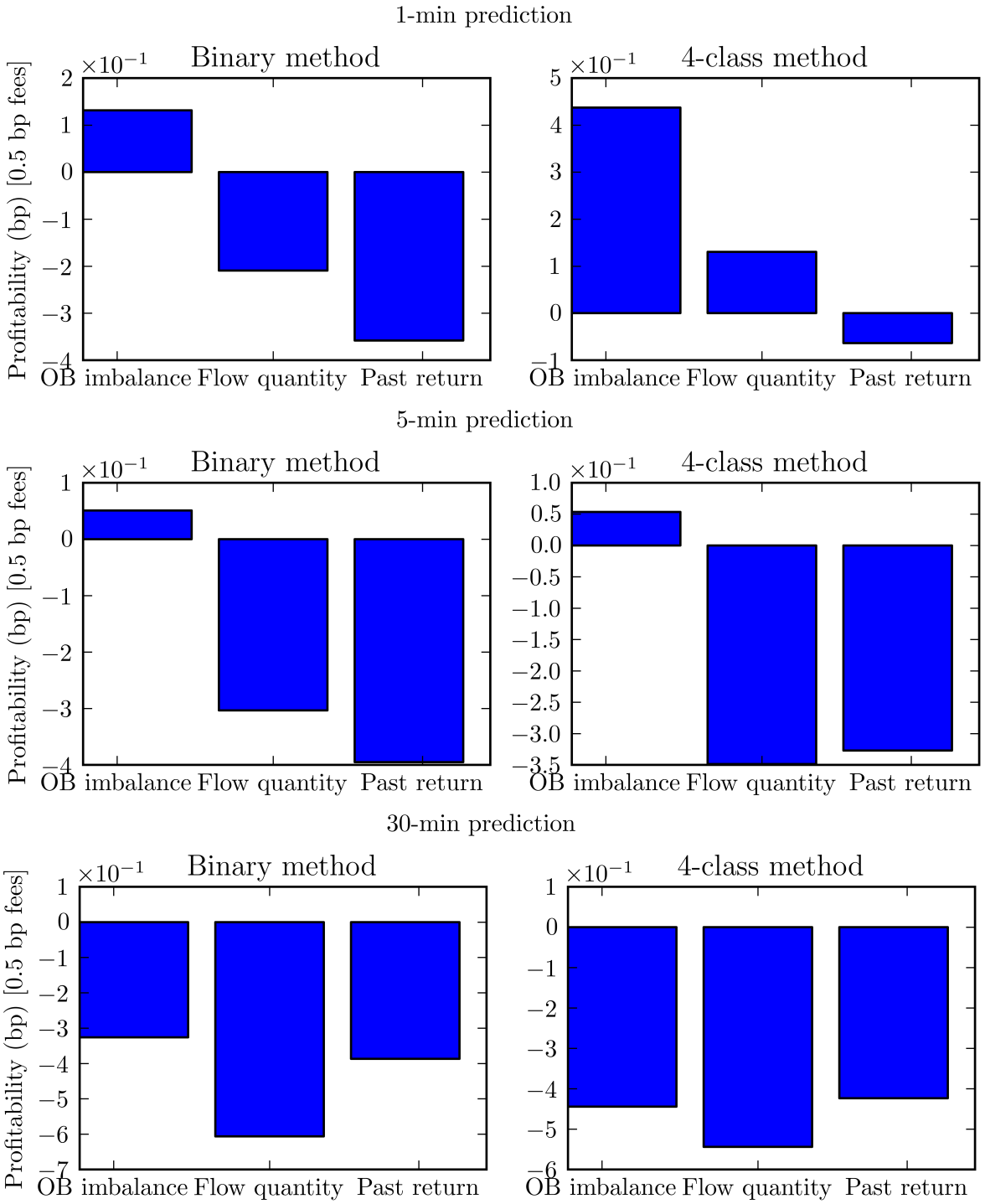


Figure 3.6: The quality of the 4-class prediction: For the 1-minute prediction, the results of the 4-class method are significantly better than the results of the binary one. For longer horizons, both strategies are not profitable when adding the trading costs.

In order to enhance the performances, the 3 indicators and their exponential moving average are combined using some classic linear methods in the next section.

3.3 Linear regression

In this section, the matrix X denotes a 30-column matrix containing the 3 indicators and their $EMA(d)$ for $d \in (1, 2, 4, 8, 16, 32, 64, 128, 256)$. The vector Y denotes the target to be predicted. Results of the previous section proved that the used indicators are informative and thus can be used to predict the target. In general, one can calibrate, on the learning sample, a function f such that $f(X)$ is "the closet possible" to Y and hope that, for some period after the learning sample, the relation between X and Y is still close enough to the function f. Hence $f(X)$ would be a "good" estimator of Y. Due to the finite number of observations in the learning sample, one can always find $f(X)$ arbitrary close to Y by increasing the number of the freedom degree. However, such perfect in-sample calibration overfits the data and the out of sample results are always irrelevant.

In the linear case, f is supposed to be linear and the model errors are supposed to be independent and identically distributed [83] (Gaussian in the standard textbook model). A more mathematical view of linear regression is that it is a probabilistic model of Y given X that assumes:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

For technical reasons, the computations are done with z-scored data (use $\frac{X_i - \overline{X_i}}{\sigma(X_i)}$ in stead of X_i).

3.3.1 Ordinary least squares (OLS)

OLS method consists of estimating the unknown parameter β by minimizing the sum of squares of the residuals between the observed variable Y and the linear approximation $X\beta$. The estimator is denoted $\widehat{\beta}$ and is defined as

$$\widehat{\beta} = argmin_\beta(J_\beta = ||Y - X\beta||_2^2)$$

This criterion is reasonable if at each time i the row X_i of the matrix X and the observation Y_i of the vector Y represent independent random sample from their populations.

The cost function J_β is quadratic on β and differentiating with respect to β gives:

$$\frac{\delta J_\beta}{\delta \beta} = 2t_X X\beta - 2t_X Y$$
$$\frac{\delta^2 J_\beta}{\delta \beta \delta \beta} = 2t_X X$$

When $t_X X$ is invertible, setting the first derivative to 0, gives the unique solution $\widehat{\beta} = (t_X X)^{-1} t_X Y$. The statistical properties of this estimator can be calculated straightforward as follows:

$$E(\widehat{\beta}|X) = (t_X X)^{-1} t_X E(Y|X) = (t_X X)^{-1} t_X X\beta = \beta$$
$$Var(\widehat{\beta}|X) = (t_X X)^{-1} t_X Var(Y|X) X (t_X X)^{-1} = (t_X X)^{-1} t_X \sigma^2 I X (t_X X)^{-1} = \sigma^2 (t_X X)^{-1}$$
$$E(||\widehat{\beta}||_2^2|X) = E(t_Y X (tXX)^{-2} tXY|X)) = Trace(X(tXX)^{-2} tX\sigma^2 I) + ||\beta||_2^2 = \sigma^2 Trace((tXX)^{-1}) + ||\beta||_2^2$$
$$MSE(\widehat{\beta}) = E(||\widehat{\beta} - \beta||_2^2|X) = E(||\widehat{\beta}||_2^2|X) - ||\beta||_2^2 = \sigma^2 Trace((tXX)^{-1}) = \sigma^2 \sum \frac{1}{\lambda_i}$$

Where MSE denotes the mean squared error and $(\lambda)_i$ denote the eigen values of $t_X X$. Notice that the OLS estimator is unbiased, but can show an arbitrary high MSE when the matrix $t_X X$ has close to 0 eigen values.

In the out of sample period, $\widehat{Y} = X\widehat{\beta}$ is used to predict the target. As seen in section 2, the corresponding trading strategy is to buy (respectively sell) 100,000 euros when $\widehat{Y} > 0$ (respectively $\widehat{Y} < 0$). To measure the quality of the predictions, the binary method based on the order book imbalance indicator is taken as a benchmark. The linear regression is computed using 30 indicators, including the order book imbalance, thus it should perform at least as well as the binary method. **Figure 3.7** compares the profitabilities of the two strategies. The detailed statistics per stock are given in **Tables 5.19, 5.20** and **5.21** of **Appendix 4**. Similar

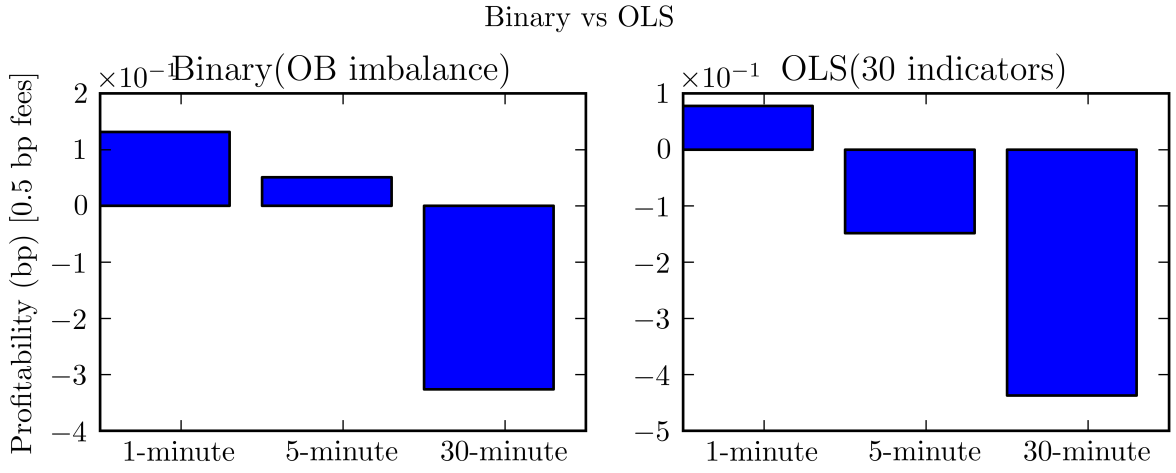


Figure 3.7: The quality of the OLS prediction: The results of the OLS method are not better than those of the binary one.

to the binary method, the performances of the OLS method decrease with an increasing horizon. Moreover, the surprising result is that when combining all the 30 indicators, the results are not better than just applying the binary method to the order book imbalance indicator. This leads to questioning the quality of the regression.

Figure 3.8 gives some example of the OLS regression coefficients. It is observed that the coefficients are not stable over the time. For example, for some period, the regression coefficient of the order book imbalance indicator is negative. This does not make any financial sense. In fact, when the imbalance is high, the order book shows more liquidity on the bid side (participants willing to buy) than the ask side (participants willing to sell). This state of the order book is observed on average before an up move -i.e. a positive return. The regression coefficient should, thus, be always positive. It is also observed that, for highly correlated indicators, the regression coefficients might be so different. This result also does not make sense, since one would expect to have close coefficients for similar indicators.

From a statistical view, this is explained by the high MSE caused by the high colinearity between the variables. In the following paragraphs, the numerical view is also addressed and some popular solutions to the OLS estimation problems are tested.

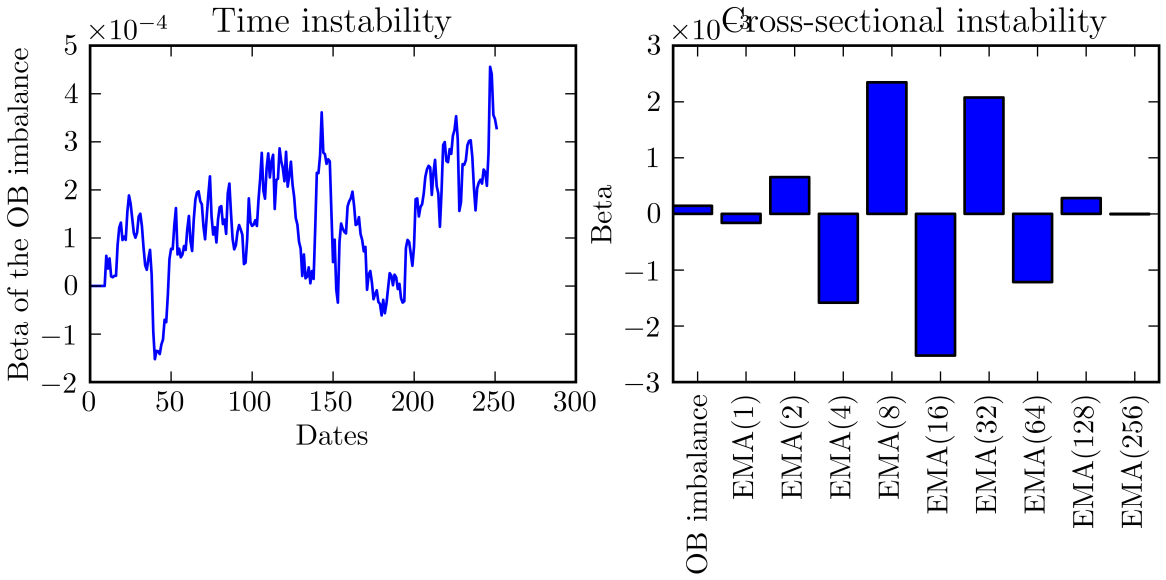


Figure 3.8: The quality of the OLS prediction: The graph on the left shows the instability of the regression coefficient of the order book imbalance indicator over the year 2013 for the stock Deutsh Telecom. The graph on the right shows, for a random day, a very different coefficients for similar indicators; the order book imbalance and its exponential moving averages.

3.3.2 Ridge regression

When solving a linear system $AX = B$, with A invertible, if a small change in the coefficient matrix (A) or a small change in the right hand side (B) results in a large change in the solution vector (X) the system is considered ill-conditioned. The resolution of the system might give a non reliable solution which seems to satisfy the system very well.

An example of an ill-conditioned system is given bellow:

$$\begin{bmatrix} 1.000 & 2.000 \\ 3.000 & 5.999 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.000 \\ 11.999 \end{bmatrix} => \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.000 \\ 1.000 \end{bmatrix}$$

When making a small change in the matrix A:

$$\begin{bmatrix} 1.001 & 2.000 \\ 3.000 & 5.999 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.000 \\ 11.999 \end{bmatrix} => \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -0.400 \\ 2.200 \end{bmatrix}$$

When making a small change in the vector B:

$$\begin{bmatrix} 1.000 & 2.000 \\ 3.000 & 5.999 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 11.999 \end{bmatrix} => \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

When dealing with experimental data, it is not reliable to have a completely different calibration because of a small change in the observations. Hence, it is mandatory to take into consideration such effects before achieving any computation.

In literature, various measures of the ill-conditioning of a matrix have been proposed [81], perhaps the most popular one [25] is $K(A) = ||A||_2||A^{-1}||_2$, where $||.||_2$ denotes the l_2-norm defined for a vector X as $||X||_2 = \sqrt{t_X X}$ and for a matrix A as $||A||_2 = \max_{||X||_2 \neq 0} \frac{||AX||_2}{||X||_2}$. The larger is $K(A)$, the more ill-conditioned is A.

The rationale behind defining the condition number $K(A)$ is to measure the sensitivity of the solution X relative to a perturbation of the matrix A or the vector B. More precisely:

- If $AX = B$ and $A(X + \delta X) = B + \delta B$ then $\frac{||\delta X||_2}{||X||_2} \leq K(A) \frac{||\delta B||_2}{||B||_2}$
- If $AX = B$ and $(A + \delta A)(X + \delta X) = B$ then $\frac{||\delta X||_2}{||X+\delta X||_2} \leq K(A) \frac{||\delta A||_2}{||A||_2}$

Proofs:

For any $A \in \mathbb{R}^{p,p}$, $X \in \mathbb{R}^p$ such $||X||_2 \neq 0$:
$\frac{||AX||_2}{||X||_2} \leq \max_{||Y||_2 \neq 0} \frac{||AY||_2}{||Y||_2} = ||A||_2$

$=> ||AX||_2 \leq ||A||_2||X||_2$ (1)

For any $A \in \mathbb{R}^{p,p}$, $B \in \mathbb{R}^{p,p}$:
$||AB||_2 = \max_{||X||_2 \neq 0} \frac{||ABX||_2}{||X||_2} = \max_{||BX||_2 \neq 0} \frac{||ABX||_2}{||BX||_2} \frac{||BX||_2}{||X||_2} \leq \max_{||Y||_2 \neq 0} \frac{||AY||_2}{||Y||_2} \max_{||X||_2 \neq 0} \frac{||BX||_2}{||X||_2} = ||A||_2||B||_2$

$=> ||AB||_2 \leq ||A||_2||B||_2$ (2)

Proof 1 :

Let A, B, X such that $AX = B$ (3) and $A(X + \delta X) = B + \delta B$ (4)

From (3) and (4) $\delta X = A^{-1}\delta B$ and using (1) $||\delta X||_2 = ||A^{-1}\delta B||_2 \leq ||A^{-1}||_2||\delta B||_2$ (5)
From (3) $||B||_2 = ||AX||$ and using (1) $||B||_2 \leq ||A||_2||X||_2$ (6)
From (5) and (6), $||\delta X||_2||B||_2 \leq ||A^{-1}||_2||\delta B||_2||A||_2||X||_2$

Thus $\frac{||\delta X||_2}{||X||_2} \leq K(A) \frac{||\delta B||_2}{||B||_2}$

Proof 2 :

Let A, B, X such that $AX = B$ (3) and $(A + \delta A)(X + \delta X) = B$ (7)

From (3) and (7), $\delta X = -A^{-1}\delta A(X + \delta X)$.
Using (1) and (2) follows $||\delta X||_2 \leq ||A^{-1}||_2||\delta A||_2||X + \delta X||_2$

Thus $\frac{||\delta X||_2}{||X+\delta X||_2} \leq K(A) \frac{||\delta A||_2}{||A||_2}$

Notice that $K(A)$ can be easily computed as the maximum singular value of A. For example, in the system above, $K(A) = 49,988$. The small perturbations can, thus, be amplified by almost 50,000, causing the previous observations.

Figure 3.9 represents the singular values of $t_X X$ used to compute the regression of the right graph of **Figure 3.8**. The graph shows a hard decreasing singular values. In particular, the condition number is higher than 80,000.

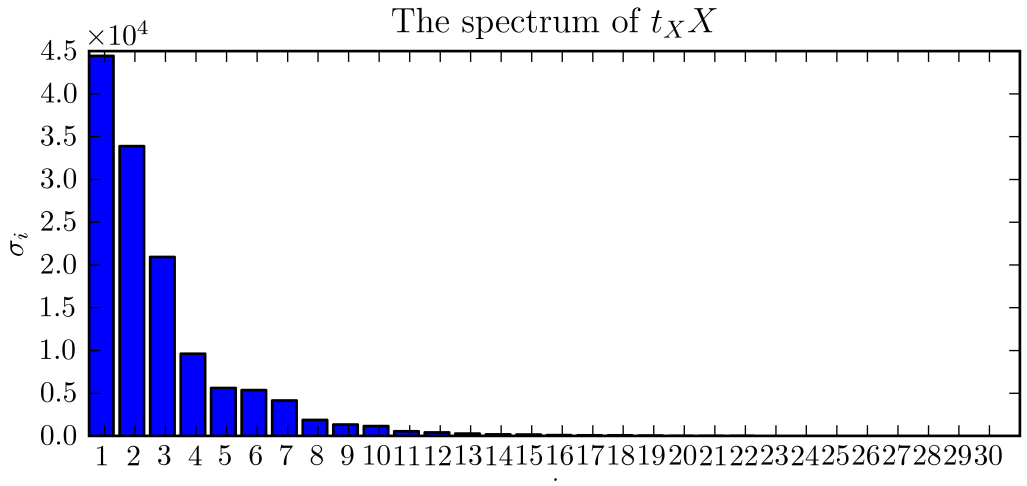


Figure 3.9: The quality of the OLS prediction: The graph shows that the matrix inverted when computing the OLS coefficient is ill-conditioned.

This finding explains the instability observed on the previous section. Moreover that the OLS estimator is statistically not satisfactory, the numerical problems due to the ill-conditioning of the matrix makes the result numerically unreliable.

One popular solution to enhance the stability of the estimation of the regression coefficients is the Ridge method. This method was introduced independently by A. Tikhonov, in the context of solving ill-posed problems [91] , around the middle of the 20th century, and by A.E. Hoerl in the context of addressing the linear regression problems by the sixteeth [50] . The Ridge regression consists of adding a regularization term to the original OLS problem:

$$\widehat{\beta_\Gamma} = argmin_\beta(||Y - X\beta||_2^2 + ||\Gamma\beta||_2^2)$$

The new term gives preference to a particular solution with desirable propreties. Γ is called the Tikhonov matrix and chosen usually as a multiple of the identity matrix; $\lambda_R I$, where $\lambda_R \geq 0$. The new estimator of the linear regression coefficients is called the Ridge estimator, denoted $\widehat{\beta}_R$, and defined as follows:

$$\widehat{\beta}_R = argmin_\beta(||Y - X\beta||_2^2 + \lambda_R||\beta||_2^2)$$

Similar to the OLS case, by straightforward calculation:

$$\widehat{\beta}_R = (t_X X + \lambda_R I)^{-1} t_X Y = Z\widehat{\beta} \qquad where \qquad Z = (I + \lambda_R (t_X X)^{-1})^{-1} = W^{-1}$$

and

$$\begin{aligned}
&E(\widehat{\beta}_R|X) = E(Z_R\widehat{\beta}|X) = Z\beta \\
&Var(\widehat{\beta}_R|X) = Var(Z\widehat{\beta}|X) = \sigma^2 Z(t_X X)^{-1} t_Z \\
&MSE(\widehat{\beta}_R) = E(t_{(Z\widehat{\beta}-\beta)}(Z\widehat{\beta} - \beta)|X) = E(t_\beta t_Z Z\beta|X) - 2t_\beta Z\beta + t_\beta\beta \\
&\qquad\qquad = Trace(t_Z Z\sigma^2 (t_X X)^{-1}) + t_\beta t_Z Z\beta - 2t_\beta Z\beta + t_\beta\beta
\end{aligned}$$

Notice that:

$$\begin{aligned}
&(t_X X)^{-1} = \frac{Z^{-1} - I}{\lambda_R} \\
&I - Z = (I - Z)WW^{-1} = (W - I)W^{-1} = \lambda_R (t_X X)^{-1} W^{-1} = \lambda_R (W t_X X)^{-1} = \lambda_R (t_X X + \lambda_R I)^{-1}
\end{aligned}$$

Thus:

$$\begin{aligned}
MSE(\widehat{\beta}_R) &= Trace(\frac{\sigma^2 Z}{\lambda_R Z}) - Trace(\frac{\sigma^2 Z}{\lambda_R Z^2}) + t_\beta (I - Z)^2 \beta \\
&= \sigma^2 \sum \frac{\lambda_i}{(\lambda_i + \lambda_R)^2} + \lambda_R^2 t_\beta (t_X X + \lambda_R I)^{-2} \beta
\end{aligned}$$

The first element of the MSE corresponds exactly to the trace of the covariance matrix of $\widehat{\beta}_R$, -i.e. the total variance of the parameters estimations. The second element is the squared distance from $\widehat{\beta}_R$ to β and corresponds to the square of the bias introduced when adding the ridge penalty. Notice that, when increasing the λ_R, the bias increases and the variance decreases. On the other hand, when decreasing the λ_R, the bias decreases and the variance increases converging to their OLS values. To enhance the stability of the linear regression, one should compute a λ_R, such that $MSE(\widehat{\beta}_R) \leq MSE(\widehat{\beta})$. As proved by Hoerl [51], this is always possible.

Theorem: There always exist $\lambda_R \geq 0$ such that $MSE(\widehat{\beta}_R) \leq MSE(\widehat{\beta})$.

From a statistical view, adding the Ridge penalty aims to reduce the MSE of the estimator, and is particularly necessary when the covariance matrix is ill-conditioned. From a numerical view, the new matrix to be inverted is $t_X X + \lambda_R I$ with as eigen values $(\lambda_i + \lambda_R)_i$. The conditional number is $K(t_X X + \lambda_R I) = \frac{\lambda_{max}+\lambda_R}{\lambda_{min}+\lambda_R} \leq \frac{\lambda_{max}}{\lambda_{min}} = K(t_X X)$. Hence, the ridge regularization enhances the conditioning of the problem and improves the numerical reliability of the result.

From the previous, it can be seen that increasing the λ_R leads to numerical stability and reduces the variance of the estimator, however it increases the bias of the estimator. One has to chose the λ_R as a tradoff between those 2 effects. Next, 2 estimators of λ_R are tested; the Hoerl-Kennard-Baldwin (HKB) estimator [49] and the Lawless-Wang (LW) estimator [58] .

In order to compare the stability of the Ridge and the OLS coefficients, **Figure 3.10** and **Figure 3.11** represent the same test of **Figure 3.8**, applied, respectively, to the Ridge HKB and the Ridge LW methods. In the 1-minute prediction case, the graphs show that the Ridge LW method gives the most coherent coefficients. In particular, the coefficient of the order book imbalance is always positive (as expected from a financial view) and the coefficients of similar indicators have the same signs.

Finally, **Figure 3.12** summarizes the profitabilities of the corresponding strategies of the 2 methods. **Tables 5.21, 5.22, 5.23, 5.24, 5.25** and **5.26** of **Appendix 5** detail the results per stock.

The coefficients of the Ridge HKB regression

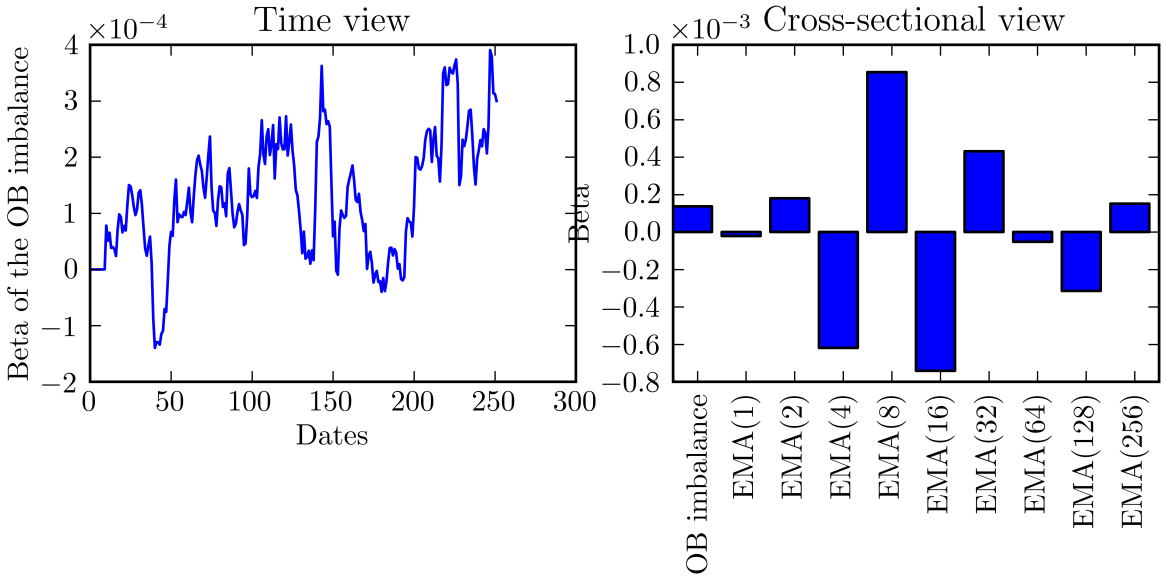


Figure 3.10: The quality of the Ridge HKB prediction: The graphs show that the results of the Ridge HKB method are not significantly different from those of the OLS method (Figure 8). In this case, the λ_R is close to 0 and the effect of the regularization is limited.

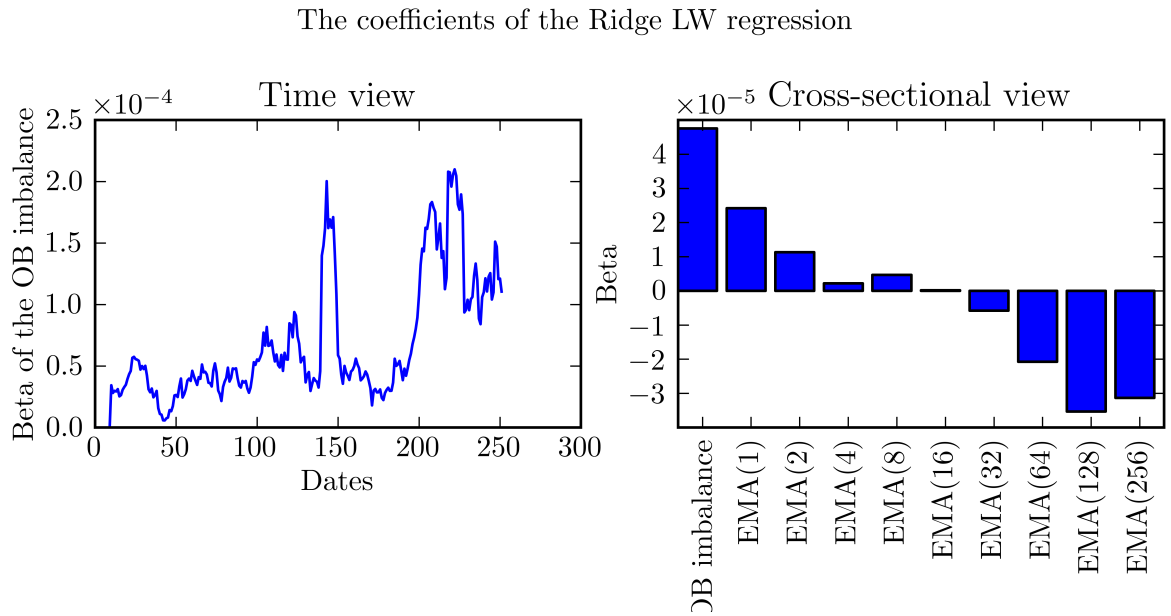


Figure 3.11: The quality of the Ridge LW prediction: The graph on the left shows the stability of the regression coefficient of the order book imbalance over the year 2013 for Deutsh Telecom. The coefficient is positive during all the period, in line with the financial view. The graph on the right shows, for a random day, a positive coefficients for the order book imbalance and its short term EMAs. The coefficients decreases with the time; -i.e. the state of the order book "long time ago" has a smaller effect than its current state. More over, for longer than a 10-second horizon, the coefficients become negative confirming the mean-reversion effect.

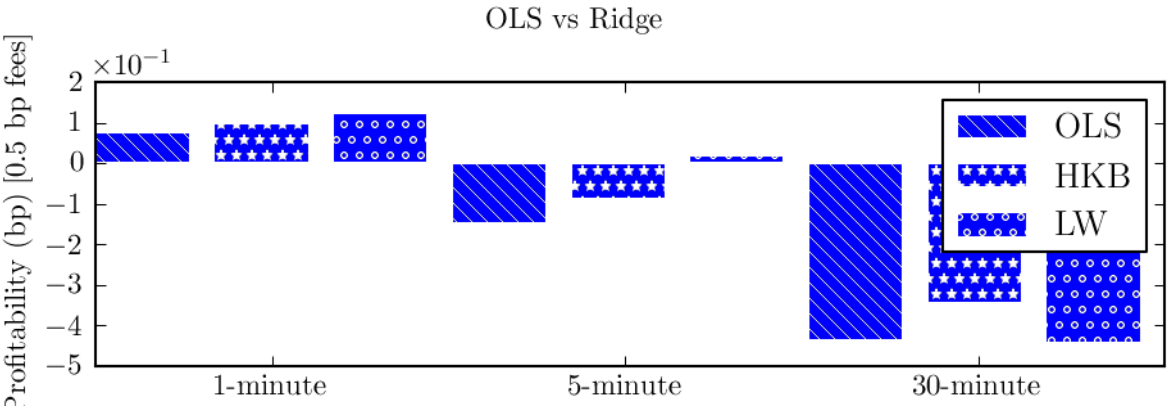


Figure 3.12: The quality of the Ridge prediction: For the 1-minute and the 5-minute horizons the LW method performs significantly better than the OLS method. However, for the 30-minute horizon, the HKB method gives the best results. Notice that for the 1-minute case, the LW method improves the performances by 58% compared to the OLS, confirming that stabilizing the regression coefficients (Figure 3.11 compared to Figure 3.8), leads to a better trading strategies.

From the previous results, it can be concluded that adding a regularization term to the regression enhances the predictions. The next section deals with an other method of regularization: the reduction of the indicators' space.

3.3.3 Least Absolute Shrinkage and Selection Operator (LASSO)

Due to the colinearity of the indicators, the eigen values spectrum of the covariance matrix might be concentrated on the largest values, leading to an ill-conditioned regression problem. The Ridge method, reduces this effect by shifting all the eigen values. This transformation leads to a more reliable results, but might introduce a bias in the estimation. In this paragraph, a simpler transformation of the original indicators' space, the LASSO regression, is presented.

The LASSO method [90] enhances the conditioning of the covariance matrix by reducing the number of the used indicators. Mathematically, the LASSO regression aims to produce a sparse regression coefficients -i.e. with some coefficients exactly equal to 0. This is possible thanks to the l_1-penalization. More precisely, the LASSO regression is to estimate the linear regression coefficient as:

$$\widehat{\beta}_L = argmin_\beta(||Y - X\beta||_2^2 + \lambda_L||\beta||_1)$$

Where $||.||_1$ denotes the l_1-norm, defined as the sum of the coordinates' absolute values. Writing $|\beta_i| = \beta_{i+} - \beta_{i-}$ and $\beta_i = \beta_{i+} + \beta_{i-}$, with $\beta_{i+} \geq 0$ and $\beta_{i-} \leq 0$, a classic quadratic problem, with a linear constraints, is obtained and can be solved by a classic solver. As far as known, there is no estimator for λ_L. In this study, the cross-validation [46] method is applied to select λ_L out of a set of parameters; $T10^{-k}$, where $k \in (2, 3, 4, 5, 6)$ and T denotes the number of the observations.

Figure 3.13 compares, graphically, the Ridge and the LASSO regularization, **Figure 3.14** addresses the instability problems observed in **Figure 3.8** and **Figure 3.15** summarizes the results of the strategies corresponding to the LASSO method. The detailed results per stock are given in **Tables 5.27, 5.28** and **5.29** of **Appendix 6**.

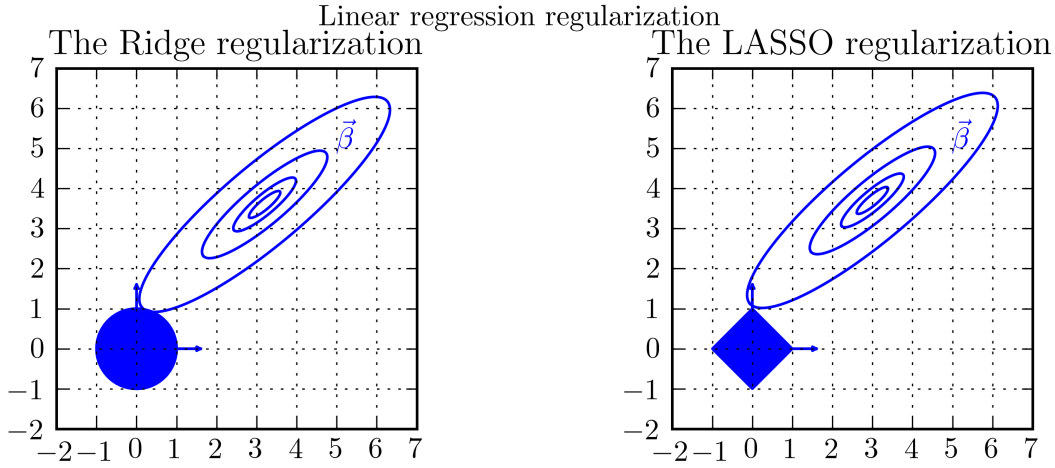


Figure 3.13: The quality of the LASSO prediction: The estimation graphs for the Ridge (on the left) and the LASSO regression (on the right). Notice that the l_1-norm leads to 0 coefficients on the less important axis.

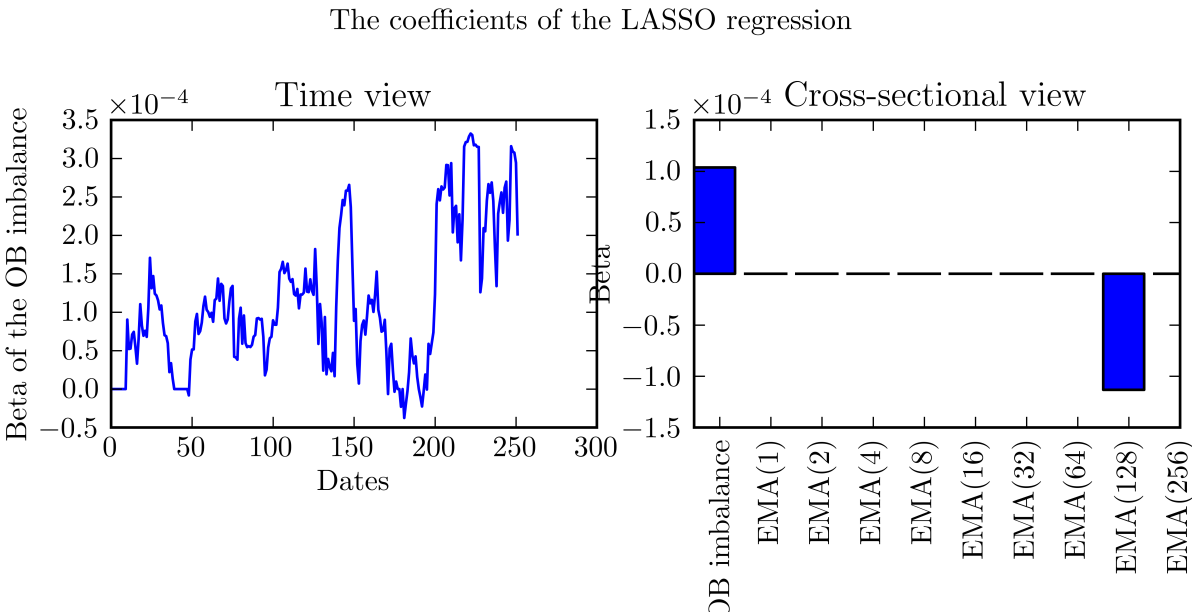


Figure 3.14: The quality of the LASSO prediction: The graphs show that the LASSO regression gives a regression coefficients in line with the financial view (similarly to **Figure 3.11**). Moreover, the coefficients are sparse and simple for the interpretation.

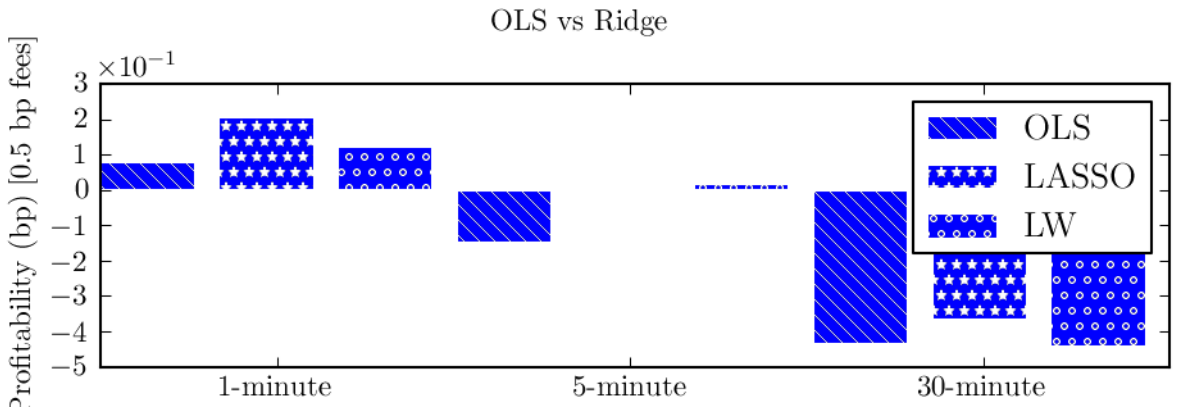


Figure 3.15: The quality of the LASSO prediction: Similar as the Ridge regression, the LASSO regression gives a better profitability than the OLS one. Notice that for the 1-minute case, the LASSO method improves the performances by 165% compared to the OLS. Eventhough the LASSO metho is using less regressors than the OLS method, (and thus less signal), the out of sample results are significantly better in the LASSO case. This result confirms the importance of the signal by noise ratio and highlights the importance of the regularization when adressing an ill-conditioned problem.

The next paragraph introduces the natural combination of the Ridge and the LASSO regression and presents this paper's conclusions concerning the market inefficiency.

3.3.4 ElASTIC NET (EN)

The EN regression aims to combine the regularization effect of the Ridge method and the selection effect of the LASSO one. The idea is to estimate the regression coefficients as:

$$\widehat{\beta}_{EN} = argmin_{\beta}(||Y - X\beta||_2^2 + \lambda_{EN_1}||\beta||_1 + \lambda_{EN_2}||\beta||_2^2)$$

The detail about the computation can be found in [101].

In this study, the estimation is computed in two steps. In the first step λ_{EN_1} and λ_{EN_2} are selected via the crossvalidation and the problem is solved same as the LASSO case. In the second step, the final coefficients are obtained by a Ridge regression ($\lambda_{EN_1} = 0$) over the selected indicators (indicators with a non-zero coefficient in the first step). The two step method avoids useless l_1-penalty effects on the selected coefficients.

Figure 3.16 shows that the coefficients obtained by the EN method are in line with the financial view and combine both regularization effects observed when using the Ridge and the LASSO methods.

The coefficients of the EN regression

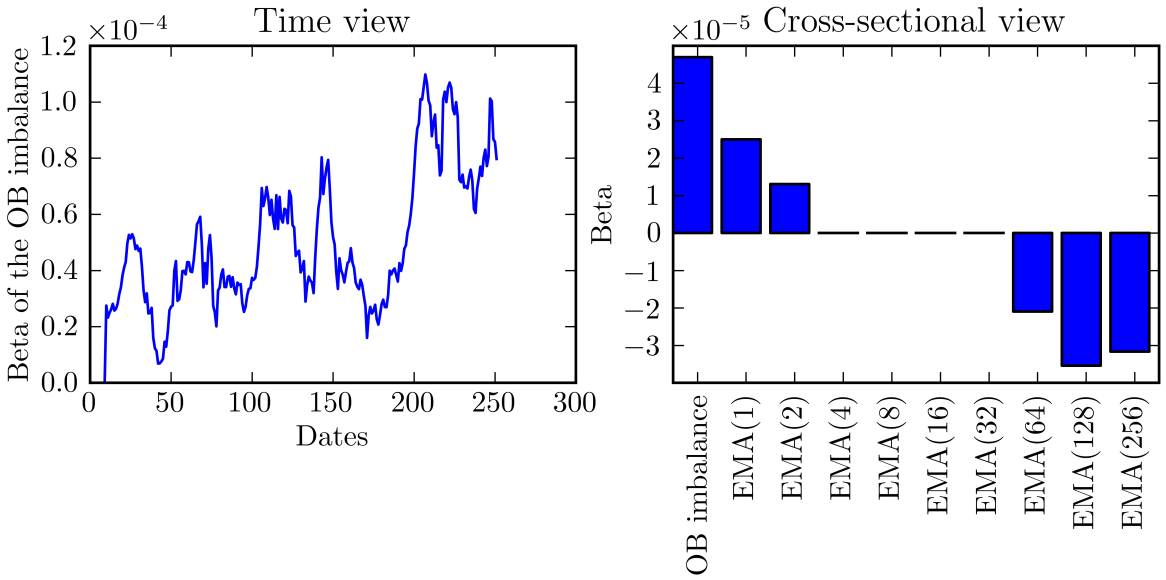


Figure 3.16: The quality of the EN prediction: The graphs show that the EN regression gives a regression coefficients in line with the financial view (similarly to **Figure 3.11** and **3.14**).

Finally, the strategy presented in **2.2** (trading only if $\widehat{Y} \geq |\theta|$) is applied to the different regression methods. **Figure 3.17** summarizes the obtained results. The results for the three horizons confirm that the predictions of all the regularized method (Ridge, LASSO, EN) are better than the OLS ones. As detailed in the previous paragraphs, this is always the case when the indicators are highly correlated. Moreover, the graphs show that the EN method gives the best results compared to the other regressions.

The 1-minute horizon results underline that, when an indicator has an obvious correlation with the target, using a simple method based exhaustively on this indicator, performs as least as well as more sophisticated methods including more indicators. Finally, the performance of the EN method for the 1-minute horizon suggest that the market is inefficient for such horizon. The conclusion is less obvious for the 5-minute horizon. On the other hand, the 30-minute horizon

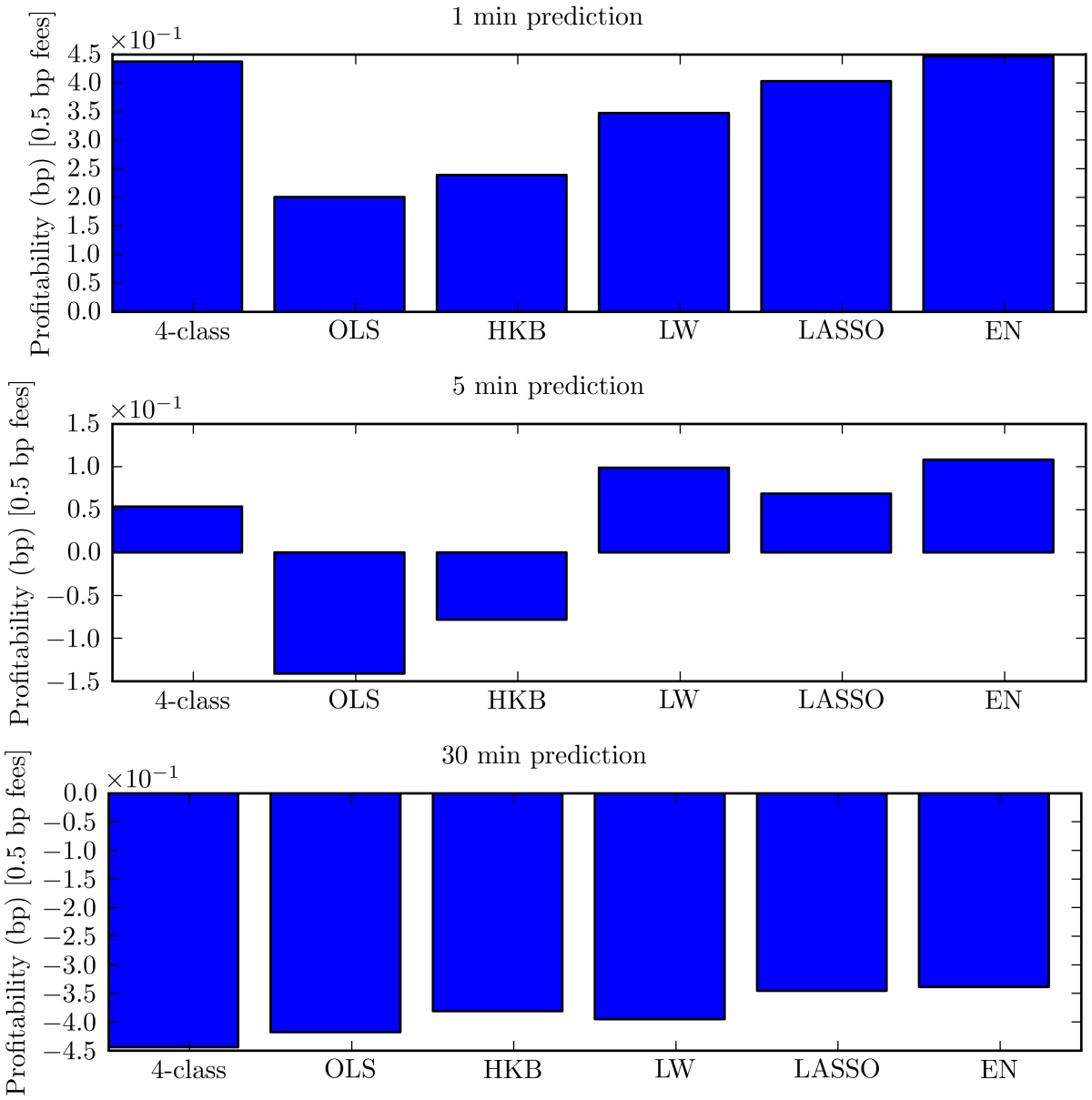


Figure 3.17: The quality of the EN prediction: The EN method gives the best results.

results show that, none of the tested methods could find any proof of the market inefficiency for such horizon. From the previous, it can be concluded that the market is inefficient in the short term, this inefficiency disappears progressively when the new information are widely diffused.

Conclusions

In this paper, a large empirical study was performed, over the stocks of the EURO STOXX 50 index, in order to test the returns predictability. The first part of the study shows that the future returns are not independent of the past dynamic and state of the order book. In particular, the order book imbalance indicator is informative and provides a reliable prediction of the returns. The second part of the study shows that combining different order book indicators using adequate regressions leads to a trading strategies with a good performances even when paying the trading costs. In particular, the obtained results show that the market is inefficient in the short term and that a few-minute period is necessary for the prices to incorporate the new information.

Chapter 4

Mathematical Modeling of the Order Book: New Approach of Predicting Single-Stock Returns

Note:
- This chapter is submitted to the journal "Market Microstructure and Liquidity".
- This chapter is presented in the conference "Market Microstructure: confronting many viewpoints", Paris, December 2014.

Abstract

This paper aims to forecast the price evolution based on modeling the order book. To design the model, the statistical properties of the order book events are empirically studied. A multivariate Poisson process is then fitted to the data and used to predict the stocks evolution. Although the Poisson model reproduces the clustering effect and the intraday seasonality correctly, the performances of the predictions are not satisfactory. To enhance the predictions, a multivariate Hawkes process is tested. This leads to a better modeling of the order book which takes into consideration the different interactions between the events. Moreover, the forecasting results are significantly enhanced, in line with the model enhancement.

Contents

Introduction **76**
4.1 Empirical study of the order book events **77**
4.1.1 Data and Framework 77
4.1.2 Introduction to the order book mechanism 77
4.1.3 Statistical properties of the order book events 79
4.1.4 Statistical dependencies between the different order book events 83
4.2 Modeling framework **88**
4.2.1 Introduction to point process 88
4.2.2 Introduction to Hawkes process 89
4.2.3 Simulation of Hawkes process 91
4.2.4 Goodness of fit 92
4.2.5 Maximum likelihood estimation of Hawkes process parameters 94
4.3 Mathematical modeling of the order book **97**
4.3.1 Poisson Model 97
4.3.2 Univariate Hawkes Model 103
4.3.3 Multivariate Hawkes Model 105

Introduction

Studying the order book dynamic has been attracting considerable attention since the rise of electronic markets. In particular, the availability and the complexity of the high frequency data make mathematical modeling necessary to understand the order book mechanism.

Many empirical studies detailed different stylized facts of the high frequency data relatively well. P. Gopikrishnan et al. [42] studied the statistical properties of the number of shares traded for a given stock in a fixed time, known as the order flow, and underlined a significant positive autocorrelation. A. Chakraborti et al. [23] computed different statistics of the order book and confirmed, in particular, that the Poisson hypothesis for the arrival of the orders is not empirically verified. Similarly, D. Challet and R. Stinchcombe [24] identified a clustering in both size and position of the orders. F. Pomponio [93] studied the particular case of trades through and observed an obvious auto-excitation of the arrival intensity. J. P. Bouchaud et al. [32] classified the order book events into twelve types and analyzed the statistical properties of the different types. In particular, this work highlights the role of the limit orders and the cancellations in price formation.

On the other hand, some theoretical studies proposed different market models able to reproduce the observed stylized facts. The most common modelization is based on Hawkes process. Bacry, Muzy et al. detailed the theoretical and technical issues of this model in different papers [5] [9] [6] [8] [7]. Their different studies show that the Hawkes process is appropriate for modeling the order book events and gives results in line with the empirical observations. Other theoretical properties of the order book models can be found in the PhD thesis of A. Jedidi [53] and the PhD thesis of B. Zheng [99] .

The goal of this paper is to fill the gap between the empirical and theoretical studies by providing a realistic application of the order book modeling. In most papers addressing order book modeling, a model is fitted to the data and well known stylized facts are reproduced. Even though this approach is necessary to validate the modeling process, it is not sufficient to use this to make conclusions about the modeling pertinence. Fitting the data using a large number of parameters might fit the noise rather than the signal. Thus, the applications of the obtained model are limited.

In this study, an order book model is said to be satisfactory if it can be used to build a profitable trading strategy. The rationale behind this criteria is that a good model should give a better view of the future than random guessing. If it is the case, forecasting based on the model should give, on average (over many dates and many stocks), positively biased performances.

This paper is organized as follows: in the first section, the statistical properties of the order book events are studied to design a mathematical model of their joint dynamic. In the second section, the potential processes that can be used to model the order book are studied. In particular, it is shown that it is possible to numerically fit such processes to the data with sufficient reliability. In the last section, the mathematical model is fitted to the data and used to design trading strategies.

4.1 Empirical study of the order book events

4.1.1 Data and Framework

This paper focuses on the DAX listed 30 stocks trading in Frankfurt Stock Exchange. Four months (Feb. to Jun. 2014) of tick-by-tick data, provided by the Chair of Quantitative Finance at Ecole Centrale Paris, are used in this study. The data, directly obtained from the exchange, are the trades and the order book states at any time a modification or a transaction occurs. In consequence, a cleaning process was done to derive the limit orders, the market orders, and the cancellations from the state of the order book and the list of the trades. Due to the large quantity of daily data, some problems such as mismatches of quantities and asynchronization were found. However, such anomalies represent less than 3% of the data and the results are thus reliable. Moreover, due to the large quantity of data, hundreds of computation cores were necessary to compute the different tests.

4.1.2 Introduction to the order book mechanism

The recent rise in electronic trading makes studying the order book mechanism necessary to understand price formation in the stock market. The historical quote-driven markets, where the market maker used to provide the liquidity for all the participants, are progressively becoming order-driven or hybrid markets, where the buy and the sell orders are matched continuously, between all the participants, with priorities subject to price and time. At each time, the list of all buy and sell limit orders with their prices and sizes constructs the current order book. An example is given in **Figure 4.1**.

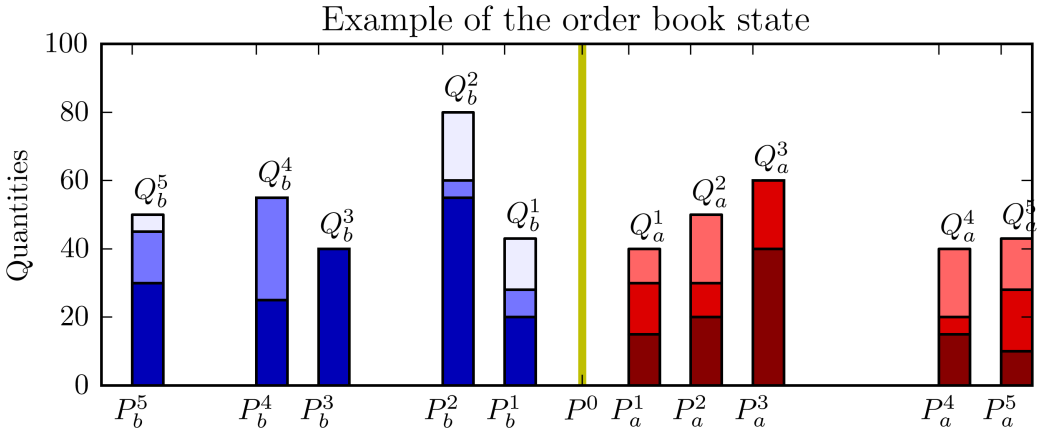


Figure 4.1: Illustrative order book. Left bars represent the buy orders with the prices and the quantities. This corresponds to the buyer side, also called the bid side or the offer side. Participants in the bid side are offering prices at which they are ready to buy some quantities of the stock. Different colors represent the arriving time of the orders, with darker bars representing older orders. Notice that the order with the best price has the priority of execution, and that at a same price level, the priority corresponds to the arriving time (first arrived, first executed). The right bars represent the sell side, commonly called the ask side, where participants willing to sell some quantities of the stock are posting their sell orders with the prices they are asking for to sell the stock. The line in the middle corresponds to the mid price level and is computed as the average between the best (highest) bid price and the best (lowest) ask price. A transaction occurs when a sell order and a buy order are at least partially matched.

In an order-driven market, participants can submit orders of three basic types: limit order, market order and cancellation:

Limit order: Order that specifies an upper/lower price limit (also called "quote") at which one (commonly called "Liquidity provider") is willing to buy/sell a certain number of shares. The advantage of the limit orders is that the transaction price is better than the instantaneous mid price. However, there is no certainty that the limit order will be executed. Notice that the priorities of limit orders are decided first by prices and then by arrival times for the same price. A limit order can be filled entirely, partly or even not executed.

Market order: Order that enforces an immediate execution of buy/sell of a number of shares at the best available opposite quote(s). The advantage is to have an immediate execution, however the price is worse than the mid price. Notice that a market order can be executed with different limit orders as counter parties. The price is not necessarily the best limit price, if the quantity demanded is so big that it has to surpass the first limit and hit the second or higher level limits.

Cancellation : Order that removes an existing limit order.

Besides the three types of orders listed above, there exists various order services provided by the electronic exchange system such as stop orders, good til' canceled etc.. However, in general, those orders can be regarded as combinations of basic orders with some predetermined conditions to execute different orders in different scenarios. For example, a stop loss order triggers a market order if the price moves out of the boundary (known as "stop price") in the undesirable direction.

Notice that other type of orders like iceberg orders are generally invisible and are, thus, difficult to be derived from the states of the order book. However, as long as the information available to all the participants is equal, the basic orders still carry enough information for the market microstructure studies.

Recall that the aim of the paper is to design a model that fits correctly the order book dynamic in order to predict the moves of the prices. The model has to be relevant from a financial view, otherwise the predictions will not be reliable. In order to design the model, it is necessary to identify the orders that change the price as well as their dynamics. It is also important to identify whether this dynamic is stock specific or is universal for all the stocks. Thus, the main basic properties of the stocks and the order types are presented in paragraph 1.3. Then, in paragraph 1.4, the dependencies between the different order types are analyzed from a statistical and a financial view.

4.1.3 Statistical properties of the order book events

In this study, any change that modifies the order book is called an "event". More precisely, an event can be a limit order, a market order, or a cancellation, and can affect the buy side or the sell side of the order book. Moreover, events will be tagged whether or not they cause a change on the mid price. **Table 4.1** summarizes the definitions and the notations of the different types of events studied in this paper:

Notation	Definition
M, L, C, O	market order, limit order, cancellation, any order.
M_{buy}, M_{sell}	buy/sell market order.
M^0_{buy}, M^0_{sell}	buy/sell market order that does not change the mid price: i.e. order quantity $<$ best ask/bid available quantity.
M^1_{buy}, M^1_{sell}	buy/sell market order that changes the mid price: ie. order quantity $\geq$ best ask/bid available quantity.
L_{buy}, L_{sell}	buy/sell limit order.
L^0_{buy}, L^0_{sell}	buy/sell limit order that does not change the mid price: i.e. order price $\leq$ / $\geq$ best bid/ask price.
L^1_{buy}, L^1_{sell}	buy/sell limit order that changes the mid price: ie. order price $>$ / $<$ best bid/ask price.
C_{buy}, C_{sell}	buy/sell cancellation.
C^0_{buy}, C^0_{sell}	buy/sell cancellation that does not change the mid price: i.e. partial cancellation at best bid/ask limit or cancellation at another limit.
C^1_{buy}, C^1_{sell}	buy/sell cancellation that changes the mid price: ie. total cancellation of best bid/ask limit order.
M^0, L^0, C^0, O^0	market order, limit order, cancellation, any order, that does not change the mid price.
M^1, L^1, C^1, O^1	market order, limit order, cancellation, any order, that changes the mid price.

Table 4.1: Event types definitions

During the trading hours, the price evolution is driven by the order arrivals. Thus, analyzing the statistical properties of the different order types would help explaining the price formation and might be informative for price predictions. The relation between the order book dynamic and the stock properties is addressed in the next paragraph.

Table 4.2 presents some basic statistics of the studied stocks. The price is the average mid price during the whole period. The volume is the daily average money exchanged on the stock. The tick size corresponds to the smallest possible change on the best bid/ask price and the spread corresponds to the average difference between the best bid and the best ask prices. The tick size is presented in Euro and in basis points ($Bp = 1\% of 1\%$). The spread is presented in Euro, in Bp and in number of tick size. The details per stock are given in the **Table 5.38** of the Appendix. Those properties are used as explanatory factors of the proportion of each type of events.

	Price (Eur)	Volume (10^6 Eur)	Tick size (Eur)	Tick size (Bp)	Spread (Eur)	Spread (Bp)	Spread (Tick)
Average	72	99	0.019	2.4	0.029	3.9	1.7
Min	9	31	0.001	1	0.003	1.9	1.1
Max	191	215	0.05	5	0.082	6.7	2.8

Table 4.2: Stocks basic properties summary

Table 4.3 summarizes the daily average numbers of the different events rounded to an integer. The statistics per stock are given in the **Table 5.39** of the Appendix.

	L_{buy}	L_{sell}	L	C_{buy}	C_{sell}	C	M_{buy}	M_{sell}	M	O
Average	24020	24219	48239	20328	20591	40919	3870	3876	7764	96904
Min	8804	8883	17687	7062	7410	14472	1575	1481	3056	36433
Max	44321	46123	90444	41296	41075	82371	7665	7321	14986	187801

Table 4.3: Event occurrences statistics summary

The statistics show that the orders are symmetric on the buy and on the sell side. The numbers of limit orders and cancellations are in the same order of magnitude and are both significantly higher than the number of market orders. The average daily number of orders is 96,904 orders, the minimum over the stock is 36,433 obtained on the stock MERCK KGAA (MRK), and the maximum is 187,801 obtained on the stock DEUTSCHE BANK (DBK).

Tables 4.2 and **Table 4.3** show that the intensity of the trading activity, represented by the volume and the total order number, varies significantly between the stocks. **Table 4.4** shows the correlation, computed over the stocks, between the trading intensity and the stocks properties defined in **Tables 4.2**.

	Price	Tick (Eu)	Tick (Bp)	Sp. (Eu)	Sp. (Bp)	Sp. (Tick)	O
O	0.13	0.02	-0.10	-0.11	**-0.51**	**-0.35**	1.00

Table 4.4: Correlation matrix

The correlations show that the spread computed in Bp or in Tick is the most relevant factor explaining the trading intensity. The higher the spread, the more costly it is to trade the stock (for the liquidity takers), which explains the observed significant negative correlation between the spread and the trading intensity.

On the other hand, a positive correlation is observed between the trading intensity and the price. This is explained by a positive correlation between the market capitalization and the price.

The Tick in Euro, the Tick in Bp and the spread in Euro are less relevant factors.

The previous results show that the total number of orders depends strongly on the stock properties. In the next paragraph, the relative proportion of each order type per stock is detailed. In line with the symmetry observed in **Table 4.2**, the proportions are computed with no distinction between buy and sell orders.

The aim of the different tests is to figure out whether the order book model can be calibrated over all the data or should be calibrated per stock. Moreover, the obtained results help in clarifying the agents behaviors when trading different kinds of stocks.

In **Table 4.5** the proportion of each type of events is computed. The details per stock are given in the **Table 5.40** of Appendix. The limit orders represent around 50% of the total orders, while the cancellations represent around 40% and the trades represent only around 10% of the total orders. Moreover, notice that O^1 events represent, on average, less than 10% of the total events. Those events are particularly interesting for the price formation, and are, thus, analyzed in detail in **Table 4.6**. The details per stock are given in the **Table 5.41** of the Appendix.

	L^0	L^1	L	C^0	C^1	C	M^0	M^1	M	O^0	O^1
Average	45.75	4.18	49.94	39.82	1.72	41.55	5.99	2.52	8.52	91.57	8.43
Min	43.09	1.18	47.64	32.58	0.42	35.34	4.05	0.76	4.82	85.07	2.37
Max	47.63	7.37	52.45	45.95	3.55	46.37	8.1	4.52	12.33	97.63	14.93

Table 4.5: Percentage of occurrences per event type

In average, around 50% of the events changing the price are limit orders. The other 50% is split up, to 20% of cancellations and 30% of trades. This result by itself is very important; in particular, it shows that studying only trade processes cannot explain the mechanism of the price formation.

	$L^1\|O^1$	$C^1\|O^1$	$M^1\|O^1$	$O^1\|O$
Average	49.64	19.77	30.59	8.43
Min	49.32	13.97	25.33	2.37
Max	50.03	24.63	36.21	14.93

Table 4.6: Repartition of events impacting the mid price

The statistics also show that the proportion of cancellations and market orders change significantly depending on the stock. This observation is analyzed in the next paragraph.

Table 4.7 represents the correlation matrix between the frequencies of different events and some stocks properties.

	Price	Volume	Tick (Eu)	Tick (Bp)	Sp. (Eu)	Sp. (Bp)	Sp. (Tick)
$O^1\|O$	-0.44	-0.36	-0.60	**-0.67**	-0.54	-0.43	**0.75**
$L^1\|O^1$	0.21	-0.35	0.35	0.33	0.38	0.42	-0.08
$C^1\|O^1$	-0.01	0.13	-0.30	**-0.74**	-0.30	**-0.80**	**0.55**
$M^1\|O^1$	-0.01	-0.10	0.28	**0.72**	0.28	**0.78**	**-0.56**

Table 4.7: Correlation matrix

The first row of the matrix shows an important negative correlation between the O^1 events proportion and the tick size of the stock. A smaller tick size leads to lower trading costs. Therefore, agents are more aggressive when trading small tick stocks. The same conclusion can be made from the high positive correlation with the spread (in tick) indicator. A small tick size, relative to the spread, results in more opportunities of scalping (making small gains on small price moves).

The other notable fact is that for stocks with high spread, the cancellation rate decreases. The priority is so important for execution of high spread stocks, so the agents cancel their orders less often to not lose their priorities. For such stocks, the quantities on the best limits are big, leading to less noisy price changes.

For high spread stocks, it is also observed that the price formation is driven by market orders. Those stocks are costly to trade, so trades changing the price are mainly initiated by informed agents.

The aim of this study is to predict the prices of the stocks. The new approach is to predict the type of the next order and to deduce the price evolution from this prediction.

In this paragraph the statistics of the different types of orders were presented. It can be concluded from those statistics that the dynamic of the order book depends on the stock properties. A good model should therefore be calibrated stock by stock, or at least by groups of similar stocks.

In the next paragraph, the time dependencies between the different types of orders are studied.This is necessary to decide which type of processes should be used to model the order book dynamic.

4.1.4 Statistical dependencies between the different order book events

In order to figure out the temporal dependencies between the occurrences of the different types of events, three tests are computed and are detailed in this paragraph.

4.1.4.1 Conditional probability of occurrence

Table 4.8 represents the historical probabilities of occurrence of an event of type j (in column) conditional to the fact that the last observed event is of type i (in row). The last row represents the unconditional probabilities of each type of events.

	L^0_{buy}	L^0_{sell}	C^0_{buy}	C^0_{sell}	M^0_{buy}	M^0_{sell}	L^1_{buy}	L^1_{sell}	C^1_{buy}	C^1_{sell}	M^1_{buy}	M^1_{sell}
$\|L^0_{buy}$	41.37	9.64	16.00	22.40	2.90	1.58	2.35	1.12	0.02	1.08	1.39	0.16
$\|L^0_{sell}$	9.61	41.79	21.95	16.12	1.61	2.96	1.02	2.29	1.05	0.02	0.15	1.44
$\|C^0_{buy}$	17.91	25.88	40.67	5.98	1.39	1.74	1.20	2.34	1.49	0.37	0.56	0.47
$\|C^0_{sell}$	25.18	17.98	6.04	41.30	1.79	1.42	2.08	1.27	0.37	1.49	0.51	0.60
$\|M^0_{buy}$	22.17	5.33	4.75	9.94	34.64	0.70	7.68	0.65	0.55	1.31	11.86	0.42
$\|M^0_{sell}$	5.60	21.14	10.61	5.01	0.72	34.32	0.53	7.19	1.48	1.10	0.42	11.88
$\|L^1_{buy}$	32.39	8.06	0.21	25.27	4.84	5.58	1.42	1.57	5.80	1.77	2.44	10.65
$\|L^1_{sell}$	7.65	29.94	26.04	0.22	5.63	5.62	1.39	1.36	1.42	5.39	12.37	2.96
$\|C^1_{buy}$	25.02	19.09	35.70	4.96	0.96	0.67	8.34	3.59	0.72	0.35	0.48	0.12
$\|C^1_{sell}$	21.48	23.28	5.42	34.70	0.76	1.16	3.20	7.88	0.63	0.75	0.18	0.57
$\|M^1_{buy}$	28.27	9.60	7.38	28.12	3.11	1.02	11.52	7.98	0.90	0.87	0.67	0.55
$\|M^1_{sell}$	11.83	23.05	33.36	7.24	1.04	3.13	6.79	9.34	1.05	1.81	0.66	0.70
$\|O$	22.82	22.93	19.80	20.03	2.99	3.00	2.07	2.12	0.85	0.88	1.27	1.26

Table 4.8: Conditional probabilities (in %) of occurrences per event type

To simplify the interpretation of the results, **Table 4.9** represents the conditional probabilities divided by the unconditional probabilities and rounded to the closest integer.

	L^0_{buy}	L^0_{sell}	C^0_{buy}	C^0_{sell}	M^0_{buy}	M^0_{sell}	L^1_{buy}	L^1_{sell}	C^1_{buy}	C^1_{sell}	M^1_{buy}	M^1_{sell}
$\|L^0_{buy}$	2	0	1	1	1	1	1	1	0	1	1	0
$\|L^0_{sell}$	0	2	1	1	1	1	0	1	1	0	0	1
$\|C^0_{buy}$	1	1	2	0	0	1	1	1	2	0	0	0
$\|C^0_{sell}$	1	1	0	2	1	0	1	1	0	2	0	0
$\|M^0_{buy}$	1	0	0	0	**12**	0	**4**	0	1	1	**9**	0
$\|M^0_{sell}$	0	1	1	0	0	**11**	0	**3**	2	1	0	**9**
$\|L^1_{buy}$	1	0	0	1	2	2	1	1	**7**	2	2	**8**
$\|L^1_{sell}$	0	1	1	0	2	2	1	1	2	**6**	**10**	2
$\|C^1_{buy}$	1	1	2	0	0	0	**4**	2	1	0	0	0
$\|C^1_{sell}$	1	1	0	2	0	0	2	**4**	1	1	0	0
$\|M^1_{buy}$	1	0	0	1	1	0	**6**	**4**	1	1	1	0
$\|M^1_{sell}$	1	1	2	0	0	1	**3**	**4**	1	2	1	1

Table 4.9: Conditional probability leverage

Results of **Table 4.9** are quite symmetric and no significant differences are observed between the buy and the sell side. Therefore, only interpretation of buy orders are detailed below:

$\boldsymbol{L^0_{buy}}$: reinforces the consensus that the stock is not moving down. This increases the probability of posting other L^0_{buy}.

$\boldsymbol{C^0_{buy}}$: decreases the available liquidity at the buy side. Other participants might feel less comfortable posting buy orders and the probability of C^0_{buy} and C^1_{buy} increases.

$\boldsymbol{M^0_{buy}}$: increases the probability of M^0_{buy}. This might be explained by order splitting and by the momentum effect (other participants following the move). The increase of the probability of M^1_{buy} and L^1_{buy} is also explained by the momentum effect.

$\boldsymbol{L^1_{buy}}$: improves the offered price to buy the stock. The first major effect observed is a big increase in the probability of M^1_{sell} -i.e. participants willing to take the newly offered liquidity and to sell taking back the price at its previous value. The second effect is a big increase in the probability of C^1_{buy} -i.e. the new liquidity is rapidly canceled. This might reflect a market manipulation where agents are posting fake orders. As far as is known, this effect has not been mentioned in other papers and should be studied in more detail in forward papers.

$\boldsymbol{C^1_{buy}}$: a total cancellation of the best buy limit increases the probability of L^1_{buy}; other participants re-offer the liquidity at the previous best buy price. It also increases the probability of L^1_{sell}, when a new consensus is concluded by the market participants at a lower price.

$\boldsymbol{M^1_{buy}}$: consumes all the offered liquidity at the best ask. This increases the probability of L^1_{sell} when some participants re-offer the liquidity at the same previous best ask price. It also increases the probability of L^1_{buy}, when a new consensus is concluded by the market participants at a higher price.

This study focuses on predicting the events that change the prices (in order to predict the stocks returns). From the dependencies observed in **Table 4.9**, it is reasonable to take into consideration in the model, in addition to the events $\{L^1_{buy}, L^1_{sell}, C^1_{buy}, C^1_{sell}, M^1_{buy}, M^1_{sell}\}$ the events $\{M^0_{buy}, M^0_{sell}\}$.

4.1.4.2 Conditional waiting time

In this paragraph, the waiting time to the next event is studied. **Table 4.10** represents the median of the waiting time (in second) to the event j (in column) since the last observed event i (in row) and **Table 4.11** represents the mean of this waiting time.

As seen in the previous paragraph, the buy and the sell case are symmetric. **Table 4.10** results interpretation is, thus, detailed for the buy events. Moreover, since the focus is in predicting the returns, only the case of O^1 events is detailed.

$\boldsymbol{L^1_{buy}}$: The median waiting time is significantly reduced after observing a M^0_{buy} or a M^1_{buy}. Participants post aggressive (-i.e. enhancing the best limit) limit orders more often when observing a market order in the same sense.

$\boldsymbol{C^1_{buy}}$: The median waiting time is reduced after observing an event of the same type.

$\boldsymbol{M^1_{buy}}$: The median waiting time is significantly reduced after observing a M^0_{buy}. This might be explained by order splitting or momentum effect.

	L^0_{buy}	L^0_{sell}	C^0_{buy}	C^0_{sell}	M^0_{buy}	M^0_{sell}	L^1_{buy}	L^1_{sell}	C^1_{buy}	C^1_{sell}	M^1_{buy}	M^1_{sell}
$\|L^0_{buy}$	0.019	0.884	0.564	0.304	13.96	17.21	10.97	13.85	43.64	37.03	18.94	25.43
$\|L^0_{sell}$	0.888	0.017	0.327	0.556	17.25	13.76	13.92	10.96	38.02	41.97	25.79	18.32
$\|C^0_{buy}$	0.398	0.130	0.015	0.987	17.02	16.15	13.06	11.55	37.23	40.06	24.41	22.23
$\|C^0_{sell}$	0.137	0.391	0.975	0.012	16.16	16.65	11.47	12.81	41.00	35.73	22.52	23.56
$\|M^0_{buy}$	0.002	0.045	0.168	0.006	0.01	8.78	0.82	6.03	31.22	21.06	0.10	15.48
$\|M^0_{sell}$	0.041	0.002	0.006	0.154	9.09	0.01	6.34	0.81	22.70	28.60	15.91	0.08
$\|L^1_{buy}$	0.005	0.084	0.258	0.009	7.48	7.45	4.29	6.58	13.59	22.27	10.39	7.93
$\|L^1_{sell}$	0.084	0.005	0.012	0.240	7.29	7.13	6.58	4.04	23.96	13.52	7.54	9.88
$\|C^1_{buy}$	0.019	0.019	0.004	0.439	14.09	16.62	2.09	6.48	11.46	28.40	18.81	20.83
$\|C^1_{sell}$	0.017	0.021	0.400	0.004	15.76	13.03	6.15	2.07	27.81	11.98	20.07	17.53
$\|M^1_{buy}$	0.003	0.033	0.158	0.003	5.47	9.35	1.21	1.99	27.82	21.73	7.24	14.96
$\|M^1_{sell}$	0.030	0.003	0.003	0.139	9.40	5.60	2.32	1.16	21.77	24.91	14.89	7.86

Table 4.10: Median conditional waiting Time

	L^0_{buy}	L^0_{sell}	C^0_{buy}	C^0_{sell}	M^0_{buy}	M^0_{sell}	L^1_{buy}	L^1_{sell}	C^1_{buy}	C^1_{sell}	M^1_{buy}	M^1_{sell}
$\|L^0_{buy}$	1.47	3.35	2.68	3.25	30.20	33.02	27.40	29.81	96.89	86.52	41.32	46.86
$\|L^0_{sell}$	3.37	1.46	3.31	2.70	33.21	29.75	30.30	27.07	91.59	91.45	47.68	40.29
$\|C^0_{buy}$	2.28	2.24	1.74	3.76	33.08	32.32	29.17	27.69	89.10	89.16	46.23	43.99
$\|C^0_{sell}$	2.24	2.27	3.73	1.72	32.52	32.55	27.91	28.57	94.05	84.18	44.79	44.96
$\|M^0_{buy}$	0.67	1.63	2.00	1.50	10.83	24.57	15.89	21.97	86.25	73.85	17.35	36.77
$\|M^0_{sell}$	1.63	0.67	1.52	1.98	25.15	10.82	22.71	15.65	79.56	79.61	37.68	17.13
$\|L^1_{buy}$	0.95	1.84	2.55	1.53	23.81	23.35	18.55	20.94	61.80	68.31	32.01	27.93
$\|L^1_{sell}$	1.82	0.96	1.58	2.50	23.20	23.41	21.50	18.13	75.24	59.76	27.80	31.44
$\|C^1_{buy}$	1.48	1.75	1.32	3.15	30.36	33.25	13.71	20.72	50.47	72.32	39.63	41.57
$\|C^1_{sell}$	1.54	1.43	2.91	1.23	32.53	29.31	20.64	13.79	76.86	49.83	41.55	38.24
$\|M^1_{buy}$	0.62	1.51	1.98	1.12	21.74	25.38	15.51	15.22	80.46	71.22	28.70	36.06
$\|M^1_{sell}$	1.45	0.59	1.09	1.91	25.45	21.65	15.92	14.81	73.89	72.67	36.10	29.11

Table 4.11: Mean conditional waiting Time

These observations confirm the results of the conditional probability test. Moreover, the fact that the means are higher than the medians underlines a clustering phenomenon. Thus, a satisfactory model to fit the arrival times should be able to reproduce the clustering and the dependencies.

4.1.4.3 Infinitesimal correlation matrix

Let $(N_t)_{t\in\mathbb{R}_+}$ be a M-dimension jumping process defined by $(N_1(t),\dots,N_M(t))$. For a duration h and a lag τ, the covariance $C_\tau^h(i,j)_{1\leq i,j\leq M}$ matrix of the process at the duration h and the lag τ can be defined by:

$$C_\tau^h(i,j) = \frac{1}{h}Cov(N_i(t+h+\tau) - N_i(t+\tau), N_j(t+h) - N_j(t))$$

E. Bacry et al. detailed [6] [7] the theoretical properties of this matrix. In particular, using the empirical estimation of this covariance, it is possible to compute a non-parametric kernel of a Hawkes process that fits the data. In this paragraph, the same concept is used to qualitatively study the time dependencies between the different types of events. In order to avoid side effects caused by the non-homogeneity of the frequencies per event type (for example, L_{buy}^0 is significantly more frequent than M_{buy}^1), results are computed using the correlation matrix Cr_τ^h defined by:

$$Cr_\tau^h(i,j) = Correlation(N_i(t+h+\tau) - N_i(t+\tau), N_j(t+h) - N_j(t))$$

Next, h is chosen as 0.1 second and $\tau \in \{0.1, 0.2, \dots 0.9\}$. The correlations are computed empirically, per day per stock, using the market data. The result is then averaged over all the stocks and the days. For each event i the function $Cr_{i,j}^h(\tau)$ describes the temporal decay of the impact function of j on i. For example, **Figure 4.2** details the impact, of the different order type occurrences, on the intensity of occurrence of an order of type M_{buy}^1.

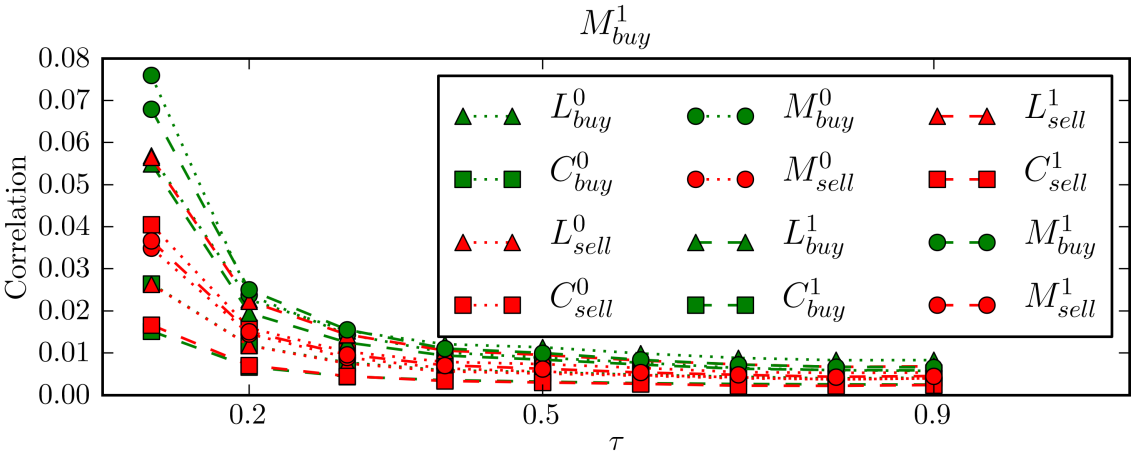


Figure 4.2: Impact functions on M_{buy}^1 arrival intensity: The graph confirms that the most relevant events to explain the instantaneous intensity of M_{buy}^1 are M_{buy}^0, M_{buy}^1 and L_{sell}^1. This is in line with the financial interpretation detailed in the previous two paragraphs.

Figure 4.3 represents the same results computed on the six events O^1. In order to plot only the most relevant information, an arbitrary threshold of 6% is chosen. The events where the highest correlation is lower than this threshold are ignored in the graphs.

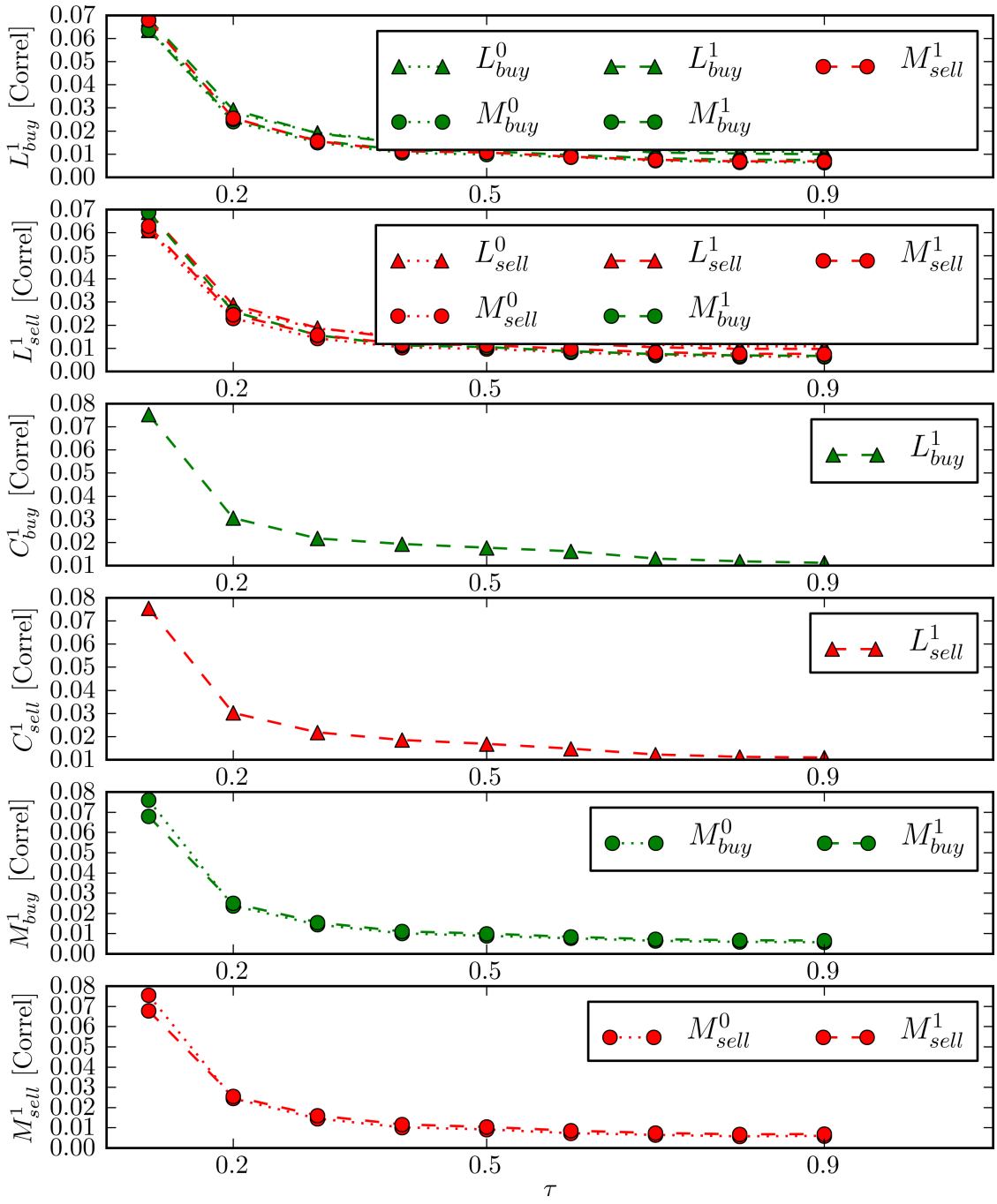


Figure 4.3: Impact functions: The intensity of L^1_{buy} event increases by the arrival of any L_{buy} or M_{buy} event. This means that liquidity providers follow on average the market consensus and provide more aggressive prices when the stock seems to move in the convenient sens. The intensity of C^1_{buy} is basically explained by L^1_{buy}. This corresponds to the suspicious case where a new limit is rapidly canceled. M^1_{buy} intensity increases by the arrival of any M_{buy}. This result is in line with the results of Tables **4.8,4.9,4.10,4.11** and with the majority of the studies addressing the order flow persisting issue. Finally, notice that the L_1 order intensities seem to increase "easier" than the C_1 and the M_1 intensities. This is in line with the results of Table 4.6.

4.2 Modeling framework

4.2.1 Introduction to point process

In the next paragraphs, some concepts useful to the rest of this paper are informally presented. More details can be found in [29] .

Point process (PP): A point process is an increasing sequence of random variables $(T_i)_{i\in\mathbb{N}}$. If $T_i < T_{i+1}$, $\forall i$, the process is called a simple point process. This may represent the times at which some events occur. If indeed, it is convenient to assume $T_0 = 0$.

Counting process: To a PP $(T_i)_{i\in\mathbb{N}}$ is associated a counting process $(N_t)_{t\in\mathbb{R}_+}$ defined by:

$$N_t = \sum_{1\leq i} \mathbb{1}_{\{T_i\leq t\}}$$

Intuitively, this process describes the number of occurrences of events and carries exactly the same information as the original process. $(N_t)_{t\in\mathbb{R}_+}$ is also called a point process.

Duration process: To a PP $(T_i)_{i\in\mathbb{N}}$ is associated a duration process $(\delta T_i)_{i\in\mathbb{N}^*}$ defined by:

$$\delta T_i = T_i - T_{i-1}$$

This process describes the waiting times between each two successive occurrences.

Intensity process: At each time t, the probability to have a jump of the PP $(N_t)_{t\in\mathbb{R}_+}$ is controlled by the intensity process $(\lambda_t)_{t\in\mathbb{R}_+}$. More precisely, the intensity process is defined by:

$$\lambda(t) = \lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}[N(t+h) - N(t)|\mathcal{F}_t]$$

$\mathcal{F}_t$ denotes the natural filtration of (N_t). Intuitively $\lambda(t)$ represents the infinitesimal rate at which events are expected to occur around a particular time t, conditional to the prior history of the point process prior to time t.

A multivariate point process is a sequence $((T_i, X_i))_{i\in\mathbb{N}^*}$, where (X_i) are some other random variables taking values in a discrete set $E = \{1, .., M\}$, and associated to the occurrences times (T_i). The X_i, called marks, contain further information about the events, and each (T_i, X_i) is said to be a marked point. Similar to the one dimensional case, a M-variate counting process $N(t) = (N_1(t), \ldots, N_M(t))$, and a M-variate intensity process $\lambda(t) = (\lambda_1(t), \ldots, \lambda_M(t))$ are associated to the marked process and are defined by:

$$N_m(t) = \sum_{1\leq i} \mathbb{1}_{\{T_i\leq t\}}\mathbb{1}_{\{X_i=m\}}$$

$$\lambda_m(t) = \lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}[N_m(t+h) - N_m(t)|\mathcal{F}_t]$$

In the rest of the paper T_k^n denotes the k^{th} arrival time of an event of type n.

Figure 4.4 shows an example of a point process with the related counting and duration process.

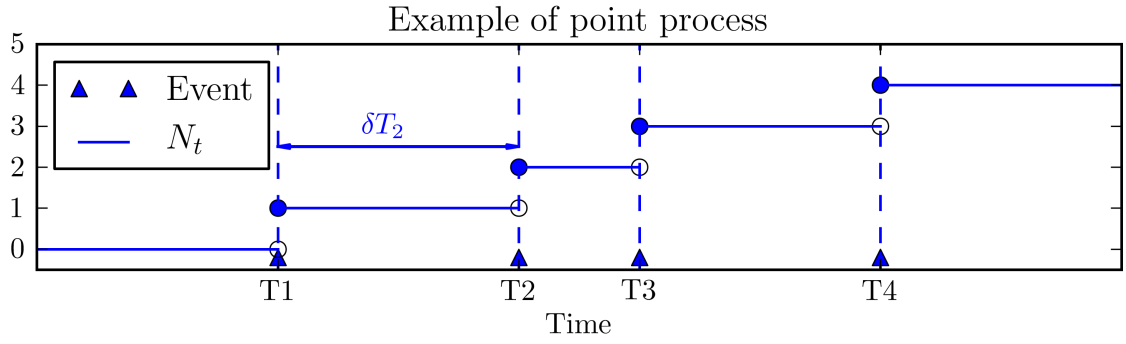


Figure 4.4: Illustrative point process.

4.2.2 Introduction to Hawkes process

4.2.2.1 Multivariate Hawkes process

A multivariate PP $((T_i, X_i))_{i\in\mathbb{N}^*}$, with a counting process $(N(t))_{t\in\mathbb{R}_+} = (N_1(t), \ldots, N_M(t))_{t\in\mathbb{R}_+}$ and an intensity process $(\lambda(t))_{t\in\mathbb{R}_+} = (\lambda_1(t), \ldots, \lambda_M(t))_{t\in\mathbb{R}_+}$, is called a multivariate Hawkes process if $\forall m \in \{1, \ldots, M\}$:

$$\lambda_m(t) = \mu_m + \sum_{n=1}^{M} \alpha_{mn} \int_0^t \omega_{mn}(t-s) dN_n(s)$$

Where μ_m and α_{mn} are positive real numbers and ω_{mn} are positive decreasing functions. The main property of such process is that the intensity is increased by the arrival of new events.

The real numbers μ_m are named base intensities and can be viewed as the background intensities. Whenever an event occurs, the intensities are increased, i.e. events arrive at a higher frequency. Such effects are controlled by ω_{mn} and α_{mn}.

The functions ω_{mn}, named decay functions, control how fast the excitation influence decreases with time. The real numbers, α_{mn}, named branching coefficients, control the amplitude of instantaneous increases in intensities.

For a multivariate Hawkes process, ω_{mm} and α_{mm} are the parameters of the self-excitation, while ω_{mn} and α_{mn} for $m \neq n$ are the parameters of the cross-excitation (the impact of the arrival of an event of type n on the probability of the arrival of an event of type m).

In this paper, the decay function is restricted to the classic case of an exponential kernel (with one exponential) $\omega_{mn} = e^{-\beta_{mn} t}$. This choice leads to an important simplification of the study, and gives a satisfactory fit of the market data. Notice that, in this case, the intensity of the Hawkes process is given by:

$$\lambda_m(t) = \mu_m + \sum_{n=1}^{M} \sum_{T_i<t} \alpha_{mn} e^{-\beta_{mn}(t-T_i)} \mathbb{1}_{\{X_i=n\}}$$

Next are detailed two important properties of this process.

4.2.2.2 Stationarity property

A point process is stationary if for all K, for all h and for all $t_1, \ldots, t_k$, the joint distribution of $\{N(t_1 + h), \ldots, N(t_k + h)\}$ does not depend on h.

In the univariate case ($M = 1$), Hawkes and Oakes [47] show that it exists a unique stationary point process, whose intensity is specified above in the exponential case, if

$$\frac{\alpha}{\beta} < 1$$

This result is generalized to the multivariate case by Bremaud and Massoulie [16]:

let

$$A_{ij} = \frac{\alpha_{ij}}{\beta_{ij}}, \qquad 1 \leq i, j \leq M$$

if

$$\rho(A) < 1$$

then it exists a unique stationary point process, whose intensity is specified above. $\rho(A)$ is the spectral radius of the matrix A (the largest absolute eigenvalue).

4.2.2.3 Markovian property

In general the Hawkes process is not Markovian; at a time t all the past path might be relevant to compute N_t and λ_t. The exponential case leads to an important simplification. Notice $I_{mn}(t)$ the impact of all the events, of type n prior to t, on the intensity λ_m. Thus for all m :

$$\lambda_m(t) = \mu_m + \sum_{1 \leq n \leq M} I_{mn}(t)$$

From straightforward calculation, for any t_1, t_2 such that $t_1 < t_2$:

$$I_{mn}(t_2) = I_{mn}(t_1) e^{-\beta_{mn}(t_2 - t_1)} + \int_{t_1}^{t_2} e^{-\beta_{mn}(t_2 - s)} dN_n(s)$$

All the impact of the events occurring before t_1 is summarized in $I_{mn}(t_1)$. The process $(N(t), I(t))$ is thus Markov. I is defined as the $(M * M)$-variate process $(I_{mn})_{1 \leq m,n \leq M}$. At the time t_1, all the path for $s < t_1$ is irrelevant.

The two previous properties are especially important for the numerical simulations and the empirical applications.

In the next paragraphs, A 2-variate Hawkes process is simulated and some numerical tests are computed. The aim is to verify whether a Hawkes model can be easily calibrated with a reliable estimator.

4.2.3 Simulation of Hawkes process

The classic method to simulate a multivariate Hawkes process is the Ogata's [76] algorithm based on the "thinning procedure" proposed in 1979 by Lewis & Shedler [62].

Ogata's proposition: Consider a multivariate PP $(N_t)_{t\in[0,T]} = (N_1(t),\dots,N_M(t))_{t\in[0,T]}$ with the intensity $(\lambda_t)_{t\in[0,T]} = (\lambda_1(t),\dots,\lambda_M(t))_{t\in[0,T]}$ and the natural filtration $\mathcal{F}_t$. Suppose one can find a one-dimensional $\mathcal{F}_t$-predictable process $\lambda^*(t)$ which is defined pathwise satisfying

$$\sum_{m=1}^{M} \lambda_m(t) \leq \lambda^*(t) \qquad 0 < t \leq T$$

Define

$$\lambda_0(t) = \lambda^*(t) - \sum_{m=1}^{M} \lambda_m(t)$$

Let $T_1^*, T_2^*, \dots, T_N^*$ be the points of the jumps of the process $N^*(t)$ associated to the intensity process $\lambda^*(t)$. For each of the points, attach a mark $X_i = m$ with probability $\lambda_m(T_i^*)/\lambda^*(T_i^*)$. Then the points with marks $X_i \neq 0$ provide a multivariate point process of intensity $(\lambda(t))$.

Figure 4.5 represents a 2-variate Hawkes process simulated with the previous procedure. Notice that $(\lambda_m(t))$ are decreasing in an inter-events time. Thus, one can choose $\lambda^*(t) = \sum_{m=1}^{M} \lambda_m(T_t)$, where T_t is the latest occurrence time prior to t. The following parameters are used:

$$\mu = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 0.2 & 0.1 \\ 0.5 & 0.1 \end{pmatrix} \qquad \beta = \begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{pmatrix}$$

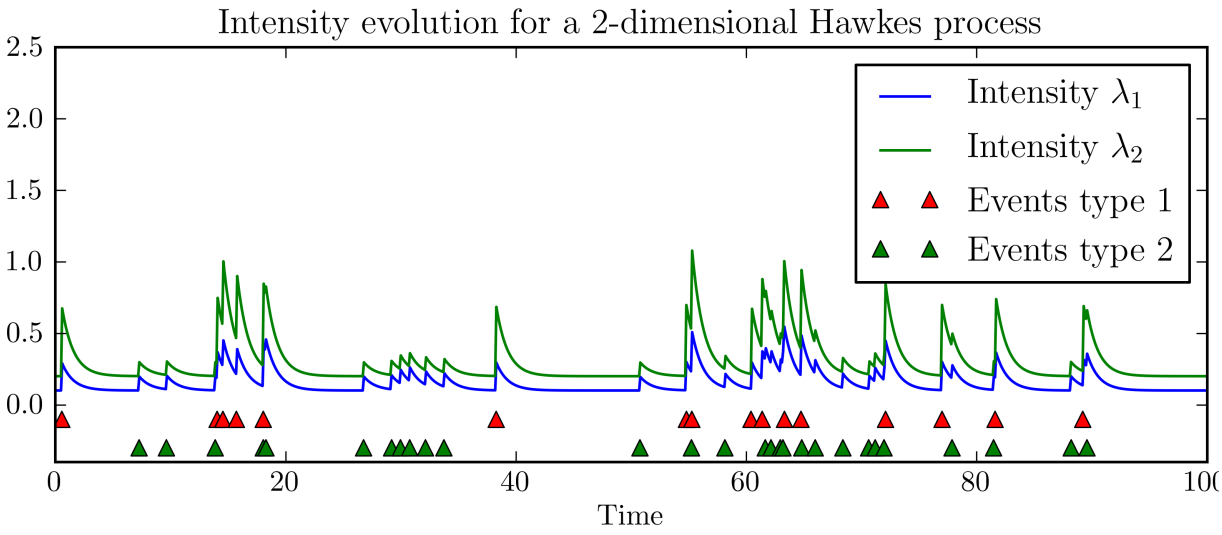


Figure 4.5: Simulated bi-variate Hawkes: Notice that the arrival of an event type 1 increases λ_1 by 0.2 and λ_2 by 0.5, where the arrival of an event of type 2 increases both intensities by 0.1. Notice also the clustering phenomenon observed when many events occur in a small interval of time and that, due to the decay, the intensities tend to their base values in non active periods.

4.2.4 Goodness of fit

To verify if some given data follow a known probability distribution, one can plot the empirical quartiles of the data vs the theoretical quartiles of the probability low. This method is called Q-Q plot and gives a graphical idea of the goodness of fit. In the Hawkes case, this test is possible thanks to the time-rescaling theorem [19] :

Time-Rescaling Theorem: Let $0 < T_1 < T_2 <, \ldots, < T_N < T$ be a realization from a point process with a conditional intensity function $\lambda(t)$ satisfying $0 < \lambda(t), \forall t$.
For $k = 1, \ldots, N$, define the transformation

$$\Lambda(T_k) = \int_0^{T_k} \lambda(t)dt$$

Assume $\Lambda(t) < \infty$ with probability one $\forall t$, the $(\Lambda(T_k))$ are a Poisson processes with unit rate.

Corollary: Let a multivariate Hawkes process $((T_i, X_i))_{i\in\mathbb{N}^*}$. With the usual notation of this paper, the random variables defined by

$$\tau_i^m = \int_{T_{i-1}^m}^{T_i^m} \lambda_m(s)ds$$

are i.i.d. exponential random variables with parameter 1.

Hawkes process with exponential decay kernel: In the particular case of an exponential decay, straightforward calculations gives:

$$\begin{aligned}\tau_i^m = \mu_m(T_i^m - T_{i-1}^m) &+ \sum_{n=1}^{M} \sum_{T_k^n < T_{i-1}^m} \frac{\alpha_{mn}}{\beta_{mn}} \left[e^{-\beta_{mn}(T_{i-1}^m - T_k^n)} - e^{-\beta_{mn}(T_i^m - T_k^n)}\right] \\ &+ \sum_{n=1}^{M} \sum_{T_{i-1}^m \le T_k^n < T_i^m} \frac{\alpha_{mn}}{\beta_{mn}} \left[1 - e^{-\beta_{mn}(T_i^m - T_k^n)}\right]\end{aligned}$$

For calculation simplification purpose, define a recursive element $A_{mn}(i)$, corresponding to the effect of all events of type n, occurring before the time T_i^m, on the intensity λ_m:

$$\begin{aligned}A_{mn}(i) &= \sum_{T_k^n < T_i^m} e^{-\beta_{mn}(T_i^m - T_k^n)} \\ &= e^{-\beta_{mn}(T_i^m - T_{i-1}^m)} A_{mn}(i-1) + \sum_{T_{i-1}^m \le T_k^n \le T_i^m} e^{-\beta_{mn}(T_i^m - T_k^n)}\end{aligned}$$

Take $A_{mn}(0) \equiv 0$, then for $\forall i \in \mathbb{N}^*$

$$\tau_i^m = \mu_m(T_i^m - T_{i-1}^m) + \sum_{n=1}^{M} \frac{\alpha_{mn}}{\beta_{mn}} \left[\left(1 - e^{-\beta_{mn}(T_i^m - T_{i-1}^m)}\right) A_{mn}(i-1) + \sum_{T_{i-1}^m \le T_k^n < T_i^m} \left(1 - e^{\beta_{mn}(T_i^m - T_k^n)}\right) \right]$$

Given a sample of marked points, and the parameters $(\mu_m, \alpha_{mn}, \beta_{mn})$ of a Hawkes process, one can compute (τ_i^m) and proceed to the Q-Q plot test to check whether the data can be satisfactorily fitted by the process.

Test on simulated data: In this test, a bi-variate Hawkes process is generated using the Ogata method and the parameters described in the previous paragraphs. The empirical quartiles of the observed $(\tau_i^m), m \in \{1, 2\}$ were computed as detailed above and were plotted, in **Figure 4.6**, vs the theoretical quartiles of an exponential law with parameter 1.

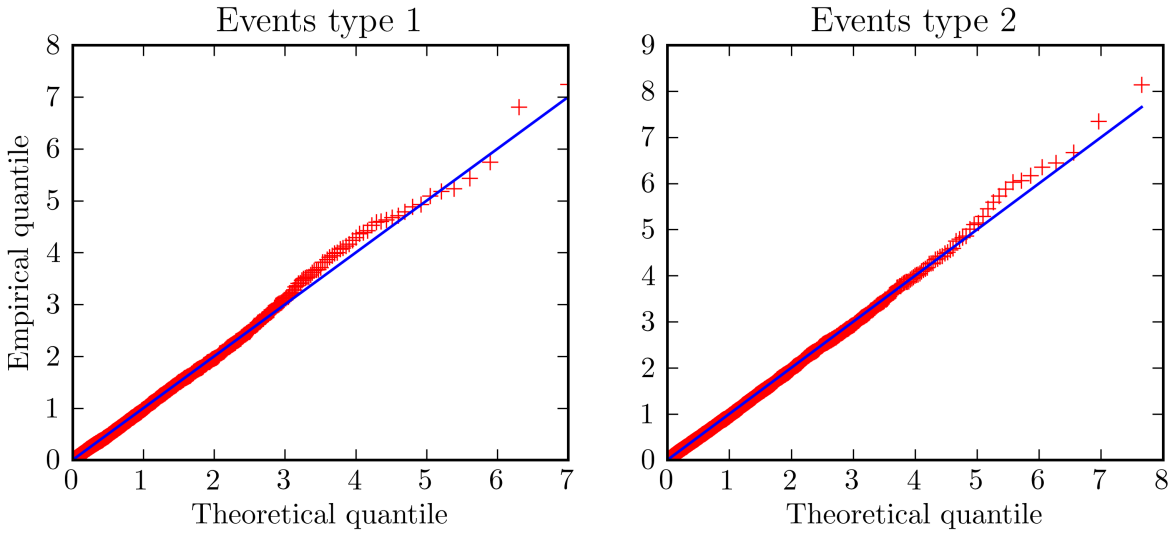


Figure 4.6: The Q-Q plots for a simulated 2-D Hawkes process: Notice that the $\tau_i^m, m \in \{1, 2\}$ are in line with an exponential distribution, the arrival process is, thus, in line with a bi-variate Hawkes process.

The simulated data follow, as expected, a bi-variate Hawkes process.

An interesting result would be to implicit the initial parameters from the data. If indeed, one can calibrate a model on any given sample that is supposed to follow a Hawkes process. This issue is addressed in the next paragraph.

4.2.5 Maximum likelihood estimation of Hawkes process parameters

Let $((T_i, X_i))_{i\in\mathbb{N}^*}$ be multivariate point process with a counting process $(N_1(t), \dots, N_M(t))$, and unknown intensities. The associated log-likelihood (see more details in [78] and [82]) of a given intensities $(\lambda_1(t), \dots, \lambda_M(t))$, and a sample of observation $\{T_i, X_i\}_{i\in\{1,\dots,D\}}$, is defined by the sum of the log-likelihood of each component:

$$\ln L(\lambda, \{T_i, X_i\}_{i\in\{1,\dots,D\}}) = \sum_{m=1}^{M} \left[\int_0^{T_D} \ln \lambda_m(s) dN_m(s) + \int_0^{T_D} (1 - \lambda_m(s)) ds \right]$$

For each component, the first term represents the probability of observing the process of intensity λ jumps accordingly to $\{T_i, X_i\}_{i\in\{1,\dots,D\}}$, where the second term represents the probability that no events occur at a different time other than $(T_i)_{i\in\{1,\dots,D\}}$. In the case of a Hawkes process with exponential decay, a straightforward calculation gives:

$$\begin{aligned}
\int_0^{T_D} \ln \lambda_m(s) dN_m(s) &= \sum_{T_i^m} \ln \left[\mu_m + \sum_{n=1}^{M} \sum_{T_k^n < T_i^m} \alpha_{mn} e^{-\beta_{mn}(T_i^m - T_k^n)} \right] \\
&= \sum_{T_i^m} \ln \left[\mu_m + \sum_{n=1}^{M} \alpha_{mn} A_{mn}(i) \right]
\end{aligned}$$

and

$$\begin{aligned}
\int_0^{T_D} \lambda_m(s) ds &= \mu_m T_D + \sum_{n=1}^{M} \int_0^{T_D} \sum_{T_k^n < s} \alpha_{mn} e^{-\beta_{mn}(s - T_k^n)} ds \\
&= \mu_m T_D + \sum_{n=1}^{M} \sum_{T_k^n} \int_{T_k^n}^{T_D} \alpha_{mn} e^{-\beta_{mn}(s - T_k^n)} ds \\
&= \mu_m T_D - \sum_{n=1}^{M} \sum_{T_k^n} \frac{\alpha_{mn}}{\beta_{mn}} (e^{-\beta_{mn}(T_D - T_k^n)} - 1)
\end{aligned}$$

thus

$$\begin{aligned}
\ln L_m(\lambda_m, \{T_i, X_i\}_{i\le D}) = T_D - \mu_m T_D + \sum_{n=1}^{M} \sum_{T_k^n} \frac{\alpha_{mn}}{\beta_{mn}} (e^{-\beta_{mn}(T_D - T_k^n)} - 1) \\
+ \sum_{T_i^m} \ln \left[\mu_m + \sum_{n=1}^{M} \alpha_{mn} A_{mn}(i) \right]
\end{aligned}$$

In practice, using the previous formula, one can estimate the unknown $(\mu_m, \alpha_{mn}, \beta_{mn})$ of a multivariate Hawkes process, given a sample of observation by maximizing the log-likelihood function.

Notice that, in this case, the problem is separable. Thus, the function is to be maximized separately on each component.

Finally, it is worth paying attention to the numerical problem concerning the maximization of the log-likelihood function. The target function is not concave so that some of the gradient descent algorithms may fail to find the optimal point, especially when no idea about the approximate value is given. This is typically the case for financial models. An efficient genetic algorithm is adopted in this study; the Differential Evolution [54] . Although the algorithm is not guaranteed to converge, experimental results with this algorithm are much more satisfactory than those with gradient descent algorithms, for example Nelder-Mead method recommended in many papers.

It is shown by Ogata in [75] that for a stationary univariate Hawkes process with an exponential decay kernel, the maximum likelihood estimator $\hat{\theta} = (\hat{\mu}, \hat{\alpha}, \hat{\beta})$ is:

Consistent, i.e. converges in probability to the true values $\theta^T = (\lambda, \alpha, \beta)$ as $T \to \infty$:

$$\forall \epsilon > 0, \quad \lim_{T \to \infty} P[|\hat{\theta} - \theta] > \epsilon] = 0$$

Asymptotically normal, i.e.

$$\sqrt{T}(\hat{\theta} - \theta) \to \mathcal{N}(0, I^{-1}(\theta))$$

where $(I^{-1}(\theta))_{i,j} = \mathbb{E}[\frac{1}{\lambda} \frac{\partial \lambda}{\partial \theta_i} \frac{\partial \lambda}{\partial \theta_j}]$

Asymptotically efficient, i.e. asymptotically reaches the lower bound of the variance.

As far as is known, theoretical properties of the maximum likelihood estimator for multivariate Hawkes process have not been concluded. In order to verify the asymptotic properties in the bivariate case, a "Montecarlo-like" method is used in this paragraph.

For each $T \in \{100, 250, 500, 1000, 2500, 5000, 10000, 25000\}$, 100 Hawkes process paths are simulated using the following parameters:

$$\mu = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 5.0 & 10.0 \\ 1.0 & 2.0 \end{pmatrix} \qquad \beta = \begin{pmatrix} 20.0 & 15.0 \\ 3.0 & 10.0 \end{pmatrix}$$

For each T, are estimated the parameters of the 100 generated processes with MLE. For each parameter, the average and the standard deviation (σ) are computed over the 100 estimations. **Figure 4.7** represents the 95% confidence intervals and **Figure 4.8** represents the log-log plot of the estimation standard deviation and the total time length T. The results are the same for both events type 1 and type 2. Thus, only results corresponding to events type 1 are plotted.

The convergence speed is calculated from a regression of $\ln(\sigma) \sim \ln(T)$. The values for $\ln(T) < 6$ in the log-log figures, which correspond to the time lengths smaller than 500, are ignored in regression. They are outliers when the time length is not significant enough. This does not influence the conclusion about the experimental asymptotic convergence of speed $T^{-0.5}$.

The results of this section show that by using a sufficient number (say thousands) of observations that are supposed to follow a Hawkes model, it is possible to "correctly" estimate the model parameters with the maximum likelihood method.

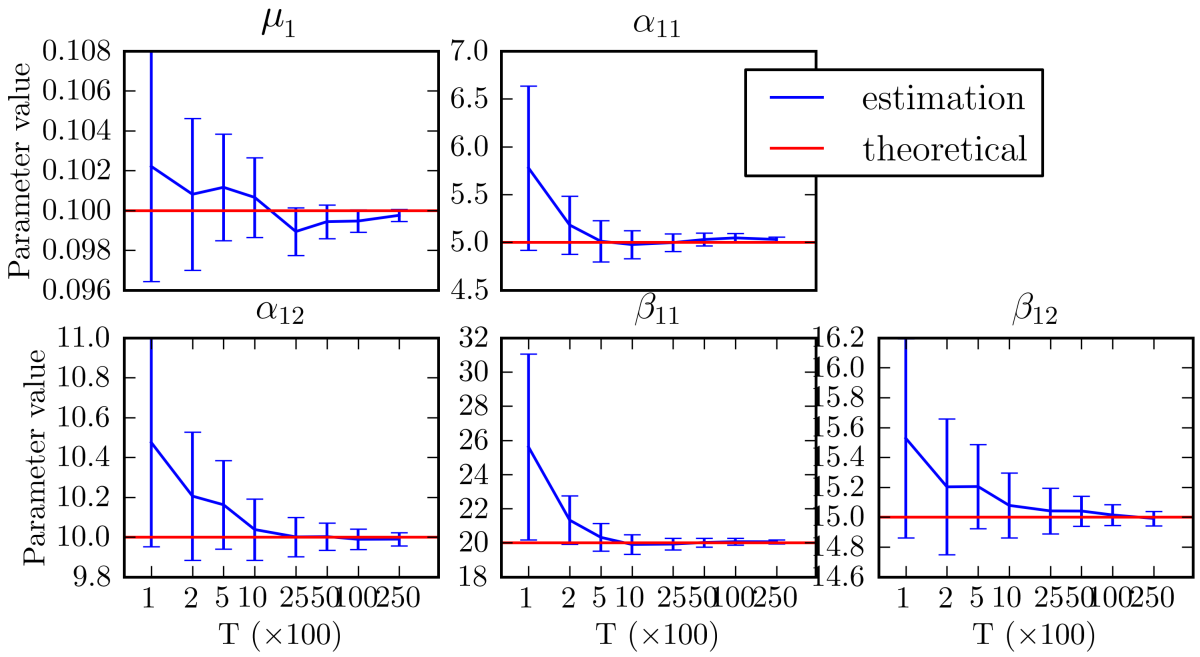


Figure 4.7: Confidence interval (95%) of parameters estimations

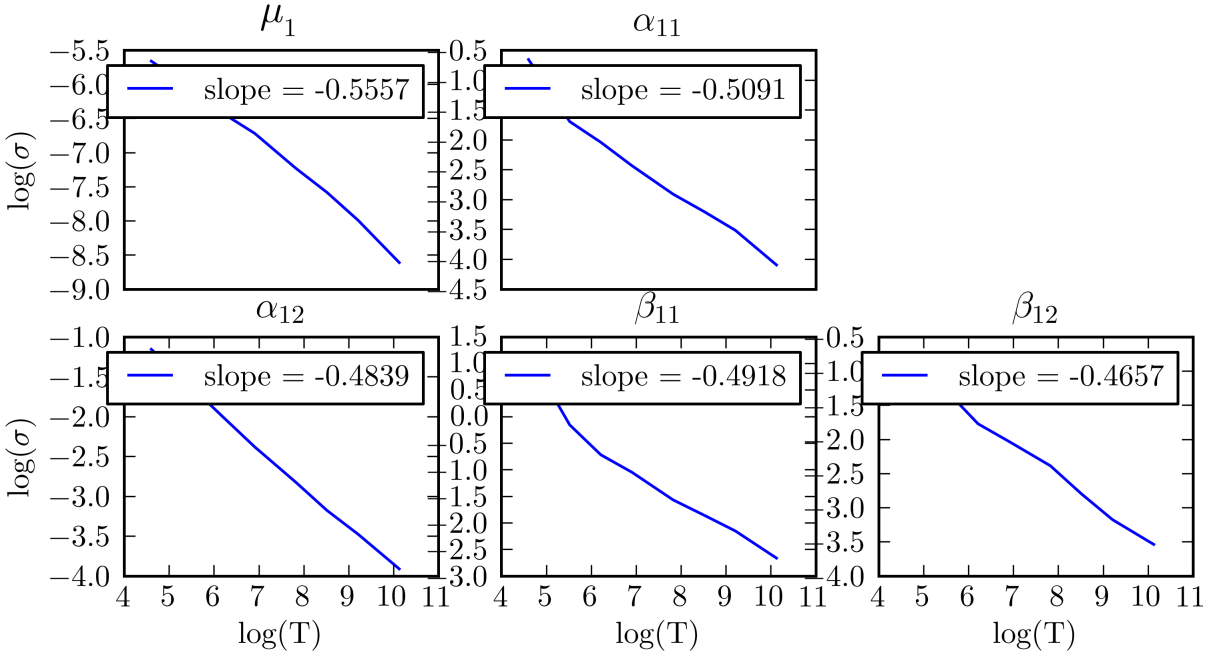


Figure 4.8: Log-log plot of the estimation error vs T

4.3 Mathematical modeling of the order book

In this study the order book events are classified into 12 types: $\{L^0_{buy}, L^0_{sell}, C^0_{buy}, C^0_{sell}, M^0_{buy}, M^0_{sell}, L^1_{buy}, L^1_{sell}, C^1_{buy}, C^1_{sell}, M^1_{buy}, M^1_{sell}\}$. Recall that the events with an upper index 1 have an immediate impact on the price. In particular, it is clear that the events $E_{up} = \{L^1_{buy}, C^1_{sell}, M^1_{buy}\}$ move the price up, where the events $E_{down} = \{L^1_{sell}, C^1_{buy}, M^1_{sell}\}$ move the price down. At any time t, a good prediction of whether the next event is in E_{up} or E_{down}, would give a good prediction about the next price move.

It is convenient to model the events arrivals by a 12-variate process $(N_1(t), \ldots, N_{12}(t))$ associated with an intensity process $(\lambda_1(t), \ldots, \lambda_{12}(t))$. At any time t, one can compute for example $\lambda_{up}(t)$ and $\lambda_{down}(t)$ as:

$$\lambda_{up}(t) = \lambda_7(t) + \lambda_{10}(t) + \lambda_{11}(t)$$

$$\lambda_{down}(t) = \lambda_8(t) + \lambda_9(t) + \lambda_{12}(t)$$

By comparing those 2 intensities, it is possible to predict the next stock return at any time t. The quality of the prediction depends strongly on the quality of the intensity model. Three models are presented and tested in this section.

4.3.1 Poisson Model

For a Poisson model, the intensities are constants -i.e. $\lambda_i(t) = \lambda_i$. In order to have a simple benchmark model, the idea is to calibrate a moving Poisson process. More precisely, for a trading day containing N events (200,000 for example), a sliding window of n events (1000 for example), containing the events $\{T_{i+1}, \ldots, T_{i+n}\}$ is used to calibrate a Poisson process and to compute $\widehat{\lambda}(T_{i+n}) = (\widehat{\lambda}_1^{\{T_{i+1},\ldots,T_{i+n}\}}, \ldots, \widehat{\lambda}_{12}^{\{T_{i+1},\ldots,T_{i+n}\}})$. Notice that $\widehat{\lambda}_j^{\{T_{i+1},\ldots,T_{i+n}\}}$ is the classic intensity estimator and defined by:

$$\widehat{\lambda}_j^{\{T_{i+1},\ldots,T_{i+n}\}} = \frac{N_j(T_{i+n}) - N_j(T_{i+1})}{T_{i+n} - T_{i+1}}$$

Finally, $\widehat{\lambda}(T_{i+n})$ is used to predict the return of the stock between the times T_{i+n} and T_{i+n+1} and a trading strategy buying/selling 100,000 euro of the stock depending on the predicted return is tested.

Figure 4.9 represents (for the example of DEUTSCHE TEL the 20^{th} Feb 2014) the intensities of the events that change the price. The graphs show a more important trading activity in the afternoon and a clustered intensities.

The Poisson model gives an interesting estimation of the intensities and is able to reproduce many known empirical facts. However two weakness can be reported; the first is that the number of events in the learning sliding window has to be fixed arbitrarily, the second is that the cross-excitation effect are not modeled using this approach.

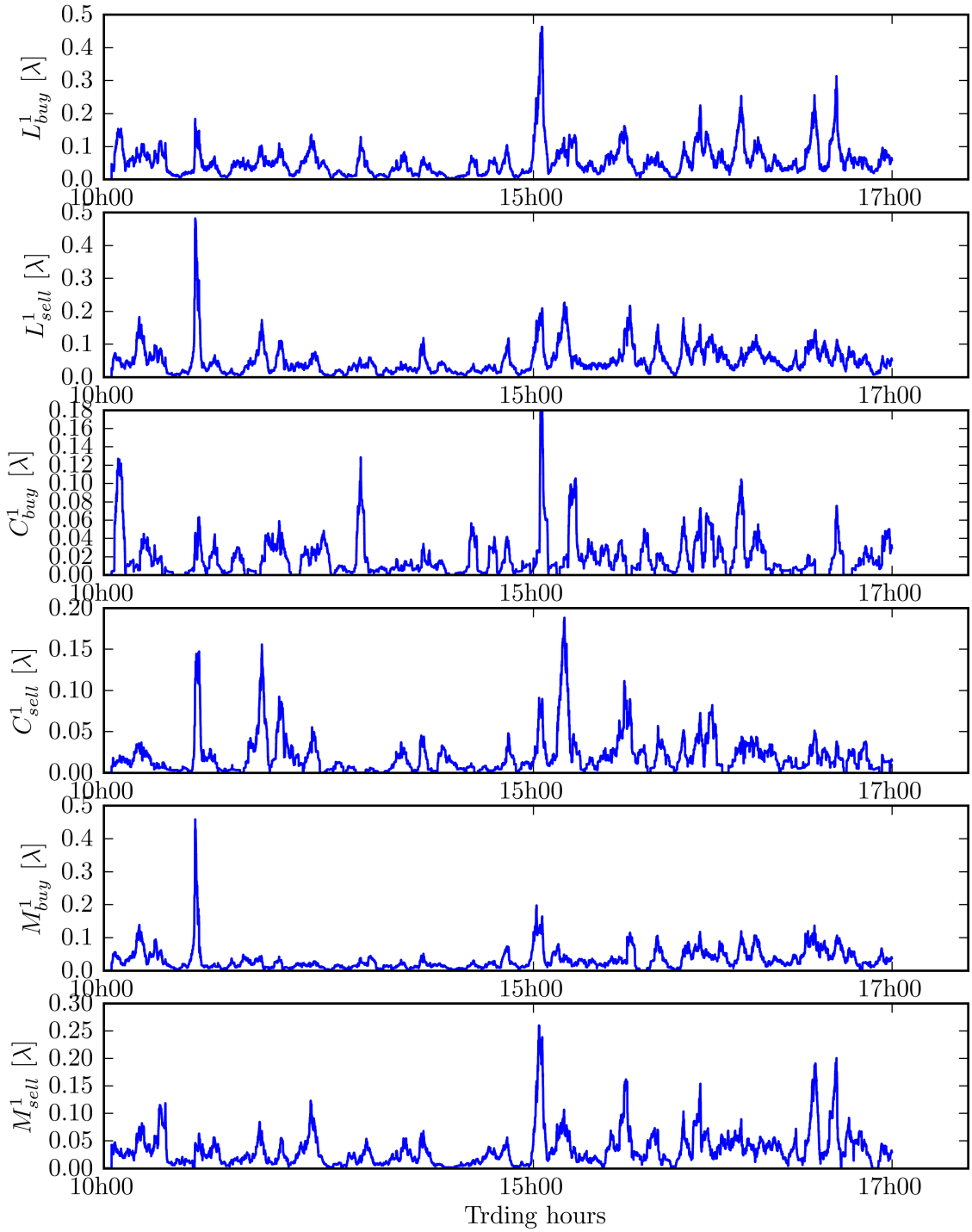


Figure 4.9: Poisson Model intensities: The model shows that the market activity seems to be more important in the afternoon. Moreover, the model is able to reproduce the clustering effect observed when comparing Table 4.10 and Table 4.11.

Next, the different results of this test are detailed.

To simplify the notations λ_j denotes the vector of $\widehat{\lambda}_j(T_i)$ for all the occurrences times during a given day. More precisely, the i^{th} value of this vector is computed using (possibly) all the available information until the time T_i. For a given stock with price S_t, R denotes the vector of the stock returns $R(T_i) = \ln(\frac{S_{T_i}}{S_{T_{i-1}}})$. R^+ denotes the vector of the shifted returns (those to be predicted). More precisely, $R^+(T_i) = \ln(\frac{S_{T_{i+1}}}{S_{T_i}})$.

Figure 4.10 presents the correlations between the returns R and the different intensities $(\lambda_j)_{1\leq j\leq 12}$. Notice that all the correlations have their intuitive expected sign. For example, as detailed previously, L_{buy}, C_{sell}, M_{buy} are supposed to move the price of the stock up, in line with this, their intensities are positively correlated with the stock return. Moreover, notice that, the absolute values of the correlations of $(L^1/C^1/M^1)$ are respectively higher than $(L^0/C^0/M^0)$. Those results confirm the intuition that using the intensity of the events that change the price might be useful for the return prediction. Finally, notice that the chosen indicators in this study $(\lambda_{up}, \lambda_{down})$ have the best in sample (synchronous) correlation with the stock returns.

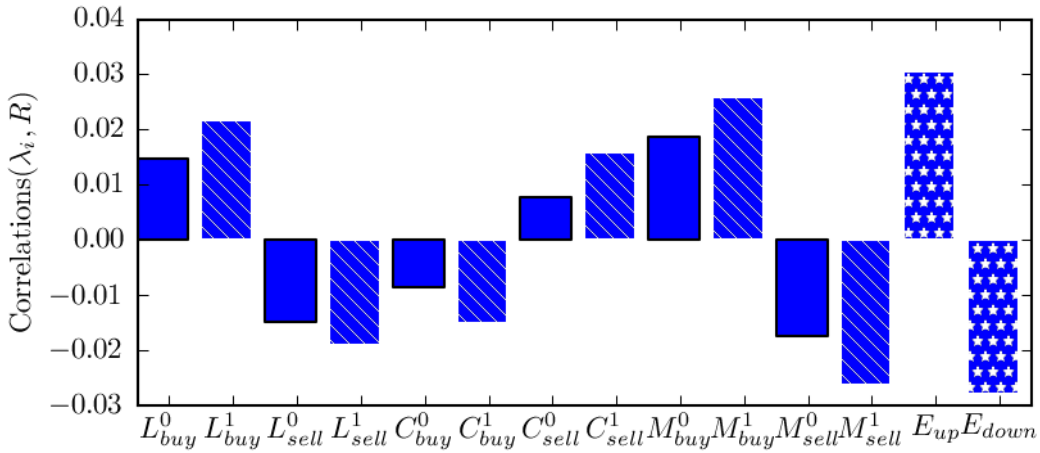


Figure 4.10: Correl.

Conceptually, the intensities λ_j are supposed to describe the future order arrivals and thus should be a good indicator to predict R^+. In the particular case of the Poisson process, the intensities are calibrated using the current observations. This explains the fact that the synchronous correlation are satisfactory. If, in real life, the order arrivals do not follow the model, the predictions do not have any reason to be better than random guessing.

In the rest of the paper, an "in sample" test denotes a test where an investment decision can be taken at the time (t) using data observed until a time $(t + dt)$. This test is not realistic and the corresponding strategy cannot be run in real life. However, it gives an idea about the intrinsic quality of an indicator, independently of the fact that this indicator can be predicted or correctly modeled. On the other hand, an "out of sample" test denotes a realistic test, where an investment decision, taken at a time (t), can be only based on the available data at any time $(s \leq t)$. More precisely, using the previous notations, if an investment based on λ makes the return R (respectively R^+), the test is called "in sample" (respectively "out of sample").

Using a Poisson model calibrated with three different sizes of the sliding learning window (10, 100 and 10000 events), λ_{up} and λ_{down} are computed from the historical data. The obtained intensities are used to run a strategy that buys (respectively sells) 100,000 euros of the stock if $\lambda_{up} > \lambda_{down}$ (respectively $\lambda_{up} < \lambda_{down}$). For both the "in sample" and the "out of sample" tests, the following performances measures are computed:

Accuracy (Acc): The proportion of the winning trades.
Gain (PnL): The average daily gain of the strategy (in Euro).
Profitability (Bps): The gain per traded notional: $\frac{\text{PnL}}{\text{Traded Notional}}$ (in basis point).
Holding period (Hp): The average holding period: the time (in seconds) a position is held.

Figure 4.11 and **4.12** gives two examples of a typical trading day using the strategy detailed above.

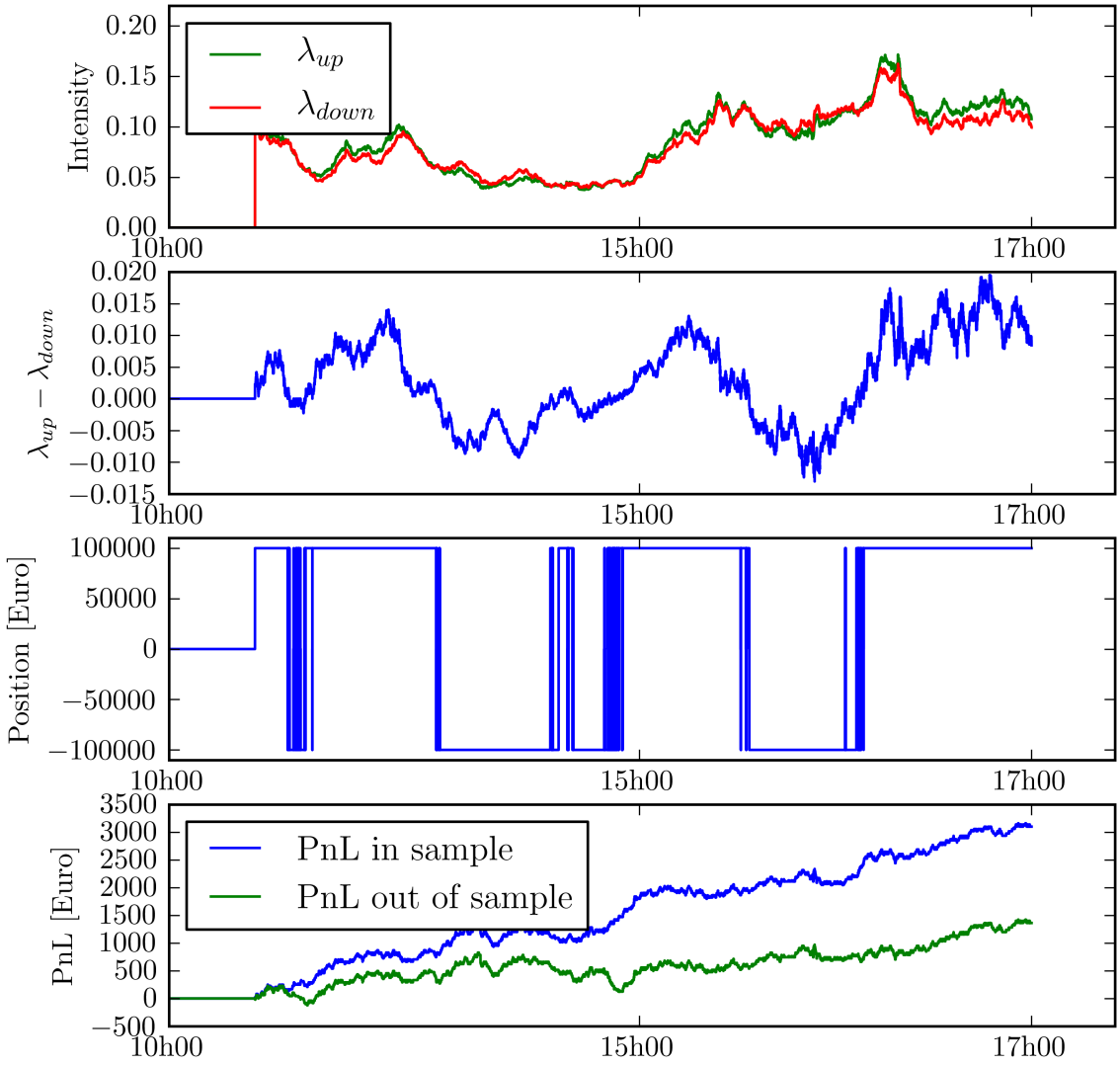


Figure 4.11: Illustration of the strategy based on a Poisson calibrated with a sliding window of 10000 events. Example of DEUTSCHE TEL the 20^{th} Feb 2014.

The first graph in **Figure 4.11** shows that the average intensities are around 0.07. This is equivalent to an arrival rate of 1.4 events (0.7 of type E_{up} and 0.7 of type E_{down}) arriving each 10 seconds. Notice that from **Table 4.3** and **Table 4.5** this arrival rate for DEUTSCHE TEL is 3495 events per 7 hours trading ,-ie 1.38 events /10 seconds, in line with the graph. The first graph also shows that the intensities increase in the afternoon (in line with classic results).

The second and the third graph of **Figure 4.11**, represent the trading signal $\lambda_{up} - \lambda_{down}$ and the investment position. The intensities change slowly due to the large sliding learning window. The investment position can, thus, represent large interval without any change. For those intervals, the in sample and the out of sample profitabilities are almost the same. In fact, having dt delay when taking the position is not important when dt is negligible compared to the holding period. However for the interval where the trading signal is around 0, the trading frequency becomes significantly high and the arrival of a single event can change the sign of the trading signal. This explains the important difference between the in sample and out of sample PnL.

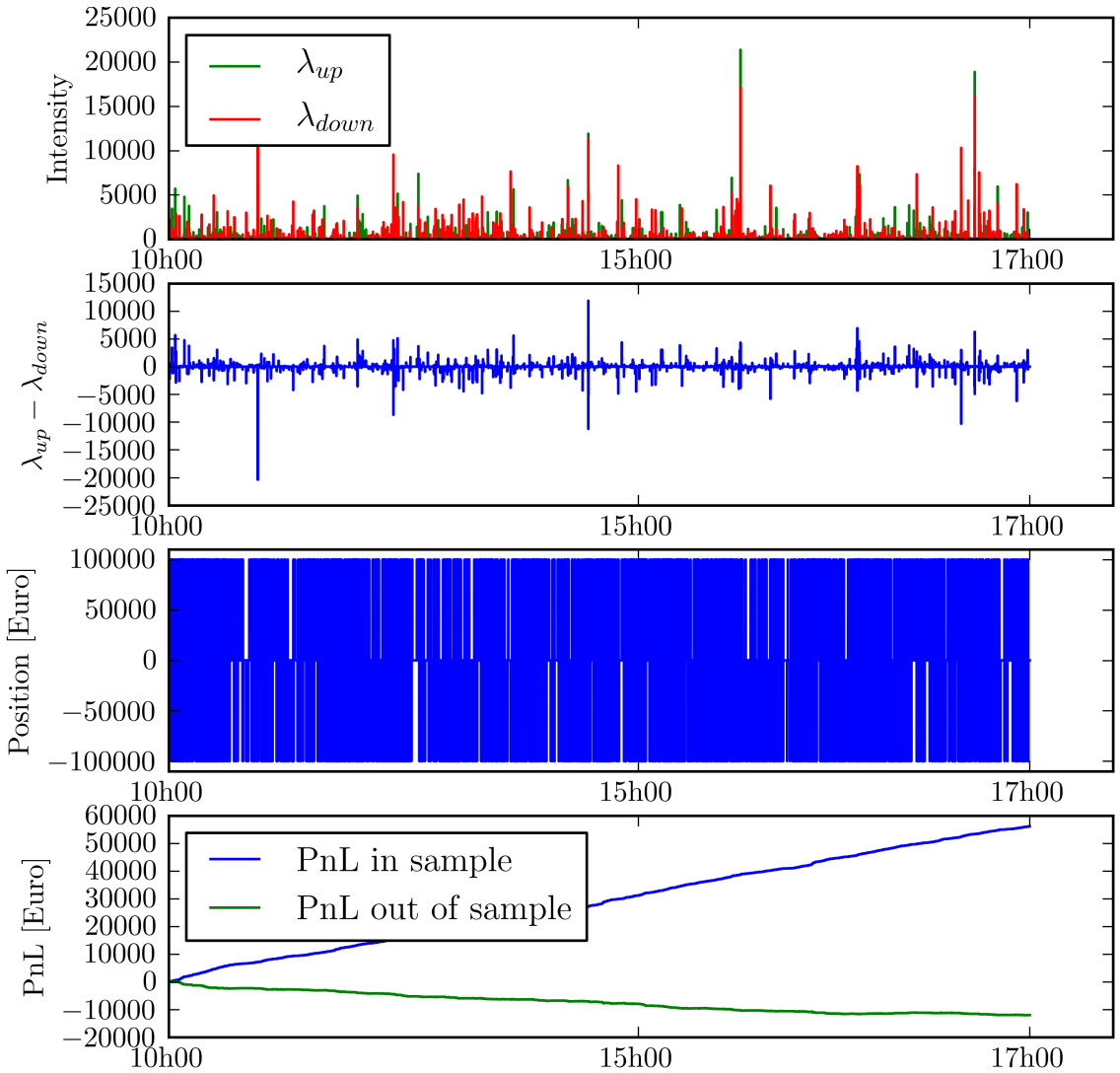


Figure 4.12: Illustration of the strategy based on a Poisson (sliding window of 10 events).

Figure 4.12 illustrates the case of a short learning window. In this case, the estimated intensity is not stable. The performance in sample is significantly increased. However, the poor performance out of sample shows that the data were overfitted and the result is not reliable.

Table 4.12, 4.13 and **4.14** summarize the results obtained with the strategies based on Poisson model using respectively 10, 100, and 10000 events for the learning windows. Details per stock are given in the Tables **Table 5.31, 5.32** and **5.33** of the Appendix.

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.92	0.48	50991	-1344	0.61	-0.03	3.99
Min	0.88	0.35	22441	-16063	0.33	-0.2	1.36
Max	0.96	0.64	89802	11209	1.03	0.12	9.76

Table 4.12: In sample and out of sample results for the strategy with 10 events learning window.

Table 4.12 shows a very good performance in sample (92% of good decisions) but a poor performance out of sample. The intensities used are not informative about the future and the model is not satisfactory. Moreover, the holding period is small (4 seconds) and the strategy might be highly dependent on the trading setup.

Table 4.13 summarizes the results of the strategy using 100 events learning window.

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.67	0.51	22369	678	0.53	0.02	8.14
Min	0.62	0.41	9429	-10629	0.30	-0.21	2.86
Max	0.78	0.56	36220	7489	0.81	0.22	16.46

Table 4.13: In sample and out of sample results for the strategy with 100 events window.

Increasing the learning window width from 10 to 100 events reduces the performance in sample (less over fit) and enhances the performance out of sample. However the profitability of the strategy is almost equal to zero (0.02 Bp).
Table 4.14 summarizes the results of the strategy using a 10000-event learning window. The results show that when taking a very large learning window, the out of sample result is close to random guessing. The in sample tests show that estimating the returns using the events arrival

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.51	0.50	1988	-187	0.22	0.00	65.91
Min	0.51	0.49	575	-2363	0.09	-0.08	6.92
Max	0.53	0.50	4415	408	0.41	0.10	312.12

Table 4.14: In sample and out of sample results for the strategy with a 10000-event window.

intensities gives satisfactory results. However, the out of sample tests show that the Poisson model is not sufficient to fit the event dynamics. In particular, notice that, in this model, an arrival of an event during a learning window has the same effect whether it occurs at the beginning or at the end of the period. In the next paragraph, Hawkes model is used to address this issue.

4.3.2 Univariate Hawkes Model

The same idea presented in the previous paragraph is applied to a multivariate Hawkes with all the cross-excitation set to zero. This process is equivalent to 12 univariate processes. More precisely, with the usual notations, each $(\lambda_j)_{1\leq j\leq 12}$ is supposed to follow the equation:

$$\lambda_j(t) = \mu_j + \sum_{T_i<t} \alpha_j e^{-\beta_j(t-T_i)} \mathbb{1}_{\{X_i=j\}}$$

Each trading day, the parameters $(\mu_j, \alpha_j, \beta_j)$ are computed applying the MLE to the previous day's data. **Figure 4.13** represents an example of Q-Q-plot corresponding to this fit.

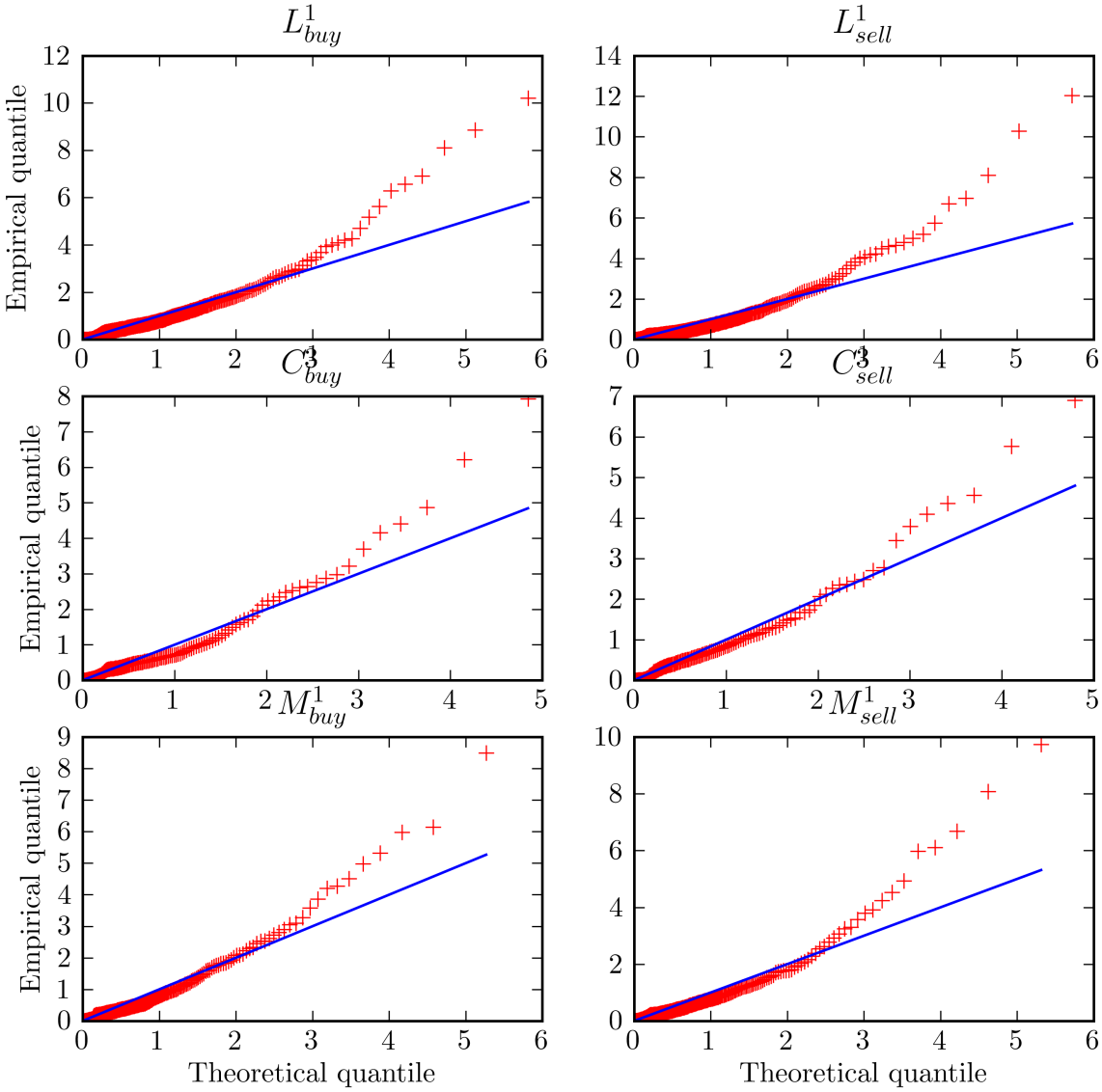


Figure 4.13: The Q-Q plots: Example of DEUTSCHE TEL, 20^{th} Feb 2014. The graphs show that the univariate Hawkes model does not perfectly fit the data.

The results of the corresponding strategy are summarized in **Table 4.15** and the detail are given in the **Table 5.34** of the Appendix.

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.89	0.52	58156	1230	0.75	0.00	8.41
Min	0.86	0.45	26666	-11901	0.41	-0.17	3.06
Max	0.91	0.59	100714	11098	1.27	0.11	22.54

Table 4.15: In sample and out of sample results for the strategy with Hawkes model.

As in the Poisson case, the model gives a very good performance in sample. Moreover, the out of sample average accuracy is higher than 50%. The average holding period is 8 seconds. This reflects a high turnover and causes a low profitability. A potential cause of this effect is the intensity instability. At each time t, the intensity describes the instantaneous probability of jumping, and might vary considerably when events occurs.

An idea, to smooth this effect, is to compute an average intensity over a time interval. This can be interpreted as the instantaneous trend of the intensity and reduces the noise caused by unconfirmed market moves. To compute an average intensity, giving more important weight to the more recent intensity, an exponential moving average (with a half life around one minute) is applied to the different intensities. The results of this test are summarized in **Table 4.16** and the detail are given in the **Table 5.35** of the Appendix.

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.52	0.50	3372	434	1.17	0.12	186.08
Min	0.51	0.49	2282	-1615	0.76	-0.63	120.77
Max	0.54	0.52	5414	1825	1.94	0.63	298.87

Table 4.16: In sample and out of sample results for the strategy with Hawkes model.

When smoothing the intensities using an exponential moving average, a reasonable average holding period of 3 minutes is obtained, resulting in a positive out of sample profitability of 0.12 bp. Notice also that over the 30 stocks only 5 show an accuracy lower than 50%. Although this result is better than all the previous, the profitability is not sufficient to cover any trading costs, and no profitable strategy can be run using this model.

The univariate Hawkes model seems to give better results than the Poisson model. However the quality of the Q-Q-plot and the low out of sample profitability show that this model is missing an important part of the order book dynamic. In order to enhance the model, a multivariate Hawkes is used in the last paragraph.

4.3.3 Multivariate Hawkes Model

The first section of this paper shows strong dependencies between the different types of events of the order book. It is, thus, reasonable to fit a model able to incorporate those different interactions. In this paragraph, the events are modeled with a 12-variate Hawkes process -i.e. each event's intensity can be impacted by the arrival of any other event. For simplification purposes, the Hawkes kernel is chosen to be one exponential. Recall that for any component $m \in \{1, \dots, 12\}$ the intensity is given by

$$\lambda_m(t) = \mu_m + \sum_{n=1}^{M} \sum_{T_i<t} \alpha_{mn} e^{-\beta_{mn}(t-T_i)} \mathbb{1}_{\{X_i=n\}}$$

To calibrate the model, for each component m, a vector θ_m of 25 parameters needs to be estimated using the maximum likelihood

$$\theta_m = (\mu_m, \alpha_{m1} \dots \alpha_{m12}, \beta_{m1} \dots \beta_{m12})$$

This leads to 300 parameters (including 150 for the events changing the price) and might overfit the data.

It is reasonable to suppose that the dynamic of the order book can be correctly fitted with fewer parameters. Using the results of the dependencies study presented in the first section, the choice is made to set the parameters related to the less important observed dependencies to zero. Recall also that this study focuses on computing the intensities of the events changing the price. Thus, the following parameters are to be estimated, the others are set to zero:

Event	Events with non 0 impact	Parameters to be estimated
L^1_{buy}	$\{M^0_{buy}, L^1_{buy}, M^1_{buy}, M^1_{sell}\}$	$\theta_7 = (\mu_7, \alpha_{7,5}, \alpha_{7,7}, \alpha_{7,11}, \alpha_{7,12}, \beta_{7,5}, \beta_{7,7}, \beta_{7,11}, \beta_{7,12})$
L^1_{sell}	$\{M^0_{sell}, L^1_{sell}, M^1_{buy}, M^1_{sell}\}$	$\theta_8 = (\mu_8, \alpha_{8,6}, \alpha_{8,8}, \alpha_{8,11}, \alpha_{8,12}, \beta_{8,6}, \beta_{8,8}, \beta_{8,11}, \beta_{8,12})$
C^1_{buy}	$\{L^1_{buy}\}$	$\theta_9 = (\mu_9, \alpha_{9,7}, \beta_{9,7})$
C^1_{sell}	$\{L^1_{sell}\}$	$\theta_{10} = (\mu_{10}, \alpha_{10,8}, \beta_{10,8})$
M^1_{buy}	$\{M^0_{buy}, M^1_{buy}\}$	$\theta_{11} = (\mu_{11}, \alpha_{11,5}, \alpha_{11,11}, \beta_{5,11}, \beta_{11,11})$
M^1_{sell}	$\{M^0_{sell}, M^1_{sell}\}$	$\theta_{12} = (\mu_{12}, \alpha_{12,6}, \alpha_{12,12}, \beta_{12,6}, \beta_{12,12})$

Table 4.17: The model dependencies matrix

This model depends on 34 parameters instead of 150 and incorporates the main observed dependencies. **Figure 4.14** shows an example of the Q-Q-plot corresponding to the fit for the same example as **Figure 4.13** (DEUTSCHE TEL, 20^{th} Feb 2014).

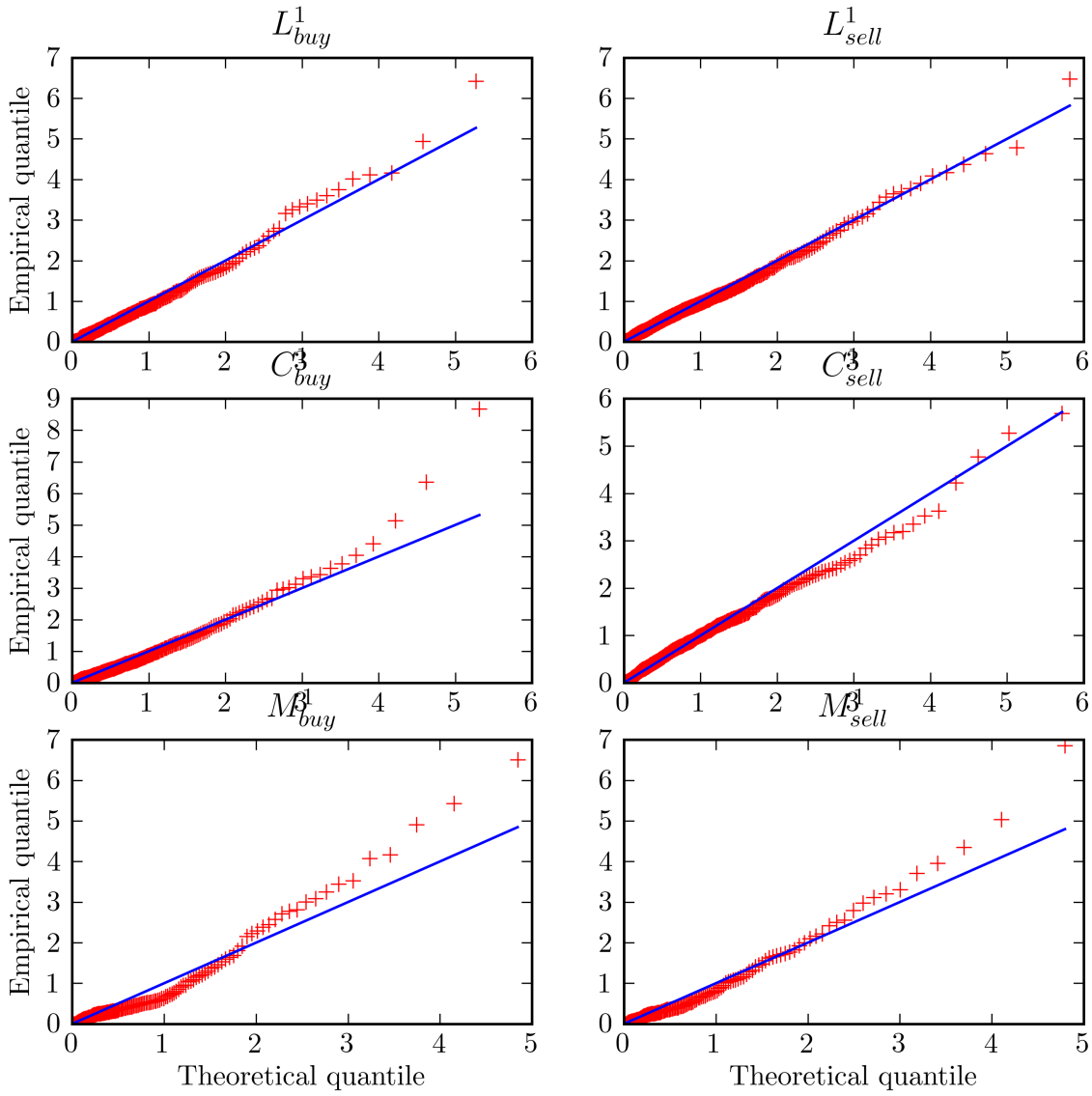


Figure 4.14: The Q-Q plots: Example of DEUTSCHE TEL, 20^{th} Feb 2014. The graphs show that the multivariate Hawkes model fit the data better than the univariate case.

The same strategy tested with the Poisson model and the univariate Hawkes model is tested with the multivariate Hawkes. Recall that for each trading day, the Hawkes parameters are calibrated based on the previous day data. The results are summarized in **Table 4.18** and the detail are given in the **Table 5.36** of the Appendix.

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.71	0.71	32,060	31,587	0.10	0.10	2.06
Min	0.59	0.69	13,281	14,665	0.05	0.06	0.88
Max	0.81	0.74	62,553	53,172	0.23	0.17	4.25

Table 4.18: In sample and out of sample results for the strategy with multivariate Hawkes model.

The results of the multivariate Hawkes model are significantly better than all the previous models. In particular, the out of sample results are as good as the in sample results reflecting a notable stability of the model. The average accuracy is of 70% and the PnL out of sample is positive for all the stocks. However, the average rentability is low and is not sufficient to cover the trading costs. Moreover the holding period is relatively short (2 seconds).

As seen in the previous paragraph, an exponential moving average (EMA) is applied to the signal in order to reduce the trading frequency. The results are summarized in **Table 4.19** and the detail are given in the **Table 5.37** of the Appendix.

	Acc		PnL		Bps		Hp
	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
Average	0.53	0.53	4,295	3,910	1.17	1.07	143.53
Min	0.52	0.51	3,162	2,805	0.67	0.60	92.98
Max	0.56	0.55	6,349	5,642	1.89	1.68	223.06

Table 4.19: In sample and out of sample results for the strategy with multivariate Hawkes model (1-minute EMA).

The model performs very well. The average profitability is higher than 1 bp and is sufficient to cover the trading costs. Moreover, for all the stocks, the PnL is positive and the holding period is reasonable. This result is surprising and leads to questioning the market efficiency hypothesis.

Finally, **Figure 4.15** and **Figure 4.16** represent the cumulative gain (in Euros) over the test period for the two strategies based on the multivariate Hawkes model.

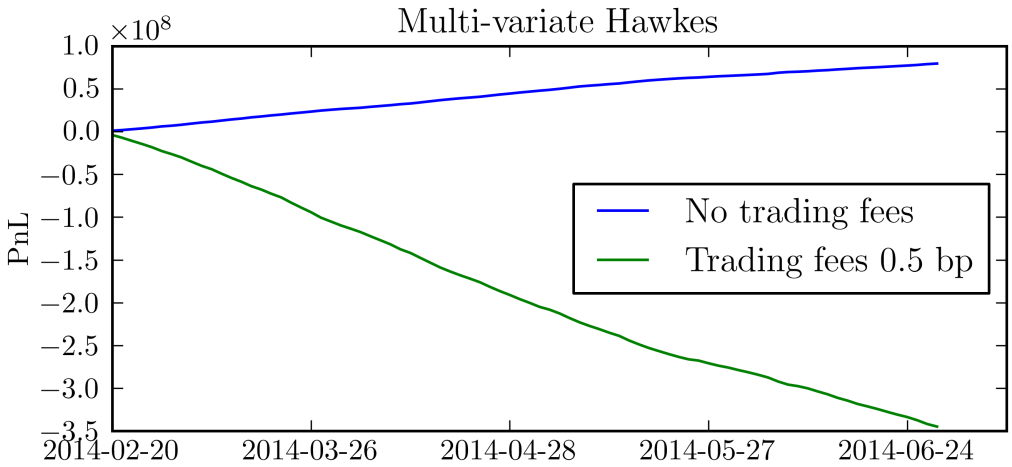


Figure 4.15: Cumulative gain over 4 months.

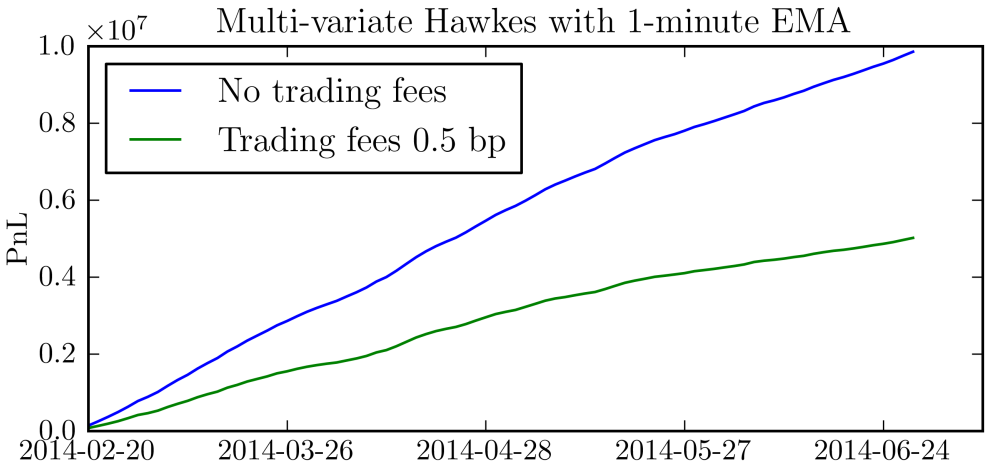


Figure 4.16: Cumulative gain over 4 months

The graphs show that both the strategies are stable over the time. The strategy without the EMA is more profitable when there are no trading fees. This is explained by the higher trading frequency and thus the higher turnover. On the other hand, when adding the fees, only the second strategy remains profitable.

Conclusion

This paper provides a large empirical study of the order book dynamic. In the first part, the classic results of the order flow persistence, the trading activity clustering, and the trading seasonality are confirmed. Moreover, a new effect is mentioned; the market manipulation using fake liquidity. In the second part, some mathematical models were fitted to the order book data. In particular, the multivariate Hawkes model reproduces the different observed statistical properties and can be used to design profitable trading strategies.

Conclusions Générales

L'objectif principal de cette thèse était d'apporter des solutions concrètes à plusieurs problématiques traitées par l'équipe Automatic Market Making de BNP Paribas.

Les résultats du premier papier ont permis de répondre quantitativement au mythe de l'apport de la latence en trading haute fréquence. De plus le protocole de test omniscient a été généralisé et mis à la disposition de tous les membres de l'équipe. Ceci permet aujourd'hui de tester le potentiel des stratégies de placement ou de couverture indépendamment du bruit engendré par la prédiction des rendements futurs.

Les résultats du deuxième papier ont prouvé que la régression Elastic Net (EN) surperforme systématiquement la régression moindres carrés classique (OLS). Ceci a conduit au remplacement des méthodes existantes qui utilisaient la régression OLS par la régression EN. Par ailleurs les bons résultats de la méthode de classification basée sur un seul indicateur ont donné suite à une validation rapide de nouvelles études menées par d'autres personnes de l'équipe sur la sélection des variables explicatives dans le cadres des modèles de microstructure. La maitrise des différentes méthodes de régression a aussi donné suite à des stratégies de gestion d'inventaire en basse fréquence. Ces stratégies , absentes du manuscrit pour des raisons de confidentialité, réalisent de très bonnes performances depuis leurs mises en production. Enfin, les performances non satisfaisantes des modèles prédictifs à un horizon de 5 minutes et de 30 minutes, ont donné suite à l'exploration des signaux multi-assets –ie basés sur les relations statistiques entre les différentes actions-. Ces signaux réalisent aussi de très bonnes performances depuis leurs mises en production.

Les résultats du troisième papier sont repris dans le cadre d'un nouveau projet qui vise à mieux modéliser le carnet d'ordre pour améliorer des stratégies de placement et de couverture. Pour le moment, aucune preuve empirique ne montre que les modèles de Hawkes surperforment une régression linéaire. Cependant cette modélisation a un très grand potentiel car ça permet de répondre à d'autres questions, outre que la prédiction court terme du prix, tel que la probabilité d'exécution d'un ordre placé dans le carnet ou la probabilité de décalage d'une limite.

Sur le plan académique, le premier papier montre l'intérêt numérique de la réécriture sparse des problèmes mathématiques. Ce résultat est général et peut, éventuellement, servir dans d'autres domaines. Les deux derniers papiers ont détaillé les méthodologies de backtest utilisées en production et peuvent servir à d'autres étudiants chercheurs pour les aider à mesurer les performances de leurs modèles.

J'estime que cette thèse a permis de montrer que la recherche académique (souvent reprise par les équipes de Quant et beaucoup moins par les équipes de Trading) peut-être très utile et directement appliquée dans des stratégies de trading. Par ailleurs, l'équipe prendra un thésard pour continuer à travailler sur les stratégies de Market Making à base d'un modèle réaliste de carnet d'ordre, et probablement un autre thésard pour explorer l'applicabilité des méthodes de Machine Learning dans le cadre du Market Making.

Au final je réitère mes chaleureux remerciements à tous ceux qui ont participé à la réalisation de ce travail.

Bibliography

[1] Aldridge I. (2009). How profitable are high-frequency strategies? *Fin Alternatives*, http://www.finalternatives.com/node/9271.

[2] Amadeo K. Stock market crash of 1929: Causes, effects and facts. *US Economy*, http://useconomy.about.com/od/glossary/g/Stock-Market-Crash-of-1929.htm.

[3] Arnuk S. and Saluzzi. J. (2009). Latency arbitrage: The real power behind predatory high frequency trading. *Themis Trading LLC*.

[4] Bachelier L. (1900). *Theorie de la speculation*. PhD thesis, Ecole normale supérieure.

[5] Bacry E., Dayri K., and Muzy J. (2011a). The nature of price returns during periods of high market activity. In *Econophysics of order-driven markets (pp. 155–172)*. Milano, Italy: Springer-Verlag Italia.

[6] Bacry E., Dayri K., and Muzy J. (2011b). Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *European Physical Journal B*, 85:157.

[7] Bacry E., Dayri K., and Muzy J. (2013a). Hawkes model for price and trades high frequency dynamic. *arXiv:1301.1135*.

[8] Bacry E., Delattre S., Hoffmann M., and Muzy J. (2012). Scaling limits for hawkes processes and application to financial statistics. *arXiv:1202.0842*.

[9] Bacry E., Delattre S., Hoffmann M., and Muzy J. (2013b). Modeling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13:65–77.

[10] Baron M., Brogaard J., and Kirilenko A. (2012). The trading profits of high frequency traders. *University of Washington*, Working Paper.

[11] Bechu T., Bertrand E., and Nebenzahl J. (2008). *L'analyse technique*. Broché.

[12] Biais B., Hillion P., and Spatt C. (1995). An empirical analysis of the limit order book and the order flow in the paris bourse. *The Journal of Finance*, 50(5):1655–1689.

[13] Bianco K. M. (2008). *The Subprime Lending Crisis: Causes and Effects of the Mortgage Meltdown*. CCH, Wolters Kluwer Law and Business.

[14] Bouchaud J., Farmer D., and Lillo F. (2009). How markets slowly digest changes in supply and demand, handbook of financial markets: Dynamics and evolution. In *Handbook of Financial Markets: Dynamics and Evolution*. Elsevier.

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

[16] Bremaud P. and Massoulie L. (1996). Stability of nonlinear hawkes processes. *The Annals of Probability*, 24(3):1563–1588.

[17] Brogaard J. (2010). High frequency trading and its impact on market quality. *Kellogg School of Management, Northwestern University*, Working Paper.

[18] Brogaard J., Hendershott T., and Riordan R. (2013). High frequency trading and price discovery. *University of California, Berkeley*, Working Paper.

[19] Brown E. N., Barbieri R., Ventura V., Kass R., and Frank L. M. (2001). The time-rescaling theorem and its application to neural spike train datra analysis. *Naval Research Logistics*, 14(2):325–346.

[20] Buffett W. (1984). The superinvestors of graham-and-doddsville. *Hermes: the Columbia Business School Magazine*, page 4–15.

[21] Carlson M. (2006). A brief history of the 1987 stock market crash. *Finance and Economics Discussion Series Divisions of Research and Statistics and Monetary Affairs Federal Reserve Board.*, 2007-13.

[22] Chakraborti A., Toke I. M., Patriarca M., and Abergel F. (2010). Econophysics: Empirical facts and agent-based models. *arXiv:0909.1979v2*.

[23] Chakraborti A., Toke I. M., Patriarca M., and Abergel F. (2011). Econophysics: Empirical facts and agent-based models. *Quantitative Finance*, 11(7).

[24] Challet D. and Stinchcombe R. (2001). Analyzing and modelling 1 + 1 d markets. *Physica*, A(300):285–299.

[25] Cheney W. and Kincaid D. (2008). *Numerical Mathematics and Computing*. Wadsworth Publishing.

[26] Cho D. and Appelbaum B. (2010). Obama's 'volcker rule' shifts power away from geithner. *Washington Post*, http://www.washingtonpost.com/wp-dyn/content/article/2010/01/21/AR2010012104935.html.

[27] Cuturi M. Linear programming and convex analysis efficiency of the simplex.

[28] Dahl J. (2012). Introduction to optimization using mosek and python.

[29] Daley D. and Vere-Jones D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods, Second Edition*. Springer.

[30] Duhigg C. (2009). Stock traders find speed pays, in milliseconds. *The New York Times*, http://www.nytimes.com/2009/07/24/business/24trading.html?_r=2&hp=&adxnnl=1&adxnnlx=1248440431-piRj8gUXNtgUHfE8g7vd7A&.

[31] Durden T. (2014). The holy grail of trading has been found: Hft firm reveals 1 losing trading day in 1238 days of trading. *http://www.zerohedge.com/news/2014-03-10/holy-grail-trading-has-been-found-hft-firm-reveals-1-losing-trading-day-1238-days-tr*.

[32] Eisler Z., Bouchaud J. P., and Kockelkoren J. (2012). The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance*, 12:1395–1419.

[33] Engel C. and Morris C. S. (1991). Challenges to stock market efficienc: Evidence from mean reversion studies. *Economic Review*, 1991(5).

[34] Fama E. F. (1969). Efficient capital markets : A review od theory and empirical work. *Journal of finance*, 25(2):383–417.

[35] Fama E. F. and French. K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

[36] Farmer J., Gerig A., Lillo F., and Mike S. (2006). Market efficiency and the longmemory of supply and demand: Is price impact variable and permanent or fixed and temporary? *Quantitative Finance*, 6:107–112.

[37] Fawcett T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

[38] Ferguson T. S. *LINEAR PROGRAMMING A Concise Introduction.*

[39] FODRA P. and LABADIE M. (2012). High-frequency market-making with inventory constraints and directional bets. *arXiv:1206.4810.*

[40] Garber P. M. (1990). Famous first bubbles. *The Journal of Economic Perspectives*, 4(2):35–54.

[41] Gerber H. U. and Pafumi G. (1998). Utility functions: From risk theory to finance. *NORTH AMERICAN ACTUARIAL JOURNAL*, 2(3):74–91.

[42] Gopikrishnan P., Plerou V., Gabaix X., and Stanley H. E. (2000). Statistical properties of share volume traded in financial markets. *Physical Review E*, 62(4).

[43] Gujarati D. N. and Porter D. C. (2009). *Basic Econometrics 5th Edition. Chap 10, Multicollinearity: What Happens If the Regressors Are Correlated?* Douglas Reiner.

[44] Gupta V. (2013). Modeling convex optimization problems.

[45] Harris L. (2003). *Trading and Exchanges: Market Microstructure for Practioners.* Oxford University Press.

[46] Hastie T., Tibshirani R., and Friedman J. (2011). *The Elements of Statistical Learning.* Springer.

[47] Hawkes A. G. and Oakes D. (1974). A cluster process representation of a self exciting process. *Journal of Applied Probability*, 11:493–503.

[48] Hendershott T., Jones C., and Menkveld A. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, (66):1–31.

[49] Hoerl A., Kennard R., and Baldwin K. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4(2):105–123.

[50] Hoerl A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:45–59.

[51] Hoerl A. E. and Kennard R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[52] Hollifield B., Miller R. A., and Sandas. P. (2004). Empirical analysis of limit order markets. *The Review of Economic Studies*, 71(4):1027–1063.

[53] Jedidi A. (2014). *Stochastic Order Book Modelling.* PhD thesis, Ecole Centrale Paris.

[54] K. R. S. and Price (1997). Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359.

[55] Kearns M., Kulesza A., and Nevmyvaka Y. (2010). Empirical limitations on high frequency trading profitability. *Journal Of Trading*, 5:50–62.

[56] Kirilenko A. A. and Lo. A. W. (2013). Moore's law versus murphy's law: Algorithmic trading and its discontents. *Journal of Economic Perspectives*, 27(2):51–72.

[57] Kissell R. (2014). *The Science of Algorithmic Trading and Portfolio Management*. Elsevier.

[58] Lawless J. F. and Wang P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, 5(4):307–323.

[59] Lelievre F. and Pilet F. (2013). *Comment les traders haute fréquence menacent de faire sauter la Bourse*. Calmann-Levy.

[60] Lewis M. (2014a). *Flash Boys*. W. W. Norton and Company.

[61] Lewis M. (2014b). The wolf hunters of wall street. *New York Times*, http://www.nytimes.com/2014/04/06/magazine/flash-boys-michael-lewis.html?ref=topics.

[62] Lewis P. A. W. and Shedler G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics*, 26:403–413.

[63] Lillo F. and Farmer J. D. (2008). The long memory of the efficient market. *arXiv:0709.0159v1*.

[64] Lima P. D. (2014). Avec le trading haute frequence, les marches jouent a un jeu immoral et dangereux. *Les Echos*, http://www.lesechos.fr/idees-debats/cercle/cercle-96814-avec-le-trading-haute-frequence-les-marches-jouent-a-un-jeu/-immoral-et-dangereux-1007482.php.

[65] Lo A. W. (2007). Efficient markets hypothesis. In *The New Palgrave: A Dictionary of Economics*. Palgrave McMillan.

[66] LO A. W., MAMAYSKY H., and WANG J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance*, LV(4):1705–1765.

[67] Malkiel B. G. (1973). *A Random Walk Down Wall Street*. W. W. Norton and Company.

[68] McNamara P. (2010). Time flies dept: Dot-com craze peaked 10 years ago. *Network World.*, http://www.networkworld.com/article/2230030/software/time-flies-dept---dot-com-craze-peaked-10-years-ago.html.

[69] Menkveld A. J. (2013). High frequency trading and the new market makers. *Journal of Financial Markets*, 16:712–740.

[70] Moinas S. (2008). Le carnet d ordres : une revue de litterature. *Finance*, 29(1):81–147.

[71] Moro E., Vicente J., Moyano L., Gerig A., Farmer J., Vaglica G., Lillo F., and Mantegna R. (2009). Market impact and trading profile of hidden orders in stock markets. *Physical Review E*, 80.

[72] Murphy J. (1999). *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York Institute of Finance.

[73] N.A (2013). La face cachée du trading haute fréquence. *L'AGEFI*, http://www.agefi.fr/fiche-academics-wikifinance/la-face-cachee-du-trading-haute-frequence-2715.html.

[74] N.A. (2014). Cette société de trading à haute fréquence n'a jamais connu les jours sans. *La Tribune*, www.latribune.fr/entreprises-finance/banques-finance/20140311trib000819317/cette-societe-de-trading-a-haute-frequence-n-a-jamais-connu-les-jours-sans-.html.

[75] Ogata Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:243–261.

[76] Ogata Y. (1981). On lewis simulation method for point processes. *IEEE Transactions on Information Theory*, 27.

[77] ORANGE M. (2014). Le trading haute frequence est un systeme de fraude de grande ampleur. *Mediapart*, http://claude-rochet.fr/trading-haute-frequence-systeme-fraude-grande-ampleur/.

[78] Ozaki T. (1978). Maximum likelihood estimation of hawkes self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 30:145–155.

[79] Philips M. (2012). Knight shows how to lose 440 million in 30 minutes. *Bloomberg Business*, http://www.bloomberg.com/bw/articles/2012-08-02/knight-shows-how-to-lose-440-million-in-30-minutes.

[80] Quiry P., Fur Y. L., and Vernimmen P. (2013). *Finance d'entreprise*. Dalloz - Dalloz Gestion.

[81] Riley J. D. (1955). Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. *Mathematical Tables and Other Aids to Computation*, 9:96–101.

[82] Rubin I. (1972). Regular point processes and their detection. *IEEE Transactions on Information Theory*, 18(5):547–557.

[83] Seber G. A. F. and Lee A. J. (2003). *Linear Regression Analysis*. Wiley.

[84] Sicard G. (1993). *Aux origines des sociétés anonymes. Les moulins de Toulouse au Moyen Age*. Addison-Wesley.

[85] Sun E. W., Kruse T., and Yu M. T. (2013). High frequency trading, liquidity, and execution cost. *Annals of Operations Research*, 223:403–432.

[86] Tajaddini R. (2013). *Momentum Trading Strategies in Financial Markets*. PhD thesis, University of Otago, Dunedin New Zealand.

[87] Taleb N. (2001). Fooled by randomness the hidden role of chance in the markets and in life.

[88] Taranto D. E., Bormetti G., and Lillo F. (2014). The adaptive nature of liquidity taking in limit order books. *arXiv preprint arXiv:1403.0842*.

[89] Tewarson R. P. (1973). *Sparse Matrices. Mathematics in Science and Engineering*. Academic Press.

[90] Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

[91] Tikhonov A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady*, 4:1035–1038.

[92] Tobin J. (1978). A proposal for international monetary reform. *Eastern Economic Journal*, 4:154–159.

[93] Toke I. M. and Pomponio F. (2012). Modelling trades-through in a limit order book using hawkes processes. *Economics:The Open-Access, Open-Assessment E-Journal*, 6.

[94] Vandenberghe L. (2010). The cvxopt linear and quadratic cone program solvers.

[95] Vidyamurthy G. (2004). *Pairs Trading: Quantitative Methods and Analysis*. Wiley.

[96] Wasserman L. (2010). *Lecture Notes for 36 707 Linear Regression.*

[97] Worse U. (2013). Modeling with mosek fusion.

[98] Zarinelli E., Treccani M., Farmer J., and Lillo F. (2014). Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *arXiv:1412.2152.*

[99] Zheng B. (2014). *Detection d'évènements rares dans les données hautes frequences et applications au trading lgorithmique.* PhD thesis, Télécom ParisTech.

[100] Zheng B., Moulines E., and Abergel F. (2013). Price jump prediction in limit order book. *Journal of Mathematical Finance*, 3(2):242–255.

[101] Zou H. and Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2):301–320.

[102] Zwick S. (2011). High-frequency trading: Good, bad or just different? *Futures*, pages 54–58.

Chapter 5

Appendix

	Order book imbalance		Flow quantity		Past return	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.54	0.54	0.53	0.53	0.51	0.51
AIR LIQUIDE	0.56	0.56	0.53	0.53	0.51	0.51
ALLIANZ	0.61	0.61	0.51	0.51	0.53	0.53
ASML Holding NV	0.56	0.56	0.53	0.53	0.51	0.51
BASF AG	0.54	0.54	0.52	0.52	0.50	0.50
BAYER AG	0.54	0.54	0.53	0.53	0.51	0.51
BBVARGENTARIA	0.54	0.54	0.53	0.53	0.51	0.51
BAY MOT WERKE	0.54	0.54	0.53	0.53	0.51	0.51
DANONE	0.56	0.56	0.53	0.53	0.51	0.51
BNP PARIBAS	0.53	0.53	0.52	0.52	0.51	0.51
CARREFOUR	0.55	0.55	0.53	0.53	0.51	0.51
CRH PLC IRLANDE	0.62	0.62	0.58	0.58	0.53	0.53
AXA	0.55	0.55	0.51	0.51	0.52	0.52
DAIMLER CHRYSLER	0.54	0.54	0.53	0.53	0.51	0.51
DEUTSCHE BANK AG	0.53	0.53	0.52	0.52	0.51	0.51
VINCI	0.54	0.54	0.53	0.53	0.51	0.51
DEUTSCHE TELEKOM	0.56	0.56	0.52	0.52	0.51	0.51
ESSILOR INTERNATIONAL	0.56	0.56	0.54	0.54	0.50	0.50
ENEL	0.63	0.63	0.51	0.51	0.55	0.55
ENI	0.64	0.64	0.51	0.51	0.56	0.56
E.ON AG	0.58	0.58	0.51	0.51	0.51	0.51
TOTAL	0.54	0.54	0.52	0.52	0.51	0.51
GENERALI ASSIC	0.62	0.62	0.50	0.50	0.54	0.54
SOCIETE GENERALE	0.52	0.52	0.51	0.51	0.51	0.51
GDF SUEZ	0.56	0.56	0.52	0.52	0.50	0.50
IBERDROLA I	0.56	0.56	0.54	0.54	0.51	0.51
ING	0.53	0.53	0.53	0.53	0.51	0.51
INTESABCI	0.60	0.60	0.51	0.51	0.53	0.53
INDITEX	0.59	0.59	0.55	0.55	0.50	0.50
LVMH	0.59	0.59	0.52	0.52	0.52	0.52
MUNICH RE	0.58	0.58	0.52	0.52	0.51	0.51
LOREAL	0.60	0.60	0.53	0.53	0.52	0.52
PHILIPS ELECTR.	0.56	0.56	0.55	0.55	0.50	0.50
REPSOL	0.57	0.57	0.54	0.54	0.51	0.51
RWE ST	0.54	0.54	0.53	0.53	0.51	0.51
BANCO SAN CENTRAL HISPANO	0.54	0.54	0.53	0.53	0.51	0.51
SANOFI	0.54	0.54	0.53	0.53	0.50	0.50
SAP AG	0.54	0.54	0.52	0.52	0.51	0.51
SAINT GOBAIN	0.54	0.54	0.53	0.53	0.51	0.51
SIEMENS AG	0.54	0.54	0.53	0.53	0.51	0.51
SCHNEIDER ELECTRIC SA	0.54	0.54	0.52	0.52	0.51	0.51
TELEFONICA	0.59	0.59	0.53	0.53	0.51	0.51
UNICREDIT SPA	0.57	0.57	0.50	0.50	0.52	0.52
UNILEVER CERT	0.56	0.56	0.52	0.52	0.51	0.51
VIVENDI UNIVERSAL	0.57	0.57	0.53	0.53	0.51	0.51
VOLKSWAGEN	0.57	0.57	0.52	0.52	0.51	0.51

Table 5.1: The quality of the binary prediction: 1-minute prediction AUC and accuracy per stock

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	1388	1201	1107	1308	174	1264
AIR LIQUIDE	1603	1112	996	1005	169	936
ALLIANZ	2775	1219	221	1107	638	1175
ASML Holding NV	1969	1278	1244	1316	190	1419
BASF AG	1156	1102	921	1311	2	1185
BAYER AG	1269	1055	1142	1251	289	1296
BBVARGENTARIA	1954	1537	1866	1700	595	1934
BAY MOT WERKE	1330	1219	1240	1325	347	1394
DANONE	1591	993	958	1143	231	1196
BNP PARIBAS	1120	1608	831	1620	526	1911
CARREFOUR	1878	1572	1461	1601	600	1665
CRH PLC IRLANDE	4144	1881	2853	1691	1496	1542
AXA	2003	1373	674	1428	582	1603
DAIMLER CHRYSLER	1380	1275	1130	1228	208	1390
DEUTSCHE BANK AG	1251	1372	905	1405	310	1672
VINCI	1410	1113	1252	1211	376	1113
DEUTSCHE TELEKOM	1586	1416	848	1196	308	1298
ESSILOR INTERNATIONAL	1762	1315	1523	1295	12	1281
ENEL	3723	1655	295	1384	1219	1307
ENI	2996	1185	321	1161	1109	1201
E.ON AG	2245	1193	481	1722	323	1445
TOTAL	1256	956	831	977	326	950
GENERALI ASSIC	3977	1764	177	1324	1210	1577
SOCIETE GENERALE	1195	1763	853	1896	643	2060
GDF SUEZ	2031	1227	934	1389	156	1355
IBERDROLA I	2220	1433	1626	1514	566	1403
ING	1511	1564	1493	1491	217	1720
INTESABCI	4019	1911	153	1787	1048	1954
INDITEX	2481	1452	1742	1525	145	1344
LVMH	2445	1220	533	1148	613	1267
MUNICH RE	1895	1107	791	1485	194	1006
LOREAL	2367	1109	894	1242	438	1220
PHILIPS ELECTR.	1978	1173	1670	1565	182	1251
REPSOL	2694	1451	1700	1607	292	1558
RWE ST	1323	1348	1475	1880	307	1747
BANCO SAN CENTRAL HISPANO	1717	1535	1393	1577	383	1684
SANOFI	1368	1040	1118	1123	107	1190
SAP AG	1225	1022	939	1071	117	1084
SAINT GOBAIN	1612	1359	1209	1449	455	1607
SIEMENS AG	1108	983	967	1196	164	1124
SCHNEIDER ELECTRIC SA	1419	1294	1014	1275	379	1436
TELEFONICA	2694	1267	1156	1341	290	1194
UNICREDIT SPA	3039	2025	382	1850	683	2002
UNILEVER CERT	1402	766	551	860	222	949
VIVENDI UNIVERSAL	2142	1223	1114	1391	244	1326
VOLKSWAGEN	2044	1440	1165	1397	225	1359

Table 5.2: The quality of the binary prediction: The daily gain average and standard deviation for the 1-minute prediction (without trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-191	1189	-788	1325	-1222	1531
AIR LIQUIDE	81	1112	-980	1057	-1211	1164
ALLIANZ	1141	1063	-1199	1309	-952	1162
ASML Holding NV	370	1179	-697	1335	-1301	1574
BASF AG	-422	1064	-955	1338	-1298	1558
BAYER AG	-363	1002	-734	1249	-1122	1503
BBVARGENTARIA	303	1477	-58	1681	-910	2027
BAY MOT WERKE	-260	1176	-530	1263	-1256	1510
DANONE	-40	963	-906	1164	-1246	1369
BNP PARIBAS	-402	1596	-1022	1618	-1115	1998
CARREFOUR	251	1486	-492	1606	-975	1690
CRH PLC IRLANDE	2971	1714	934	1612	-27	1549
AXA	313	1299	-1064	1488	-1152	1560
DAIMLER CHRYSLER	-231	1243	-748	1235	-1206	1529
DEUTSCHE BANK AG	-394	1368	-959	1423	-1277	1819
VINCI	-170	1072	-656	1224	-1093	1324
DEUTSCHE TELEKOM	50	1407	-949	1225	-1128	1516
ESSILOR INTERNATIONAL	185	1265	-389	1296	-1104	1575
ENEL	2151	1456	-1069	1610	-329	1198
ENI	1513	971	-1136	1375	-281	1046
E.ON AG	583	1096	-1108	1887	-1047	1592
TOTAL	-362	934	-1058	1024	-1278	1206
GENERALI ASSIC	2369	1565	-1403	1539	-484	1490
SOCIETE GENERALE	-405	1718	-846	1901	-968	2002
GDF SUEZ	402	1140	-951	1438	-1249	1513
IBERDROLA I	762	1332	-312	1503	-1094	1475
ING	-186	1519	-450	1470	-1186	1890
INTESABCI	2333	1715	-1081	1822	-517	1820
INDITEX	1110	1375	-195	1535	-1155	1457
LVMH	831	1119	-1183	1296	-928	1235
MUNICH RE	366	1011	-1019	1490	-1260	1177
LOREAL	816	985	-797	1274	-982	1236
PHILIPS ELECTR.	377	1113	-272	1575	-1255	1490
REPSOL	1233	1308	-184	1585	-1188	1713
RWE ST	-182	1251	-399	1864	-1122	1960
BANCO SAN CENTRAL HISPANO	205	1431	-492	1566	-1064	1822
SANOFI	-279	998	-720	1127	-1382	1454
SAP AG	-340	1000	-944	1093	-1428	1277
SAINT GOBAIN	-48	1326	-694	1463	-1060	1655
SIEMENS AG	-472	966	-898	1209	-1353	1363
SCHNEIDER ELECTRIC SA	-162	1263	-872	1296	-1339	1493
TELEFONICA	1124	1130	-686	1342	-1044	1257
UNICREDIT SPA	1434	1940	-896	1953	-738	2067
UNILEVER CERT	-253	730	-1246	938	-1344	1142
VIVENDI UNIVERSAL	547	1113	-804	1386	-1186	1452
VOLKSWAGEN	446	1373	-785	1408	-979	1584

Table 5.3: The quality of the binary prediction: The daily gain average and standard deviation for the 1-minute prediction (with trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.50	0.50	0.50	0.50	0.51	0.51
AIR LIQUIDE	0.53	0.53	0.50	0.50	0.50	0.51
ALLIANZ	0.54	0.54	0.51	0.51	0.52	0.52
ASML Holding NV	0.51	0.51	0.51	0.51	0.51	0.51
BASF AG	0.50	0.50	0.50	0.50	0.50	0.50
BAYER AG	0.51	0.51	0.51	0.51	0.51	0.51
BBVARGENTARIA	0.50	0.50	0.50	0.50	0.50	0.50
BAY MOT WERKE	0.50	0.50	0.51	0.51	0.50	0.50
DANONE	0.51	0.51	0.51	0.51	0.50	0.50
BNP PARIBAS	0.50	0.50	0.50	0.50	0.50	0.50
CARREFOUR	0.51	0.51	0.50	0.51	0.51	0.51
CRH PLC IRLANDE	0.56	0.56	0.53	0.53	0.51	0.51
AXA	0.50	0.50	0.49	0.49	0.50	0.50
DAIMLER CHRYSLER	0.51	0.51	0.50	0.50	0.51	0.51
DEUTSCHE BANK AG	0.50	0.50	0.50	0.50	0.50	0.50
VINCI	0.51	0.51	0.51	0.51	0.50	0.51
DEUTSCHE TELEKOM	0.52	0.52	0.50	0.51	0.50	0.50
ESSILOR INTERNATIONAL	0.51	0.51	0.51	0.51	0.50	0.50
ENEL	0.55	0.55	0.51	0.51	0.51	0.51
ENI	0.56	0.56	0.51	0.51	0.51	0.51
E.ON AG	0.52	0.52	0.50	0.50	0.50	0.50
TOTAL	0.51	0.51	0.50	0.51	0.51	0.51
GENERALI ASSIC	0.56	0.56	0.50	0.50	0.51	0.51
SOCIETE GENERALE	0.50	0.50	0.50	0.50	0.50	0.50
GDF SUEZ	0.52	0.52	0.50	0.50	0.50	0.50
IBERDROLA I	0.52	0.52	0.51	0.51	0.51	0.51
ING	0.50	0.50	0.50	0.50	0.50	0.50
INTESABCI	0.54	0.54	0.49	0.49	0.50	0.50
INDITEX	0.53	0.53	0.50	0.50	0.52	0.52
LVMH	0.53	0.53	0.50	0.50	0.51	0.51
MUNICH RE	0.53	0.53	0.51	0.51	0.51	0.51
LOREAL	0.53	0.53	0.50	0.50	0.51	0.51
PHILIPS ELECTR.	0.52	0.52	0.51	0.51	0.50	0.50
REPSOL	0.53	0.53	0.51	0.51	0.50	0.50
RWE ST	0.51	0.51	0.50	0.50	0.50	0.50
BANCO SAN CENTRAL HISPANO	0.51	0.51	0.51	0.51	0.51	0.51
SANOFI	0.51	0.51	0.50	0.50	0.50	0.50
SAP AG	0.51	0.51	0.51	0.51	0.50	0.50
SAINT GOBAIN	0.50	0.50	0.50	0.50	0.51	0.51
SIEMENS AG	0.51	0.51	0.50	0.50	0.51	0.51
SCHNEIDER ELECTRIC SA	0.50	0.50	0.50	0.50	0.50	0.50
TELEFONICA	0.53	0.53	0.51	0.51	0.51	0.51
UNICREDIT SPA	0.52	0.52	0.51	0.51	0.50	0.50
UNILEVER CERT	0.52	0.52	0.50	0.50	0.50	0.50
VIVENDI UNIVERSAL	0.52	0.52	0.50	0.50	0.50	0.50
VOLKSWAGEN	0.52	0.52	0.50	0.50	0.51	0.51

Table 5.4: The quality of the binary prediction: 5-minute prediction AUC and accuracy per stock

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	45	978	112	958	44	1010
AIR LIQUIDE	308	752	40	741	42	798
ALLIANZ	479	1073	74	871	182	906
ASML Holding NV	146	1027	83	1029	-39	1143
BASF AG	-7	976	107	969	22	987
BAYER AG	195	972	161	1016	50	963
BBVARGENTARIA	129	1529	107	1307	83	1332
BAY MOT WERKE	67	1005	196	1010	43	969
DANONE	203	1008	52	938	-65	845
BNP PARIBAS	65	1327	1	1350	-61	1376
CARREFOUR	134	1238	193	1214	32	1267
CRH PLC IRLANDE	1167	1433	567	1378	310	1523
AXA	112	1230	-68	1254	-120	1246
DAIMLER CHRYSLER	79	1037	-72	1088	27	1059
DEUTSCHE BANK AG	-13	1362	-35	1287	6	1277
VINCI	226	877	195	927	147	892
DEUTSCHE TELEKOM	319	857	195	837	31	980
ESSILOR INTERNATIONAL	103	990	114	977	-14	968
ENEL	700	1227	-4	1183	108	1117
ENI	556	822	39	841	71	815
E.ON AG	279	1005	78	1158	23	1022
TOTAL	139	738	71	842	150	845
GENERALI ASSIC	853	1233	-22	1126	-30	1257
SOCIETE GENERALE	121	1523	-72	1542	-75	1587
GDF SUEZ	328	993	61	1105	105	964
IBERDROLA I	443	1173	169	1165	66	1085
ING	49	1342	250	1521	-18	1341
INTESABCI	757	1549	-102	1540	-75	1536
INDITEX	333	1108	160	1099	138	1078
LVMH	367	915	1	882	71	927
MUNICH RE	362	917	100	930	135	903
LOREAL	345	920	5	860	124	955
PHILIPS ELECTR.	308	1053	268	1087	52	980
REPSOL	548	1138	182	1190	41	1175
RWE ST	209	1229	252	1668	104	1627
BANCO SAN CENTRAL HISPANO	246	1309	190	1289	58	1136
SANOFI	171	891	78	951	-26	860
SAP AG	45	799	76	787	-0	846
SAINT GOBAIN	134	1135	26	1106	153	1149
SIEMENS AG	161	927	42	755	84	896
SCHNEIDER ELECTRIC SA	140	1015	83	993	109	1075
TELEFONICA	443	924	192	1028	141	927
UNICREDIT SPA	383	1738	156	1594	43	1697
UNILEVER CERT	169	734	-11	677	-13	704
VIVENDI UNIVERSAL	324	1000	5	971	14	1018
VOLKSWAGEN	219	1185	38	1002	46	1087

Table 5.5: The quality of the binary prediction: The daily gain average and standard deviation for the 5-minute prediction (without trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-182	973	-128	974	-212	1028
AIR LIQUIDE	54	746	-214	764	-184	800
ALLIANZ	156	1059	-172	894	-113	878
ASML Holding NV	-115	1022	-167	1032	-278	1138
BASF AG	-190	997	-180	980	-228	992
BAYER AG	-46	960	-73	1032	-202	967
BBVARGENTARIA	-95	1528	-152	1314	-137	1330
BAY MOT WERKE	-110	997	-71	1004	-183	952
DANONE	-62	995	-225	962	-309	867
BNP PARIBAS	-154	1327	-240	1348	-309	1371
CARREFOUR	-119	1229	-50	1213	-228	1270
CRH PLC IRLANDE	852	1407	235	1360	-7	1477
AXA	-149	1203	-310	1269	-346	1261
DAIMLER CHRYSLER	-119	1032	-288	1107	-184	1078
DEUTSCHE BANK AG	-246	1368	-249	1293	-249	1279
VINCI	35	887	9	937	-41	904
DEUTSCHE TELEKOM	52	857	-53	844	-205	1000
ESSILOR INTERNATIONAL	-116	1002	-134	967	-251	948
ENEL	395	1196	-201	1208	-97	1111
ENI	226	783	-181	837	-138	807
E.ON AG	19	996	-157	1173	-166	1051
TOTAL	-68	766	-162	860	-122	856
GENERALI ASSIC	496	1205	-251	1155	-323	1241
SOCIETE GENERALE	-82	1526	-282	1562	-291	1591
GDF SUEZ	64	977	-182	1114	-134	978
IBERDROLA I	175	1155	-96	1165	-205	1073
ING	-186	1359	-12	1506	-287	1349
INTESABCI	419	1505	-302	1560	-311	1520
INDITEX	70	1074	-142	1097	-119	1060
LVMH	71	900	-206	912	-166	911
MUNICH RE	74	888	-133	935	-145	880
LOREAL	83	901	-184	889	-134	957
PHILIPS ELECTR.	-1	1058	-30	1083	-222	967
REPSOL	256	1116	-93	1183	-173	1177
RWE ST	12	1237	19	1668	-114	1635
BANCO SAN CENTRAL HISPANO	6	1306	-79	1281	-184	1150
SANOFI	-59	892	-165	955	-207	901
SAP AG	-188	794	-184	794	-255	850
SAINT GOBAIN	-51	1152	-206	1112	-81	1155
SIEMENS AG	-65	927	-144	782	-162	889
SCHNEIDER ELECTRIC SA	-107	1007	-155	1002	-127	1065
TELEFONICA	134	906	-29	1041	-79	930
UNICREDIT SPA	120	1703	-50	1590	-197	1701
UNILEVER CERT	-88	715	-229	713	-215	730
VIVENDI UNIVERSAL	38	989	-222	983	-266	1029
VOLKSWAGEN	-20	1178	-175	1021	-173	1088

Table 5.6: The quality of the binary prediction: The daily gain average and standard deviation for the 1-minute prediction (with 0.5 bp trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.50	0.50	0.50	0.50	0.49	0.49
AIR LIQUIDE	0.50	0.50	0.50	0.50	0.49	0.49
ALLIANZ	0.51	0.51	0.51	0.51	0.49	0.49
ASML Holding NV	0.49	0.49	0.50	0.50	0.51	0.51
BASF AG	0.49	0.49	0.50	0.50	0.49	0.49
BAYER AG	0.51	0.51	0.49	0.49	0.49	0.50
BBVARGENTARIA	0.50	0.50	0.51	0.51	0.51	0.51
BAY MOT WERKE	0.50	0.50	0.51	0.51	0.50	0.50
DANONE	0.50	0.50	0.51	0.51	0.51	0.51
BNP PARIBAS	0.48	0.49	0.49	0.49	0.51	0.51
CARREFOUR	0.52	0.52	0.51	0.51	0.50	0.50
CRH PLC IRLANDE	0.51	0.51	0.51	0.51	0.50	0.50
AXA	0.49	0.49	0.49	0.49	0.49	0.49
DAIMLER CHRYSLER	0.51	0.52	0.50	0.51	0.50	0.50
DEUTSCHE BANK AG	0.52	0.52	0.49	0.49	0.51	0.51
VINCI	0.51	0.52	0.50	0.51	0.52	0.52
DEUTSCHE TELEKOM	0.49	0.50	0.50	0.51	0.51	0.51
ESSILOR INTERNATIONAL	0.50	0.50	0.50	0.50	0.51	0.51
ENEL	0.50	0.50	0.50	0.50	0.51	0.51
ENI	0.51	0.51	0.50	0.50	0.51	0.51
E.ON AG	0.49	0.49	0.50	0.50	0.51	0.51
TOTAL	0.51	0.51	0.52	0.52	0.51	0.51
GENERALI ASSIC	0.50	0.50	0.51	0.51	0.50	0.50
SOCIETE GENERALE	0.50	0.50	0.51	0.51	0.51	0.51
GDF SUEZ	0.51	0.51	0.50	0.50	0.52	0.52
IBERDROLA I	0.52	0.52	0.50	0.50	0.52	0.52
ING	0.50	0.50	0.50	0.50	0.49	0.49
INTESABCI	0.50	0.50	0.50	0.50	0.49	0.49
INDITEX	0.50	0.50	0.49	0.49	0.51	0.51
LVMH	0.50	0.50	0.50	0.50	0.49	0.49
MUNICH RE	0.50	0.50	0.51	0.51	0.50	0.50
LOREAL	0.50	0.51	0.51	0.51	0.50	0.50
PHILIPS ELECTR.	0.49	0.49	0.50	0.50	0.51	0.51
REPSOL	0.49	0.49	0.50	0.50	0.51	0.51
RWE ST	0.50	0.51	0.50	0.51	0.51	0.51
BANCO SAN CENTRAL HISPANO	0.50	0.50	0.48	0.48	0.50	0.50
SANOFI	0.50	0.50	0.51	0.51	0.49	0.49
SAP AG	0.51	0.51	0.50	0.50	0.51	0.51
SAINT GOBAIN	0.50	0.50	0.51	0.52	0.50	0.50
SIEMENS AG	0.52	0.52	0.50	0.50	0.50	0.50
SCHNEIDER ELECTRIC SA	0.50	0.50	0.48	0.48	0.51	0.51
TELEFONICA	0.52	0.52	0.49	0.49	0.50	0.50
UNICREDIT SPA	0.52	0.52	0.50	0.50	0.51	0.51
UNILEVER CERT	0.52	0.52	0.50	0.50	0.50	0.50
VIVENDI UNIVERSAL	0.50	0.50	0.50	0.50	0.50	0.50
VOLKSWAGEN	0.52	0.51	0.50	0.50	0.50	0.50

Table 5.7: The quality of the binary prediction: 30-minute prediction AUC and accuracy per stock

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-24	952	-9	832	-54	936
AIR LIQUIDE	-1	676	-15	682	-68	737
ALLIANZ	102	866	15	911	-9	855
ASML Holding NV	-69	1009	31	934	-27	1046
BASF AG	-88	855	57	905	-20	859
BAYER AG	7	989	-65	917	-68	884
BBVARGENTARIA	-76	1271	19	1185	86	1214
BAY MOT WERKE	-87	941	57	919	-50	867
DANONE	-85	813	-46	806	-11	812
BNP PARIBAS	-87	1273	-63	1252	54	1257
CARREFOUR	56	1153	-45	1208	-86	1068
CRH PLC IRLANDE	79	1125	84	1292	-26	1263
AXA	-128	1067	-94	1095	-44	1188
DAIMLER CHRYSLER	106	906	-53	964	-17	950
DEUTSCHE BANK AG	168	1090	-137	1120	50	1121
VINCI	123	837	36	821	109	801
DEUTSCHE TELEKOM	10	855	47	866	13	796
ESSILOR INTERNATIONAL	-26	932	29	922	11	976
ENEL	19	1044	10	1039	72	1015
ENI	38	746	-59	767	-2	775
E.ON AG	-9	968	-29	979	57	971
TOTAL	72	752	106	707	47	743
GENERALI ASSIC	-107	1067	-8	1124	-57	1207
SOCIETE GENERALE	1	1554	38	1454	35	1466
GDF SUEZ	61	908	-3	891	49	881
IBERDROLA I	78	1114	-28	1033	111	1075
ING	-48	1348	-52	1258	-34	1324
INTESABCI	-77	1457	-18	1431	17	1437
INDITEX	8	975	-62	984	57	900
LVMH	-5	857	-20	873	-81	807
MUNICH RE	-17	787	-17	754	-31	744
LOREAL	14	842	72	804	45	877
PHILIPS ELECTR.	-69	845	-25	844	14	903
REPSOL	-66	1011	-32	1022	47	999
RWE ST	60	1242	113	1259	63	1228
BANCO SAN CENTRAL HISPANO	-26	1227	-109	1205	10	1180
SANOFI	-60	924	34	952	-34	890
SAP AG	48	776	-16	863	25	725
SAINT GOBAIN	12	1016	93	1072	-40	1075
SIEMENS AG	137	912	-39	909	-41	893
SCHNEIDER ELECTRIC SA	17	985	-122	917	72	940
TELEFONICA	135	927	38	906	56	879
UNICREDIT SPA	188	1709	-3	1592	41	1625
UNILEVER CERT	29	605	-54	639	-14	661
VIVENDI UNIVERSAL	-18	945	43	935	-27	933
VOLKSWAGEN	100	1110	-6	1113	43	1135

Table 5.8: The quality of the binary prediction: The daily gain average and standard deviation for the 30-minute prediction (without trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-56	952	-33	835	-88	937
AIR LIQUIDE	-30	679	-46	686	-99	735
ALLIANZ	67	866	-22	910	-38	858
ASML Holding NV	-107	1008	-9	934	-63	1045
BASF AG	-117	857	28	903	-51	862
BAYER AG	-29	991	-95	919	-95	886
BBVARGENTARIA	-104	1273	-8	1186	53	1212
BAY MOT WERKE	-115	942	22	918	-78	868
DANONE	-117	811	-82	806	-50	810
BNP PARIBAS	-113	1274	-96	1254	15	1253
CARREFOUR	20	1153	-79	1209	-122	1065
CRH PLC IRLANDE	48	1121	48	1290	-58	1263
AXA	-159	1067	-120	1093	-84	1185
DAIMLER CHRYSLER	74	906	-79	966	-47	950
DEUTSCHE BANK AG	129	1091	-171	1120	17	1120
VINCI	95	839	8	827	86	802
DEUTSCHE TELEKOM	-18	856	16	866	-15	794
ESSILOR INTERNATIONAL	-51	931	-1	922	-26	973
ENEL	-11	1044	-16	1041	46	1015
ENI	5	745	-88	767	-27	774
E.ON AG	-36	968	-57	983	23	967
TOTAL	41	756	78	706	13	744
GENERALI ASSIC	-139	1066	-39	1124	-92	1208
SOCIETE GENERALE	-28	1556	4	1453	-2	1464
GDF SUEZ	25	909	-34	888	17	880
IBERDROLA I	47	1112	-60	1035	75	1073
ING	-82	1348	-86	1257	-69	1325
INTESABCI	-103	1459	-48	1429	-15	1438
INDITEX	-20	976	-88	986	24	899
LVMH	-38	856	-55	871	-112	806
MUNICH RE	-49	787	-50	758	-61	746
LOREAL	-14	842	43	805	17	879
PHILIPS ELECTR.	-97	847	-57	844	-20	898
REPSOL	-97	1012	-65	1023	8	998
RWE ST	34	1243	81	1263	33	1227
BANCO SAN CENTRAL HISPANO	-53	1228	-137	1208	-23	1177
SANOFI	-93	928	1	953	-63	890
SAP AG	15	776	-54	864	-20	716
SAINT GOBAIN	-14	1020	57	1075	-73	1074
SIEMENS AG	101	912	-68	911	-71	893
SCHNEIDER ELECTRIC SA	-15	984	-155	917	37	938
TELEFONICA	107	926	8	910	24	881
UNICREDIT SPA	154	1708	-33	1593	10	1628
UNILEVER CERT	-8	606	-89	639	-51	659
VIVENDI UNIVERSAL	-53	947	3	935	-66	932
VOLKSWAGEN	75	1111	-32	1116	10	1138

Table 5.9: The quality of the binary prediction: The daily gain average and standard deviation for the 30-minute prediction (with 0.5 bp trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.58	0.59	0.50	0.42	0.50	0.50
AIR LIQUIDE	0.71	0.72	nan	nan	0.50	0.58
ALLIANZ	0.69	0.69	0.50	0.54	0.61	0.61
ASML Holding NV	0.60	0.60	0.50	0.54	0.48	0.48
BASF AG	0.60	0.60	nan	nan	0.49	0.50
BAYER AG	0.53	0.55	0.50	0.59	0.50	0.56
BBVARGENTARIA	0.57	0.57	0.55	0.55	0.55	0.56
BAY MOT WERKE	0.57	0.58	0.55	0.55	0.54	0.55
DANONE	0.60	0.60	nan	nan	0.58	0.58
BNP PARIBAS	0.58	0.59	0.50	0.50	0.52	0.53
CARREFOUR	0.59	0.60	0.50	0.56	0.56	0.56
CRH PLC IRLANDE	0.70	0.70	0.64	0.64	0.55	0.56
AXA	0.58	0.60	nan	nan	0.56	0.56
DAIMLER CHRYSLER	0.57	0.57	0.50	0.51	0.54	0.54
DEUTSCHE BANK AG	0.55	0.55	0.54	0.56	0.52	0.52
VINCI	0.60	0.60	0.55	0.56	0.56	0.56
DEUTSCHE TELEKOM	0.71	0.72	nan	nan	0.51	0.51
ESSILOR INTERNATIONAL	0.60	0.60	0.50	0.55	0.52	0.56
ENEL	0.73	0.73	nan	nan	0.57	0.60
ENI	0.76	0.76	nan	nan	0.61	0.61
E.ON AG	0.64	0.64	nan	nan	0.53	0.53
TOTAL	0.54	0.59	nan	nan	0.50	0.46
GENERALI ASSIC	0.68	0.68	nan	nan	0.60	0.60
SOCIETE GENERALE	0.55	0.56	0.50	0.54	0.52	0.54
GDF SUEZ	0.62	0.62	nan	nan	0.53	0.53
IBERDROLA I	0.63	0.63	0.56	0.56	0.57	0.57
ING	0.55	0.55	0.54	0.55	0.52	0.55
INTESABCI	0.67	0.67	nan	nan	0.58	0.58
INDITEX	0.68	0.68	0.58	0.58	0.55	0.55
LVMH	0.65	0.66	nan	nan	0.58	0.58
MUNICH RE	0.66	0.66	0.55	0.55	0.54	0.54
LOREAL	0.67	0.67	nan	nan	0.58	0.58
PHILIPS ELECTR.	0.61	0.62	0.50	0.51	0.52	0.54
REPSOL	0.63	0.63	0.53	0.58	0.57	0.57
RWE ST	0.58	0.58	0.53	0.55	0.52	0.52
BANCO SAN CENTRAL HISPANO	0.57	0.56	0.52	0.51	0.58	0.58
SANOFI	0.60	0.60	nan	nan	0.50	0.60
SAP AG	0.52	0.61	0.50	0.56	0.52	0.54
SAINT GOBAIN	0.56	0.58	0.54	0.58	0.54	0.55
SIEMENS AG	0.56	0.61	0.55	0.56	0.59	0.59
SCHNEIDER ELECTRIC SA	0.57	0.58	nan	nan	0.56	0.57
TELEFONICA	0.68	0.68	0.53	0.57	0.56	0.56
UNICREDIT SPA	0.64	0.65	0.50	0.54	0.57	0.57
UNILEVER CERT	0.50	0.63	nan	nan	nan	nan
VIVENDI UNIVERSAL	0.63	0.63	nan	nan	0.51	0.52
VOLKSWAGEN	0.62	0.62	0.49	0.49	0.52	0.53

Table 5.10: The quality of the 4-class prediction: 1-minute prediction AUC and accuracy per stock

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	137	388	-6	98	4	131
AIR LIQUIDE	306	577	0	0	3	42
ALLIANZ	1363	779	4	47	68	276
ASML Holding NV	440	651	5	63	-2	132
BASF AG	87	287	0	0	-2	48
BAYER AG	21	128	14	137	14	99
BBVARGENTARIA	390	665	273	669	208	582
BAY MOT WERKE	107	281	47	276	52	238
DANONE	168	366	0	0	4	47
BNP PARIBAS	171	428	3	66	44	453
CARREFOUR	486	715	11	139	136	469
CRH PLC IRLANDE	2534	1240	1364	1077	202	560
AXA	594	786	0	0	55	320
DAIMLER CHRYSLER	93	289	2	24	16	191
DEUTSCHE BANK AG	34	224	38	212	12	291
VINCI	154	451	13	111	27	147
DEUTSCHE TELEKOM	488	827	0	0	3	66
ESSILOR INTERNATIONAL	351	596	17	164	10	106
ENEL	2219	1056	0	0	193	503
ENI	2000	773	0	0	110	300
E.ON AG	651	680	0	0	10	168
TOTAL	10	93	0	0	1	38
GENERALI ASSIC	2520	1420	0	0	249	756
SOCIETE GENERALE	184	503	2	25	56	410
GDF SUEZ	504	692	0	0	21	171
IBERDROLA I	738	951	155	512	115	409
ING	109	373	59	296	7	138
INTESABCI	2512	1248	0	0	185	731
INDITEX	1039	914	151	587	44	223
LVMH	930	847	0	0	64	277
MUNICH RE	370	533	26	145	3	50
LOREAL	800	674	0	0	22	112
PHILIPS ELECTR.	440	613	6	94	11	116
REPSOL	1234	1013	142	445	110	555
RWE ST	192	556	85	380	29	364
BANCO SAN CENTRAL HISPANO	228	501	4	158	168	635
SANOFI	26	127	0	0	6	90
SAP AG	50	196	24	187	6	200
SAINT GOBAIN	210	519	30	186	88	362
SIEMENS AG	26	139	31	198	28	162
SCHNEIDER ELECTRIC SA	123	434	0	0	37	214
TELEFONICA	1402	825	36	232	34	205
UNICREDIT SPA	1316	1393	17	197	247	835
UNILEVER CERT	16	104	0	0	0	0
VIVENDI UNIVERSAL	583	826	0	0	5	141
VOLKSWAGEN	530	745	-0	78	1	215

Table 5.11: The quality of the 4-class prediction: The daily gain average and standard deviation for the 1-minute prediction (without trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	22	263	-9	150	-38	183
AIR LIQUIDE	128	329	0	0	1	31
ALLIANZ	586	559	-1	16	8	194
ASML Holding NV	125	408	-0	32	-25	168
BASF AG	15	189	0	0	-7	51
BAYER AG	-14	105	1	86	-2	77
BBVARGENTARIA	107	507	31	465	16	474
BAY MOT WERKE	-1	193	1	199	-12	184
DANONE	21	210	0	0	-4	42
BNP PARIBAS	34	271	-12	126	-65	481
CARREFOUR	116	506	-8	131	18	362
CRH PLC IRLANDE	1848	1102	518	844	18	442
AXA	174	550	0	0	-23	274
DAIMLER CHRYSLER	-7	245	-1	14	-32	199
DEUTSCHE BANK AG	-23	204	8	122	-33	311
VINCI	38	281	-3	73	-5	111
DEUTSCHE TELEKOM	241	526	0	0	-10	79
ESSILOR INTERNATIONAL	88	388	-14	157	-4	91
ENEL	1338	881	0	0	-18	443
ENI	1082	613	0	0	-25	211
E.ON AG	185	475	0	0	-22	173
TOTAL	-5	72	0	0	-3	49
GENERALI ASSIC	1518	1179	0	0	58	636
SOCIETE GENERALE	2	412	-2	24	-41	394
GDF SUEZ	142	464	0	0	-8	126
IBERDROLA I	340	722	28	331	18	292
ING	-13	329	6	209	-12	147
INTESABCI	1514	1096	0	0	-20	658
INDITEX	547	702	3	400	-10	198
LVMH	372	581	0	0	-11	169
MUNICH RE	111	322	-3	62	-6	46
LOREAL	285	443	0	0	-5	85
PHILIPS ELECTR.	113	417	-6	96	-8	105
REPSOL	611	809	40	254	27	437
RWE ST	38	450	-2	299	-42	372
BANCO SAN CENTRAL HISPANO	20	392	-31	203	49	463
SANOFI	1	69	0	0	-0	79
SAP AG	2	120	-4	137	-30	207
SAINT GOBAIN	25	403	-1	114	-7	289
SIEMENS AG	2	74	-2	89	-3	141
SCHNEIDER ELECTRIC SA	16	317	0	0	-14	195
TELEFONICA	656	663	6	139	-7	183
UNICREDIT SPA	693	1159	-5	173	19	628
UNILEVER CERT	1	56	0	0	0	0
VIVENDI UNIVERSAL	214	617	0	0	-27	175
VOLKSWAGEN	171	545	-7	115	-45	246

Table 5.12: The quality of the 4-class prediction: The daily gain average and standard deviation for the 1-minute prediction (with trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.51	0.51	0.50	0.50	0.51	0.51
AIR LIQUIDE	0.55	0.55	0.50	0.50	0.51	0.51
ALLIANZ	0.56	0.56	0.51	0.51	0.54	0.54
ASML Holding NV	0.52	0.52	0.51	0.51	0.52	0.52
BASF AG	0.51	0.51	0.51	0.51	0.52	0.52
BAYER AG	0.51	0.52	0.50	0.51	0.50	0.50
BBVARGENTARIA	0.52	0.52	0.51	0.51	0.52	0.52
BAY MOT WERKE	0.51	0.51	0.50	0.50	0.51	0.51
DANONE	0.52	0.52	0.51	0.51	0.50	0.50
BNP PARIBAS	0.49	0.49	0.51	0.51	0.50	0.50
CARREFOUR	0.52	0.52	0.50	0.50	0.51	0.51
CRH PLC IRLANDE	0.57	0.57	0.54	0.54	0.52	0.52
AXA	0.51	0.50	0.49	0.49	0.50	0.50
DAIMLER CHRYSLER	0.52	0.52	0.49	0.50	0.49	0.50
DEUTSCHE BANK AG	0.49	0.49	0.49	0.49	0.50	0.50
VINCI	0.51	0.52	0.50	0.51	0.52	0.53
DEUTSCHE TELEKOM	0.55	0.55	0.51	0.51	0.49	0.50
ESSILOR INTERNATIONAL	0.52	0.52	0.52	0.52	0.51	0.51
ENEL	0.58	0.58	0.51	0.51	0.52	0.53
ENI	0.60	0.60	0.52	0.52	0.52	0.52
E.ON AG	0.55	0.55	0.51	0.51	0.51	0.51
TOTAL	0.51	0.52	0.50	0.50	0.52	0.53
GENERALI ASSIC	0.58	0.58	0.49	0.49	0.53	0.53
SOCIETE GENERALE	0.50	0.50	0.49	0.49	0.50	0.50
GDF SUEZ	0.52	0.52	0.50	0.50	0.51	0.51
IBERDROLA I	0.54	0.54	0.51	0.51	0.51	0.51
ING	0.50	0.50	0.50	0.50	0.49	0.49
INTESABCI	0.55	0.55	0.50	0.50	0.51	0.51
INDITEX	0.56	0.56	0.51	0.51	0.53	0.53
LVMH	0.55	0.55	0.49	0.49	0.52	0.52
MUNICH RE	0.57	0.57	0.53	0.53	0.54	0.54
LOREAL	0.57	0.57	0.50	0.50	0.52	0.52
PHILIPS ELECTR.	0.53	0.53	0.50	0.51	0.51	0.52
REPSOL	0.54	0.54	0.52	0.52	0.51	0.51
RWE ST	0.51	0.51	0.52	0.52	0.51	0.51
BANCO SAN CENTRAL HISPANO	0.51	0.51	0.51	0.51	0.51	0.51
SANOFI	0.50	0.50	0.49	0.49	0.49	0.49
SAP AG	0.52	0.51	0.52	0.52	0.48	0.48
SAINT GOBAIN	0.51	0.51	0.51	0.51	0.50	0.50
SIEMENS AG	0.51	0.52	0.50	0.50	0.52	0.52
SCHNEIDER ELECTRIC SA	0.52	0.52	0.51	0.51	0.52	0.52
TELEFONICA	0.56	0.56	0.52	0.52	0.51	0.51
UNICREDIT SPA	0.54	0.54	0.50	0.50	0.51	0.51
UNILEVER CERT	0.54	0.54	0.51	0.51	0.52	0.52
VIVENDI UNIVERSAL	0.53	0.53	0.50	0.50	0.49	0.49
VOLKSWAGEN	0.53	0.53	0.50	0.50	0.51	0.50

Table 5.13: The quality of the 4-class prediction: 5-minute prediction AUC and accuracy per stock

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	56	392	7	362	-5	557
AIR LIQUIDE	73	317	-5	214	4	364
ALLIANZ	298	629	22	388	68	379
ASML Holding NV	112	573	26	538	39	547
BASF AG	-1	328	55	506	71	441
BAYER AG	81	501	50	446	19	388
BBVARGENTARIA	86	993	117	868	123	856
BAY MOT WERKE	5	404	28	516	34	452
DANONE	83	522	16	344	7	282
BNP PARIBAS	-46	691	26	823	6	642
CARREFOUR	87	614	59	740	5	655
CRH PLC IRLANDE	982	1141	366	998	264	987
AXA	26	685	-58	618	-8	708
DAIMLER CHRYSLER	87	480	-7	509	-20	588
DEUTSCHE BANK AG	-26	715	-24	696	31	651
VINCI	30	432	22	409	57	494
DEUTSCHE TELEKOM	132	410	27	394	3	421
ESSILOR INTERNATIONAL	50	513	74	518	61	517
ENEL	515	877	-0	625	58	641
ENI	296	471	29	313	19	354
E.ON AG	217	667	29	698	8	570
TOTAL	52	311	8	353	100	389
GENERALI ASSIC	528	951	-64	575	89	677
SOCIETE GENERALE	100	1026	-66	928	35	863
GDF SUEZ	133	604	25	566	-3	425
IBERDROLA I	209	620	84	603	52	561
ING	64	819	96	948	-8	734
INTESABCI	504	1016	-35	916	-17	962
INDITEX	264	661	85	539	86	612
LVMH	152	503	-21	380	41	375
MUNICH RE	183	416	74	373	63	399
LOREAL	221	527	29	462	40	416
PHILIPS ELECTR.	85	472	84	560	71	480
REPSOL	254	701	88	733	44	607
RWE ST	21	661	139	785	108	990
BANCO SAN CENTRAL HISPANO	68	791	77	840	23	760
SANOFI	27	418	-12	369	-32	406
SAP AG	14	342	25	292	-21	312
SAINT GOBAIN	23	532	102	609	47	736
SIEMENS AG	9	338	19	358	65	337
SCHNEIDER ELECTRIC SA	121	618	24	428	64	560
TELEFONICA	314	632	72	517	65	498
UNICREDIT SPA	331	1222	30	1032	29	1113
UNILEVER CERT	31	274	5	193	13	191
VIVENDI UNIVERSAL	171	575	0	406	-17	449
VOLKSWAGEN	115	645	24	616	-4	608

Table 5.14: The quality of the 4-class prediction: The daily gain average and standard deviation for the 5-minute prediction (without trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-42	387	-86	378	-90	576
AIR LIQUIDE	-3	306	-56	238	-69	360
ALLIANZ	117	610	-72	400	-29	389
ASML Holding NV	-29	566	-95	544	-58	537
BASF AG	-64	341	-56	511	-41	438
BAYER AG	-46	511	-57	438	-74	389
BBVARGENTARIA	-65	994	-69	864	-49	842
BAY MOT WERKE	-87	415	-99	526	-72	453
DANONE	-20	506	-69	349	-44	287
BNP PARIBAS	-180	708	-131	838	-120	645
CARREFOUR	-60	612	-104	733	-133	667
CRH PLC IRLANDE	730	1111	146	974	88	962
AXA	-122	695	-206	643	-131	722
DAIMLER CHRYSLER	-27	475	-133	524	-131	596
DEUTSCHE BANK AG	-163	723	-188	713	-99	657
VINCI	-75	440	-82	419	-39	479
DEUTSCHE TELEKOM	33	398	-51	396	-86	439
ESSILOR INTERNATIONAL	-64	521	-48	510	-49	521
ENEL	298	845	-132	639	-70	642
ENI	126	441	-54	330	-58	360
E.ON AG	56	660	-114	712	-110	582
TOTAL	-25	310	-80	366	-1	380
GENERALI ASSIC	291	918	-178	603	-44	678
SOCIETE GENERALE	-80	1015	-222	947	-125	859
GDF SUEZ	-16	593	-101	576	-109	438
IBERDROLA I	57	605	-70	595	-56	545
ING	-94	815	-109	948	-141	738
INTESABCI	262	1000	-210	923	-164	970
INDITEX	113	629	-38	533	-34	613
LVMH	8	482	-100	397	-54	372
MUNICH RE	56	397	4	343	-30	400
LOREAL	81	512	-69	462	-62	420
PHILIPS ELECTR.	-40	482	-78	564	-52	478
REPSOL	86	684	-76	722	-72	605
RWE ST	-119	679	-48	783	-17	988
BANCO SAN CENTRAL HISPANO	-83	785	-102	842	-108	767
SANOFI	-73	429	-106	386	-102	422
SAP AG	-77	352	-54	296	-90	324
SAINT GOBAIN	-106	545	-66	605	-101	725
SIEMENS AG	-75	349	-44	358	-25	331
SCHNEIDER ELECTRIC SA	-5	616	-102	435	-54	560
TELEFONICA	134	608	-58	513	-55	496
UNICREDIT SPA	113	1207	-149	1042	-140	1112
UNILEVER CERT	-38	275	-27	201	-24	198
VIVENDI UNIVERSAL	24	562	-104	423	-109	457
VOLKSWAGEN	-17	641	-95	621	-124	615

Table 5.15: The quality of the 4-class prediction: The daily gain average and standard deviation for the 1-minute prediction (with 0.5 bp trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.50	0.50	0.50	0.50	0.49	0.49
AIR LIQUIDE	0.48	0.48	0.50	0.50	0.51	0.51
ALLIANZ	0.50	0.50	0.51	0.51	0.50	0.50
ASML Holding NV	0.48	0.48	0.51	0.51	0.49	0.49
BASF AG	0.51	0.51	0.50	0.49	0.48	0.48
BAYER AG	0.50	0.50	0.49	0.50	0.50	0.50
BBVARGENTARIA	0.50	0.50	0.51	0.51	0.50	0.51
BAY MOT WERKE	0.51	0.51	0.50	0.50	0.51	0.51
DANONE	0.50	0.50	0.52	0.52	0.51	0.51
BNP PARIBAS	0.48	0.48	0.50	0.50	0.51	0.51
CARREFOUR	0.52	0.52	0.51	0.51	0.50	0.50
CRH PLC IRLANDE	0.50	0.51	0.49	0.49	0.50	0.50
AXA	0.50	0.50	0.48	0.48	0.51	0.51
DAIMLER CHRYSLER	0.52	0.52	0.50	0.50	0.51	0.52
DEUTSCHE BANK AG	0.51	0.51	0.50	0.50	0.49	0.49
VINCI	0.52	0.53	0.50	0.51	0.51	0.52
DEUTSCHE TELEKOM	0.51	0.51	0.51	0.52	0.50	0.51
ESSILOR INTERNATIONAL	0.51	0.51	0.52	0.52	0.51	0.51
ENEL	0.50	0.50	0.49	0.49	0.51	0.51
ENI	0.51	0.51	0.51	0.51	0.50	0.50
E.ON AG	0.48	0.48	0.50	0.50	0.51	0.52
TOTAL	0.52	0.52	0.51	0.51	0.52	0.52
GENERALI ASSIC	0.51	0.51	0.51	0.51	0.51	0.51
SOCIETE GENERALE	0.51	0.51	0.51	0.51	0.50	0.50
GDF SUEZ	0.50	0.50	0.50	0.50	0.51	0.51
IBERDROLA I	0.50	0.50	0.49	0.49	0.50	0.50
ING	0.49	0.49	0.51	0.51	0.49	0.49
INTESABCI	0.51	0.51	0.49	0.49	0.50	0.50
INDITEX	0.50	0.50	0.48	0.48	0.51	0.51
LVMH	0.50	0.50	0.51	0.51	0.49	0.48
MUNICH RE	0.51	0.51	0.50	0.50	0.51	0.51
LOREAL	0.50	0.50	0.51	0.51	0.50	0.50
PHILIPS ELECTR.	0.49	0.49	0.51	0.52	0.50	0.50
REPSOL	0.49	0.49	0.51	0.51	0.50	0.50
RWE ST	0.50	0.50	0.50	0.50	0.50	0.50
BANCO SAN CENTRAL HISPANO	0.50	0.50	0.49	0.49	0.51	0.51
SANOFI	0.50	0.51	0.51	0.51	0.50	0.50
SAP AG	0.51	0.51	0.50	0.50	0.51	0.51
SAINT GOBAIN	0.50	0.50	0.50	0.50	0.49	0.49
SIEMENS AG	0.52	0.53	0.48	0.48	0.50	0.50
SCHNEIDER ELECTRIC SA	0.50	0.50	0.49	0.49	0.50	0.50
TELEFONICA	0.50	0.50	0.50	0.50	0.51	0.51
UNICREDIT SPA	0.51	0.51	0.50	0.50	0.51	0.51
UNILEVER CERT	0.51	0.51	0.50	0.50	0.51	0.51
VIVENDI UNIVERSAL	0.50	0.50	0.50	0.50	0.51	0.51
VOLKSWAGEN	0.51	0.51	0.51	0.51	0.51	0.51

Table 5.16: The quality of the 4-class prediction: 30-minute prediction AUC and accuracy per stock

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-11	887	-6	845	-41	823
AIR LIQUIDE	-57	669	-17	633	-21	624
ALLIANZ	-14	762	69	689	-41	729
ASML Holding NV	-87	862	43	1075	-29	897
BASF AG	-20	807	-3	781	-67	722
BAYER AG	38	759	-93	774	-46	765
BBVARGENTARIA	-16	1263	-63	1138	16	1084
BAY MOT WERKE	-25	783	-23	923	-13	901
DANONE	-61	744	19	726	-18	745
BNP PARIBAS	-28	998	-2	1179	-9	1151
CARREFOUR	4	1108	-135	1082	-52	972
CRH PLC IRLANDE	75	962	-105	1161	-6	1117
AXA	12	1054	6	1055	49	1111
DAIMLER CHRYSLER	75	872	-51	825	9	961
DEUTSCHE BANK AG	54	1054	-89	1152	-35	996
VINCI	110	761	80	742	100	743
DEUTSCHE TELEKOM	27	722	81	700	-14	718
ESSILOR INTERNATIONAL	29	830	43	827	41	872
ENEL	27	991	-40	971	55	959
ENI	7	628	-18	651	-16	645
E.ON AG	-70	911	-4	963	65	826
TOTAL	49	660	108	689	73	669
GENERALI ASSIC	18	1011	2	1094	11	1085
SOCIETE GENERALE	53	1413	67	1253	-5	1335
GDF SUEZ	59	906	-24	847	25	823
IBERDROLA I	3	1017	-73	960	51	949
ING	-21	1138	105	1205	-80	1142
INTESABCI	-128	1359	-54	1329	85	1288
INDITEX	-8	894	-161	912	17	860
LVMH	-36	831	15	725	-26	675
MUNICH RE	29	641	-25	688	-7	727
LOREAL	-19	671	31	755	15	727
PHILIPS ELECTR.	-24	844	24	789	-29	841
REPSOL	-87	878	-5	920	3	925
RWE ST	32	1132	61	1217	46	1140
BANCO SAN CENTRAL HISPANO	2	1150	-60	1072	48	1090
SANOFI	-29	810	25	856	7	794
SAP AG	4	683	-52	709	-15	682
SAINT GOBAIN	-66	996	22	994	-51	945
SIEMENS AG	127	771	-35	802	-59	725
SCHNEIDER ELECTRIC SA	-31	896	-79	837	8	838
TELEFONICA	-12	759	42	918	111	912
UNICREDIT SPA	130	1529	58	1498	81	1357
UNILEVER CERT	5	543	31	546	-26	508
VIVENDI UNIVERSAL	21	874	-15	899	6	859
VOLKSWAGEN	71	929	120	994	75	1055

Table 5.17: The quality of the 4-class prediction: The daily gain average and standard deviation for the 30-minute prediction (without trading costs)

	Order book imbalance		Flow quantity		Past return	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-55	887	-51	845	-84	824
AIR LIQUIDE	-96	672	-61	635	-62	625
ALLIANZ	-57	764	23	687	-80	731
ASML Holding NV	-132	863	-6	1072	-73	896
BASF AG	-61	809	-47	780	-108	724
BAYER AG	-7	758	-136	777	-84	767
BBVARGENTARIA	-58	1265	-108	1137	-25	1082
BAY MOT WERKE	-65	784	-69	923	-53	902
DANONE	-101	743	-25	726	-60	742
BNP PARIBAS	-71	997	-46	1180	-51	1149
CARREFOUR	-39	1110	-182	1085	-94	972
CRH PLC IRLANDE	31	960	-152	1163	-48	1116
AXA	-31	1052	-37	1054	4	1109
DAIMLER CHRYSLER	36	874	-93	825	-31	961
DEUTSCHE BANK AG	9	1053	-138	1151	-77	997
VINCI	72	763	40	742	65	743
DEUTSCHE TELEKOM	-12	722	36	702	-53	720
ESSILOR INTERNATIONAL	-9	830	-1	828	-2	869
ENEL	-17	993	-81	974	17	959
ENI	-36	627	-58	652	-57	642
E.ON AG	-106	911	-45	965	22	824
TOTAL	10	661	66	690	34	666
GENERALI ASSIC	-26	1011	-44	1096	-32	1087
SOCIETE GENERALE	10	1415	19	1252	-51	1336
GDF SUEZ	14	905	-70	847	-16	818
IBERDROLA I	-40	1016	-117	962	5	947
ING	-63	1137	58	1207	-122	1144
INTESABCI	-172	1359	-97	1327	47	1290
INDITEX	-48	896	-204	913	-22	859
LVMH	-82	830	-30	725	-68	675
MUNICH RE	-13	641	-66	691	-49	728
LOREAL	-57	674	-9	754	-22	728
PHILIPS ELECTR.	-65	845	-23	788	-71	839
REPSOL	-128	877	-52	920	-41	920
RWE ST	-7	1130	15	1218	5	1140
BANCO SAN CENTRAL HISPANO	-37	1149	-103	1073	6	1089
SANOFI	-67	810	-21	856	-34	797
SAP AG	-37	683	-100	709	-60	680
SAINT GOBAIN	-105	997	-23	995	-93	946
SIEMENS AG	84	772	-77	805	-98	725
SCHNEIDER ELECTRIC SA	-73	896	-123	836	-34	838
TELEFONICA	-49	760	-4	919	68	913
UNICREDIT SPA	84	1529	15	1499	40	1359
UNILEVER CERT	-39	543	-14	545	-67	509
VIVENDI UNIVERSAL	-24	874	-61	900	-37	856
VOLKSWAGEN	33	929	76	995	38	1058

Table 5.18: The quality of the binary prediction: The daily gain average and standard deviation for the 30-minute prediction (with 0.5 bp trading costs)

Notice that the nans on the tables of the **Appendix 3** correspond to the cases where $|\widehat{Y}|$ is always lower than θ thus no positions are taken.

	1-min horizon		5-min horizon		30-min horizon	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.54	0.54	0.50	0.50	0.50	0.50
AIR LIQUIDE	0.57	0.57	0.52	0.52	0.49	0.49
ALLIANZ	0.61	0.61	0.53	0.53	0.50	0.50
ASML Holding NV	0.55	0.55	0.51	0.51	0.51	0.51
BASF AG	0.54	0.54	0.52	0.52	0.50	0.50
BAYER AG	0.54	0.54	0.51	0.51	0.51	0.51
BBVARGENTARIA	0.54	0.54	0.51	0.51	0.49	0.49
BAY MOT WERKE	0.55	0.55	0.51	0.51	0.49	0.49
DANONE	0.56	0.56	0.51	0.51	0.49	0.49
BNP PARIBAS	0.53	0.53	0.51	0.51	0.50	0.50
CARREFOUR	0.55	0.55	0.51	0.51	0.52	0.52
CRH PLC IRLANDE	0.62	0.62	0.56	0.56	0.52	0.52
AXA	0.55	0.55	0.51	0.51	0.50	0.50
DAIMLER CHRYSLER	0.54	0.54	0.51	0.51	0.50	0.50
DEUTSCHE BANK AG	0.53	0.53	0.51	0.51	0.51	0.51
VINCI	0.55	0.55	0.52	0.52	0.51	0.51
DEUTSCHE TELEKOM	0.56	0.56	0.52	0.52	0.50	0.51
ESSILOR INTERNATIONAL	0.56	0.56	0.51	0.51	0.51	0.51
ENEL	0.62	0.62	0.53	0.53	0.48	0.48
ENI	0.64	0.64	0.54	0.54	0.50	0.50
E.ON AG	0.57	0.57	0.52	0.52	0.48	0.48
TOTAL	0.54	0.54	0.51	0.51	0.50	0.50
GENERALI ASSIC	0.61	0.61	0.54	0.54	0.50	0.50
SOCIETE GENERALE	0.53	0.53	0.50	0.50	0.52	0.52
GDF SUEZ	0.56	0.56	0.51	0.51	0.50	0.50
IBERDROLA I	0.57	0.57	0.52	0.52	0.51	0.51
ING	0.53	0.53	0.51	0.51	0.49	0.49
INTESABCI	0.59	0.59	0.51	0.51	0.50	0.50
INDITEX	0.59	0.59	0.53	0.53	0.52	0.52
LVMH	0.59	0.59	0.52	0.52	0.52	0.52
MUNICH RE	0.58	0.58	0.53	0.53	0.50	0.50
LOREAL	0.60	0.60	0.52	0.52	0.51	0.51
PHILIPS ELECTR.	0.56	0.56	0.51	0.51	0.50	0.50
REPSOL	0.57	0.57	0.52	0.52	0.51	0.51
RWE ST	0.54	0.54	0.51	0.51	0.49	0.49
BANCO SAN CENTRAL HISPANO	0.54	0.54	0.51	0.51	0.49	0.49
SANOFI	0.54	0.54	0.51	0.51	0.49	0.49
SAP AG	0.54	0.54	0.51	0.51	0.51	0.51
SAINT GOBAIN	0.54	0.54	0.51	0.51	0.52	0.52
SIEMENS AG	0.54	0.54	0.51	0.51	0.50	0.50
SCHNEIDER ELECTRIC SA	0.54	0.54	0.52	0.52	0.51	0.51
TELEFONICA	0.59	0.59	0.52	0.52	0.50	0.50
UNICREDIT SPA	0.56	0.56	0.51	0.51	0.49	0.49
UNILEVER CERT	0.56	0.56	0.51	0.51	0.50	0.50
VIVENDI UNIVERSAL	0.57	0.57	0.51	0.51	0.51	0.51
VOLKSWAGEN	0.56	0.56	0.52	0.52	0.51	0.51

Table 5.19: The quality of the OLS prediction: The AUC and the accuracy per stock for the different horizons

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	1410	1151	89	1022	-54	1022
AIR LIQUIDE	1756	1028	237	775	-22	707
ALLIANZ	2832	1332	355	907	-81	935
ASML Holding NV	1693	1237	55	1208	156	1080
BASF AG	1220	1109	143	883	5	877
BAYER AG	1412	1086	129	948	-32	853
BBVARGENTARIA	2297	1759	315	1518	-27	1178
BAY MOT WERKE	1749	1243	124	984	-22	904
DANONE	1729	1045	143	843	-121	791
BNP PARIBAS	1362	1580	263	1386	-21	1246
CARREFOUR	2108	1465	211	1205	69	1242
CRH PLC IRLANDE	4302	1924	1121	1352	94	1239
AXA	2139	1450	239	1334	-34	1101
DAIMLER CHRYSLER	1380	1325	139	1006	-6	1139
DEUTSCHE BANK AG	1431	1493	118	1302	105	1106
VINCI	1803	1192	340	950	31	736
DEUTSCHE TELEKOM	1780	1380	218	858	25	784
ESSILOR INTERNATIONAL	1934	1244	299	1041	-34	910
ENEL	3632	1526	298	1115	-35	984
ENI	3095	1170	369	887	8	742
E.ON AG	2119	1412	182	1247	-126	976
TOTAL	1336	1054	220	852	-20	780
GENERALI ASSIC	3937	1763	537	1260	37	996
SOCIETE GENERALE	1499	1787	98	1627	155	1448
GDF SUEZ	2115	1279	175	1084	-72	964
IBERDROLA I	2499	1587	450	1123	83	1046
ING	1358	1477	135	1351	38	1159
INTESABCI	3829	1878	152	1458	-41	1482
INDITEX	2729	1515	486	1043	77	957
LVMH	2552	1236	203	901	133	870
MUNICH RE	2019	1171	355	795	-21	812
LOREAL	2447	1107	196	966	55	828
PHILIPS ELECTR.	2152	1174	264	952	-58	943
REPSOL	2952	1678	426	1240	117	989
RWE ST	1729	1559	297	1571	-10	1158
BANCO SAN CENTRAL HISPANO	1754	1675	187	1218	-46	1091
SANOFI	1258	1045	93	978	-36	982
SAP AG	1351	1096	77	863	59	793
SAINT GOBAIN	1800	1414	154	1053	73	1021
SIEMENS AG	1192	1019	83	866	-45	810
SCHNEIDER ELECTRIC SA	1668	1297	310	1118	92	937
TELEFONICA	2768	1317	269	934	27	970
UNICREDIT SPA	2924	2062	332	1643	-256	1459
UNILEVER CERT	1385	878	89	654	-24	632
VIVENDI UNIVERSAL	2259	1363	129	1073	65	1046
VOLKSWAGEN	2052	1316	168	1085	64	1041

Table 5.20: The quality of the OLS prediction: The daily gain average and standard deviation for the different horizons (without trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-410	1115	-247	1027	-110	1020
AIR LIQUIDE	-45	971	-118	766	-78	705
ALLIANZ	988	1166	5	898	-137	933
ASML Holding NV	-165	1177	-292	1205	100	1078
BASF AG	-581	1102	-209	877	-48	877
BAYER AG	-414	1042	-220	946	-88	852
BBVARGENTARIA	546	1652	-34	1508	-85	1177
BAY MOT WERKE	-110	1150	-229	978	-75	904
DANONE	-163	965	-212	835	-179	790
BNP PARIBAS	-410	1505	-81	1383	-78	1245
CARREFOUR	236	1351	-139	1206	14	1240
CRH PLC IRLANDE	2815	1765	775	1328	40	1239
AXA	273	1347	-109	1334	-91	1103
DAIMLER CHRYSLER	-399	1279	-207	1004	-63	1138
DEUTSCHE BANK AG	-344	1433	-224	1302	48	1106
VINCI	-14	1113	-1	939	-24	736
DEUTSCHE TELEKOM	-7	1287	-126	848	-30	782
ESSILOR INTERNATIONAL	80	1126	-44	1014	-90	909
ENEL	1745	1340	-56	1099	-88	983
ENI	1244	952	18	881	-45	740
E.ON AG	278	1310	-169	1243	-180	974
TOTAL	-474	998	-127	848	-78	778
GENERALI ASSIC	2094	1570	182	1248	-12	995
SOCIETE GENERALE	-324	1712	-257	1615	99	1448
GDF SUEZ	259	1185	-173	1071	-129	963
IBERDROLA I	714	1462	118	1107	27	1046
ING	-414	1425	-219	1339	-20	1158
INTESABCI	1936	1710	-210	1456	-100	1480
INDITEX	968	1429	143	1036	24	960
LVMH	692	1069	-145	890	79	870
MUNICH RE	202	1076	1	786	-78	812
LOREAL	581	966	-150	958	0	826
PHILIPS ELECTR.	288	1074	-91	946	-115	943
REPSOL	1139	1532	82	1223	62	987
RWE ST	-86	1509	-48	1567	-68	1158
BANCO SAN CENTRAL HISPANO	-40	1585	-159	1205	-101	1090
SANOFI	-560	1021	-253	978	-91	982
SAP AG	-456	1008	-276	861	3	793
SAINT GOBAIN	-65	1339	-194	1051	16	1019
SIEMENS AG	-595	1003	-259	866	-98	810
SCHNEIDER ELECTRIC SA	-193	1213	-42	1108	38	934
TELEFONICA	952	1172	-75	919	-31	969
UNICREDIT SPA	1083	1958	-32	1639	-314	1459
UNILEVER CERT	-491	795	-267	654	-81	634
VIVENDI UNIVERSAL	409	1277	-220	1073	9	1043
VOLKSWAGEN	239	1219	-186	1076	7	1039

Table 5.21: The quality of the OLS prediction: The daily gain average and standard deviation for the different horizons (with 0.5 bp trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.54	0.54	0.50	0.50	0.50	0.50
AIR LIQUIDE	0.57	0.57	0.52	0.52	0.50	0.50
ALLIANZ	0.61	0.61	0.53	0.53	0.49	0.49
ASML Holding NV	0.55	0.55	0.51	0.51	0.51	0.51
BASF AG	0.54	0.54	0.52	0.52	0.50	0.50
BAYER AG	0.54	0.54	0.51	0.51	0.50	0.50
BBVARGENTARIA	0.54	0.54	0.51	0.51	0.50	0.50
BAY MOT WERKE	0.55	0.55	0.51	0.51	0.50	0.50
DANONE	0.56	0.56	0.51	0.51	0.50	0.50
BNP PARIBAS	0.53	0.53	0.51	0.51	0.50	0.50
CARREFOUR	0.55	0.55	0.51	0.51	0.52	0.52
CRH PLC IRLANDE	0.62	0.62	0.56	0.56	0.52	0.52
AXA	0.56	0.55	0.51	0.51	0.50	0.50
DAIMLER CHRYSLER	0.54	0.54	0.51	0.51	0.50	0.50
DEUTSCHE BANK AG	0.53	0.53	0.51	0.51	0.51	0.51
VINCI	0.55	0.55	0.51	0.52	0.51	0.51
DEUTSCHE TELEKOM	0.56	0.56	0.52	0.52	0.51	0.52
ESSILOR INTERNATIONAL	0.56	0.56	0.51	0.51	0.51	0.51
ENEL	0.62	0.62	0.53	0.53	0.48	0.48
ENI	0.65	0.65	0.54	0.54	0.50	0.50
E.ON AG	0.57	0.57	0.52	0.52	0.48	0.48
TOTAL	0.54	0.54	0.51	0.51	0.50	0.50
GENERALI ASSIC	0.62	0.62	0.54	0.54	0.51	0.51
SOCIETE GENERALE	0.53	0.53	0.50	0.50	0.53	0.52
GDF SUEZ	0.57	0.57	0.52	0.52	0.50	0.50
IBERDROLA I	0.57	0.57	0.53	0.53	0.52	0.52
ING	0.53	0.53	0.51	0.50	0.50	0.50
INTESABCI	0.60	0.60	0.52	0.52	0.50	0.50
INDITEX	0.59	0.59	0.53	0.53	0.52	0.52
LVMH	0.59	0.59	0.52	0.52	0.50	0.50
MUNICH RE	0.59	0.59	0.53	0.53	0.50	0.50
LOREAL	0.60	0.60	0.52	0.52	0.51	0.51
PHILIPS ELECTR.	0.56	0.56	0.51	0.51	0.49	0.49
REPSOL	0.58	0.58	0.52	0.52	0.52	0.52
RWE ST	0.54	0.54	0.51	0.51	0.50	0.50
BANCO SAN CENTRAL HISPANO	0.54	0.54	0.51	0.51	0.50	0.50
SANOFI	0.54	0.54	0.51	0.51	0.51	0.51
SAP AG	0.55	0.55	0.51	0.51	0.51	0.51
SAINT GOBAIN	0.54	0.54	0.51	0.51	0.52	0.52
SIEMENS AG	0.54	0.54	0.51	0.51	0.51	0.51
SCHNEIDER ELECTRIC SA	0.55	0.55	0.52	0.52	0.50	0.50
TELEFONICA	0.59	0.59	0.52	0.52	0.51	0.51
UNICREDIT SPA	0.57	0.57	0.51	0.51	0.49	0.49
UNILEVER CERT	0.56	0.56	0.51	0.51	0.49	0.49
VIVENDI UNIVERSAL	0.57	0.57	0.51	0.51	0.51	0.51
VOLKSWAGEN	0.57	0.57	0.52	0.52	0.51	0.51

Table 5.22: The quality of the Ridge HKB prediction: The AUC and the accuracy per stock for the different horizons

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	1476	1155	108	973	-22	979
AIR LIQUIDE	1793	976	264	805	6	703
ALLIANZ	2884	1343	399	958	-96	871
ASML Holding NV	1817	1233	141	1155	189	1037
BASF AG	1244	1138	218	942	-3	899
BAYER AG	1475	1096	125	921	-24	872
BBVARGENTARIA	2429	1763	288	1321	-63	1206
BAY MOT WERKE	1784	1208	143	1006	29	978
DANONE	1762	1032	151	794	-36	840
BNP PARIBAS	1548	1554	258	1403	18	1201
CARREFOUR	2159	1485	239	1285	62	1277
CRH PLC IRLANDE	4325	1976	1181	1378	105	1185
AXA	2293	1471	254	1292	-48	1131
DAIMLER CHRYSLER	1439	1354	94	1052	20	1083
DEUTSCHE BANK AG	1469	1497	127	1305	43	1055
VINCI	1903	1289	313	1002	72	846
DEUTSCHE TELEKOM	1826	1420	203	846	70	767
ESSILOR INTERNATIONAL	2002	1268	238	1010	50	915
ENEL	3733	1545	330	1158	-60	996
ENI	3158	1196	413	852	-14	734
E.ON AG	2253	1392	249	1222	-112	1016
TOTAL	1341	1024	237	855	-25	767
GENERALI ASSIC	4025	1839	576	1257	1	1020
SOCIETE GENERALE	1521	1793	131	1617	202	1504
GDF SUEZ	2206	1290	222	1048	-59	921
IBERDROLA I	2532	1573	466	1161	119	1023
ING	1487	1473	174	1298	77	1177
INTESABCI	3982	1882	280	1463	-82	1498
INDITEX	2816	1492	566	1042	114	958
LVMH	2606	1270	269	870	28	847
MUNICH RE	2119	1157	407	752	10	739
LOREAL	2549	1127	220	965	42	771
PHILIPS ELECTR.	2176	1180	293	926	-70	939
REPSOL	3016	1648	502	1227	207	946
RWE ST	1812	1551	336	1486	91	1146
BANCO SAN CENTRAL HISPANO	1829	1603	221	1198	19	1125
SANOFI	1358	1014	138	939	3	933
SAP AG	1388	1149	67	885	60	781
SAINT GOBAIN	1899	1433	242	954	87	979
SIEMENS AG	1281	1081	118	904	-55	951
SCHNEIDER ELECTRIC SA	1744	1315	387	1166	-16	902
TELEFONICA	2835	1336	336	979	93	928
UNICREDIT SPA	3060	2056	317	1605	-260	1482
UNILEVER CERT	1459	863	107	669	-63	637
VIVENDI UNIVERSAL	2311	1337	201	1022	30	956
VOLKSWAGEN	2171	1341	227	1205	21	1085

Table 5.23: The quality of the Ridge HKB prediction: The daily gain average and standard deviation for the different horizons (without trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-334	1110	-221	973	-76	976
AIR LIQUIDE	-12	911	-90	792	-47	701
ALLIANZ	1042	1167	60	945	-150	870
ASML Holding NV	-36	1164	-191	1148	135	1036
BASF AG	-552	1119	-126	935	-56	899
BAYER AG	-338	1047	-214	916	-77	870
BBVARGENTARIA	683	1653	-49	1302	-119	1206
BAY MOT WERKE	-71	1104	-201	994	-21	976
DANONE	-128	951	-194	787	-90	840
BNP PARIBAS	-217	1463	-73	1396	-39	1199
CARREFOUR	290	1373	-102	1286	9	1274
CRH PLC IRLANDE	2851	1822	837	1350	51	1186
AXA	420	1361	-82	1288	-103	1130
DAIMLER CHRYSLER	-345	1302	-239	1052	-35	1080
DEUTSCHE BANK AG	-279	1422	-205	1304	-13	1057
VINCI	80	1213	-18	993	18	846
DEUTSCHE TELEKOM	45	1335	-129	835	17	765
ESSILOR INTERNATIONAL	142	1152	-92	986	-4	913
ENEL	1841	1355	-14	1145	-110	996
ENI	1306	964	71	846	-65	733
E.ON AG	406	1284	-89	1214	-164	1014
TOTAL	-468	967	-101	850	-79	766
GENERALI ASSIC	2174	1648	230	1245	-47	1020
SOCIETE GENERALE	-286	1711	-206	1606	149	1503
GDF SUEZ	337	1199	-114	1035	-112	919
IBERDROLA I	753	1442	147	1146	66	1021
ING	-282	1403	-169	1285	21	1175
INTESABCI	2087	1698	-78	1450	-138	1497
INDITEX	1056	1399	227	1034	65	959
LVMH	736	1110	-66	855	-24	847
MUNICH RE	299	1049	60	743	-43	738
LOREAL	678	980	-117	953	-10	767
PHILIPS ELECTR.	308	1074	-54	920	-125	939
REPSOL	1210	1499	168	1214	155	945
RWE ST	-6	1492	-2	1480	38	1144
BANCO SAN CENTRAL HISPANO	48	1500	-110	1181	-36	1124
SANOFI	-455	995	-194	936	-48	932
SAP AG	-421	1055	-278	875	6	781
SAINT GOBAIN	27	1339	-95	941	33	977
SIEMENS AG	-510	1051	-210	906	-106	951
SCHNEIDER ELECTRIC SA	-112	1214	46	1149	-69	901
TELEFONICA	1015	1176	-1	961	39	925
UNICREDIT SPA	1222	1939	-40	1596	-317	1482
UNILEVER CERT	-409	772	-237	662	-118	638
VIVENDI UNIVERSAL	459	1231	-139	1012	-23	952
VOLKSWAGEN	366	1240	-120	1198	-35	1084

Table 5.24: The quality of the Ridge HKB prediction: The daily gain average and standard deviation for the different horizons (with 0.5 bp trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.55	0.55	0.52	0.52	0.50	0.50
AIR LIQUIDE	0.57	0.57	0.53	0.53	0.49	0.49
ALLIANZ	0.61	0.61	0.54	0.54	0.50	0.50
ASML Holding NV	0.56	0.56	0.52	0.52	0.52	0.52
BASF AG	0.54	0.54	0.52	0.52	0.50	0.50
BAYER AG	0.55	0.55	0.51	0.51	0.50	0.50
BBVARGENTARIA	0.54	0.54	0.51	0.51	0.50	0.50
BAY MOT WERKE	0.55	0.55	0.51	0.51	0.50	0.50
DANONE	0.56	0.56	0.51	0.51	0.49	0.49
BNP PARIBAS	0.54	0.54	0.52	0.52	0.50	0.50
CARREFOUR	0.55	0.55	0.51	0.51	0.51	0.51
CRH PLC IRLANDE	0.62	0.62	0.57	0.57	0.51	0.51
AXA	0.56	0.56	0.51	0.51	0.51	0.51
DAIMLER CHRYSLER	0.54	0.54	0.52	0.52	0.51	0.51
DEUTSCHE BANK AG	0.53	0.53	0.51	0.51	0.52	0.52
VINCI	0.56	0.56	0.52	0.53	0.51	0.52
DEUTSCHE TELEKOM	0.57	0.57	0.52	0.52	0.52	0.52
ESSILOR INTERNATIONAL	0.56	0.56	0.51	0.51	0.50	0.50
ENEL	0.63	0.63	0.54	0.54	0.50	0.50
ENI	0.65	0.65	0.55	0.55	0.50	0.50
E.ON AG	0.58	0.58	0.52	0.52	0.50	0.51
TOTAL	0.54	0.54	0.52	0.52	0.51	0.51
GENERALI ASSIC	0.62	0.62	0.55	0.55	0.49	0.49
SOCIETE GENERALE	0.53	0.53	0.50	0.50	0.52	0.52
GDF SUEZ	0.57	0.57	0.52	0.52	0.50	0.50
IBERDROLA I	0.57	0.57	0.53	0.53	0.52	0.52
ING	0.53	0.53	0.51	0.51	0.50	0.50
INTESABCI	0.60	0.60	0.53	0.53	0.48	0.48
INDITEX	0.60	0.60	0.54	0.54	0.51	0.51
LVMH	0.59	0.59	0.52	0.52	0.50	0.50
MUNICH RE	0.59	0.59	0.54	0.54	0.50	0.50
LOREAL	0.60	0.60	0.53	0.53	0.51	0.51
PHILIPS ELECTR.	0.57	0.57	0.52	0.52	0.50	0.50
REPSOL	0.58	0.58	0.53	0.53	0.51	0.51
RWE ST	0.55	0.55	0.51	0.51	0.49	0.49
BANCO SAN CENTRAL HISPANO	0.54	0.54	0.52	0.52	0.51	0.51
SANOFI	0.55	0.55	0.51	0.51	0.50	0.50
SAP AG	0.55	0.55	0.51	0.51	0.51	0.51
SAINT GOBAIN	0.55	0.55	0.51	0.51	0.52	0.52
SIEMENS AG	0.55	0.55	0.52	0.52	0.51	0.52
SCHNEIDER ELECTRIC SA	0.55	0.55	0.52	0.52	0.50	0.50
TELEFONICA	0.60	0.60	0.53	0.53	0.51	0.51
UNICREDIT SPA	0.57	0.57	0.52	0.52	0.49	0.49
UNILEVER CERT	0.57	0.57	0.51	0.51	0.51	0.51
VIVENDI UNIVERSAL	0.58	0.58	0.52	0.52	0.51	0.51
VOLKSWAGEN	0.57	0.57	0.52	0.52	0.50	0.50

Table 5.25: The quality of the Ridge LW prediction: The AUC and the accuracy per stock for the different horizons

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	1651	1145	288	882	-49	898
AIR LIQUIDE	1848	1048	337	809	-21	713
ALLIANZ	2925	1362	382	987	21	802
ASML Holding NV	1963	1221	253	1023	194	1096
BASF AG	1401	1177	177	1022	-58	847
BAYER AG	1621	1109	157	869	13	804
BBVARGENTARIA	2488	1790	324	1511	-87	1149
BAY MOT WERKE	1853	1283	151	939	-46	912
DANONE	1753	1035	152	852	-123	790
BNP PARIBAS	1544	1683	242	1417	11	1311
CARREFOUR	2334	1468	244	1331	19	1254
CRH PLC IRLANDE	4428	1976	1387	1378	39	1074
AXA	2356	1448	181	1245	21	1238
DAIMLER CHRYSLER	1614	1495	198	1013	13	1113
DEUTSCHE BANK AG	1482	1556	247	1373	139	1170
VINCI	1958	1301	386	1143	108	851
DEUTSCHE TELEKOM	1916	1380	272	894	60	836
ESSILOR INTERNATIONAL	2118	1243	306	1075	-22	912
ENEL	3826	1584	520	1219	4	1009
ENI	3230	1264	479	840	-66	770
E.ON AG	2277	1255	265	1142	-55	1023
TOTAL	1428	1079	224	798	75	787
GENERALI ASSIC	4044	1813	686	1323	-94	1054
SOCIETE GENERALE	1487	1932	93	1550	118	1412
GDF SUEZ	2307	1296	282	976	-32	1004
IBERDROLA I	2721	1542	522	1194	107	969
ING	1565	1569	127	1383	11	1251
INTESABCI	4060	1876	504	1504	-218	1541
INDITEX	2928	1509	640	1119	52	937
LVMH	2700	1264	363	894	-39	842
MUNICH RE	2225	1220	407	857	7	829
LOREAL	2606	1100	354	919	-7	827
PHILIPS ELECTR.	2284	1254	320	962	4	889
REPSOL	3053	1694	537	1173	67	968
RWE ST	1988	1641	258	1637	27	1180
BANCO SAN CENTRAL HISPANO	1981	1535	415	1326	83	1282
SANOFI	1466	1047	101	967	15	926
SAP AG	1522	1207	89	841	-13	809
SAINT GOBAIN	2060	1473	194	1088	64	979
SIEMENS AG	1425	1132	192	790	13	993
SCHNEIDER ELECTRIC SA	1841	1359	281	1093	-26	907
TELEFONICA	2859	1269	458	1000	111	932
UNICREDIT SPA	3178	2159	411	1603	-98	1563
UNILEVER CERT	1539	841	158	681	-0	630
VIVENDI UNIVERSAL	2486	1370	295	999	-4	1021
VOLKSWAGEN	2280	1432	288	1047	-90	992

Table 5.26: The quality of the Ridge LW prediction: The daily gain average and standard deviation for the different horizons (without trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-169	1072	-30	866	-98	895
AIR LIQUIDE	59	961	8	790	-69	709
ALLIANZ	1091	1178	49	978	-28	800
ASML Holding NV	100	1133	-66	1010	145	1093
BASF AG	-358	1098	-140	1011	-108	844
BAYER AG	-182	1069	-154	853	-33	802
BBVARGENTARIA	706	1682	9	1493	-132	1148
BAY MOT WERKE	-1	1172	-167	923	-92	911
DANONE	-153	956	-167	840	-170	789
BNP PARIBAS	-205	1577	-69	1400	-37	1309
CARREFOUR	459	1334	-70	1325	-27	1252
CRH PLC IRLANDE	2967	1810	1053	1341	-5	1073
AXA	451	1325	-126	1239	-27	1235
DAIMLER CHRYSLER	-136	1410	-113	1014	-34	1112
DEUTSCHE BANK AG	-267	1454	-57	1362	89	1169
VINCI	96	1218	73	1128	63	848
DEUTSCHE TELEKOM	140	1294	-34	872	16	833
ESSILOR INTERNATIONAL	223	1106	-10	1059	-71	909
ENEL	1925	1384	194	1200	-41	1003
ENI	1380	1036	152	824	-109	768
E.ON AG	430	1137	-46	1127	-99	1022
TOTAL	-375	1027	-82	787	28	785
GENERALI ASSIC	2174	1611	349	1292	-140	1052
SOCIETE GENERALE	-298	1845	-213	1537	70	1408
GDF SUEZ	421	1175	-31	957	-78	1000
IBERDROLA I	925	1393	224	1178	61	968
ING	-201	1468	-193	1369	-38	1248
INTESABCI	2152	1690	157	1483	-265	1541
INDITEX	1138	1403	312	1105	6	938
LVMH	811	1108	49	871	-86	840
MUNICH RE	399	1097	75	849	-37	829
LOREAL	717	926	28	902	-52	825
PHILIPS ELECTR.	391	1145	-10	951	-45	885
REPSOL	1245	1557	218	1152	19	968
RWE ST	188	1554	-45	1634	-19	1177
BANCO SAN CENTRAL HISPANO	198	1429	95	1300	36	1279
SANOFI	-336	1005	-197	954	-31	923
SAP AG	-281	1103	-234	836	-62	810
SAINT GOBAIN	137	1373	-111	1077	14	975
SIEMENS AG	-382	1087	-112	789	-33	992
SCHNEIDER ELECTRIC SA	-9	1250	-24	1076	-74	903
TELEFONICA	1021	1123	128	983	64	928
UNICREDIT SPA	1361	2027	82	1587	-150	1561
UNILEVER CERT	-340	751	-166	666	-49	628
VIVENDI UNIVERSAL	593	1245	-18	978	-54	1021
VOLKSWAGEN	462	1328	-36	1039	-136	991

Table 5.27: The quality of the Ridge LW prediction: The daily gain average and standard deviation for the different horizons (with 0.5 bp trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
INTERBREW	0.54	0.54	0.51	0.51	0.50	0.50
AIR LIQUIDE	0.58	0.58	0.52	0.52	0.49	0.49
ALLIANZ	0.61	0.61	0.54	0.54	0.52	0.52
ASML Holding NV	0.56	0.56	0.52	0.52	0.51	0.51
BASF AG	0.53	0.53	0.51	0.51	0.51	0.51
BAYER AG	0.54	0.54	0.51	0.51	0.50	0.50
BBVARGENTARIA	0.54	0.54	0.51	0.51	0.49	0.49
BAY MOT WERKE	0.55	0.55	0.51	0.51	0.49	0.49
DANONE	0.56	0.56	0.51	0.51	0.50	0.50
BNP PARIBAS	0.54	0.54	0.51	0.51	0.49	0.49
CARREFOUR	0.55	0.55	0.51	0.51	0.50	0.50
CRH PLC IRLANDE	0.62	0.62	0.56	0.56	0.52	0.52
AXA	0.55	0.55	0.51	0.51	0.49	0.49
DAIMLER CHRYSLER	0.53	0.53	0.52	0.52	0.50	0.50
DEUTSCHE BANK AG	0.53	0.53	0.51	0.51	0.51	0.51
VINCI	0.55	0.55	0.52	0.53	0.52	0.52
DEUTSCHE TELEKOM	0.58	0.58	0.52	0.52	0.52	0.52
ESSILOR INTERNATIONAL	0.56	0.56	0.51	0.51	0.50	0.50
ENEL	0.62	0.62	0.53	0.53	0.50	0.50
ENI	0.64	0.64	0.55	0.55	0.49	0.49
E.ON AG	0.57	0.57	0.52	0.52	0.49	0.50
TOTAL	0.54	0.54	0.52	0.52	0.51	0.51
GENERALI ASSIC	0.62	0.62	0.54	0.54	0.51	0.51
SOCIETE GENERALE	0.53	0.53	0.50	0.50	0.52	0.52
GDF SUEZ	0.56	0.56	0.52	0.52	0.51	0.51
IBERDROLA I	0.56	0.56	0.53	0.53	0.53	0.53
ING	0.52	0.52	0.51	0.51	0.50	0.50
INTESABCI	0.60	0.60	0.53	0.53	0.50	0.50
INDITEX	0.59	0.59	0.53	0.53	0.52	0.52
LVMH	0.59	0.59	0.52	0.52	0.51	0.51
MUNICH RE	0.58	0.58	0.54	0.54	0.50	0.50
LOREAL	0.60	0.60	0.53	0.53	0.50	0.50
PHILIPS ELECTR.	0.56	0.56	0.52	0.52	0.50	0.50
REPSOL	0.57	0.57	0.52	0.52	0.51	0.51
RWE ST	0.54	0.54	0.51	0.51	0.50	0.50
BANCO SAN CENTRAL HISPANO	0.54	0.54	0.52	0.52	0.50	0.50
SANOFI	0.54	0.54	0.51	0.51	0.50	0.50
SAP AG	0.53	0.53	0.52	0.52	0.50	0.50
SAINT GOBAIN	0.54	0.54	0.51	0.51	0.52	0.52
SIEMENS AG	0.54	0.54	0.51	0.51	0.50	0.50
SCHNEIDER ELECTRIC SA	0.54	0.54	0.51	0.51	0.49	0.49
TELEFONICA	0.59	0.59	0.53	0.53	0.51	0.51
UNICREDIT SPA	0.57	0.57	0.52	0.52	0.48	0.48
UNILEVER CERT	0.57	0.57	0.51	0.51	0.51	0.51
VIVENDI UNIVERSAL	0.57	0.57	0.52	0.52	0.52	0.52
VOLKSWAGEN	0.56	0.56	0.52	0.52	0.49	0.49

Table 5.28: The quality of the LASSO prediction: The AUC and the accuracy per stock for the different horizons

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	621	1094	246	916	-9	991
AIR LIQUIDE	952	1241	294	822	21	665
ALLIANZ	2758	1232	368	959	13	846
ASML Holding NV	1509	1514	232	1007	100	1003
BASF AG	294	771	74	979	-1	870
BAYER AG	563	1116	157	864	-61	829
BBVARGENTARIA	1780	1847	397	1307	-74	1107
BAY MOT WERKE	1158	1429	127	989	-67	921
DANONE	926	1167	132	942	-51	763
BNP PARIBAS	867	1664	199	1462	-21	1171
CARREFOUR	1738	1663	110	1299	63	1189
CRH PLC IRLANDE	4301	1951	1293	1421	85	1142
AXA	1861	1610	98	1261	-37	1217
DAIMLER CHRYSLER	610	1399	199	1045	36	999
DEUTSCHE BANK AG	657	1311	190	1288	112	1125
VINCI	693	1172	377	1088	114	860
DEUTSCHE TELEKOM	943	1448	176	922	72	790
ESSILOR INTERNATIONAL	1419	1477	223	1043	35	969
ENEL	3631	1654	365	1198	14	1105
ENI	3010	1190	491	867	-78	787
E.ON AG	2009	1316	327	1129	-91	1017
TOTAL	304	778	234	767	72	727
GENERALI ASSIC	3923	1787	642	1283	65	1201
SOCIETE GENERALE	783	1859	92	1515	78	1478
GDF SUEZ	1821	1340	280	1046	34	1001
IBERDROLA I	2340	1640	495	1170	139	1002
ING	819	1545	129	1427	23	1189
INTESABCI	3966	1850	488	1435	-151	1525
INDITEX	2359	1670	559	1160	122	935
LVMH	2490	1255	321	925	71	836
MUNICH RE	1657	1341	421	853	37	792
LOREAL	2320	1089	326	925	6	857
PHILIPS ELECTR.	1640	1298	304	1008	-8	941
REPSOL	2770	1671	489	1225	83	1000
RWE ST	989	1515	161	1335	77	1158
BANCO SAN CENTRAL HISPANO	1229	1658	368	1269	-46	1099
SANOFI	513	960	206	942	-40	900
SAP AG	313	801	130	831	25	809
SAINT GOBAIN	1059	1544	195	1153	37	973
SIEMENS AG	334	941	99	733	-101	1010
SCHNEIDER ELECTRIC SA	674	1201	222	1051	21	930
TELEFONICA	2647	1293	386	1031	109	893
UNICREDIT SPA	2859	2055	331	1540	-243	1499
UNILEVER CERT	344	713	126	724	2	677
VIVENDI UNIVERSAL	1991	1341	284	1015	75	1048
VOLKSWAGEN	1709	1480	232	1158	-32	953

Table 5.29: The quality of the LASSO prediction: The daily gain average and standard deviation for the different horizons (without trading costs)

	1-min horizon		5-min horizon		30-min horizon	
Stock	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$	$\overline{Gain}$	$\sigma(Gain)$
INTERBREW	-99	839	-48	905	-60	988
AIR LIQUIDE	181	857	3	792	-30	662
ALLIANZ	1136	1028	55	942	-38	843
ASML Holding NV	354	1100	-57	992	48	1002
BASF AG	-114	614	-215	979	-55	869
BAYER AG	-65	806	-135	849	-110	829
BBVARGENTARIA	368	1553	90	1282	-128	1105
BAY MOT WERKE	73	1012	-158	973	-117	920
DANONE	27	809	-153	933	-102	762
BNP PARIBAS	-32	1412	-100	1443	-74	1169
CARREFOUR	322	1377	-180	1297	13	1188
CRH PLC IRLANDE	2965	1760	967	1381	34	1142
AXA	371	1367	-201	1256	-91	1217
DAIMLER CHRYSLER	-291	1277	-95	1038	-17	997
DEUTSCHE BANK AG	-240	1166	-83	1276	58	1125
VINCI	25	792	88	1066	62	857
DEUTSCHE TELEKOM	292	1011	-108	891	22	788
ESSILOR INTERNATIONAL	94	1257	-69	1029	-16	966
ENEL	2051	1450	52	1182	-37	1102
ENI	1507	965	169	851	-127	786
E.ON AG	429	1172	12	1112	-144	1015
TOTAL	-77	548	-32	745	20	724
GENERALI ASSIC	2278	1562	316	1250	15	1200
SOCIETE GENERALE	-197	1650	-198	1503	25	1476
GDF SUEZ	336	1124	-18	1033	-17	997
IBERDROLA I	856	1425	216	1147	88	1001
ING	-165	1346	-175	1408	-31	1186
INTESABCI	2267	1662	153	1407	-206	1523
INDITEX	1030	1435	246	1148	72	935
LVMH	845	1120	36	900	22	833
MUNICH RE	312	1119	107	838	-13	790
LOREAL	751	977	20	917	-47	855
PHILIPS ELECTR.	186	1079	-7	999	-60	938
REPSOL	1198	1502	192	1195	31	1001
RWE ST	-96	1313	-133	1344	26	1156
BANCO SAN CENTRAL HISPANO	87	1351	62	1244	-95	1098
SANOFI	-63	620	-77	932	-88	898
SAP AG	-163	672	-178	822	-28	808
SAINT GOBAIN	27	1234	-90	1135	-17	969
SIEMENS AG	-135	839	-175	735	-150	1011
SCHNEIDER ELECTRIC SA	-122	957	-71	1029	-30	926
TELEFONICA	1073	1163	71	1023	57	889
UNICREDIT SPA	1306	1911	13	1527	-298	1499
UNILEVER CERT	-29	328	-148	714	-50	675
VIVENDI UNIVERSAL	459	1143	6	997	20	1046
VOLKSWAGEN	294	1265	-83	1147	-84	952

Table 5.30: The quality of the LASSO prediction: The daily gain average and standard deviation for the different horizons (with 0.5 bp trading costs)

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.91	0.57	45243	4115	0.41	0.04	2.55
ALV	0.96	0.37	59189	-9247	0.93	-0.15	4.53
BAS	0.92	0.55	68118	4351	0.37	0.02	1.56
BAYN	0.92	0.56	64548	-2556	0.48	-0.02	3.67
BEI	0.91	0.59	22441	2814	0.48	0.06	6.19
BMW	0.90	0.58	63962	8453	0.39	0.05	1.80
CBK	0.93	0.38	82099	-16063	0.93	-0.18	3.18
CON	0.94	0.39	54048	-8604	0.69	-0.11	3.58
DAI	0.92	0.51	77912	-169	0.41	-0.00	1.47
DB1	0.91	0.55	33872	2308	0.56	0.04	4.80
DBK	0.93	0.48	89802	-4239	0.43	-0.02	1.36
DPW	0.93	0.47	53527	-2957	0.49	-0.03	2.55
DTE	0.95	0.35	52334	-9908	0.94	-0.18	5.16
EOAN	0.95	0.37	58718	-9605	0.83	-0.14	4.00
FME	0.89	0.62	29628	5153	0.45	0.08	4.35
FRE	0.94	0.36	30474	-5873	1.03	-0.20	9.76
HEI	0.90	0.59	41610	5866	0.56	0.08	4.00
HEN3	0.90	0.58	36881	5164	0.42	0.06	3.27
IFX	0.89	0.64	45874	11209	0.48	0.12	3.00
LHA	0.92	0.46	56646	-4957	0.71	-0.06	3.65
LIN	0.95	0.36	34460	-6286	0.76	-0.14	6.21
LXS	0.88	0.59	37615	4907	0.66	0.09	5.36
MRK	0.94	0.35	28130	-5582	0.93	-0.18	9.59
MUV2	0.95	0.36	39933	-6960	0.71	-0.12	5.38
RWE	0.91	0.56	58488	5397	0.55	0.05	2.64
SAP	0.95	0.41	53789	-6408	0.43	-0.05	2.26
SDF	0.88	0.57	42052	4152	0.73	0.07	5.03
SIE	0.92	0.56	62315	5446	0.33	0.03	1.54
TKA	0.92	0.49	43915	-1783	0.67	-0.03	4.35
VOW3	0.95	0.39	62122	-8477	0.62	-0.08	2.91
Average	0.92	0.48	50991	-1344	0.61	-0.03	3.99
Min	0.88	0.35	22441	-16063	0.33	-0.2	1.36
Max	0.96	0.64	89802	11209	1.03	0.12	9.76

Table 5.31: In sample and out of sample results for the strategy with 10 events learning window.

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.65	0.53	17812	3626	0.34	0.07	5.78
ALV	0.78	0.41	35716	-10629	0.71	-0.21	5.68
BAS	0.64	0.53	25797	4634	0.31	0.06	3.62
BAYN	0.66	0.53	29903	-2188	0.46	-0.03	5.41
BEI	0.66	0.56	9429	3576	0.51	0.19	16.39
BMW	0.63	0.54	23534	6918	0.34	0.10	4.67
CBK	0.66	0.48	35482	-5651	0.74	-0.12	6.13
CON	0.68	0.50	25421	-818	0.60	-0.02	6.79
DAI	0.65	0.52	30363	2858	0.33	0.03	3.21
DB1	0.64	0.53	12810	2431	0.46	0.09	11.39
DBK	0.65	0.51	36220	680	0.35	0.01	2.86
DPW	0.67	0.50	22885	510	0.40	0.01	5.04
DTE	0.76	0.43	30554	-8182	0.74	-0.20	7.06
EOAN	0.73	0.45	31014	-7194	0.65	-0.15	5.93
FME	0.64	0.56	11662	4827	0.48	0.20	12.80
FRE	0.70	0.46	15327	-2956	0.81	-0.16	15.35
HEI	0.64	0.55	16183	5277	0.52	0.17	10.29
HEN3	0.63	0.53	13860	3477	0.35	0.09	8.11
IFX	0.63	0.56	17244	7489	0.47	0.20	8.60
LHA	0.65	0.52	23631	2984	0.66	0.08	8.64
LIN	0.73	0.44	18545	-4316	0.59	-0.14	9.10
LXS	0.63	0.55	14114	4518	0.61	0.19	13.73
MRK	0.71	0.48	14336	-1686	0.80	-0.09	16.46
MUV2	0.74	0.44	21983	-4896	0.56	-0.13	7.92
RWE	0.64	0.53	21865	3972	0.44	0.08	6.01
SAP	0.70	0.49	25547	-1258	0.37	-0.02	4.17
SDF	0.62	0.54	15721	4920	0.69	0.22	14.03
SIE	0.65	0.54	24341	6327	0.30	0.08	3.74
TKA	0.65	0.53	18225	3208	0.63	0.11	10.34
VOW3	0.72	0.49	31555	-2128	0.53	-0.04	4.99
Average	0.67	0.51	22369	678	0.53	0.02	8.14
Min	0.62	0.41	9429	-10629	0.30	-0.21	2.86
Max	0.78	0.56	36220	7489	0.81	0.22	16.46

Table 5.32: In sample and out of sample results for the strategy with 100 events window.

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.51	0.50	1489	173	0.18	0.02	50.37
ALV	0.53	0.49	4415	-2363	0.09	-0.05	6.92
BAS	0.51	0.50	2418	133	0.15	0.01	20.86
BAYN	0.51	0.50	3084	-648	0.18	-0.04	18.17
BEI	0.51	0.50	575	89	0.22	0.03	189.95
BMW	0.51	0.50	2066	408	0.20	0.04	35.58
CBK	0.51	0.50	3058	-581	0.37	-0.07	35.48
CON	0.52	0.50	2102	-216	0.23	-0.02	35.11
DAI	0.51	0.50	2968	118	0.18	0.01	19.22
DB1	0.51	0.50	923	35	0.25	0.01	129.15
DBK	0.51	0.50	3330	-143	0.16	-0.01	15.75
DPW	0.52	0.50	2182	3	0.18	0.00	28.39
DTE	0.52	0.49	3056	-932	0.15	-0.05	15.82
EOAN	0.52	0.49	3130	-1008	0.16	-0.05	15.77
FME	0.51	0.50	847	245	0.26	0.08	174.98
FRE	0.52	0.49	1027	-362	0.23	-0.08	58.15
HEI	0.51	0.50	1172	249	0.27	0.06	93.24
HEN3	0.51	0.50	1094	138	0.19	0.02	62.55
IFX	0.51	0.50	1167	392	0.28	0.09	83.42
LHA	0.51	0.50	1824	244	0.36	0.05	72.03
LIN	0.52	0.49	1721	-555	0.15	-0.05	36.66
LXS	0.51	0.50	873	217	0.39	0.10	312.12
MRK	0.52	0.50	1118	-185	0.30	-0.05	92.36
MUV2	0.52	0.49	2105	-736	0.13	-0.05	23.87
RWE	0.51	0.50	1936	273	0.25	0.03	38.03
SAP	0.52	0.50	2359	-552	0.11	-0.03	15.06
SDF	0.51	0.50	1028	140	0.41	0.06	159.24
SIE	0.51	0.50	2112	206	0.14	0.01	22.14
TKA	0.51	0.50	1364	297	0.34	0.07	102.45
VOW3	0.52	0.50	3109	-679	0.14	-0.03	14.39
Average	0.51	0.50	1988	-187	0.22	0.00	65.91
Min	0.51	0.49	575	-2363	0.09	-0.08	6.92
Max	0.53	0.50	4415	408	0.41	0.10	312.12

Table 5.33: In sample and out of sample results for the strategy with a 10000-event window.

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.90	0.55	53516	5470	0.50	0.05	5.19
ALV	0.87	0.46	61801	-7359	1.10	-0.13	10.24
BAS	0.90	0.55	79601	8809	0.45	0.05	3.20
BAYN	0.89	0.55	70429	759	0.58	0.01	8.42
BEI	0.90	0.57	26666	3840	0.59	0.09	12.19
BMW	0.90	0.57	76759	10820	0.47	0.07	3.47
CBK	0.87	0.46	93468	-11901	1.20	-0.15	7.07
CON	0.87	0.47	58915	-5036	0.87	-0.07	8.07
DAI	0.89	0.53	88952	5077	0.52	0.03	3.17
DB1	0.89	0.55	40286	3473	0.69	0.06	9.54
DBK	0.88	0.51	100714	2150	0.55	0.01	3.06
DPW	0.89	0.51	60375	854	0.61	0.01	5.56
DTE	0.86	0.45	55121	-8198	1.12	-0.17	11.49
EOAN	0.88	0.46	64667	-6883	1.00	-0.11	8.65
FME	0.90	0.58	36180	5307	0.51	0.07	7.90
FRE	0.86	0.46	32695	-3577	1.27	-0.14	22.54
HEI	0.90	0.57	49458	7317	0.67	0.10	7.71
HEN3	0.91	0.56	46321	5593	0.49	0.06	5.84
IFX	0.91	0.59	57987	11098	0.55	0.11	5.21
LHA	0.88	0.50	64853	-1384	0.93	-0.02	8.18
LIN	0.88	0.46	37596	-4047	0.92	-0.10	13.56
LXS	0.89	0.56	46852	5506	0.78	0.09	9.95
MRK	0.87	0.46	30898	-3902	1.16	-0.15	21.49
MUV2	0.88	0.46	43045	-4732	0.86	-0.09	11.68
RWE	0.90	0.55	69063	7855	0.66	0.08	5.24
SAP	0.87	0.49	56116	-1788	0.54	-0.02	5.28
SDF	0.89	0.55	53894	4781	0.87	0.08	9.11
SIE	0.90	0.55	72221	8510	0.41	0.05	3.24
TKA	0.88	0.52	49754	1348	0.85	0.02	9.50
VOW3	0.88	0.49	66485	-2861	0.77	-0.03	6.69
Average	0.89	0.52	58156	1230	0.75	0.00	8.41
Min	0.86	0.45	26666	-11901	0.41	-0.17	3.06
Max	0.91	0.59	100714	11098	1.27	0.11	22.54

Table 5.34: In sample and out of sample results for the strategy with Hawkes model.

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.52	0.51	3177	1038	0.96	0.32	158.07
ALV	0.51	0.49	2282	-1615	0.89	-0.63	204.19
BAS	0.52	0.50	3480	904	0.82	0.21	122.10
BAYN	0.52	0.50	2841	-190	0.85	-0.06	183.50
BEI	0.54	0.52	2684	1034	1.16	0.45	225.41
BMW	0.52	0.51	3977	1483	0.92	0.34	121.93
CBK	0.52	0.50	5414	-104	1.94	-0.04	186.35
CON	0.52	0.50	3853	333	1.39	0.12	187.64
DAI	0.52	0.50	3893	837	0.90	0.19	120.77
DB1	0.53	0.51	3016	724	1.20	0.29	206.79
DBK	0.51	0.50	3801	466	0.89	0.11	123.25
DPW	0.52	0.50	3094	262	0.97	0.08	162.60
DTE	0.52	0.49	2773	-996	1.19	-0.43	220.63
EOAN	0.52	0.49	2990	-1098	1.14	-0.42	201.41
FME	0.53	0.51	2968	1164	1.04	0.41	183.09
FRE	0.53	0.49	2437	-847	1.33	-0.46	283.95
HEI	0.53	0.51	3680	1350	1.32	0.49	187.46
HEN3	0.53	0.51	3274	1156	0.99	0.35	159.92
IFX	0.53	0.51	4099	1825	1.16	0.52	148.39
LHA	0.53	0.51	4476	862	1.66	0.32	195.08
LIN	0.52	0.50	2370	-544	1.11	-0.26	245.63
LXS	0.53	0.51	3704	1373	1.53	0.57	215.38
MRK	0.53	0.49	2556	-520	1.42	-0.29	298.87
MUV2	0.52	0.50	2558	-521	1.04	-0.21	218.05
RWE	0.52	0.51	4178	1311	1.23	0.39	153.48
SAP	0.52	0.50	2566	91	0.80	0.03	166.65
SDF	0.53	0.51	4659	1669	1.75	0.63	197.17
SIE	0.52	0.51	3258	1032	0.76	0.24	123.72
TKA	0.53	0.51	3870	856	1.54	0.34	207.78
VOW3	0.52	0.50	3232	-304	1.06	-0.10	173.23
Average	0.52	0.50	3372	434	1.17	0.12	186.08
Min	0.51	0.49	2282	-1615	0.76	-0.63	120.77
Max	0.54	0.52	5414	1825	1.94	0.63	298.87

Table 5.35: In sample and out of sample results for the strategy with Hawkes model.

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.76	0.72	35811	28428	0.10	0.08	1.46
ALV	0.67	0.70	28843	33436	0.06	0.07	1.13
BAS	0.81	0.73	62381	43995	0.11	0.08	1.00
BAYN	0.78	0.73	43897	38894	0.09	0.08	1.38
BEI	0.72	0.73	14638	14665	0.10	0.10	3.70
BMW	0.77	0.72	53356	41168	0.11	0.09	1.25
CBK	0.61	0.69	25993	48038	0.09	0.17	1.88
CON	0.59	0.74	14658	37682	0.05	0.12	1.68
DAI	0.75	0.71	57401	48337	0.09	0.08	0.88
DB1	0.73	0.73	23316	22699	0.13	0.13	3.14
DBK	0.73	0.70	62553	53172	0.10	0.08	0.88
DPW	0.73	0.72	35910	33775	0.09	0.08	1.34
DTE	0.63	0.70	19413	29932	0.06	0.09	1.59
EOAN	0.67	0.71	28022	34662	0.08	0.09	1.48
FME	0.79	0.71	26511	18334	0.14	0.10	2.94
FRE	0.65	0.69	13281	17525	0.09	0.12	3.84
HEI	0.73	0.73	29203	28147	0.13	0.12	2.59
HEN3	0.78	0.73	32835	24911	0.12	0.09	2.11
IFX	0.77	0.73	38947	30362	0.14	0.11	1.99
LHA	0.66	0.70	27077	33421	0.12	0.15	2.53
LIN	0.68	0.72	18017	21490	0.07	0.08	2.09
LXS	0.76	0.71	31492	23976	0.21	0.16	3.67
MRK	0.68	0.69	15406	15869	0.11	0.12	4.25
MUV2	0.70	0.71	22705	24105	0.07	0.08	1.86
RWE	0.75	0.72	45135	37955	0.13	0.11	1.52
SAP	0.67	0.72	25553	32530	0.05	0.06	1.06
SDF	0.74	0.70	34076	26084	0.23	0.17	3.64
SIE	0.76	0.72	47104	39092	0.08	0.07	0.94
TKA	0.66	0.71	20997	26506	0.11	0.13	2.82
VOW3	0.66	0.72	27259	38411	0.06	0.08	1.16
Average	0.71	0.71	32,060	31,587	0.10	0.10	2.06
Min	0.59	0.69	13,281	14,665	0.05	0.06	0.88
Max	0.81	0.74	62,553	53,172	0.23	0.17	4.25

Table 5.36: In sample and out of sample results for the strategy with multivariate Hawkes model.

	Acc		PnL		Bps		Hp
Ticker	In sam.	Out of sam.	In sam.	Out of sam.	In sam.	Out of sam.	
ADS	0.53	0.52	3788	3329	0.88	0.77	123.18
ALV	0.52	0.52	3458	3207	1.00	0.93	151.88
BAS	0.52	0.52	4171	3547	0.74	0.63	92.98
BAYN	0.52	0.52	3919	3554	0.94	0.85	142.89
BEI	0.56	0.55	3626	3415	1.20	1.13	173.17
BMW	0.52	0.52	4617	4160	0.88	0.80	99.80
CBK	0.52	0.52	4897	4678	1.34	1.28	143.93
CON	0.53	0.53	4891	4835	1.41	1.40	150.96
DAI	0.52	0.52	4276	3815	0.80	0.71	97.34
DB1	0.54	0.54	4157	3802	1.25	1.14	154.90
DBK	0.52	0.51	4517	3918	0.87	0.75	99.01
DPW	0.52	0.52	3912	3498	0.98	0.87	131.50
DTE	0.52	0.52	3493	3205	1.08	0.99	160.18
EOAN	0.52	0.52	3674	3365	1.09	1.00	156.70
FME	0.55	0.54	4122	3649	1.08	0.95	136.88
FRE	0.54	0.54	3713	3443	1.53	1.42	215.00
HEI	0.54	0.54	5255	4824	1.45	1.33	146.71
HEN3	0.54	0.53	4324	3882	1.05	0.94	126.85
IFX	0.54	0.54	5589	5056	1.24	1.12	115.77
LHA	0.53	0.53	5568	5282	1.59	1.51	146.69
LIN	0.53	0.53	3162	2805	1.05	0.93	174.51
LXS	0.55	0.54	5363	4894	1.65	1.51	161.18
MRK	0.54	0.54	3746	3497	1.58	1.48	223.06
MUV2	0.53	0.53	3399	2957	1.10	0.96	169.38
RWE	0.53	0.53	4946	4439	1.18	1.06	124.06
SAP	0.52	0.52	3299	3036	0.78	0.72	123.63
SDF	0.55	0.54	6349	5642	1.89	1.68	155.53
SIE	0.52	0.52	3638	3266	0.67	0.60	97.18
TKA	0.54	0.54	5083	4800	1.63	1.54	167.35
VOW3	0.52	0.52	3894	3509	1.06	0.96	143.64
Average	0.53	0.53	4,295	3,910	1.17	1.07	143.53
Min	0.52	0.51	3,162	2,805	0.67	0.60	92.98
Max	0.56	0.55	6,349	5,642	1.89	1.68	223.06

Table 5.37: In sample and out of sample results for the strategy with multivariate Hawkes model (1-minute EMA).

Ticker	Price (Eur)	Volume (10^6 Eur)	Tick size (Eur)	Tick size (Bp)	Spread (Eur)	Spread (Bp)	Spread (Tick)
ADS	78	90	0.01	1.3	0.019	2.5	1.9
ALV	123	184	0.05	4.1	0.057	4.6	1.1
BAS	82	176	0.01	1.2	0.018	2.1	1.8
BAYN	100	163	0.05	5.0	0.040	4.0	1.1
BEI	72	31	0.01	1.4	0.025	3.5	2.5
BMW	89	141	0.01	1.1	0.021	2.4	2.1
CBK	13	111	0.005	4.0	0.008	6.1	1.5
CON	171	81	0.05	2.9	0.071	4.2	1.4
DAI	68	201	0.01	1.5	0.018	2.7	1.8
DB1	56	35	0.01	1.8	0.021	3.8	2.1
DBK	31	215	0.005	1.6	0.009	2.9	1.8
DPW	27	104	0.005	1.9	0.007	2.7	1.5
DTE	12	131	0.005	4.1	0.006	5.2	1.3
EOAN	14	112	0.005	3.6	0.006	4.5	1.3
FME	49	44	0.01	2.0	0.014	2.9	1.4
FRE	110	39	0.05	4.6	0.073	6.7	1.5
HEI	62	49	0.01	1.6	0.026	4.1	2.6
HEN3	81	51	0.01	1.2	0.022	2.7	2.2
IFX	9	54	0.001	1.2	0.003	3.2	2.8
LHA	18	69	0.005	2.7	0.009	4.7	1.7
LIN	149	66	0.05	3.4	0.068	4.6	1.4
LXS	53	37	0.01	1.9	0.025	4.7	2.5
MRK	123	34	0.05	4.1	0.082	6.7	1.6
MUV2	159	97	0.05	3.1	0.074	4.7	1.5
RWE	29	81	0.005	1.7	0.009	3.3	1.9
SAP	57	147	0.01	1.8	0.015	2.6	1.5
SDF	24	40	0.005	2.0	0.010	3.9	1.9
SIE	97	190	0.01	1.0	0.019	1.9	1.9
TKA	21	45	0.005	2.4	0.010	4.7	1.9
VOW3	191	168	0.05	2.6	0.073	3.8	1.5
Average	72	99	0.019	2.4	0.029	3.9	1.7
Min	9	31	0.001	1	0.003	1.9	1.1
Max	191	215	0.05	5	0.082	6.7	2.8

Table 5.38: Stocks basic properties summary

Ticker	L_{buy}	L_{sell}	L	C_{buy}	C_{sell}	C	M_{buy}	M_{sell}	M	O
ADS	25376	25544	50920	19459	18651	38110	4729	4322	9051	98081
ALV	41186	41365	82551	39301	39122	78423	4149	4000	8149	169123
BAS	39236	40392	79628	33163	34925	68088	5980	6105	12085	159801
BAYN	36273	36900	73173	29827	30202	60029	5592	5554	11146	144348
BEI	10851	10785	21636	8104	7937	16041	2062	2036	4098	41775
BMW	34188	34751	68939	24636	26116	50752	6138	6370	12508	132199
CBK	17750	17843	35593	15938	15593	31531	3797	3771	7568	74692
CON	20608	20550	41158	17878	17740	35618	2959	2983	5942	82718
DAI	40772	40626	81398	34144	34341	68485	6716	6846	13562	163445
DB1	12317	12549	24866	9547	9443	18990	2415	2301	4716	48572
DBK	44321	46123	90444	41296	41075	82371	7665	7321	14986	187801
DPW	27523	27603	55126	23779	23813	47592	4175	4256	8431	111149
DTE	27457	26961	54418	25525	25573	51098	3491	3736	7227	112743
EOAN	27932	27706	55638	25558	26132	51690	3356	3659	7015	114343
FME	13886	13458	27344	9910	9762	19672	3163	3073	6236	53252
FRE	9048	9151	18199	8149	8170	16319	1575	1593	3168	37686
HEI	16778	16226	33004	12964	12305	25269	2995	2922	5917	64190
HEN3	17890	17742	35632	13692	13587	27279	3214	3256	6470	69381
IFX	18791	19634	38425	13375	14453	27828	3487	3820	7307	73560
LHA	14973	15748	30721	13110	13426	26536	3644	3586	7230	64487
LIN	17788	17740	35528	16584	16578	33162	2137	2001	4138	72828
LXS	10873	11127	22000	7184	7640	14824	2537	2584	5121	41945
MRK	8804	8883	17687	7726	7964	15690	1575	1481	3056	36433
MUV2	22801	23371	46172	21877	22237	44114	2589	2403	4992	95278
RWE	24074	24214	48288	17949	18775	36724	4204	4356	8560	93572
SAP	38442	39211	77653	34939	36111	71050	5556	5504	11060	159763
SDF	10126	10055	20181	7062	7410	14472	2396	2478	4874	39527
SIE	41812	41961	83773	33809	34534	68343	6991	7043	14034	166150
TKA	13659	13507	27166	11537	12237	23774	2548	2717	5265	56205
VOW3	35069	34839	69908	31813	31882	63695	4265	4217	8482	142085
Average	24020	24219	48239	20328	20591	40919	3870	3876	7764	96904
Min	8804	8883	17687	7062	7410	14472	1575	1481	3056	36433
Max	44321	46123	90444	41296	41075	82371	7665	7321	14986	187801

Table 5.39: Event occurrences statistics summary

Ticker	L^0	L^1	L	C^0	C^1	C	M^0	M^1	M	O^0	O^1
ADS	47.03	4.89	51.92	36.77	2.08	38.86	6.33	2.89	9.23	90.14	9.86
ALV	47.63	1.18	48.81	45.95	0.42	46.37	4.05	0.76	4.82	97.63	2.37
BAS	45.11	4.72	49.83	40.45	2.16	42.61	4.91	2.65	7.56	90.47	9.53
BAYN	46.93	3.76	50.69	39.93	1.66	41.59	5.52	2.20	7.72	92.38	7.62
BEI	46.82	4.97	51.79	36.37	2.03	38.40	6.77	3.05	9.81	89.96	10.04
BMW	46.36	5.79	52.15	35.79	2.60	38.39	6.14	3.32	9.46	88.29	11.71
CBK	43.65	4.00	47.65	41.02	1.19	42.21	7.26	2.88	10.13	91.93	8.07
CON	46.59	3.16	49.76	41.69	1.37	43.06	5.35	1.83	7.18	93.63	6.37
DAI	45.37	4.43	49.80	40.01	1.89	41.90	5.64	2.66	8.30	91.02	8.98
DB1	45.79	5.40	51.20	36.80	2.30	39.10	6.57	3.13	9.71	89.17	10.83
DBK	44.04	4.12	48.16	42.00	1.86	43.86	5.61	2.36	7.98	91.65	8.35
DPW	46.06	3.54	49.60	41.40	1.42	42.82	5.39	2.19	7.58	92.85	7.15
DTE	46.72	1.55	48.27	44.84	0.48	45.32	5.33	1.08	6.41	96.90	3.10
EOAN	46.70	1.96	48.66	44.52	0.69	45.21	4.84	1.29	6.13	96.06	3.94
FME	44.82	6.53	51.35	34.22	2.72	36.94	7.83	3.88	11.71	86.87	13.13
FRE	45.75	2.54	48.29	42.59	0.71	43.30	6.55	1.85	8.40	94.89	5.11
HEI	46.03	5.38	51.42	36.89	2.47	39.37	6.16	3.06	9.22	89.09	10.91
HEN3	45.33	6.03	51.36	36.41	2.91	39.32	6.18	3.15	9.32	87.92	12.08
IFX	45.04	7.20	52.24	34.28	3.55	37.83	6.29	3.65	9.93	85.61	14.39
LHA	43.09	4.55	47.64	39.63	1.52	41.15	8.10	3.11	11.21	90.82	9.18
LIN	46.82	1.97	48.79	44.81	0.72	45.53	4.43	1.25	5.68	96.07	3.93
LXS	45.58	6.87	52.45	32.58	2.76	35.34	8.08	4.13	12.21	86.24	13.76
MRK	45.85	2.70	48.55	42.25	0.81	43.07	6.49	1.90	8.39	94.59	5.41
MUV2	46.62	1.84	48.46	45.63	0.67	46.30	4.06	1.18	5.24	96.31	3.69
RWE	46.65	4.96	51.60	37.23	2.01	39.25	6.11	3.04	9.15	89.99	10.01
SAP	46.04	2.56	48.61	43.52	0.95	44.47	5.27	1.65	6.92	94.83	5.17
SDF	43.69	7.37	51.06	33.57	3.04	36.61	7.81	4.52	12.33	85.07	14.93
SIE	45.59	4.83	50.42	39.01	2.12	41.13	5.63	2.82	8.45	90.23	9.77
TKA	43.90	4.43	48.33	40.60	1.70	42.30	6.55	2.82	9.37	91.05	8.95
VOW3	46.94	2.26	49.20	43.93	0.90	44.83	4.57	1.40	5.97	95.45	4.55
Average	45.75	4.18	49.94	39.82	1.72	41.55	5.99	2.52	8.52	91.57	8.43
Min	43.09	1.18	47.64	32.58	0.42	35.34	4.05	0.76	4.82	85.07	2.37
Max	47.63	7.37	52.45	45.95	3.55	46.37	8.1	4.52	12.33	97.63	14.93

Table 5.40: Percentage of occurrences per event type

Ticker	$L^1\|O^1$	$C^1\|O^1$	$M^1\|O^1$	$O^1\|O$
ADS	49.56	21.11	29.34	9.86
ALV	49.99	17.78	32.23	2.37
BAS	49.48	22.69	27.83	9.53
BAYN	49.38	21.75	28.87	7.62
BEI	49.45	20.23	30.31	10.04
BMW	49.45	22.20	28.35	11.71
CBK	49.58	14.80	35.63	8.07
CON	49.66	21.56	28.79	6.37
DAI	49.36	21.06	29.58	8.98
DB1	49.87	21.20	28.94	10.83
DBK	49.34	22.33	28.33	8.35
DPW	49.48	19.83	30.69	7.15
DTE	49.84	15.45	34.71	3.10
EOAN	49.77	17.46	32.77	3.94
FME	49.69	20.74	29.57	13.13
FRE	49.82	13.97	36.21	5.11
HEI	49.32	22.66	28.02	10.91
HEN3	49.90	24.06	26.03	12.08
IFX	50.03	24.63	25.33	14.39
LHA	49.54	16.57	33.89	9.18
LIN	49.95	18.36	31.69	3.93
LXS	49.94	20.05	30.01	13.76
MRK	49.87	15.02	35.11	5.41
MUV2	49.90	18.24	31.86	3.69
RWE	49.52	20.10	30.38	10.01
SAP	49.61	18.42	31.97	5.17
SDF	49.35	20.39	30.27	14.93
SIE	49.44	21.74	28.82	9.77
TKA	49.54	18.99	31.47	8.95
VOW3	49.62	19.72	30.65	4.55
Average	49.64	19.77	30.59	8.43
Min	49.32	13.97	25.33	2.37
Max	50.03	24.63	36.21	14.93

Table 5.41: Repartition of events impacting the mid price